VNIVERSITAT DE VALÈNCIA

Facultat d'Economia

DEPARTAMENT DE COMPTABILITAT

DOCTORADO EN CONTABILIDAD Y FINANZAS CORPORATIVAS

(R.D. 99/2011)

TESIS DOCTORAL

CONTABILIDAD FORENSE Y BLANQUEO DE CAPITALES: APLICACIÓN DEL APRENDIZAJE AUTOMÁTICO EN UN PROCESO JUDICIAL ESPAÑOL

Presentada por:
Elena Badal Valero

Directores:
Dr. José Manuel Pavía Miralles
Dra. Begoña Giner Inchausti
Dr. José Antonio Álvarez Jareño

Valencia, septiembre de 2017

A mis padres y abuelos, pilares de todo lo que soy.

A toda mi familia, por su enorme cariño e incondicional apoyo.

AGRADECIMIENTOS

Me gustaría que estas líneas sirvieran para expresar mi más profundo y sincero agradecimiento a todas aquellas personas que con su ayuda han colaborado en la realización de esta Tesis Doctoral.

En primer lugar, al Dr. Don José Manuel Pavía, al Dr. Don José Antonio Álvarez y a la Dra. Doña Begoña Giner, como directores de esta investigación y mentores durante estos años, por la orientación, el seguimiento y la supervisión continua, pero sobre todo por la motivación y el apoyo recibido.

Expresar mi gratitud al Cuerpo Superior de la Policía Judicial Española, por confiar en este ambicioso proyecto y por su entrega diaria en la lucha contra el blanqueo de capitales. Un reconocido agradecimiento a Roberto, Pepe y Eugenio.

También agradecer a los compañeros del Departament d´Economia Aplicada sus continuas muestras de apoyo y sugerencias recibidas, especialmente a Belén, Fran y Prudencio.

Asimismo, expresar mi gratitud al profesor Don Dionysios S. Demetis, por su tutela durante mi estancia de investigación en la University of Hull, así como a los profesores del Department of Management Systems por su cálida acogida.

Mi agradecimiento, también, a la Generalitat Valenciana y a la Universitat de València, por la confianza que mostraron en mí al concederme la Ayuda VALi+d con la que fue posible aventurarme en esta travesía.

Finalmente, agradecer a toda mi familia las infinitas muestras de cariño recibidas y su amparo incondicional, sin los que no hubiera sido capaz de realizar este trabajo. A mis padres y abuelos, a quienes especialmente dedico esta Tesis Doctoral, por su esfuerzo, dedicación y confianza ciega. A mi hermana, por enseñarme el valor de la constancia. A Cristina, por sus constantes muestras de ánimo en los momentos más necesitados. Y a todos mis amigos y compañeros que, aunque no nombre, forman parte de este proyecto.

A todos, muchas gracias.

PRÓLOGO

La Tesis Doctoral aquí presentada se corresponde en buena medida con los resultados obtenidos en el desarrollo de un peritaje forense sobre datos contables anonimizados obtenidos de información proveniente de los registros de las operaciones de compra-venta de una empresa núcleo investigada por la Policía Judicial Española, Grupo de Blanqueo de Capitales, Cuerpo Superior de la Policía Nacional, en el transcurso de una investigación por delito de blanqueo de capitales.

En el proceso judicial la autora de esta Tesis Doctoral ha colaborado como forense contable, y en ella se han implementado de forma precursora técnicas de aprendizaje automático y sistemas expertos que en este trabajo se exponen para la detección de patrones de blanqueo de capitales. Dichas técnicas ofrecen una nueva herramienta que les permite a los investigadores policiales priorizar los recursos de investigación disponibles hacia aquellas empresas sospechosas que presentan potenciales patrones de fraude.

El caso objeto de estudio representa una de las investigaciones policiales de blanqueo de capitales más importantes en Europa hasta la fecha, tanto en cuantía monetaria defraudada, como en número de empresas involucradas. El entramado empresarial estaba formado por más de 640 empresas que, junto con la empresa núcleo, habrían estado blanqueando una cifra desorbitada de cientos de millones de euros en los últimos años. La base de datos contiene más de doce millones de registros extraídos de la contabilidad interna de la empresa núcleo de esta estructura empresarial potencialmente defraudadora.

Lo anterior permite valorar la complejidad del proceso de investigación y la gran utilidad de la implementación de los modelos estadísticos más vanguardistas.

La importancia del caso proviene del hecho de que la empresa acaparó, comenzando desde cero y en pocos años, aproximadamente el 50% de la cuota total nacional del mercado en el que operó, hundiendo a la competencia legal al comercializar bienes a precios por debajo de costes. Estos hechos representan claros indicios de blanqueo de capitales, y han sido descubiertos y probados gracias a las técnicas de aprendizaje

automático aplicadas de forma pionera en el trabajo forense, análisis que complementan los estudios de detección de patrones que aquí se presentan.

Por motivos de confidencialidad, en este trabajo se procura no ofrecer detalles que permitan identificar ni el proceso judicial, todavía sub júdice, ni la empresa de la que provienen los datos examinados. Particularmente, se evita ofrecer detalles de los productos que se comercializaron y dar información relativa a importes económicos para dificultar que cualquier lector pueda "descifrar" a qué proceso judicial concreto corresponden los resultados presentados.

En el marco de esta investigación se ha creado el Grupo de Investigación de Delitos Financieros y Blanqueo de Capitales, adscrito al Departament d´Economia Aplicada de la Universitat de València, con el objetivo de implementar las avanzadas técnicas de aprendizaje automático (Statistical/Machine Learning) para detectar y analizar las operativas de delitos financieros y desarrollar protocolos que ayuden a prevenirlos. Como se examina en esta Tesis Doctoral, la aplicación de estas técnicas permite predecir patrones de comportamiento de los agentes económicos y detectar operaciones anómalas que pasarían inadvertidas con los métodos tradicionales.

Este grupo de investigación ya ha colaborado como forense contable en tres procesos judiciales de blanqueo de capitales a petición de la Policía Nacional, Brigada de Blanqueo de Capitales. Y dados los satisfactorios resultados obtenidos, la Universitat de València y la Policía Nacional están pendientes de firmar, en unos meses, el primer convenio marco que facilitará la incorporación de expertos docentes e investigadores en procesos judiciales, no solo en el ámbito de la contabilidad forense sino también en otras muchas especialidades.

ÍNDICE

ÍNDICE DE CONTENIDO

ÍNDICE DE GRÁFICOS, FIGURAS Y TABLAS

INTRODUCTION

Council of Bars and Law Societies of Europe (*CCBE*), 2014, p. 2.

"*Money laundering and terrorist financing represent serious threats to life and society and result in violence, fuel further criminal activity, and threaten the foundations of the rule of law (in its broadest sense).*"

INTRODUCTION

Money laundering is a financial crime which has evolved over time and is implemented at different levels and to different degrees. The defrauded amounts range from the traditional laundering of small amounts of money from retail and local drug trafficking to large amounts (billions of euros) from business macro-structures that emerged in recent decades and which operate on an international scale (Khac and Kechadi, 2010).

With respect to the high socio-economic impact of this crime, money laundering has many negative effects on society as a whole, directly and indirectly (money laundering offences and crimes giving rise to money laundering), from its effect on the individual victims to the wider scale world economy, both in the short and long term (Bartlett, 2002; Quirk, 1996; UNODC, 2011; Unger, 2007). On an international scale, these negative effects are seen at all levels: in economic systems, in financial institutions, in public bodies and in companies (Unger, 2007).

The fight against financial crime and money laundering has been intensified by public administrations in all developed countries, given the direct relationship of both crimes to the financing of terrorism and weapons of mass destruction (FATF, 2012). The threat posed by financial crimes and money laundering to any country makes having all the necessary economic and intelligent resources essential in order to combat these illicit activities and those they promote. The main objective of the anti-money laundering and counter terrorism financing regime (AML/CFT) is to reduce crime rates related to professional crime, organized crime and terrorism, and in turn to protect society as a whole (CCBE, 2014).

All agents involved in a criminal organization, with few exceptions, carry out illegal activities for the sole purpose of making a profit (Lopez, 2015). Understanding money laundering as the "Achilles' heel" of any criminal organization is the key to combating this crime and preventing criminals from carrying out illegal activities and their enjoyment of illicit capital (Alhosani, 2016).

The current global scenario, shaped by free movement, globalization and the liberalization of world trade, has led to a network of financial and business structures

which take advantage of this network to integrate capital from their illegal activities into the financial market (Demetis, 2011; FATF, 2013).

At the European level, economic and political circumstances pushed the Economic and Monetary Union (EMU) to adapt to a changing and increasingly globalized world. In a first phase, the EMU introduced full freedom for the movement of capital, as outlined in Council Directive 88/361/EEC. Subsequently, in a second phase, with the signing of the Maastricht Treaty in 1992, it was envisaged that all restrictions on capital movements and payments, both between Member States and between Member States and third parties (Galán, 2016), would be prohibited. At the moment, the exceptions are mainly limited to movements of capital related to third parties, Articles 64 and 65 of the Treaty of Functioning of the European Union.

The new order in world trade, emanating from the globalization of the economy, makes it very difficult to exercise effective control over the movement of money (UNODC, 2011). Furthermore, trying to put in place mechanisms of obstruction would go against the liberalization of world trade which, in addition to most companies and governments, defends institutions such as the World Bank Group and the World Trade Organization.

In this context, efforts to combat these economic crimes are implemented at all levels. At international level, mechanisms to combat money laundering and financing of terrorism are structured in a coordinated manner among international institutions, such as the International Financial Action Task Force (FATF), the International Monetary Fund, the United Nations or the European Union and Financial Intelligence Units (FIUs) of the different countries (Unger, 2009). This suggests the existence of a coordinated global regime to combat money laundering and the financing of terrorism (FATF, 2012). The International Financial Action Task Force has developed 40 recommendations against money laundering which, together with the 9 recommendations against the financing of terrorism, set out the basic framework for the detection, prevention and elimination of money laundering and financing of terrorism (FATF, 2012).

These recommendations constitute a set of principles that countries must translate into their regulations. The European Union has disseminated these recommendations to its

Member States in a number of Directives (91/308/EEC, 2005/60/EC, 2006/70/EC, 2015/849/EC).

At national level, FIUs are an important component of these strategies in that these institutions are responsible for receiving, analysing and transmitting suspicious money laundering transactions to competent authorities, all of which is coordinated by the Egmont Group of Financial Intelligence Units (Forget and Hočevar, 2004). In Spain, the Financial Intelligence Unit is *La Comisión de Prevención de Blanqueo de Capitales e Infracciones Monetarias* (SEPBLAC), which is responsible for the development of the anti-money laundering policy in our country.

However, despite coordinated efforts, there is currently not even a universally accepted methodology for estimating the amount of money from illicit sources that is integrated into financial systems around the world and, thus, the effectiveness of the International system against such crimes (AML/CFT) is also an unknown (Unger, 2007; Barone and Masciandaro, 2011). A consensus of different "guesstimates" places it between 2% to 5% of the Global Gross Domestic Product (Camdessus, 1998; Unger, 2007; UNODC, 2011; Unger, 2013). In Spain, estimates of money laundering differ considerably and range from €36 million (Fernandez, 2012) to $56 US billion per year (Walker, 1999).

The context described in the previous paragraphs is where the current research is focused. As a starting point, an investigation has been carried out, together with two agents of the Judicial Police, the Money Laundering Group and the Economic Delinquency Group, to analyse the evolution of money laundering offenses and arrests in Spain between 1998 and 2012[1].

The main results, presented in Section I.6, show how during the period 1998-2012 the rate of the number of money laundering offenses in Spain increased almost 244%, while the rate of the number of people arrested for money laundering increased by 431%. Conversely, the rate of the number of predicate offenses of money laundering decreased by almost 30%, and the rate of the number of persons detained in predicate offenses also declined by 16% in the period under review.

[1] Full article in *ANEXO II*.- Publicaciones. "*Money laundering trend in Spain: offences and arrests over 15 years*".

The study would indicate that the anti-money laundering regime's efforts in prevention and legislation are having positive results in the short-term; however, factors are known to exist which could limit its efficiency[2].

Currently, there is an increase in the number of cases of money laundering and their degree of sophistication. This makes it difficult to identify them and, consequently, increases the negative effects on international economic systems and global development (Zhongfei *et al.*, 2003). In addition, current technological advances, as well as communication systems, have equipped criminals with tools which aid the creation of large, complex and coordinated corporate structures in order to impede capital traceability (financial engineering and numerical engineering) (Petrucelli, 2012).

For this reason, researchers specializing in this area face a double challenge: on the one hand, they must identify large money laundering networks and, on the other hand, they must select from a huge amount of information derived from the companies involved those patterns of behaviour of individuals concealing money laundering. However, traditional techniques make it very difficult to meet these goals (Sremack, 2015).

Specifically, judicial money laundering processes require investigators to have advanced knowledge of data analysis, economics and accounting, as well as knowledge of the most advanced software in data management and statistical techniques (Dutta, 2013). In this sense, the most important contributions of the forensic accountant in judicial processes of money laundering are: (1) the detection of suspicious operations (fraud patterns), (2) the analysis of possible financial crimes and fraud (traceability of the capitals) and (3) the development of techniques that help detect emerging criminal behaviours (Owojori and Asaolu, 2009; Dutta, 2013).

The demands of the work carried out by the forensic accountant, and particularly in the judicial processes of money laundering, require the use of avant-garde techniques for the analysis of the information and its rapid application to the databases available. These techniques are embedded in the concept of machine learning, and offer a wide range of possibilities and resources that grow exponentially (Sremack, 2015).

[2]Section I.6.

Despite the large bibliography on pattern detection in many subject areas, there are still insufficient studies that apply machine learning techniques to detect patterns of delinquency, especially when it comes to studies applying the avant-garde methodologies of machine learning using information obtained from court proceedings relating to money laundering (Hassani *et al.*, 2016; Pearsall, 2010). Thus, most research focuses on developing self-learning techniques to process Suspicious Activity Reports (SARs) received by different financial intelligence units or to address the problem from a general perspective (Nath, 2006; Zhang *et al.*, 2003; Bolton and Hand, 2002; Tang and Yin, 2005).

This research adds two new results in detecting signs of financial fraud: (1) the application of automated learning techniques to internal accounting databases of companies to detect money laundering, and (2) the offer of information to the investigating authorities on how the money laundering network is organized, with the objective of orientating the judicial investigation towards those companies or physical persons who present signs of suspicious patterns.

Thus, in the context of a real macro-case on money laundering in which the author has collaborated as forensic accountant, this study analyses the database available of the operations carried out between a core company and a set of 643 supplier companies, 26 of which had already been identified *a priori* by the Judicial Police as fraudulent. Faced with a well-founded suspicion that other suppliers within the network might have committed criminal acts, and in order to better manage the scarce police resources available, machine learning techniques are proposed with two different approaches to detect patterns of fraud.

The first proposed approach (Section IV.12) is the implementation of Neural Network models to the expert-assisted work for the detection of fraud operations. For this purpose, based on machine learning techniques, the network structure used is that proposed by Hastie *et al.* (2008): The Back-Propagation Network[3].

[3] Full article in *ANEXO III*.- Publicaciones. "*Detección de fraude financiero mediante Redes Neuronales de clasificación en un caso real español*".

In the second approach (Section IV.2), it is proposed a more ambitious procedure to pattern detection than the previous one, in which Benford's Law (Nigrini and Mittermaider, 1997), a tool to characterize accounting records of the commercial operations between the main company and its supplier, is combined with four models of classification: Ridge Logistic Regression (LG) (Le Cessie and van Houwelingen, 1992), Artificial Neural Networks (NN) (Hastie *et al.*, 2008), Decision Tree C4.5 (DT) (Quinlan, 1993 and 1996) and Random Forest (RF) (Breiman, 2001)[4].

Machine learning techniques are employed, using the *a priori* information provided by the Judicial Police on suppliers that are fraudulent (fraud demonstrated) and other supply companies for which information is not available (suspected fraud), for the detection and recognition of patterns in a supervised process of two phases: the training phase (learning) and the classification phase (verification).

The design of an automatic pattern detection system essentially addresses the following three aspects which are covered in successive chapters of this paper: access to information and pre-processing of data (Chapter II), application of methodology and presentation of results (Chapter III) and decision making (Chapter IV) (Li, 2005).

In the context of the expert-assisted study carried out, the pre-processing of data is especially important, as this prior analysis of the content makes up about 65% of the time consumed by forensic work. This process improves the accuracy of the classification of the models, reducing both the amount of data needed to obtain the desired level of performance and the computational resources required as well as incorporating as much information as possible into the study (García *et al.*, 2015).

Although the related literature in this area still is not enough due to the heterogeneity of the different subject areas in which it is applied (García *et al.*, 2015), this work complies enough information to structure the process of management of the databases in different steps that describe the different processes required for preparation of the data.

[4] Full article in *ANEXO IV*.-Publicaciones. "*Combining Benford's Law and Machine Learning to detect Money Laundering. An actual Spanish Court case*".

As such, Data Pre-processing has been divided into six stages: access to information (Data Engineering), data collection (Data Scraping), Data Cleansing, Missing data processing (Missing Values), the detection of atypical values (Outliers) and the transformation of these values (Feature Engineering).

Once the process of management and purification of the databases is carried out, the proposed learning methodologies are applied to the available sample. Chapter III describes the four methodologies of machine learning and ensemble models finally used for the precise detection of patterns: (1) Ridge Logistic Regression (LG), (2) Decision Tree (DT), (3) Neural Network (NN) and (4) Random Forest (RF). In Chapter III we also justify their choice in this research.

From an overall perspective, the outcomes offered by the new ensemble model methodologies, especially the random forest methodology, have been especially successful compared to the independent classification models. In the results of this research, the Random Forest C4.5 applied to Benford's Law estimators achieved the best results, reaching a 96.15% True Negative Rate and a 94.98% True Positive Rate[5].

The application of Benford's Law is used as proposed by Nigrini (1996), as a measure of detection of anomalous data in combination with the 4 methodologies listed above. This law states that some sets of numerical data have a non-uniform distribution of the different digits. Specifically, Benford's Law postulates that there is a pattern of behaviour for each of the digits, these following a particular logarithmic law (Benford, 1938). As economic data largely follow Benford's Law, this offers a tool for making comparisons and for detecting potential data anomalies (Nigrini, 2011; Torres *et al.*, 2007; Bologna and Lindquist, 1995; Nigrini, 1996; Thomas, 1989; York, 2000). However, failure to comply with Benford's Law is not a crime in itself; it is only evidence that data could present irregularities.

In this case, to classify companies into legal or illegal depending on the anomalies detected in their accounting, contrasts are established to determine the degree of compliance of the law in each company. In this sense a new statistic is proposed, the

[5] Random Forest C4.5 using the SMOTE technique.

OverBenford Test; this will be used in parallel with the Statistic Z (Nigrini, 2012) to find the p-values of adjustment to this law of the first and second digits of the variable "Value of transactions".

When machine learning techniques are applied to real databases, a significant problem is the existence of unbalanced data sets (He and García, 2009; Chawla, 2005). Unbalanced data sets are quite common in real cases presented in scientific literature and, as in the case in hand, normally the category that is most relevant for the analysis is the one with the lower proportion of instances (Japkowicz, 2000b; Chawla *et al.* 2003b; Chawla *et al.*, 2004; Dietterich *et al.*, 2003).

In the judicial process under scrutiny, the database is clearly unbalanced[6]. Of the 643 companies that are part of the laundering network, there is only certainty that 26 of them are fraudulent, that is, the operations they carry out do not fall within a legal framework. Therefore, from the information available *a priori* only 4% of the companies can be clearly identified as fraudulent. In this sense, a method of synthetic data generation, called Synthetic Minority Oversampling Technique (SMOTE), and another based on Cost-Sensitive Learning (Cost Matrix) are used to study the improvement that occurs in relation to an analysis without any transformation of the original data (Raw Data).

Finally, to evaluate the sensitivity and performance of classifiers, different assessment strategies are proposed. In the first approach, the confusion matrix is used to evaluate the performance and sensitivity of the network over 100 training sets randomly extracted. In the second approach, cross-validation is used to determine the accuracy of the models over 10 training sets. In this approach, the measures selected to evaluate the precision are: the area of the ROC curve, the Kappa statistic and the RMSE statistic (Root Mean Squared Error).

The evaluation techniques are used to compare the different methodologies of machine learning and to analyse the differences in results between the training data (explanation) and the data of verification (prediction). This determines the capacity of the models,

[6] Information provided by police investigators in the context of the judicial process.

which provides enough information to analyse the patterns of behaviour followed by each company under scrutiny.

Thus, the machine learning techniques proposed in this paper represent an efficient and objective new tool for detecting fraudulent patterns of behaviour for the investigation of money laundering offences, allowing police investigators to focus the limited economic and human resources available in the judicial processes on those companies under suspicion who present a pattern of behaviour similar to that of previously recognized fraudulent companies.

This PhD Dissertation is structured in two parts. On the first part, composed of three Chapters, establishes the theoretical framework on which the research is based. The first Chapter outlines the concept of money laundering and studies the tendency of this crime in Spain. Chapter II describes the process of management and access to information prior to the application of the proposed techniques (Data Pre-processing). Next, Chapter III specifies the methodology applied based on machine learning techniques for the detection of money laundering pattern.

The second part is devoted to the presentation of the judicial process and the analysis of the results. After the presentation of the judicial process and the description of the sample, on the Chapter IV are presented the results obtained in the application of the machine learning techniques proposed in the two approaches. The PhD Dissertation ends with the conclusions and with proposals for further research.

PARTE 1

CONTABILIDAD FORENSE Y APRENDIZAJE AUTOMÁTICO PARA LA DETECCIÓN DE PATRONES DE BLANQUEO DE CAPITALES

CAPÍTULO I

EL DELITO DE BLANQUEO DE CAPITALES Y LA CONTABILIDAD FORENSE

CAPÍTULO I: EL DELITO DE BLANQUEO DE CAPITALES Y LA CONTABILIDAD FORENSE

I.1.- Definición legal de delito de blanqueo de capitales

El Artículo 1.2 de la Ley 10/2010, de 28 de abril, de prevención del blanqueo de capitales y de la financiación del terrorismo, define las siguientes actividades por las cuales se origina el hecho:

"*A los efectos de la presente Ley, se considerarán blanqueo de capitales las siguientes actividades:*

a) La conversión o la transferencia de bienes, a sabiendas de que dichos bienes proceden de una actividad delictiva o de la participación en una actividad delictiva, con el propósito de ocultar o encubrir el origen ilícito de los bienes o de ayudar a personas que estén implicadas a eludir las consecuencias jurídicas de sus actos.

b) La ocultación o el encubrimiento de la naturaleza, el origen, la localización, la disposición, el movimiento o la propiedad real de bienes o derechos sobre bienes, a sabiendas de que dichos bienes proceden de una actividad delictiva o de la participación en una actividad delictiva.

c) La adquisición, posesión o utilización de bienes, a sabiendas, en el momento de la recepción de los mismos, de que proceden de una actividad delictiva o de la participación en una actividad delictiva.

d) La participación en alguna de las actividades mencionadas en las letras anteriores, la asociación para cometer este tipo de actos, las tentativas de perpetrarlas y el hecho de ayudar, instigar o aconsejar a alguien para realizarlas o facilitar su ejecución.

Existirá blanqueo de capitales aun cuando las conductas descritas en las letras precedentes sean realizadas por la persona o personas que cometieron la actividad delictiva que haya generado los bienes."

I.2.- Concepto de blanqueo de capitales

Durante la última década se ha intentado unificar el Derecho Penal en toda la Unión Europea a través del "*Corpus Iuiris* de disposiciones penales para la protección de los intereses financieros de la Unión Europea" (Mallada, 2012). Por otra parte, no existe una definición internacionalmente aceptada del concepto de blanqueo de capitales, de ahí que sean numerosos los autores que realizan definiciones del concepto más o menos amplias.

En este contexto, y ante el gran desafío ligado al fenómeno de blanqueo de capitales, esto es, aspectos de índole internacional y que influyen en el estado de bienestar de los países, como la financiación del terrorismo o el tráfico de armas y de personas, una de las definiciones más empleada por las distintas Unidades de Inteligencia Financiera de los países (FIUs) es la siguiente:

"(…)"*cualquier acto o intento de ocultar o disimular la identidad de los ingresos obtenidos ilegalmente de manera que parezcan haber sido originados de fuentes legítimas*[7]".

Como se aprecia en la definición anterior, el concepto de blanqueo de capitales incorpora intrínsecamente definiciones adicionales.

Relacionado con este hecho delictivo está asociado el concepto de "dinero B", que se refiere a los capitales generados por alguna actividad legal que no han sido declarados a la Hacienda Pública y que pretenden ser integrados en el sistema financiero para poder ser utilizados con normalidad. En este caso a pesar de cometerse un fraude contra la Hacienda Pública, y en su extensión un delito financiero, no se puede hablar de un delito de blanqueo de capitales ya que los rendimientos son generados a través de una actividad legal.

Sin embargo, cuando nos referimos al "dinero sucio" (también comúnmente llamado "dinero negro*"*) (Cordero, 2002), entendido como las cuantías monetarias que proceden de alguna actividad ilícita las cuales pretenden ser integradas en el sistema financiero con

[7] https://www.interpol.int/Crime-areas/Financial-crime/Money-laundering.

apariencia de dinero legal, este concepto sí atiende a capitales que proceden de actividades ilegales. Por tanto, es en este hecho delictivo cuando hablamos directamente de un delito por blanqueo de capitales. Y es el que incorpora los dos aspectos remarcados en la definición de blanqueo de capitales anterior, por un lado, la generación de riqueza procedente de actividades ilícitas y, por otro, la reintroducción de esos bienes en la economía legal.

El "dinero sucio" es riqueza derivada de actividades ilegales tan perjudiciales para el bienestar de la sociedad como la trata de personas, el tráfico de drogas, el fraude, la evasión fiscal, el soborno o la piratería entre otros. También está directamente relacionado con la financiación del terrorismo y con la financiación de armas de destrucción masiva (FATF, 2013).

A pesar de que los delincuentes cometen delitos por muchas razones, la principal motivación del crimen organizado es el beneficio económico (FATF, 2012). Existe una gran variedad de instrumentos que les permiten disfrutar de esa riqueza, por ejemplo recibir dinero en efectivo o en propiedades, y puede hacerse de forma personal o mediante empresas (López, 2015). En este contexto, la legislación aplicable al blanqueo de capitales tiene dos importantes objetivos. En primer lugar, evitar que los delincuentes lleven a cabo las actividades ilegales que generan las ganancias, es decir, los delitos antecedentes del blanqueo de capitales. En segundo lugar, impedir que los delincuentes puedan disfrutar de sus ganancias ilícitas (Alhosani, 2016).

Atendiendo a la necesidad de la existencia de un delito antecedente (actividades ilícitas) para que exista delito de blanqueo de capitales aparece una indexación entre el concepto de blanqueo de capitales y el delito financiero.

Mayoritariamente, en el proceso de integración de las cuantías monetarias ilícitas en el sistema legal, los delincuentes directa o indirectamente defraudan a la Hacienda Pública, e incluso intentan incrementar su patrimonio por medio de las devoluciones de impuestos (SEPBLAC, 2008). Por tanto, al tratar el concepto de blanqueo de capitales en un sentido amplio, se debe reflexionar sobre las repercusiones económicas y sociales que este delito genera.

Por otra parte, la estrecha relación entre el blanqueo de capitales y la financiación del terrorismo hace que este delito esté cobrando una gran importancia a nivel internacional. Cuando hablamos de blanqueo de capitales estamos también refiriéndonos al sistema por el que se financia el terrorismo (FATF, 2013). En efecto, los terroristas combinan actividades legítimas e ilícitas para financiar sus organizaciones, generando de este modo entramados empresariales muy complejos. Por lo que actualmente, cuando lo que se persigue es detener la financiación del terrorismo, la lucha y prevención del blanqueo de capitales es un reto mayúsculo a escala internacional (Demetis, 2011).

I.3.- Tipologías de blanqueo de capitales

En contra de lo que se suele pensar, el proceso de blanqueo de capitales es un fenómeno muy antiguo que fue documentado por primera vez por el historiador americano Sterling Seagrave en su libro *Lords of the Rim* (1995) (Seagrave, 1995).

En este libro, Sterling describe como hace más de tres mil años los comerciantes en la antigua China, por miedo a ser desterrados, ya "blanqueaban" su riqueza trasladando el dinero en efectivo fuera de su propia jurisdicción comprando a precios por encima de mercado o convirtiendo el dinero en bienes muebles. El concepto de blanqueo de capitales no es un proceso reciente y ha ocupado las mentes de los políticos y gobernantes durante siglos (Chong y López-de-Silanes, 2007).

Sin embargo, el concepto actual de delito de blanqueo de capitales ha sido empleado desde que el famoso gánster *Al Capone* usara su red de lavanderías para esconder los ingresos ilegales procedentes de la comercialización de alcohol en Estados Unidos durante la prohibición en los años treinta (Duyne, 2003; Unger, 2009).

Aunque el blanqueo suponga el último escalón para que los criminales puedan hacer uso de los ingresos de las actividades ilícitas, es una acción fundamental de toda organización criminal que cobra una relevancia importante en tanto que garantiza el disfrute de los ingresos ilícitos y elimina cualquier indicio sospechoso de su origen (Unger, 2007). Sin el beneficio económico que deriva del hecho de blanquear el dinero al introducirlo en el sistema económico/financiero, la mayoría de crímenes no tienen ningún sentido. Por ello, las autoridades ponen especial atención a la investigación de este proceso que conlleva el análisis de la trazabilidad del dinero que les permitirá identificar a los autores y sus fondos ilícitos (UNODC, 2011).

Entendiendo el proceso de blanqueo de capitales como el "Talón de Aquiles*"* de cualquier organización criminal (Bernasconi, 1995; Guiora y Feeld, 2007), la experiencia policial expone que combatir el blanqueo de capitales es una buena manera de terminar con las organizaciones criminales transnacionales, y así coordinar acciones para proteger la integridad del sistema financiero internacional (Unger, 2007; UNODC, 2011; FMI, 2014).

Debido a su relevancia, durante la última década ha habido un esfuerzo global de fortalecimiento en el desarrollo de la regulación contra el blanqueo de capitales (Unger y Hertog, 2012). En el *ANEXO I* se expone la abundante regulación y los esfuerzos aunados a nivel internacional existentes en la lucha contra estos delitos. Por un lado, el Grupo Intergubernamental de Acción Financiera Internacional (FATF), que cuenta con cuarenta recomendaciones contra el blanqueo de capitales y nueve recomendaciones contra la financiación del terrorismo (FATF, 2012) y, por otro, en las sucesivas directivas de la Unión Europea (91/308/CEE, 2005/60/CE, 2006/70/CE, 2015/849/CE). Ambas organizaciones establecen las normas internacionales que los países miembros de la UE deben transponer a sus legislaciones nacionales, y que marcan una guía de acción común para fortalecer los controles.

Si bien, el endurecimiento de la regulación, tanto a nivel nacional como internacional, persigue la disminución de estos delitos, y acotan las posibilidades de los delincuentes de enriquecerse (Becker, 1968; Chong y López-de-Silanes, 2007; Ferwerda, 2009), los procesos por los cuales se puede blanquear dinero son cada vez más sofisticados y prácticamente imperceptibles cuando hablamos de organizaciones de empresas grandes y/o internacionales, lo que complica notoriamente realizar la trazabilidad del dinero. Es por ello, que existen autores que defienden que en tanto las actividades ilegales sean más sofisticadas y más beneficiosas existirá un incremento de la cantidad de dinero blanqueado y ello derivará en realizarlo por procesos casi indetectables por los controles de las instituciones (Rahn, 2001).

El delito de blanqueo de capitales desencadena directa e indirectamente numerosos efectos negativos a la sociedad en su conjunto, desde la víctima del delito que genera el beneficio económico (delito antecedente de blanqueo de capitales) hasta su efecto en la economía mundial en el corto y en el largo plazo. Dichos efectos negativos impactan a nivel socioeconómico, a nivel espacial y a nivel temporal (Bartlett, 2002; Unger, 2007; UNODC, 2011). El blanqueo de capitales estimula la competencia desleal, las salidas de dinero ilegal, la corrupción política y policial, e incrementa el descontento social con las instituciones (incluidos los tribunales de justicia). Además, los efectos negativos, que se derivan de los sectores inmobiliario, financiero y público, no se mantienen restringidos a

escala nacional, se trasladan a través del sistema financiero internacional a otras economías (Unger, 2007).

A nivel internacional, los efectos negativos afectan a todos los niveles, a los sistemas económicos, a las instituciones financieras, a los organismos públicos y en las empresas (Unger, 2007). De hecho, la incidencia del blanqueo de capitales en un país está relacionada tanto con su grado de desarrollo tecnológico como con la efectividad de su marco legal y de su sistema fiscal (Chong y López-de-Silanes, 2007; Vaithilingam y Nair, 2009).

El proceso por el cual se lleva a cabo el blanqueo de capitales es un proceso complicado, por lo que se requiere de metodología altamente sofisticada para detectar los patrones de blanqueo y reconocer la trazabilidad de los capitales. El Grupo de Acción Financiera[8] resume este proceso delictivo en tres etapas o fases secuenciales: colocación, encubrimiento e integración. Etapas que en numerosas ocasiones se solapan.

1. *Fase de colocación*. En esta primera etapa se trata de introducir en el sistema el dinero procedente de actividades delictivas y hacer desaparecer el rastro original para que se desvincule de su origen. Se realiza eludiendo las obligaciones de información e identificación, generalmente mediante el fraccionamiento de los importes y su colocación e inversión a través de depósitos bancarios múltiples, oficinas de cambio de moneda, etc. Es la fase más próxima a la actividad delictiva y normalmente se lleva a cabo cerca del lugar en donde se cometen las primeras acciones delictivas.

2. *Fase de encubrimiento, estratificación o diversificación*. En esta fase se intenta desvincular los ingresos procedentes de la actividad delictiva de su origen, mediante la utilización de diversas operaciones financieras o no financieras. Consiste en el fraccionamiento, acumulación, ocultación, traslado de los importes hacia países con legislaciones menos rigurosas o a cuentas donde el dinero pueda tener una apariencia legal.

[8] http://www.fatf-gafi.org/faq/moneylaundering/.

3. *Fase de integración*. Es la fase final consistente en la integración al sistema financiero de los importes fraudulentos con el objetivo de aparentar dinero procedente de actividades legales.

Generalmente, la banca minorista es utilizada en la fase de colocación y en la fase final de integración, mientras que la banca de corresponsales es utilizada fundamentalmente en la fase de estratificación. Es muy importante la actuación y detección de las actividades fraudulentas y operaciones sospechosas en un primer momento, las posibilidades de éxito serán mayores cuanto antes se identifiquen (FATF, 2003; SEPBLAC, 2008).

Dada la heterogeneidad de los procesos de blanqueo, debido a las peculiaridades de cada organización criminal por su tamaño, localización geográfica y frecuencia de sus ingresos entre otros, no se puede detallar un listado concreto de tipos de blanqueo, pero sí hay una serie de escenarios en los cuales se ha identificado un alto grado de blanqueo, entre los que destacan los siguientes (SEPBLAC, 2008):

1. Operaciones en el sector inmobiliario, por sus características están localizadas en prácticamente todos los territorios dado el marcado carácter subjetivo que tiene la valoración de los bienes inmuebles y cobra especial interés en los países en desarrollo. Además, la titularidad de bienes inmuebles admite muchas figuras jurídicas distintas, tanto de carácter nacional como internacional, incluidas las formas de copropiedad temporal o espacial. Este sector está muy relacionado con actividades de corrupción.

2. Organizaciones empresariales con sistemas de compensación, que han desarrollado una tupida red bancaria que cubre la práctica totalidad del mundo y que produce una total opacidad de la parte comercial que justifica estas operaciones compensatorias.

3. La utilización de dinero en efectivo como medio de pago, lo que está experimentando incrementos anuales significativos, implicando toda la gama de medios disponibles, que incluyen desde los más clásicos procedimientos como "hawala"[9] hasta

[9] El método "hawala" es uno de los métodos más conocidos que existen en lo que se denomina sistemas de transferencia alternativo e informal de fondos y consiste en la transferencia informal de capitales sin pasar por instituciones financieras, por ejemplo, a través de negocios familiares.

los más sofisticados y modernos montajes (utilización de transportes específicos aéreos, marítimos y terrestres).

Por otro lado, las sociedades gestoras de transferencias, al contrario de lo que ocurre con las entidades financieras, suelen actuar a través de agentes que adquieren unos elevados niveles de autonomía en cuanto a la capacidad de adulterar la información que transmiten a su casa matriz, lo que dificulta detectar las operaciones más sofisticadas. Por esa razón, el sector de envío de fondos a través de circuitos no bancarios es siempre uno de los canales especialmente sensibles al blanqueo de capitales y a la financiación del terrorismo.

4. El empleo de "carruseles[10] de Impuesto de Valor Añadido (I.V.A.)", con la consecuente pérdida de los ingresos que debería producir la venta y consumo de los bienes de las empresas afectadas. Esta técnica genera cantidades de dinero negro obtenido a partir de la comisión de delitos de naturaleza fiscal, que debe ser blanqueado e introducido nuevamente en los circuitos formales mediante complejas operaciones de blanqueo de capitales.

5. La banca corresponsal que ha hecho posible la universalización de las transacciones financieras con origen o destino en entidades bancarias de diferentes países, a su vez permitiendo el desarrollo de una tupida y opaca red de nodos que permiten que los fondos discurran con rapidez y seguridad, cualesquiera que sean los países de origen y destino. Esto ha derivado en una escasez de controles que incrementa el acceso a las transacciones de dinero proveniente de actividades ilícitas.

A pesar de que el resto de agentes que operan en los sistemas financieros busquen reducir la complejidad del sistema para obtener una mayor eficiencia, en las organizaciones donde aparece el blanqueo de capitales se observa un movimiento sistémico en la dirección opuesta (Demetis, 2011). Este hecho incrementa la complejidad de las interacciones en los niveles institucionales, regulatorios y tecnológicos con el objetivo

[10] Defraudación a la Hacienda Pública por medio del régimen de tributación por IVA de las operaciones intracomunitarias. Véase el Apartado 1.2 de la PARTE 2.

final de aumentar los problemas de coordinación, comunicación y prevención de este delito.

I.4.- Mecanismos de lucha contra el blanqueo de capitales

La lucha contra el blanqueo de capitales y la financiación del terrorismo se ha reforzado notablemente en los últimos veinte años, muy especialmente después del 11 de septiembre de 2001, cuando ocurrieron los ataques terroristas contra las Torres Gemelas en Estados Unidos. Con ello, el blanqueo de capitales es un tema que está presente en las agendas de seguridad de las organizaciones internacionales, como el Grupo de Acción Financiera Internacional (FATF), el Fondo Monetario Internacional, las Naciones Unidas o la Unión Europea (Unger, 2009). Esto ha llevado a la comunidad internacional a establecer un régimen mundial coordinado de lucha contra el blanqueo de capitales y la financiación del terrorismo (FATF, 2012).

A nivel institucional, se refleja principalmente en la constitución, por parte del *G8*[11] en 1989, del Grupo de Acción Financiera Internacional (FATF) para desarrollar políticas de lucha contra el blanqueo de capitales. Desde 2001 el objetivo se expandió para actuar también sobre la financiación del terrorismo.

Como ya se ha indicado en la introducción, el FATF ha desarrollado 40 recomendaciones para combatir el blanqueo de capitales, que requiere la transposición a los países de las siguientes directrices: (1) implementar los convenios internacionales pertinentes, (2) criminalizar el blanqueo de capitales y permitir a las autoridades confiscar los rendimientos, (3) implementar la debida diligencia del cliente (*Custom Due Diligence*), el registro y la notificación de transacciones sospechosas (STRs) por parte de las instituciones financieras, empresas, profesionales y otras instituciones no financieras designadas, (4) establecer una unidad de inteligencia financiera nacional (FIU) para recibir y enviar informes de transacciones sospechosas (STRs) y (5) cooperar internacionalmente en su investigación.

Además, la FATF especifica otras nueve recomendaciones contra la financiación del terrorismo que, combinadas con las 40 recomendaciones relativas al blanqueo de

[11] G7 desde 2007 debido a la suspensión de Rusia.

capitales, establecen el marco básico para la detección, prevención y eliminación de la financiación del terrorismo y los actos terroristas (FATF, 2012).

En base a las recomendaciones de la FATF, la Unión Europea ha introducido una serie de Directivas (91/308/CEE, 2005/60/CE, 2006/70/CE y la directiva 2015/849/CE), difundiendo así las recomendaciones en sus estados miembros. Por otro lado, las Unidades de Inteligencia Financiera (FIUs) constituyen un componente importante de estas estrategias. Siguiendo la definición de Forget y Hočevar (2004, pág. 1), una Unidad de Inteligencia Financiera "(…) *es una agencia central nacional encargada de recibir, analizar y transmitir las transacciones sospechosas a las autoridades competentes*".

Hoy 152 países ya han sido admitidos en el Grupo Egmont, una institución de carácter internacional creada en 1878 que coordina las unidades de inteligencia financiera de todo el mundo. Los miembros que son admitidos en el Grupo Egmont-Unidades de Inteligencia Financiera comparten las mismas funciones básicas de recibir, analizar y difundir información financiera relativa al blanqueo de capitales y la financiación del terrorismo.

Asimismo en Europa se ha creado una red informática descentralizada (FIU.net) para apoyar y coordinar a las FIUs de la Unión Europea en su lucha contra el blanqueo de capitales y la financiación del terrorismo. Actualmente, los 28 países miembros de la Unión Eurpea son colaboradores de esta organización.

A pesar de ello, la heterogeneidad entre las estructuras de los sistemas de lucha contra el blanqueo de capitales que cada país posee es muy visible (Demetis, 2011; Forget y Hočevar, 2004). En efecto, existen múltiples mecanismos para aunar los esfuerzos para combatir este delito. A pesar de la incorporación de la serie de iniciativas legislativas, la heterogeneidad en las estructuras AML/CFT en los distintos estados todavía demuestra una falta de estandarización real que, solo si es corregida, incrementaría la eficacia de las estructuras de lucha contra estos graves delitos (Forget y Hočevar, 2004).

La estandarización es importante para coordinar al número creciente de agentes interesados dentro de la lucha internacional y es fundamental para establecer confianza y compartir información de forma eficiente. Sin embargo, existen países que han

establecido acuerdos bilaterales (un proceso que consume mucho tiempo y que ha tardado más de dos décadas en desarrollarse) en lugar de trabajar para coordinar estándares únicos en su conjunto (Demetis, 2011; Alhosani, 2016). Dado que estas relaciones implican a muchos países, se está introduciendo un número cada vez mayor de entidades informantes al cumplimiento del régimen AML (por ejemplo, compañías de seguros, casinos, oficinas de cambio, abogados, economistas, etc.), por lo que la coordinación de todo el proceso supone un reto mayúsculo.

Además, la relación del blanqueo de capitales con la financiación del terrorismo añade gran complejidad al tema, sobretodo en términos de prevención y control (Alhosani, 2016). Así, Demetis (2011, pág. 65) destaca que:

"(…) *la diversidad y la naturaleza multifacética de la actividad financiera de los terroristas lo convierten en un desafío. Los terroristas utilizan métodos legítimos e ilegales para financiar sus actividades organizativas y operacionales, por lo que la información financiera por sí sola puede no ser suficiente para identificar la actividad de financiación del terrorismo"*.

Por ello, existe la necesidad de aplicación de las técnicas más vanguardistas para tratar de combatirlo.

La detección de la participación de terroristas en actividades financieras legítimas requiere que las instituciones financieras apliquen las normas del FATF mediante una aplicación sólida del principio "conozca a su cliente" (KYC[12]), y de las políticas y procedimientos de "debida diligencia del cliente" (CDD[13]) (FATF, 2008).

También es fundamental la notificación de transacciones sospechosas (SAR) que permitan identificar operaciones dudosas relacionadas con terrorismo, y la aplicación de programas específicos de sanciones financieras, así como la protección de sectores vulnerables, incluido el sector caritativo y las empresas de servicios monetarios; particularmente las empresas de servicios de moneda virtual (Bitcoin) (D'Souza, 2011; FATF, 2012)

[12] *Know Your Customer*.
[13] *Custom Due Diligence.*

A nivel internacional, se está incrementando el esfuerzo para evaluar la experiencia operacional de los organismos de lucha contra el terrorismo, las llamadas Unidades Financieras de Inteligencia (FIU), y establecer "alertas" e "indicadores" que notifiquen cuando alguna transacción sospechosa de blanqueo de capitales presenta particularmente algún riesgo de financiación del terrorismo (FATF, 2008).

Concretamente en España, la Unidad Financiera de Inteligencia es la Comisión de Blanqueo de Capitales e Infracciones Monetarias (SEPBLAC), dependiente del Ministerio de Economía, y que es el máximo responsable del desarrollo de la política preventiva y de lucha contra el blanqueo de capitales. En España la lucha contra el blanqueo de capitales se fundamenta en dos pilares: la aplicación de ley y la prevención.

Desde 1995, el Código Penal español se ha ido reformando progresivamente en las siguientes Leyes Orgánicas: LO 10/1995, LO 15/2003 y LO 5/2010 (Carpio Delgado, 2011). Estas leyes han ido ampliando la definición de delitos antecedentes de blanqueo de capitales, mediante los cuales se genera el capital ilícito, como delitos penados con casi cinco años de prisión (1996-2004) a cualquier delito incluido en el Código Penal (2005-2010) y a cualquier actividad sospechosa de ser criminal (desde 2011). A pesar de los cambios, las penas se han mantenido constantes con el encarcelamiento de entre seis meses y seis años. Finalmente, con la aprobación en mayo de 2014 del Real Decreto 304/2014, quedó aprobado en España el Reglamento de la Ley 10/2010, de 28 de abril, de prevención del blanqueo de capitales y de la financiación del terrorismo.

Desde la perspectiva de la prevención, esta se ha desarrollado a través de leyes administrativas, las cuales han sido continuamente modificadas con el objetivo de limitar las cuantías máximas de los pagos en efectivo y de identificar a los propietarios reales de los capitales y de propiedades inmobiliarias (Cordero, 2009). El desarrollo de medidas de cumplimiento ha incrementado la obligación de las instituciones financieras y profesionales de informar a las autoridades de actividades sospechosas.

Además, en la práctica totalidad de sectores profesionales, especialmente en el sector bancario e inmobiliario, las empresas están obligadas a establecer mecanismos de control interno con el objeto de "(…) *conocer, prevenir e impedir la realización de operaciones*

sospechosas de blanqueo de capitales" [14], estableciéndose protocolos para ello, y existiendo sanciones para quienes no lo apliquen. Algunos de estos protocolos son: la inscripción de la empresa en el registro del SEPBLAC, establecer procedimientos necesarios para el control interno (manuales y procedimientos para efectuar dicho control), la realización de cursos formativos sobre la normativa de prevención (obligatorios para todos los trabajadores de los sujetos obligados que exige la Ley) o la realización de auditorías periódicas por parte de un consultor externo experto.

Algunos organismos internacionales, como la FATF, han alabado las mejoras puestas en marcha por las autoridades españolas. La Unidad de Inteligencia Financiera Española (SEPBLAC) ha incrementado desde 2012 en un 55% el número de investigaciones por indicio de delitos detectados a través de entidades financieras y profesionales liberales, un tercio de todos ellos en Madrid, donde se concentran los cuerpos estatales que coordinan la lucha contra el blanqueo (SEPBLAC, 2013).

Como se observa en el Gráfico I.1, los esfuerzos de las autoridades españolas por combatir este grave delito se han ido intensificando muy notablemente durante las dos últimas décadas. Tanto es así que actualmente prácticamente todos los intervinientes en una transacción financiera están obligados a informar si sospechan del origen del capital o de cualquier variable que interceda. Esto es, se ha extendido la obligatoriedad desde las instituciones financieras a otras instituciones y profesionales como aseguradoras, abogados, inmobiliarias e incluso se está prestando una atención especial a los denominados "mineros informáticos"[15] que albergan las transacciones realizadas mediante dinero virtual (*BlockChain*). Así, la primera línea (Figura C) del Gráfico I.1 muestra el crecimiento, entre 1997 y 2012, del número de instituciones obligadas a informar de operaciones sospechosas de blanqueo de capitales.

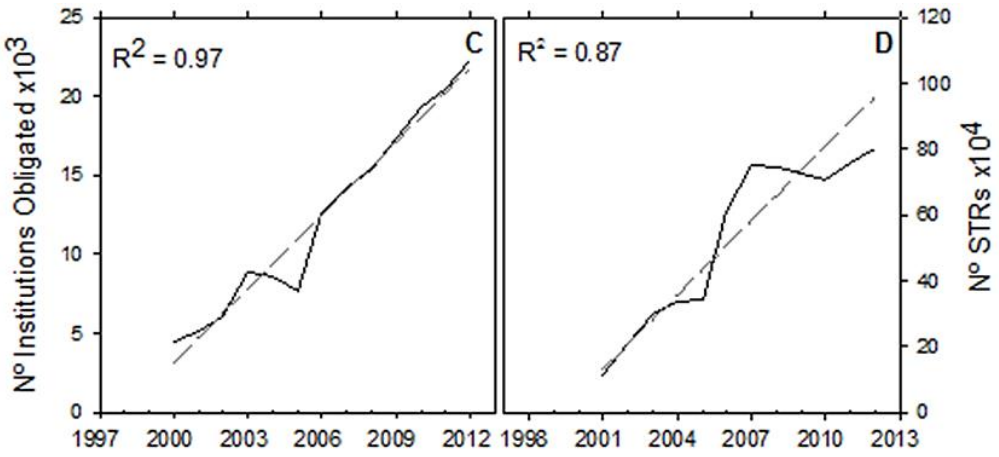
Además, si centramos la atención en el instrumento que se encuentra en el núcleo de cualquier sistema nacional de lucha contra el blanqueo de capitales o financiación del terrorismo: el Informe de Transacciones Sospechosas (STRs) o el Informe de Actividades

[14]http://www.sepblac.es/espanol/informes_y_publicaciones/documento%20_recomendaciones_sobre_medidas%20_control_interno_PBCFT.pdf

[15] Los "mineros informáticos", en el contexto de la moneda virtual, son gestores de información encargados de generar los bloques de cada cadena de transacciones en moneda virtual. Cada bloque es un paquete que contiene el *hash* de todas las transacciones realizadas durante un mismo periodo de tiempo, generalmente, unos 10 minutos.

Sospechosas (SARs) (Figura D), comprobamos que el número de transacciones sospechosas informadas ha crecido notoriamente durante las últimas décadas, llegando a alcanzar las 80 transacciones por cada 100.000 habitantes en España.

Gráfico I.1
Evolución anual del número de instituciones españolas obligadas a informar y número de operaciones sospechosas informadas en España.



Fuente: Elaboración Propia. Notas: (1) La línea discontinua representa el ajuste a un modelo lineal con su varianza explicada (R2), (2) Artículo completo en el *ANEXO II*. Publicaciones. "*Money laundering trend in Spain: offences and arrests over 15 years*".

En función del sector de actividad el SEPBLAC establece ciertos indicadores que orientan a los sujetos obligados a detectar las operaciones/actividades sospechosas (STRs/SARs). Por ejemplo, la detección de clientes/proveedores de bienes o servicios que muestran una especial preferencia porque se realicen los cobros/pagos empleando efectivo, eludiendo otros medios de pago más habituales en el tráfico comercial habitual, o el empleo de intervinientes en operaciones comerciales que operan bajo una determinada sociedad o denominación que posteriormente, y sin una clara explicación lícita, siguen actuando con las mismas características pero a través de otras personas o sociedades[16].

Por otro lado, los recursos destinados a la eliminación del fraude financiero y el blanqueo de capitales en España no sólo se concentran en la Unidad de Inteligencia Financiera, sino

[16] http://www.sepblac.es/espanol/informes_y_publicaciones/otra_documentacion.htm.

que también se extiende a otras instituciones públicas, como el incremento de personal destinado a investigar casos de blanqueo de capitales de la Agencia Tributaria o de la Guardia Civil.

I.5.- Tendencia del blanqueo de capitales en España[17]

Actualmente, no existe una metodología universalmente aceptada para estimar la cuantía de dinero de procedencia ilícita que se integra en los sistemas financieros de todo el mundo (Barone y Masciandaro, 2011; Unger, 2007). La falta de datos empíricos y de modelos para cuantificar este delito sólo permite detectar "la sombra" del dinero realmente blanqueado (Levi y Reuter, 2006; Unger y Hertog, 2012; Unger, 2013).

Un consenso de diferentes "*guesstimates*[18]" lo sitúa en el rango entre el 2% y el 5% del Producto Interior Bruto Global (Camdessus, 1998; Unger, 2007; Unger, 2013; UNODC, 2011). Esto representaría aproximadamente 2 billones de dólares americanos por año (Camdessus, 1998; UNODC, 2011; Walker y Unger, 2009). Desafortunadamente, los informes han cuantificado que sólo es interceptado el 1% del total de dinero de procedencia ilícita cuando es blanqueado a través del sistema financiero (UNODC, 2011).

En España, las estimaciones del dinero blanqueado difieren considerablemente y oscilan entre los 36 millones de euros, estimación de 2011 (Fernández, 2012), y los 56 mil millones de dólares americanos por año calculado a partir de un modelo teórico, estimación de 1997 (Walker, 1999). Según el estudio de Unger (2007), España se encontraba en el puesto 51 de 288 en la escala de países que más dinero ilegal habría integrado en su sistema económico, situando a Luxemburgo como el país que pudo blanquear más dinero.

Siguiendo esta línea de investigación, y dada la escasez de información relativa a la cuantía de dinero que podría estar efectivamente blanqueándose en España, se ha llevado a cabo una investigación, junto con dos agentes de la Policía Judicial (Grupo de Blanqueo de Capitales y Grupo de Delincuencia Económica), para analizar la tendencia del banqueo de capitales en España, entre los años 1998 y 2012, y la eficiencia de las instituciones

[17]Artículo completo en el *ANEXO II*.-Publicaciones. "*Money laundering trend in Spain: offences and arrests over 15 years"*.

[18]Se conoce como "guesstimate" a una estimación hecha sin utilizar información adecuada o completa, e incluso como una estimación obtenida por conjeturas (Merriam-Webster, On-line Dictionary).

españolas para combatirlo. Los siguientes sub-apartados de este Capítulo detallan los principales resultados obtenidos.

I.5.1.- Descripción de la muestra

Los datos de delitos de blanqueo y número de arrestados en casos de blanqueo de capitales para el período 1998-2012 provienen de los informes estadísticos anuales publicados cada año por el Ministerio del Interior del Gobierno Español, que pueden descargarse gratuitamente desde su página web institucional: http://www.interior.gob.es/. Los datos incluyen registros de las fuerzas de seguridad españolas, compuestas principalmente por la Policía Nacional y la Guardia Civil.

Los conjuntos de datos incluyen también las estadísticas recopiladas por otras tres fuerzas policiales creadas en los últimos veinte años: Mossos d'Esquadra en Cataluña, Ertzaintza en el País Basco y Policía Foral en Navarra, que están reemplazando progresivamente a la Policía Nacional y a la Guardia Civil en sus respectivos territorios. Aunque los registros de las fuerzas policiales regionales todavía son escasos en comparación con los proporcionados por la Policía Nacional y la Guardia Civil, se han tenido en cuenta cuando están disponibles. Los datos de arrestos en casos de blanqueo de capitales de 2011 y 2012 fueron excluidos del análisis porque en esos años los informes sólo presentan a los acusados por las fuerzas de seguridad, sin detallar a los detenidos.

Los delitos antecedentes se clasifican en el Código Penal español como delitos contra el patrimonio y la administración pública, el delito de corrupción y el tráfico de drogas.

Las estadísticas de los informes españoles sobre delitos contra el patrimonio incluyen: robo, fraude, falsificación y daños. Sin embargo, los delitos de daños fueron finalmente excluidos porque no necesariamente todos ellos conducirían a un delito de blanqueo de capitales. Los delitos de corrupción incluyen: soborno, tráfico de influencias, malversación de fondos públicos, desobediencia al derecho penal y prevaricación.

Se definen la tasa de delitos por blanqueo de capitales, la tasa de arrestos por blanqueo de capitales, la tasa de delitos antecedentes y la tasa de arrestados en delitos antecedentes

del blanqueo de capitales, todas ellas en función de datos relativos a la población. Los datos demográficos han sido obtenidos del Instituto Nacional de Estadística (http://www.ine.es/).

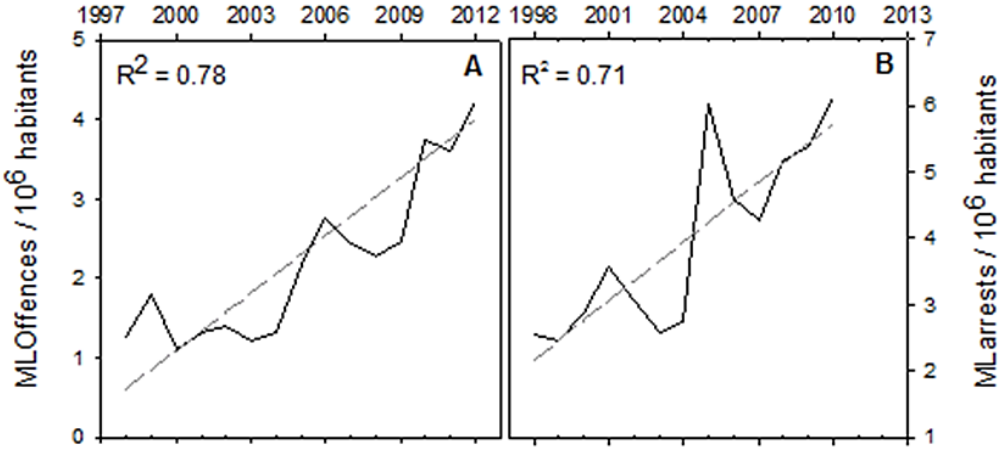
I.5.2 Principales resultados obtenidos[19]

Los resultados presentados en este apartado se centran en la evolución de las tasas de delitos por blanqueo de capitales y de la tasa de arrestos por blanqueo de capitales, e incluye análisis de la evolución del número de delitos antecedentes y del número de arrestados en dichos delitos. Adicionalmente (*ANEXO II*), se abordan análisis complementarios respecto de la eficacia de las estrategias seguidas por las autoridades españolas en materia de actuación y prevención contra el blanqueo de capitales desde la perspectiva de la criminología.

El Gráfico I.2 muestra la evolución de la tasa del número de delitos de blanqueo de capitales (Figura A) y del número de arrestados por delito de blanqueo de capitales (Figura B); ambas tasas en función de datos relativos a la población.

[19] Para profundizar en los métodos estadísticos utilizados y conocer resultados adicionales véase *ANEXO II*.

Gráfico I.2
Evolución del número de delitos de blanqueo de capitales y del número de arrestos en casos de blanqueo de capitales en España.



Fuente: Elaboración Propia. Notas: (1) La línea discontinua representa el ajuste a un modelo lineal con su varianza explicada (R2), (2) Artículo completo en el *ANEXO II*.-Publicaciones. "*Money laundering trend in Spain: offences and arrests over 15 year*".

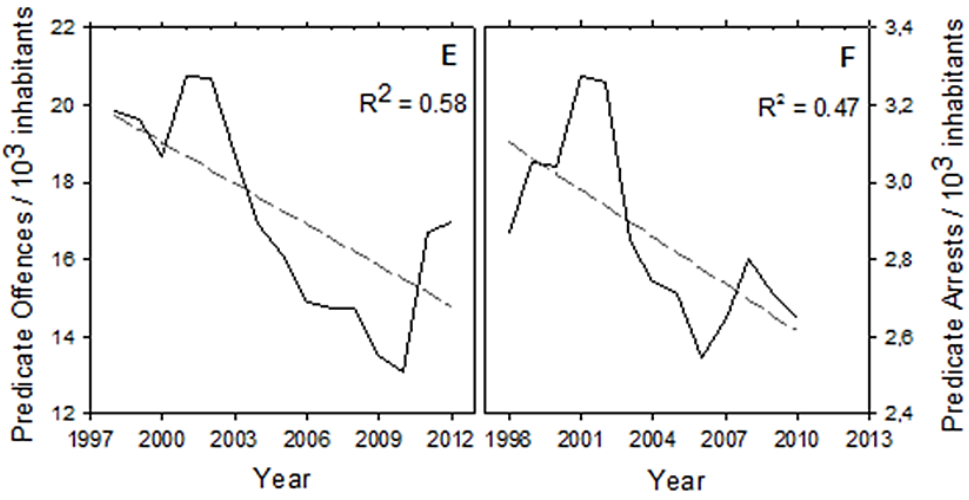
Entre las conclusiones más llamativas del estudio, destaca el gran aumento de las tasas de delincuencia por delito de blanqueo de capitales durante los últimos 15 años en España. Para el período 1998-2012, la tasa que recoge el número de delitos de blanqueo de capitales creció casi un 244%, aumentando cada año de media 0,2 unidades por millón de habitantes, mientras que la tasa del número de arrestados por delito de blanqueo de capitales aumentó un 431%, lo que supone un incremento anual de 0,3 unidades por millón de habitantes.

El Gráfico I.2 reforzaría las hipótesis que establecen que el endurecimiento de las políticas de prevención AML/CFT habría motivado la detección de un número cada vez mayor de casos de blanqueo de capitales (Schneider, 2005; Zdanowicz, 2009; Ferwerda *et al.*, 2013; Unger, 2013).

Dichas políticas se han llevado a cabo a través del desarrollo de un marco de lucha contra el blanqueo que ha modificado las normas administrativas para reducir el flujo de efectivo en el sistema económico, y que ha mejorado la identificación de los verdaderos dueños de los capitales. Uno de sus efectos directos más destacables es el incremento del número de operaciones sospechosas informadas al SEPBLAC (Gráfico I.1, pág. 29).

Por su parte, y en contra de la acentuada tendencia creciente que marca la evolución del delito de blanqueo de capitales (Gráfico I.2), la tasa de delitos antecedentes refleja un decrecimiento de la línea de tendencia anual (Gráfico I.3).

Gráfico I.3
Evolución anual del número de delitos subyacentes de blanqueo de capitales y del número de arrestos por blanqueo de capitales en España.



Fuente: Elaboración Propia. Notas: (1) La línea discontinua representa el ajuste a un modelo lineal con su varianza explicada (R2), (2) Artículo completo en el ANEXO II.-Publicaciones. "*Money laundering trend in Spain: offences and arrests over 15 years*".

El número de delitos antecedentes o subyacentes disminuyó casi un 30% entre 1997 y 2012. Cada año descendieron, de media, en 436,7 unidades por millón de habitantes, y la tasa que representa el número de personas detenidas en delitos antecedentes o subyacentes también disminuyó un 16% en el periodo analizado, registrando una minoración de 40,8 unidades por millón de habitantes. Esta tendencia decreciente podría indicar efectos positivos del régimen AML/CFT aplicado en España. Los controles existentes en los sistemas financieros estarían dificultando la integración de los capitales ilícitos y habrían provocado un descenso de la rentabilidad final obtenida por los delincuentes.

Desde una perspectiva global, los resultados de los Gráficos I.2 y I.3 sugieren que el régimen AML/CFT español estaría alcanzando resultados positivos a medio plazo. Las actuaciones en materia de regulación y prevención habrían precipitado un aumento de la detección del número de casos de blanqueo de capitales, y consecuentemente del número

de arrestados. Asimismo, las estrictas políticas implementadas estarían disuadiendo la delincuencia profesional.

Sin embargo, a pesar de que los datos apuntan a que el régimen AML/CFT aplicado en España, bajo el mandato de la Unión Europea durante más de 20 años (91/308/CEE, 2005/60/CE, 2006/70/CE, 2015/849/CE), estaría siendo eficiente en la lucha contra el blanqueo de capitales, hay ciertos factores que podrían estar enmascarando las causas reales de los cambios observados.

Por un lado, la disminución de las tasas de delitos antecedentes se ha observado también en otros países de Europa, América del Norte y Australia. Teniendo en cuenta la heterogeneidad de las políticas AML/CFT implementadas en cada país desde los años 90, y los diferentes plazos de vigencia de la normas, esta tendencia decreciente podría responder a una interacción compleja de factores de control por parte de las autoridades.

Así, los avances tecnológicos han promovido medidas de seguridad para prevenir los delitos clave, los cuales han precipitado una disminución del número de delitos antecedentes como respuesta a un aumento de la seguridad más que a las medidas de lucha contra el blanqueo de capitales (Farrell *et al.*, 2014).

A su vez, Europa ha centrado su estrategia de lucha AML/CFT en la regla "sigue el dinero" (*follow the money*), que persigue conocer la trazabilidad de beneficios obtenidos en actividades ilícitas e identificar a los verdaderos dueños de los capitales. Pero debido a que los delincuentes combinan sus actividades entre los mercados ilícitos y la economía legal, y que existen países regulados por estrictas normas y decretos que impiden la transferencia de información (paraísos fiscales), se habría generado un escenario que favorece la "ocultación" de dinero de procedencia ilícita. Lo que dificulta seriamente la identificación de la trazabilidad de los capitales ocultos y limita la eficiencia del régimen (Levi, 2015; Verhage, 2009).

La necesidad de cartas rogatorias o solicitudes de información entre países, y de colaboraciones internacionales, que se dilatan en el tiempo, disuadiría a las autoridades de continuar la investigación para perseguir el "dinero sucio". Actualmente, las barreras AML/CFT pueden ser superadas por los delincuentes con la creación de complejas

organizaciones empresariales, donde las cuentas bancarias se controlan a través de Internet.

Con ello, las tasas de delincuencia de delitos antecedentes no podrían verse directamente afectadas por el actual régimen AML/CFT, ya que los delincuentes estarían cometiendo los delitos en distintos países y blanqueando sus beneficios en diversas localizaciones (Varese, 2011).

A este respecto, una subestimación del censo de "blanqueadores" y de la capacidad de las autoridades encargadas de hacer cumplir la ley también podría estar limitando la eficiencia del régimen. Desafortunadamente, no hay datos realistas sobre el personal especializado dedicado a combatir el blanqueo en España, a fin de estimar la capacidad reactiva del régimen.

Por otro lado, teniendo en cuenta que los esfuerzos contra el blanqueo de capitales han venido acompañados de un aumento en la complejidad de los procedimientos de blanqueo, es decir, los "blanqueadores" son cada vez más especializados y hay más delincuentes involucrados en el proceso (Unger, 2007; Fernández, 2012). Daría la sensación que las estrictas regulaciones implementadas han promovido una carrera entre los delincuentes para superar las dificultades legales a través de la diversidad y la innovación de metodologías (Unger y Hertog, 2012).

Las nuevas tecnologías como Internet, las transferencias electrónicas y los servicios de pago mediante plataforma *on-line*, son excelentes alternativas de "*cyber laundering methodologies*" que todavía no están reconocidas en el régimen AML/CFT (Souto, 2013). Ciertos efectos indirectos del marco AML/CFT también habrían favorecido la sofisticación de los procesos de blanqueo. Por ejemplo, un aumento en la regulación del sector bancario obliga a los delincuentes a buscar nuevas formas de inversión menos controladas (por ejemplo, monedas virtuales) o a trasladar al exterior del sector financiero los capitales ilegales mediante la adopción de formas alternativas y modernas de blanqueo (Woda, 2006; Unger y Hertog, 2012). Ello tiene un alcance que va más allá de la regulación de un país en concreto.

La migración a nuevos sectores económicos explicaría por qué los resultados de algunos estudios económicos establecen que el blanqueo de capitales únicamente se vería

fuertemente reducido por un mayor esfuerzo global en el desarrollo de la regulación AML/CFT y una mejor divulgación en los países (Chong y López-De-Silanes, 2007; Chong y López-De-Silanes, 2015; Ferwerda, 2009).

En este contexto, a corto plazo se esperaría un aumento de la tasa del número de casos de blanqueo (Rahn, 2001). Sin embargo, a largo plazo se podría prever una disminución del número de delitos de blanqueo de capitales; ya sea porque la regulación ALM/CFT tiene efecto o porque los delincuentes han mejorado sus procedimientos y no son identificados por las autoridades.

I.6.- Contabilidad forense y aprendizaje automático

Como ya se ha señalado, los investigadores encargados de detectar las operaciones de fraude financiero y de blanqueo de capitales se enfrentan a un complejo cometido, caracterizado por el desarrollo global de la economía, los avances tecnológicos y los novedosos sistemas de negocios *on-line*, especialmente por los bancos *on-line* (e-banking) y monedas virtuales (Bitcoin) (FIU- The Netherlands, 2015). Dada la importancia de estas actividades para los delincuentes, las ricas y solventes organizaciones criminales cuentan con expertos profesionales para que les ayuden a definir los sofisticados procesos de blanqueo que les permitan ocultar la procedencia ilícita de sus ingresos. Además, también es de suponer que estos cuenten con los mejores analistas, contables e ingenieros para que les ayuden a crear registros contables que "maquillen" la evidencia de la procedencia de los capitales (ingeniería financiera e ingeniería numérica) (Malm y Bichler, 2013; Soudijn, 2012 y 2014).

El proceso de investigación está bajo la premisa de que los delincuentes habrían creado un entramado de operaciones complejo y sofisticado a fin de ocultar las operaciones fraudulentas y moldear los datos para aparentar actividades legales.

En efecto, en el proceso de investigación los investigadores obtienen ingentes cantidades de información contable y financiera, en gran medida proveniente de la contabilidad interna de las empresas, pero también, información bancaria de las operaciones, de la información aduanera e incluso de registros internos de entrada de mercancía de un país, además de cualquier otra información relevante relativa a las actividades de las personas físicas y jurídicas investigadas (Sremack, 2015).

El tratamiento de estas ingentes cantidades de datos conlleva procesos de análisis de información de todo tipo, de gestión de datos, análisis descriptivos, análisis cuantitativos, de integración de información, de predicción e incluso de ingeniería numérica. Ello obliga a los investigadores a conocer las técnicas más vanguardistas de aprendizaje automático (*machine learning* en terminología inglesa). Estas técnicas complementan modelos estadísticos empleados para el análisis con el manejo de los softwares más avanzados a fin de agilizar el proceso de investigación y reducir el consumo de recursos (humano y

computacional). Actualmente resulta casi imposible realizar los procesos de análisis que requieren estos casos empleando únicamente programas informáticos tradicionales, como son programas del paquete de Microsoft Office u otros como Visual Basic Aplicado (Sremack, 2015).

El aprendizaje automático se puede definir como el estudio, diseño y desarrollo de los algoritmos que dotan a los sistemas informáticos de la capacidad de "aprender" sin ser programados de forma explícita (Samuel, 1967). Por tanto, el aprendizaje automático utiliza algoritmos para modelar y encontrar automáticamente patrones en los datos, generalmente con el objetivo de predecir algún resultado o respuesta. Estos algoritmos están basados en la estadística y en la optimización matemática. En los modelos de aprendizaje automático la información se va actualizando con los nuevos datos que se van generando, de tal forma que al convertir el proceso de modelización en un proceso continuo las predicciones se puedan adaptar a los cambios coyunturales o estructurales que se produzcan en el problema objeto de análisis (Sremack, 2015).

Las principales categorías del aprendizaje automático son el aprendizaje supervisado y el aprendizaje no supervisado, aunque también existen otras categorías como el aprendizaje semi-supervisado o por refuerzo, entre otros (Bishop, 2006). En el aprendizaje supervisado el algoritmo produce una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema. Un ejemplo de este tipo de algoritmo es el problema de clasificación, donde el sistema de aprendizaje trata de etiquetar (clasificar) una serie de casos utilizando una entre varias categorías (clases). Por otro lado, en el aprendizaje no supervisado todo el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formado tan sólo por entradas al sistema. No se tiene información sobre las categorías de esos ejemplos. Por lo tanto, en este caso, el sistema tiene que ser capaz de reconocer patrones para poder etiquetar las nuevas entradas.

La idiosincrasia de los casos de blanqueo de capitales requiere que los responsables del análisis de los datos tengan, por un lado, conocimientos avanzados de análisis de datos, de economía y contabilidad, y por otro, conocimientos del manejo de los softwares más avanzados en esta área, como son el programa STATA o los software libres R, Phyton o Weka. En esta Tesis Doctoral los análisis han sido realizados mediante los softwares libres R (versión 3.1.3) y Weka.

Dado que la mayor parte de la información procede de las organizaciones supuestamente delictivas, resulta de especial relevancia la función del experto contable, quien se conoce como contable forense. Según señalan Owojori y Asaolu (2009, pág.184) se conoce como contable forense, "(…) *al experto que se encarga de analizar e interpretar la compleja información contable y financiera a fin de ofrecer un informe detallado que ofrezca información sintetizada de la actividad económica de los investigados en los procesos judiciales y ayude a detectar indicios de fraude financiero*"

Concretamente, en los casos judiciales que este tipo de investigación aborda, el forense contable es el experto encargado de dar una visión preliminar de la información para comenzar a orientar las investigaciones y servir de guía y ayuda a las autoridades competentes para definir el proceso de investigación. Asimismo, es el responsable de realizar informes periciales para aportar información descriptiva del caso al proceso judicial (indicios de fraude) que complemente la investigación de las autoridades competentes. Se centra también en la selección de aquellos documentos e información que podrá ser presentada como indicio de fraude y selecciona aquellos datos que serán objeto de los análisis. En este sentido, cobra especial interés la selección de la información que será empleada para argumentar la tesis que se presentará ante el juez. Toda ella presentada en un leguaje simplificado y conciso, propio del lenguaje contable.

Además, el experto contable forense deberá estudiar la documentación ofrecida por la parte contraria a fin de poner en valor la tesis que se pretende defender y contraponerla a los indicios de fraude que ha identificado. En este sentido el forense contable es un asesor en materia de fraude financiero que se encarga también de ofrecer información a las autoridades para crear protocolos de prevención de delitos económicos (Owojori y Asaolu, 2009). Las exigencias del trabajo que desempeña el forense contable hacen que indudablemente sea necesario el empleo de técnicas vanguardistas para el análisis de datos y su rápida aplicación en las bases de datos disponibles (Yadav y Yadav, 2013).

A pesar de que las técnicas disponibles a emplear son casi ilimitadas y las aplicaciones muy heterogéneas, las aportaciones del forense contable en los procesos judiciales de blanqueo de capitales más destacables son: (1) la detección de operativas sospechosas, (2) el análisis de los posibles delitos financieros y de fraude y (3) el desarrollo de técnicas que ayuden a detectar comportamientos delictivos incipientes. Además, (4) permite a las

instituciones que luchan contra estos delitos combatirlo en todas las fases del blanqueo de capitales: ubicación, ocultamiento e integración de capitales provenientes de actividades ilegales[20].

Por tanto, en procesos judiciales que manejan ingentes cantidades de datos se utilizan procesos estandarizados de predicción para detectar, dentro de los complejos entramados empresariales, los individuos anómalos e indicios de actividades u operaciones sospechosas. Se destacan a los sospechosos potenciales para un posterior análisis individual y exhaustivo por parte de la autoridad policial (Dutta, 2013).

Fundamentándose en la información obtenida por los investigadores policiales en las investigaciones previas del caso judicial, y basándose en técnicas de aprendizaje automático, la contabilidad forense ofrece a las autoridades judiciales, nueva información sobre las posibles estrategias utilizadas por los criminales para ubicar, ocultar e integrar los capitales procedentes de actividades ilegales en los sistemas financieros.

El objetivo final del trabajo del forense contable y de la detección de patrones de delincuencia es "(…) *predecir la delincuencia, anticipar la actividad criminal y prevenirla"* (Wang *et al.* (2013), pág. 1).

La necesidad urgente de nuevas técnicas y herramientas que puedan extraer información útil y conocimiento de volúmenes masivos de datos, ha incentivado el crecimiento exponencial de las técnicas de aprendizaje automático en los últimos años. En este ámbito destaca la técnica conocida como Big Data.

Se denomina Big Data[21] al conjunto de técnicas vanguardistas de gestión de datos para avanzar las tendencias en la tecnología que abre la puerta a un nuevo enfoque para entender el mundo y tomar decisiones (Lohr, 2012). El Big Data se fundamenta en 7 dimensiones, denominadas "las 7 Vs del Big Data": Volumen, Variedad, Velocidad, Veracidad, Valor, Visibilidad y Visualización (Sremack, 2015).

[20]Esta afirmación se basa en la experiencia de trabajo de los casos de blanqueo de capitales en los que la autora ha colaborado como forense contable.

[21] Anglicismo ampliamente empleado en el castellano.

Las técnicas de Big Data se han aplicado a una amplia variedad de dominios, incluyendo el marketing, la evaluación de créditos y préstamos, la detección de fraudes financieros, la predicción de enfermedades o para desarrollar la visión artificial, también se ha aplicado muy exitosamente para el reconocimientos de patrones (Gao y Ye, 2007; García *et al.*, 2015; Hastie *et al.*, 2008; Khac y Kechadi, 2010; Bishop, 2006). Generalmente, los sistemas de reconocimiento de patrones son modelos estadísticos que siguen un modelo de aprendizaje supervisado. Aunque cuando no se dispone de datos clasificados se pueden utilizar otros algoritmos para descubrir patrones previamente desconocidos (aprendizaje no supervisado) (Bishop, 2006). El diseño de un sistema automático de detección de patrones esencialmente trata los siguientes tres aspectos: (1) acceso a la información y pre-procesamiento de datos (Capítulo II), (2) aplicación de la metodología y representación de los resultados (Capítulo III) y (3) la toma de decisiones (Capítulo IV).

Esta disciplina también se conoce como el Descubrimiento de Conocimiento en Bases de Datos (KDD-*Knowledge Discovery in Databases*), a razón de englobar el conjunto de técnicas y algoritmos disponibles que permiten detectar los patrones ocultos en las grandes bases de datos (Dutta, 2013).

Los términos Big Data, aprendizaje automático, reconocimiento de patrones y descubrimiento de conocimiento en bases de datos (KDD) son difíciles de separar, ya que en gran parte se superponen en su alcance. El aprendizaje automático es el término común para los métodos de aprendizaje supervisado y se origina en la inteligencia artificial, mientras que las técnicas KDD y la minería de datos tienen un mayor enfoque en métodos no supervisados (Bishop, 2006). En su conjunto estas técnicas también son denominadas sistemas expertos, en tanto que emplean sistemas informáticos que tratan de emular la capacidad de toma de decisiones humanas (inteligencia artificial) (Jackson, 1998).

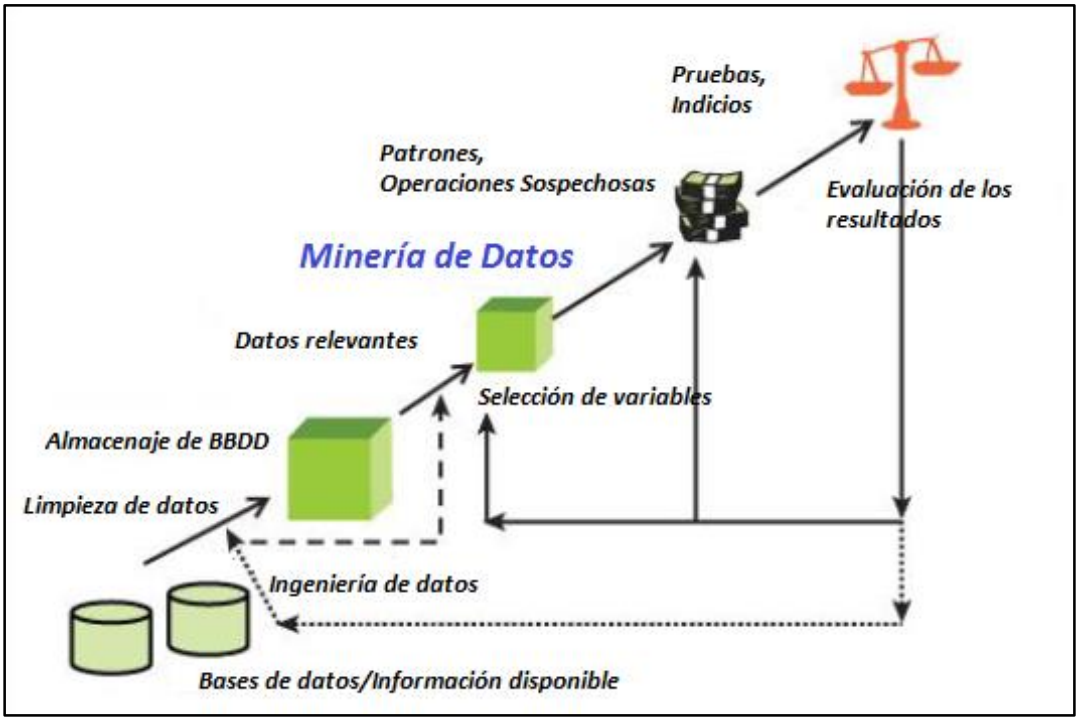
En esta Tesis Doctoral se ilustra el poder de las poderosas herramientas de aprendizaje automático para descubrir patrones de comportamiento en la detección de indicios de blanqueo los cuales puedan ser empleados en otros casos judiciales. Esto ayudaría a relacionar conceptos comunes del aprendizaje automático con la inferencia probabilística y la toma de decisiones.

La singularidad de cada caso de investigación, junto con las ingentes cantidades de datos que se derivan de la información contable y financiera obtenida, hacen que los forenses contables tengan que diseñar procedimientos de análisis *ad-hoc* para cada proceso judicial. Sin embargo, la interpretación tradicional, realizada de forma manual, es una práctica que consume una cantidad de recursos desproporcionada y obstaculiza la detección de los patrones de fraude (Yadav y Yadav, 2013).

La siguiente Figura I.1 describe de forma visual el proceso básico de un análisis basado en técnicas de aprendizaje automático por parte de un forense contable. El primer paso es la selección, depuración, transformación e integración de las bases de datos, este proceso es el más heterogéneo debido a las diversas fuentes de información y suele suponer más del 60% del trabajo del forense (García *et al.*, 2015). Una vez realizado el pre-procesamiento de la información se realiza un análisis descriptivo de los datos que ofrece información a los investigadores de donde enfocar los análisis. *A posteriori,* se seleccionarán los datos que serán analizados con técnicas de *machine learning* en un proceso supervisado, los cuales se dividen en dos conjuntos, el conjunto de entrenamiento y el conjunto de evaluación o comprobación.

Los datos del conjunto de entrenamiento se extraen utilizando diversas herramientas y técnicas, y la cantidad de datos seleccionados varían en función de la estructura del mismo. Por último, los resultados son evaluados de acuerdo al conjunto de comprobación, de este modo se obtiene un modelo ajustado y dirigido para detectar los patrones que han quedado "revelados" en el gran volumen de información (Dutta, 2013).

Figura I.1:
Etapas del reconocimiento de patrones (KDD) en Contabilidad Forense.



Fuente: Adaptado de Dutta (2013).

Estas herramientas de análisis son de gran alcance y potencialmente generan muchos patrones, el rol del contable forense en este sentido es ofrecer su experiencia y conocimientos para identificar qué patrones deben investigarse más a fondo (Dutta, 2013). No todos los patrones ofrecen información clarificada y argumentable, sin embargo, un patrón interesante describe una tesis o hipótesis, y por tanto, deben ser seleccionados únicamente aquellos que aportan más valor para la investigación.

Para responder rápidamente a los cambios y tomar decisiones lógicas, los investigadores necesitan un rápido acceso a la información para evaluar sucesos pasados e identificar tendencias relevantes (Lee e*t al.*, 1999).

CAPÍTULO II

ACCESO Y GESTIÓN DE LA INFORMACIÓN

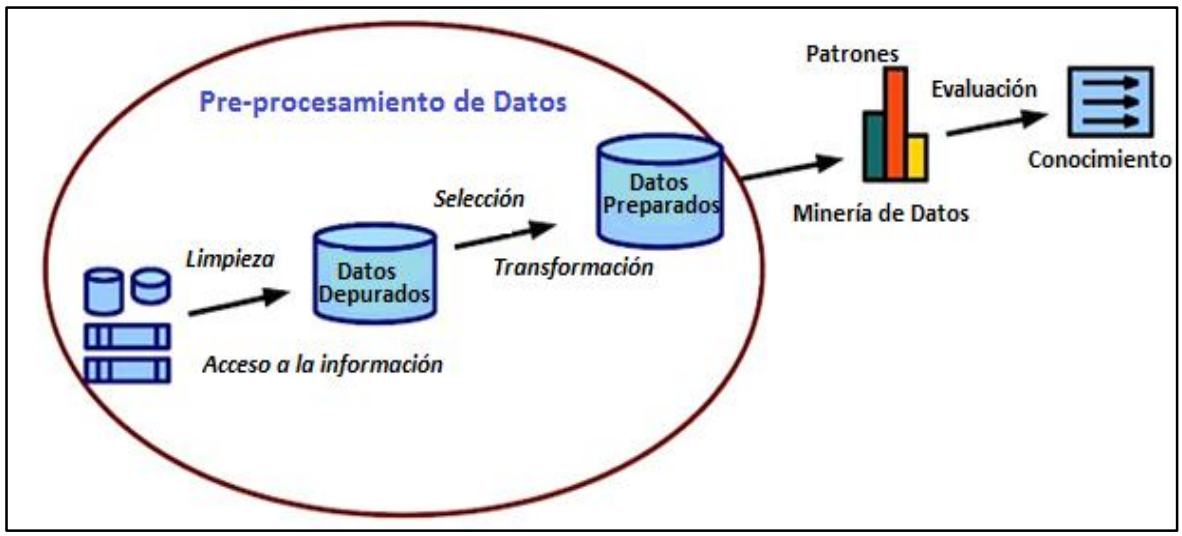
CAPÍTULO II: ACCESO Y GESTIÓN DE LA INFORMACIÓN

II.1.- Pre-procesamiento de datos

Con el objetivo de analizar las bases de datos de cada caso de investigación a fin de detectar posibles indicios de fraude, las técnicas de localización, gestión y limpieza de datos cobran vital importancia. El pre-procesamiento de datos mejora la exactitud de la clasificación y reduce la cantidad de datos necesarios para obtener el nivel de rendimiento deseado de los modelos, reduciendo los recursos computacionales e incorporando mayor información. Sin embargo, la particularidad de cada investigación en la que se trabaja, junto con la ingente cantidad de datos procedentes del área contable, complica considerablemente el diseño de procesos estándar para abordar un análisis contable forense (Dutta, 2013).

Este proceso de gestión de datos o de pre-procesamiento puede dividirse en cuatro etapas, las cuales son heterogéneas en función de las necesidades de cada caso: (1) el acceso a la información, (2) la recopilación de los datos, (3) la limpieza de datos y (4) la transformación de los mismos. Si bien la bibliografía relacionada en esta área todavía profundiza poco en la heterogeneidad de las distintas áreas de conocimiento en las que se aplica, se ha recopilado suficiente información para estructurar el proceso de gestión de las bases de datos en diferentes pasos que recogen y describen los diferentes procesos que requiere la preparación de los datos (ver Figura II.1).

Figura II.1
Pre-procesamiento de datos.



Fuente: Adaptado de Dutta (2013).

Además, dado el sucesivo avance tecnológico que se experimenta, las técnicas que se describen en este apartado necesitarán ser reevaluadas periódicamente para que incorporen las nuevas técnicas que pudieran mejorar los procesos y los adaptaran a los nuevos métodos de almacenamiento de datos.

Por tanto, el presente Capítulo ofrece una exposición de los pasos necesarios para llevar a cabo una gestión eficiente de las bases de datos antes de implementar los modelos de detección de patrones propuestos.

Este proceso es tan importante para los analistas como la implementación de los modelos propuestos en sí. Cobra relevancia en tanto que todas las conclusiones del análisis de los datos están sujetas a la calidad de los datos. En inglés este problema es conocido como *GIGO*, del inglés *Garbage In, Garbage Out*, y que se puede aplicar a todas las áreas de conocimiento. Como David Hand (1999) que apuntaba que "(…) *los analistas de datos de cualquier área no pueden realizar milagros y extraer joyas de la basura"*.

Especialmente, en los casos en los que opera el contable forense, con bases de datos muy grandes y, en particular, cuando se buscan patrones pequeños y sutiles o desviaciones de la regularidad, los problemas son particularmente agudos.

Para la gestión de la información disponible en la presente investigación se llevó a cabo un interrogatorio con el informático de la empresa núcleo intervenida (responsable de la gestión de la base de datos) en el que la autora de este trabajo también participó. A través de ello se consiguió una minuciosa definición de cada una de las variables que quedaron registradas en la gran base de datos a analizar.

La efectiva preparación de los datos antes de ser utilizados para su análisis garantiza la calidad de los resultados que se obtendrán al aplicar las técnicas de aprendizaje automático. Destacar que este proceso de preparación de datos incorpora también técnicas de aprendizaje automático en sí mismo, pues para aplicar muchas de las técnicas que en este apartado se presentan se requieren dichas habilidades.

II.2.- Acceso a la Información (*Data Engineer*).

En la contabilidad forense la selección de los datos viene dirigida por el caso de estudio en el que se va a trabajar. En los trabajos de investigación abiertos la selección de las fuentes de datos es un aspecto fundamental que condiciona los resultados finales. Normalmente, las fuentes de información son facilitadas por el contratante, instituciones públicas o empresas privadas, acotando la investigación a los datos que ellos estiman oportunos.

Dada la tendencia recurrente de las estructuras empresariales relacionadas con blanqueo de capitales en crear y cerrar multitud de empresas en poco tiempo para obstaculizar la trazabilidad de las operaciones, y también en radicarlas en diferentes países, la obtención de la información necesaria para investigarlos por parte de las autoridades es muy compleja y dilatada en el tiempo (Demetis, 2010).

La información disponible para analizar los casos de blanqueo es la que la autoridad competente estime oportuno considerar y quedará acotada a la misma, y así aparecerá especificado en las diligencias del proceso judicial. Sin embargo, es conveniente realizar análisis macroeconómicos complementarios que requieran la selección de datos publicados por las instituciones.

Además, si el informe forense es requerido para un juicio sólo se podrán emplear datos adicionales que garanticen la certeza de los mismos, esto es, datos publicados por los ministerios o fuentes públicas, o datos publicados por mercados regulados y supervisados (Sremack, 2015). En este ámbito el concepto de "Veracidad de los datos" es uno de los siete principios que se incluyen en la definición de "Big Data".

II.3.- Recogida y captura de datos (*Data Scraping*)

Un aspecto fundamental que agiliza la realización del informe pericial es la estrategia de recogida y captura de datos (raspado de datos) o *Data Scraping* en su terminología inglesa, lo cual se refiere al proceso por el cual se extraen los datos de las distintas fuentes de información disponibles en la red. Este proceso está en constante evolución y los métodos de importación de información varían del mismo modo que lo hacen las páginas de internet que albergan la información.

Usualmente el raspado de información se realizará de forma puntual y aislada. La recogida de información en fuentes públicas de internet se utilizaría para realizar comparativas entre los investigados y otros que pudieran ser de interés para el proceso. En los casos en los que el análisis se pueda extrapolar a otras investigaciones el uso de este tipo de técnicas de importación cobra mayor interés.

Así para poder utilizar métodos automáticos de raspado de información se deben tener en cuenta los siguientes aspectos, lo cuales se adaptarán del tipo de información con la que se vaya a trabajar:

1. La frecuencia con la que se van a extraer los datos.

Cuando se recurre a este tipo de técnicas el objetivo es mantener actualizados los datos, por tanto, los métodos que se emplearán serán los métodos automáticos que de forma periódica extraigan datos.

Por el contrario, para migraciones de información puntuales los métodos que se pueden utilizar podrían ser semi-automáticos y aplicables de forma más genérica.

2. La accesibilidad a las fuentes de información.

La accesibilidad de las fuentes de información tiene que ver con la estructura de los datos, si los datos se encuentran en forma de tabla o ficha, si está en una página, en varias páginas, etc. La forma en que la fuente de información ofrece los datos, es decir, si la propia página web ofrece la información, o si simplemente ofrece el enlace o si es una

combinación de ambos. De forma contraria los datos se pueden ofrecer desde la misma página web cargándose dinámicamente y presentando los datos usando *javascript*[22] y/o técnicas *AJAX*[23].

Una vez definidas las necesidades de información se tendrá que tener en cuenta la legalidad de la obtención de los datos. Las empresas y los particulares propietarias recelan de ofrecer de forma masiva la información, a medida que la información es más valiosa mayores son los controles para evitar que se pueda extraer. Algunas técnicas que las páginas emplean para evitar estos procesos de captura de datos y que se deben tener en cuenta para redirigir el tipo de estrategia para conseguir la información, esto es, hacer peticiones formales o compra de información, serían las siguientes (Hirschey, 2014):

- Bloqueos en las páginas web que obstaculizan los procesos automáticos y aplican métodos de identificación que deben ser respondidos por *acciones humanas*.

- Uso de *JavaScript* en el diseño de las páginas que obligan a integrar en las técnicas de raspado mecanismos para adaptar los programas a este tipo de páginas.

- Acceso de sesión, mediante el cual se obliga a programar un inicio de sesión para confirmar la autorización al acceso de los datos.

Una vez analizado cada caso están al acceso de los investigadores numerosas técnicas y programas gratuitos que ofrecen un mecanismo automático, que de forma manual supondría un consumo de tiempo desorbitado y una mayor probabilidad de cometer errores, que se traduce en un ahorro de recursos y garantizan la actualización periódica de la información, motivos que obligan a los usuarios a implementarlos en sus investigaciones. En función de la complejidad del "raspado" de datos, los tipos de mecanismos se pueden sintetizar en tres grandes grupos. En la siguiente Tabla II.1 se exponen las distintas metodologías que se pueden emplear según los requerimientos de cada caso.

[22] Lenguaje de programación.
[23] Técnica de desarrollo web para crear aplicaciones interactivas.

Tabla II.1:
Metodologías de *Data Scraping.*

Tipología	Descripción	Ventajas	Inconvenientes	Programas
Servicio Web Scraping	Extracción de los datos de la página web. Se extraen los datos de forma automática.	Conocimientos técnicos mínimos.	No se tiene el control del procedimiento de extracción, se limita a la funcionalidad que implemente el *servicio web de scraping* elegido.	*80legs, Import.IO, Scrapinghub*
Scraping Local	Desde el PC se crean las reglas para los enlaces.	Opciones gratuitas y sin límite de tiempo.	Fácil bloqueo por parte de las fuentes de información.	*FMiner, OutWit Hub, Scraper, USCIS Scraper, Web Scraper, Screen-Scraper*
Programación del Algoritmo	Utilizando lenguaje de programación con un *framework se* crea el algoritmo.	*Scrapings* complejos, muy frecuentes o de mucho volumen de datos.	Puede ser bloqueado por la fuente de información.	*DOMXPath, Guzzle, CasperJS / pjscrape (PhantomJS, SlimerJS), Scrapy*

Fuente: Elaboración Propia.

Para acceder a la información de un caso judicial a investigar es necesaria la designación oficial de una de las partes como perito forense. El documento de designación es público e informado a todos los intervinientes. Concretamente, en el caso judicial en el que se basa este trabajo de tesis, la información fue facilitada por la Policía Judicial, Brigada de Blanqueo de Capitales, previa encriptación de los datos para mantener el anonimato de los implicados y su correspondiente designación judicial.

II.4.- Limpieza de datos (*Data Cleansing*)

Una vez tomada la posesión de la información para investigar, y previo al proceso de análisis de la misma, cobra importancia el proceso por el cual se eliminan los valores que pudieran estar duplicados (Lee *et al.*, 1999).

Dada la sensibilidad de los datos que se disponen cuando se investiga un caso de blanqueo de capitales (registros contables y de operaciones y cuentas bancarias entre otros), y la repercusión de los resultados obtenidos, se deben definir las estrategias que se van a llevar a cabo para realizar la detección de valores que pudiesen estar, por ejemplo, duplicados, incompletos, incorrectos, inexactos o sean no pertinentes, para evitar que los resultados de los cuales pudieran depender las penas de los investigados se alejen de la realidad.

A pesar de la complejidad de este proceso, se pueden diferenciar diferentes pasos o técnicas mediante los cuales se puede construir la estrategia de limpieza de las bases de datos. En este apartado se expone el modelo propuesto por Lee y sus compañeros (1999) y las aportaciones del trabajo de Müller y Freytag (2005).

1. *Limpieza de errores y abreviaturas*. Se deben detectar los errores tipográficos y las abreviaturas para que permita más adelante realizar una detección eficiente de valores anómalos.

Para llevar a cabo este proceso se pueden utilizar diferentes programas específicos diseñados en el contexto del aprendizaje automático que permiten una mejor detección de este tipo de errores, como por ejemplo el programa Data Cleansing ofrecido por *Oracle Corporation*[24]. En estos problemas un algoritmo decide si una consecución de datos es aceptable dentro de la especificación de datos permitida. Proceso similar al modo en que un analizador gramatical trabaja con gramáticas y lenguajes (Rahm, 2000).

De forma paralela a los programas diseñados para este fin, la versatilidad de los softwares libres disponibles ofrece una amplia gama de paquetes y metodologías en esta área. Centrándonos en el software empleado para llevar a cabo los resultados de los apartados

[24] *Oracle Corporation* es una compañía multinacional de software que desarrolla bases de datos (*Oracle Database*) y sistemas de gestión de bases de datos.

posteriores, el software libre *R* cuenta con paquetes programados para tal fin como son el paquete "*tidyr*[25]" o el paquete "*lubridate*[26]".

2. *Agrupación de datos por campos*. Una vez se dispone de una base de datos con aparente grado de unidad en los registros, nula existencia de abreviaturas u errores de entrada, se lleva a cabo el proceso de agrupación de campos para comparar los registros.

Los diferentes grupos pueden ser características o variables recogidas en la base de datos e incluso pueden ser agrupados simbólicamente por su nomenclatura. De este modo podremos hacer una comparativa masiva de los registros y se permitirá la detección de valores erróneos. Este proceso se suele realizar *ad-hoc* para cada caso dependiendo del grado de similitud entre ellos. Cobra especial relevancia el conocimiento de cada una de las variables recogidas en la base de datos.

Para aplicar este proceso, en el trabajo que nos ocupa se realizó un reconocimiento exhaustivo de la base de datos que llevó a participar en un interrogatorio con el informático responsable de la gestión de la base de datos de la empresa núcleo intervenida para definir perfectamente cada una de las variables recogidas y su interpretación.

3. *Ordenación de los grupos o variables obtenidas de la base de datos*. Mediante las herramientas de *machine learning* se facilita la ordenación masiva de las bases de datos permitiendo la comparativa de los valores de cada registro.

4. *Comparación de los registros*. Se implementan en la base de datos mecanismos de comparación de registros que cuantifican el grado de similitud entre determinados registros. Para ello se fija previamente el número de registros a comparar, de ese modo se puede realizar un minucioso análisis de los registros que contengan un alto grado de similitud.

En los informes periciales llevados a cabo por el forense contable existen numerosas ocasiones donde los registros poseen un alto grado de similitud e incluso podrían ser iguales, esto es debido a las características de estrategias de registro de las operaciones

[25] Hadley Wickham (2017). tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions. R package version 0.6.1. http://CRAN.R-project.org/package=tidyr

[26] Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.

que lleva a cabo cada empresa. Si existe un alto grado de similitud entre los registros de las operaciones será necesaria una mayor exhaustividad en este proceso y el proceso tendrá menor grado de automatismo.

5. *Unión de los registros*. Una vez clasificados y ordenados los registros se podrán unir de forma automática los registros que consideremos sospechosos de estar erróneos para realizar comparativas eficientes y eficaces.

A pesar de aplicar técnicas de aprendizaje automático este proceso consume muchos recursos porque se requiere un minucioso e independiente análisis sobre cada registro sospechoso de estar erróneos.

II.5.- Tratamiento de datos faltantes (*Missing Data*)

Uno de los inconvenientes más evidentes en el análisis de bases de datos es el gran número de vacíos u observaciones perdidas que contienen y que deterioran la calidad de los resultados posteriores. Debido a errores tipográficos de entrada y, principalmente, al no registro[27], encontramos datos perdidos de diferentes tipos, tanto en registros de origen cuantitativo como cualitativo.

Los motivos por los cuales la base de datos presenta datos faltantes se pueden agrupar en errores en la extracción de los datos y errores en la recogida de los datos.

1. *Los errores producidos en la extracción de los datos*. Cuando no se realiza correctamente la extracción de los datos de las fuentes que albergan la información en las bases de datos extraídas pueden aparecer datos faltantes.

Este problema se puede subsanar incorporando controles en el proceso de extracción y realizando comprobaciones aleatorias para cotejar los datos extraídos.

2. *Los errores producidos en el proceso de la recogida de los datos*. Si los datos faltantes son generados por un problema en la recolección o de registro de los datos el error que conlleva es difícil de subsanar y las estrategias para tratar el problema dependerá del tipo de error que se haya generado.

Donald Rubin (1976) fue pionero en abordar de manera sistemática el análisis de bases de datos incompletos y fue el primer autor en proponer una terminología específica. Su estudio ha servido de base a prácticamente la totalidad de publicaciones y es por ello que las clasificaciones de las diferentes metodologías emplean su terminología. Una de las clasificaciones más generalizada de datos faltantes es la siguiente (Rivero, 2011):

- *Mecanismo Completamente Aleatorio (MCAR-Missing Completely At Random)*, cuando los datos perdidos no dependen de ninguna información contenida en la base de datos, es decir no poseen ningún patrón común, por tanto los datos perdidos son una

[27] En el área del análisis de encuestas este problema se define como *No Respuesta* del entrevistado. En el caso de la evaluación o análisis de datos contables la falta de datos suele originarse principalmente por fallos en la introducción de los registros.

muestra aleatoria simple de la base de datos original. Se trata del caso menos dañino para el análisis ya que las propiedades de la base de datos completa se conservan en la submuestra que contenga únicamente los casos completos.

- Mecanismo Aleatorio (MAR- *Missing At Random*), cuando los datos perdidos muestren un patrón en el que la no observación de los datos de una variable depende de los valores de otra u otras variables observadas.

- Mecanismo No Aleatorio (MNAR- *Missing Not At Random*), la probabilidad de que un valor haya sido observado depende de información no observada, bien porque se trata de valores de la propia variable con información incompleta o bien porque se trata de variables con las que no contamos en la base de datos. También serán MNAR aquellos casos en los que la pérdida de datos dependa de información que no ha sido incluida en el estudio.

Teniendo en cuenta el tipo de mecanismo observado, existen numerosas metodologías y técnicas para solventar, en cierta manera, el problema y que permiten continuar con el análisis, pero frecuentemente conllevan grandes sesgos.

Una de las prácticas más utilizadas en la actualidad es proseguir ignorando el problema, ya que implementar una estrategia adecuada es complicado y en pocas ocasiones se disponen de los recursos necesarios para ello. Es decir, la complejidad de las técnicas y de las metodologías asociadas ha hecho que la estrategia habitual frente a los *huecos* en la base de datos haya sido, simplemente, ignorarlos. De hecho, un estudio de Gary King (2002) concluye que sólo el 19% de los artículos analizados en ciencias sociales da información sobre el análisis de los datos incompletos y que del 79% restante más del 94% de los estudios simplemente omite los casos incompletos en el análisis y realizaba su investigación basándose únicamente en los casos completos. Somos conscientes que en estadística ignorar u obviar datos es muy pocas veces una solución eficiente.

Dado que la falta de datos representa un serio problema, el creciente desarrollo de técnicas por parte de la comunidad científica para tratar de corregirlo hace que nos encontramos con modelos muy heterogéneos. Los agruparemos en las siguientes dos grandes estrategias (Scheffer, 2002).

1. *Eliminación de los valores faltantes de registros o de pares*. Estos modelos son utilizados cuando la naturaleza de los datos es "Valores faltantes completamente al azar", de lo contrario, los valores faltantes no aleatorios pueden sesgar el modelo estimado.

En la eliminación por registros se quitarán todas las observaciones en las que hay algún valor faltante en cualquiera de las variables. A pesar de que por su simplicidad es ampliamente utilizado, esta estrategia reduce la capacidad predictiva de los modelos al reducir el tamaño de la muestra. No podría emplearse esta técnica cuando se registrara un elevado número de valores faltantes ya que la muestra disminuiría afectando muy seriamente la capacidad de los modelos.

En la eliminación por pares se realiza un análisis con todos los casos en los que las variables de interés están presentes. La ventaja de este método es que mantiene todos los casos disponibles para el análisis. Sin embargo, afectará a la eficiencia del modelo porque utilizará diferentes tamaños para diferentes variables.

2. *Modelos de imputación simple y modelos de imputación múltiple*. Los métodos de imputación de datos se pueden dividir en dos grandes grupos, los métodos de imputación simple y los métodos de imputación múltiple. Las técnicas simples presentan como gran ventaja una implantación más sencilla pero sufren una gran pérdida de eficiencia en comparación con las técnicas múltiples. Así mismo, las técnicas sencillas se subdividen en técnicas aleatorias o determinísticas. Las técnicas determinísticas a pesar de subestimar en mayor medida que las aleatorias las varianzas suelen dar lugar a estadísticos aparentemente más precisos.

La imputación múltiple parte de la construcción de $M \geq 2$ conjuntos de datos completos, los cuales son obtenidos mediante el reemplazo de cada dato faltante por M valores imputados. Este proceso lo convierte en una técnica muy potente, pero que genera problemas a tener en cuenta, tales como conducir a estimadores de la varianza inconsistentes en el caso de bases multietápicas estratificadas (Muñoz, 2009) o a la gran complejidad de los cálculos que dificulta su implantación (Rao y Shao, 1992).

En este sentido, la evolución de las herramientas de cálculo y software especializados ha supuesto un gran avance en esta área, pero en la actualidad la imputación automática de

datos faltantes basada en un proceso múltiple o de máxima verosimilitud con información completa no es un proceso estándar generalizado. Por otra parte, los resultados obtenidos mediante bases de datos con datos faltantes pueden reducir la capacidad predictiva de los modelos implementados, así como llevar a modelos sesgados que deriven en una mala interpretación de los resultados (Little y Rubin, 2014).

Particularmente al trabajo de esta Tesis Doctoral, y atendiendo a los requisitos de transparencia de los informes periciales, se ha optado por la eliminación de los registros de las operaciones que presentaban en alguna de sus variables datos faltantes.

A pesar del avance de la metodología que se puede emplear para gestionar los datos faltantes, todavía supone un gran reto presentar en un caso judicial metodologías estadísticas tan avanzadas, ya que si los intervinientes del proceso judicial no pueden comprender correctamente los algoritmos del modelo puede conllevar el rechazo del informe pericial.

Si los cálculos y modelos de detección de operaciones sospechosas se aplican a datos que no son gestionados desde una perspectiva conservadora puede verse comprometida la objetividad del informe y, por tanto, su valía en un proceso judicial.

II.6.- Detección de valores atípicos (*Outliers*)

Los conocidos como valores atípicos, *Outliers* en inglés, es otro de los problemas que pueden presentar las bases de datos disponibles y que necesitan ser tratados eficientemente. De acuerdo con la definición de Hawkins (1980) un valor atípico es "(…) *una observación que se desvía mucho de otras observaciones y despierta sospechas de ser generada por un mecanismo diferente*", sin embargo, a pesar de que estos valores pueden aparentar ser inválidos podrían ser correctos. Estos valores también se han denominado como valores que son "*dudosos a los ojos del investigador*" (Dixon, 1950) o "*contaminantes*" (Wainer, 1976).

Entre los efectos más comunes que los valores atípicos pueden tener en los análisis estadísticos destaca el aumento de la varianza del error, donde para los casos en el que el error no se distribuye aleatoriamente pueden disminuir la normalidad alterando las probabilidades de producir los errores[28]. También pueden sesgar las estimaciones o influir en los resultados (Rasmussen, 1988; Schwager y Margolin, 1982; Zimmerman, 1995 y 1998). Los valores atípicos se pueden dividir en dos grupos (Anscombe, 1960); univariados y multivariados. Los valores atípicos multivariados son los errores que se localizan en un espacio n-dimensional. Para afrontar la detección de estos valores es importante conocer el origen del evento, ya que aunque un registro sea atípico no necesariamente ha sido generado por un error. Se llevarán a cabo metodologías muy dispares en función del origen y de las características del mismo. De forma genérica las causas que provocan estos valores pueden ser de dos tipos: los valores atípicos generados de forma natural y los valores atípicos generados artificialmente por error en la generación del registro.

El primer caso no supondría un gran problema ya que, a pesar de su valor atípico en comparación con el resto de registros, este no es generado de forma artificial, sino que de forma natural este registra el valor real que debe representar. Por otro lado, en los casos los que se detecten datos atípicos generados de forma artificial, deberán ser subsanados mediante alguna de las estrategias que se definen más adelante.

[28] Errores tanto de Tipo I como de Tipo II.

Conviene remarcar que cuando la causa del valor atípico no está clara el problema se vuelve más complejo (Judd y McClelland, 1989). En estos casos existe controversia en la estrategia a seguir, pero son numerosos los autores que defienden la eliminación de los mismos a fin de obtener la estimación más honesta de los parámetros de la población (Barnett y Lewis, 1994). Por lo tanto, para tratar estos casos los investigadores deben utilizar su formación, la intuición, el razonamiento lógico y una consideración cuidadosa en la toma de la decisión a seguir.

Según la agrupación propuesta por Osborne y Overbay (2004), los valores atípicos más comunes se pueden clasificar en los siguientes grupos:

1. *Errores de entrada de datos*. La mayoría de los errores son errores humanos que se cometen durante la recogida, digitalización o entrada de datos y pueden generar outliers.

2. *Errores de medida*. Es la fuente más habitual de valores atípicos. Esto suele ocurrir cuando el instrumento de medida está defectuoso.

3. *Errores experimentales*. Otra fuente de outliers son los errores experimentales. Estos valores atípicos se encuentran comúnmente en las medidas autoinformadas referidas a datos sensibles.

4. *Error en el procesado de los datos*. Cuando se realiza el aprendizaje automático, los datos utilizados suelen proceder de diferentes fuentes. Es muy posible que la manipulación, extracción y procesamiento de los datos pueda generar valores atípicos.

5. *Error de muestreo*. Otra fuente de error es el error de muestreo al incluir en el conjunto de encuestados elementos que no pertenecen a la muestra.

6. *Valores atípicos naturales*. Los valores atípicos que no están generados por errores son *Outliers* naturales. Puede ocurrir que en una muestra haya dos subconjuntos, uno grande y otro pequeño que por las características de sus elementos generen outliers naturales. Al realizar una labor de minería de datos se debe tomar la decisión de analizar los dos conjuntos separada o conjuntamente.

Para su detección se pueden usar metodologías más o menos automáticas, pero por su simplicidad el método más utilizado para la detección de valores atípicos es la visualización de los datos. Para ello, se pueden emplear diferentes tipos de gráficos, por ejemplo, el gráfico de caja y bigotes (*Box-Plot*), los histogramas y los gráficos de dispersión (*Scatter Plot*), los cuales se utilizaran en función de las características de los registros.

No obstante, suele ser más efectivo utilizar un contraste formal estadístico para detectarlos. Por ejemplo el test de Dixon (Rorabacher, 1991) o el test de Grubbs (Grubbs, 1950), en los que se entiende que un valor es atípico cuando registra un *p-valor* menor de 0,05, lo que implicaría la presencia de valores atípicos con una confianza del 95%.

Siguiendo las reglas analíticas de detección de valores anómalos, destacan también la regla del rango intercuartílico (Hoaglin *et al.* 1986), en la que cualquier valor que quede fuera de ella será detectado como atípico (IQR = $Q_3 - Q_1$) o la regla fuera del rango de percentiles 5% y 95%. También podrían ser considerados los valores que se encuentren a una distancia superior a 3 desviaciones típicas de la media muestral.

Por otro lado, los valores atípicos bivariantes o multivariantes se podrán medir utilizando un índice de influencia o de distancias. Los índices más populares para detectar valores atípicos serían la distancia de Mahalanobis (Mahalanobis, 1936) y la distancia de Cook (Aguinis *et al.*, 2013)

El software libre *R* dispone de varios paquetes para la detección de valores atípicos, entre los que se destacan el paquete "*outliers*[29]" y el paquete "*mvoutlier*[30]".

Una vez detectados los valores atípicos para gestionarlos se emplearán métodos similares a los presentados en el apartado anterior referente a los valores faltantes. Por tanto, los métodos estadísticos más utilizados son suprimirlos, transformarlos, clasificarlos, o aplicar metodologías de imputación.

[29] Lukasz Komsta (2011). outliers: Tests for outliers. R package version 0.14. http://CRAN.Rproject.org/package=outliers.

[30] Peter Filzmoser and Moritz Gschwandtner (2017). mvoutlier: Multivariate Outlier Detection Based on Robust Methods. R package version 2.0.8.

En el caso objeto de estudio se detectaron, por un lado, numerosos valores atípicos generados por errores en la importación de los datos que fueron transparente subsanados una vez contrastado su origen con las autoridades judiciales[31]. También se detectaron valores atípicos de los cuales no se pudo conocer a ciencia cierta su origen y que fueron finalmente eliminados atendiendo a la objetividad que un informe pericial requiere.

[31] En el informe pericial presentado al proceso judicial se detalla minuciosamente el proceso de gestión de datos atípicos.

II.7.- Transformación de variables (*Feature Engineering*)

Este proceso es fundamental en la aplicación de técnicas de gestión de datos y es difícil y costoso, puede ser abordado de forma manual o empleando técnicas de aprendizaje automático, y resulta difícil y costoso. Su objetivo es extraer información adicional de los datos mediante transformaciones de las variables. No se añade más información a los datos, sino que se obtiene de ella información más beneficiosa. La extracción de información de las variables suele consistir en un proceso de reducción automática de la dimensionalidad de las mismas a un conjunto mucho más pequeño para que pueda ser aprovechado. También se requerirá la ingeniería de variables para gestionar algunas observaciones demasiado voluminosas y que de este modo puedan ser directamente introducidas en los algoritmos predictivos. Algunas de las técnicas más empleadas son:

La transformación numérica. La aplicación de logaritmos es un método de transformación comúnmente utilizado para cambiar la forma de la distribución de una variable en un gráfico de distribución. Se utiliza generalmente para reducir el sesgo por la derecha de las variables. Se puede aplicar a valores nulos o negativos. Con ello, la transformación logarítmica puede conseguir estacionariedad en media y en varianza para los datos. También se puede incorporar la raíz cuadrada o cúbica, aunque con resultados menos efectivos.

Por otro lado, el denominado "*binning*" que es una técnica de pre-procesamiento de datos utilizada para reducir los efectos de errores menores de observación. Los valores de datos originales que caen en un intervalo pequeño dado, un *bin*, son reemplazados por un valor representativo de ese intervalo, a menudo el valor central que se incorpora sobre los valores originales como en percentiles o frecuencias.

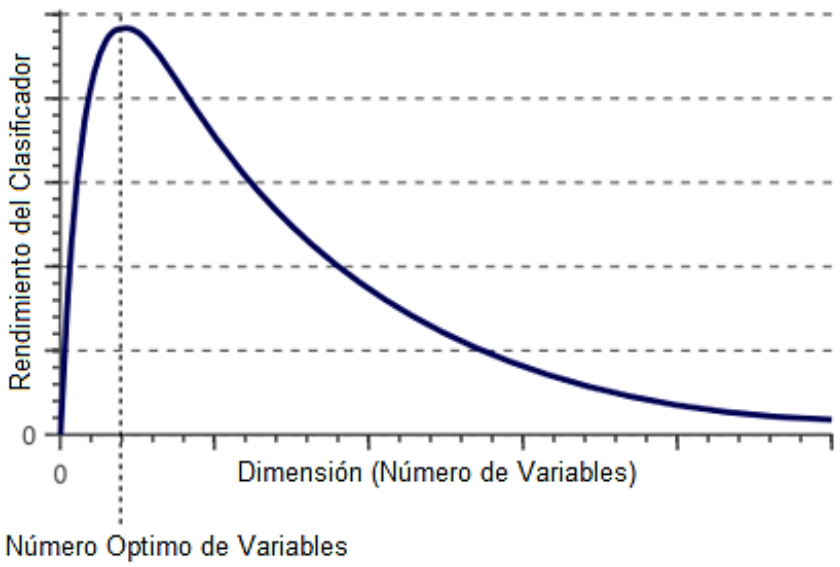
El proceso de ingeniería de variables en el caso objeto de estudio se llevó a cabo mediante la descomposición de atributos categóricos en las variables "almacén" y "producto", y en la descomposición de la variable "fecha". De ese modo se obtuvo información más detallada que permitió una mejor predicción de los modelos aplicados.

II.8.- Selección de variables

En el contexto de la contabilidad forense, y especialmente cuando se trabaja en la detección de patrones, la selección eficaz de variables es una de las tareas previas esenciales para que la predicción de los modelos a emplear sea más eficiente y precisa (Forman, 2003).

Llegados a este punto de desarrollo de esta Tesis Doctoral se debe reflexionar sobre los problemas que generan la gestión de ingentes cantidades de datos. No en cómo se van a llevar a cabo las estrategias de análisis para obtener los resultados, sino de cómo hacer frente al gran reto computacional que supone trabajar con grandes bases de datos. El problema de la dimensionalidad implica que hay un número máximo de variables a partir del cual la eficiencia del modelo disminuye en vez de aumentar

Gráfico II.1
Relación rendimiento/dimensionalidad



Fuente: Elaboración propia.

Como se representa en el Gráfico II.1, a medida que aumenta la dimensionalidad, el rendimiento del modelo de clasificación aumenta hasta que se alcanza el número óptimo de características. Aumentar aún más la dimensionalidad sin aumentar el número de

muestras en el subconjunto de entrenamiento da como resultado una disminución en el rendimiento del modelo.

Concretamente, en el caso aquí expuesto se tuvo que realizar una selección exhaustiva de variables debido a que las metodologías de detección de patrones propuestas requerían un espacio computacional grande que superaba los límites de los softwares utilizados[32]. En los Apartados IV.1.1 y IV.2.1 se exponen las estrategias llevadas a cabo en la selección de variables, y las variables que finalmente fueron introducidas en cada modelo.

La selección de variables es necesaria para gestionar eficientemente los problemas computacionales, conservando las propiedades de computación, almacenamiento y red para la fase de entrenamiento. Además, una selección eficaz de variables puede mejorar sustancialmente la exactitud de la clasificación o reducir la cantidad de datos necesarios para obtener el nivel de rendimiento deseado. A pesar de que existen distintas opiniones sobre por qué ocurre, los beneficios de la selección de variables son frecuentemente reconocidos (Forman, 2003).

De nuevo, el paralelismo con la principal regla del aprendizaje automático es evidente, *GIGO* (*Garbage In, Garbage Out*).

Por tanto, las técnicas de selección de variables, independientemente de cualquier metodología de aprendizaje automático que se vaya a aplicar, mejoran el aprendizaje de los métodos estadísticos utilizados, y sirven también para categorizar, dirigir, filtrar y buscar la información relevante.

Existen numerosas técnicas disponibles para llevar a cabo la selección de variables, siendo las más importantes las siguientes:

1. *Selección basada en la correlación*. Estos métodos analizan la correlación de las variables para eliminar información redundante y que elimine el ruido que incorporan. Se filtran las variables en función de grado de correlación que presentan. La aplicación de esta técnica no elimina la multicolinealidad, por lo que será necesario tratar los problemas de colinealidad de las variables antes de la modelización. El *coeficiente de correlación*

[32] El software libre R (32-bit) tiene un límite de espacio por defecto de 4GB.

de Pearson es el más utilizado para cuantificar la dependencia lineal entre dos variables cuantitativas (Benesty *et al.*, 2009).

También se pueden aplicar el Análisis Discriminante Lineal (*ADL*) que se utiliza para determinar una combinación lineal de variables que caracteriza o separa a dos o más clases de una variable cualitativa (Izenman, 2013). Por otro lado, se puede emplear el Análisis de la Varianza (ANOVA) que es similar al análisis ADL pero con la diferencia que la variable dependiente es cuantitativa, y las variables independientes son factores. El contraste ANOVA proporciona un estadístico para determinar si existen diferencias en la media entre los diferentes factores.

Por último, también es muy común el contraste Chi-cuadrado (X^2) (Nikulin, 1973) que se utiliza para determinar el grado de asociación entre dos variables cualitativas o atributos basándose en la distribución de frecuencias.

2. *Selección basada en la información y en el aprendizaje*. Estos modelos a diferencia de los modelos de selección por filtrado, que emplean técnicas estadísticas para la evaluación de un subconjunto de características, utilizan validación cruzada (Phuong, 2005).

Los modelos de "envoltura", o métodos *Wrapper* en su denominación en inglés, seleccionan un subconjunto de variables que será utilizado como subconjunto de entrenamiento. El algoritmo de selección de variables lleva a cabo una búsqueda utilizando un subconjunto de entrenamiento como parte de la función de evaluación. La precisión de los clasificadores inducidos se estima utilizando técnicas de estimación de exactitud (Kohavi y John, 1995).

La técnica *Wrapper* puede ser aplicada siguiendo diferentes estrategias, algunas de las más empleadas son las siguientes:

- La selección hacia delante, que es un método iterativo en el que se empieza sin ninguna variable y en cada iteración, se va añadiendo la variable que mejora más el modelo.

- La técnica análoga a la anterior, la eliminación hacia atrás, donde se comienza con todas las variables y se van eliminado iterativamente la variable menos significativa.

- La eliminación de variables recursivas se efectúa con un algoritmo que busca encontrar el subconjunto de variables con mejor rendimiento. Este modelo es el que suele tener mayor grado de eficacia pero requiere mayor esfuerzo computacional.

Estas técnicas puedes llevarse a cabo mediante el software *R* utilizando alguno de los paquetes creados para ello como son "Boruta[33]" o "Caret[34]" entre otros.

Debe remarcarse que estas técnicas incrementan el riesgo de sobreajuste cuando el número de observaciones es insuficiente. Además, el tiempo de cálculo es significativo cuando el número de variables es grande.

Actualmente, también están disponibles técnicas de selección de variables que utilizan de manera complementaria las técnicas de filtrado y las técnicas *Wrapper*, y que se denominan métodos integrados. En este sentido destacan los métodos que tienen funciones de penalización intrínsecas para reducir el sobreajuste: el método LASSO introducido por Tibshirani en 1996, el método Ridge introducido por Tikhonov en 1969 y Fu en 1998, el método Bridge por Friedman en 1993, y el método SCAD introducido por Fan y Li (2001). Ninguno de estos modelos de regresión domina uniformemente sobre los otros (Zou y Hastie, 2005).

Antes de implementar los modelos propuestos para la identificación de patrones de fraude se realizó una selección de variables de forma que se garantizase la incorporación de la máxima información en este trabajo.

Dado que se proponen dos enfoques distintos para tratar la detección de patrones, uno basado en la detección del fraude financiero mediante Redes Neuronales de clasificación y un segundo enfoque combinando la Ley de Benford y metodologías de aprendizaje

[33] Miron B. Kursa, Witold R. Rudnicki (2010). Feature Selection with the Boruta Package. Journal of Statistical Software, 36(11), 1-13. URL http://www.jstatsoft.org/v36/i11/.

[34] Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2016). caret: Classification and Regression Training. R package version 6.0-73. http://CRAN.R-project.org/package=caret

automático, también se llevaron a cabo dos estrategias diferentes en la selección de variables.

CAPÍTULO III

TÉCNICAS DE APRENDIZAJE AUTOMÁTICO PARA LA DETECCIÓN DE PATRONES

CAPÍTULO III: TÉCNICAS DE APRENDIZAJE AUTOMÁTICO PARA LA DETECCIÓN DE PATRONES

La información que se deriva de las complejas redes criminales es a menudo difícil de desentrañar porque suele estar geográficamente difusa y por períodos de tiempo muy cortos. Este hecho provoca que los procesos de investigación se dilaten en el tiempo, dificultando seriamente la eficacia de los procesos judiciales.

Con ello, existe una creciente necesidad de empleo de técnicas vanguardistas para mejorar la capacidad de análisis de la información obtenida y garantizar una lucha eficiente contra los defraudadores. Desde hace más de una década se están haciendo visibles las peticiones para la inclusión de estas técnicas en procesos judiciales y para incorporar metodologías de predicción estadísticas o sistemas expertos. En Estados Unidos hace ya varios años que se incorporan técnicas de contabilidad forense en procesos judiciales y encabeza las iniciativas de inclusión de estas técnicas. Por ejemplo, el Instituto Nacional de Justicia de Estados Unidos ha lanzado en los últimos años varias iniciativas en apoyo del uso de modelos de predicción para detectar patrones de delincuencia (Pearsall, 2010). También son numerosos los autores que reclaman la incorporación de técnicas de contabilidad forense para la investigación y lucha contra el blanqueo de capitales (Singleton *et al.*, 2006; Owojori y Asaolu, 2009).

Las técnicas de aprendizaje automático son una herramienta poderosa de información para explorar rápida y eficientemente las ingentes bases de datos que se analizan en los casos investigados. La capacidad de procesamiento disponible en los softwares aplicados (sistemas expertos) en el aprendizaje automático permite procesar miles de datos en segundos. Además, los procesos automáticos que se utilizan son menos propensos a errar que los procesos no automáticos que puedan desempeñar personalmente los investigadores, especialmente en aquellos procesos repetitivos. Sin embargo, el trabajo fundamental de los investigadores sigue siendo difícil y a menudo manual, y, dada la heterogeneidad de los casos, los patrones específicos de delincuencia en numerosas ocasiones no son fáciles de encontrar mediante herramientas automatizadas.

En este contexto, se entiende por reconocimiento de patrones a "(…) *la ciencia que estudia como las máquinas pueden analizar el entorno, aprender a detectar patrones y ofrecer decisiones razonables sobre las categorías de patrones*" (Jain *et al.*, 2000, pág. 2). Desde hace más de 60 años el reconocimiento de patrones es uno de los grandes temas de investigación, sin embargo, todavía hoy el mejor detector de patrones sigue siendo la mente humana, y parece imposible descifrarlo (Ripley, 2007).

"*Se considera que los procesos de toma de decisiones de un ser humano están ligados en cierta forma al reconocimiento de patrones,.., el objetivo del reconocimiento de patrones es aclarar estos complicados mecanismos de toma de decisiones y automatizar estas funciones utilizando ordenadores. Sin embargo, debido a la naturaleza compleja del problema, la mayoría de investigaciones de reconocimiento de patrones se han concentrado en problemas más realistas, como el reconocimiento de caracteres de texto latinos y la clasificación de formas de onda*" (Fukunaga (2013), pág. 1).

En el reconocimiento de patrones automático coexisten los siguientes tres aspectos a tener en cuenta (Li, 2005): (1) la tecnología disponible, (2) las características de los datos y (3) el enfoque del problema. Estos aspectos originan que la detección de patrones sea un área con un entorno propicio para aplicar las técnicas de aprendizaje automático. En la mayoría de los problemas de reconocimiento de patrones, los atributos y características de los datos originales y procesados son de tipo principalmente numérico. Por lo tanto, las técnicas utilizadas para construir modelos de apoyo a las decisiones deben ser capaces de tratar con los atributos numéricos de manera efectiva.

En ocasiones el comportamiento de los datos de reconocimiento de patrones no puede ser descrito por distribuciones estadísticas (como la distribución normal o binomial) lo que implica que los enfoques estadísticos paramétricos tradicionales no sean eficaces. Por ello, son ampliamente utilizadas las técnicas de aprendizaje automático, especialmente las basadas en la teoría del aprendizaje estadístico, en prácticamente todas las áreas de conocimiento en la detección de patrones[35].

35 Algunos de los ejemplos de más éxito se han obtenido en el área de la visión artificial, en el reconocimiento de patrones de voz, en la bioinformática o en la valoración del rendimiento deportivo (Chen, 2015; Albus et al., 2012; Wright et al., 2010; Jatoba et al., 2008; Zeng et al., 2009).

Tomando como referencia los artículos publicados por el Instituto de Ingeniería Eléctrica y Electrónica (IEEE)[36] se observa como desde la primera publicación en enero de 1979 realizada por Reaven y Miller (1979), sólo durante el siglo XX, más de 350 trabajos tratan del reconocimiento de patrones. Además, aproximadamente 300 de estos trabajos adoptaron un enfoque de aprendizaje automático basado en técnicas estadísticas (Jain *et al.*, 2000).

Actualmente, existen excelentes trabajos sobre detección de patrones como el de Duda y Hart (1973), Fukunaga (2013), Devijver y Kittler (1982), Bishop (1995), Ripley (2007), Chen (2015) o Schürmann (1996). También destaca el estudio de Nagy (1968), que marcó las primeras raíces del reconocimiento de patrones, que en ese momento se compartieron con los avances en inteligencia artificial.

Las técnicas predictivas que se proponen en este trabajo son técnicas de aprendizaje automático (sistemas expertos) enfocadas a la modelización y clasificación *ad-hoc*, en las que se especifican los modelos basándose en un conocimientos teórico previo (investigación policial previa del caso). Por otro lado, enmarcadas en las técnicas de aprendizaje automático propiamente dicho, también existen técnicas descriptivas de datos más enfocadas a la clasificación *post-hoc*. Estas técnicas realizan una clasificación sin especificación previa de los grupos, por ejemplo las técnicas de agrupación, o técnicas de *Cluster*. Las técnicas de *Cluster* han sido aplicadas con finalidad descriptiva en el informe pericial aportado al proceso judicial en el que se basa este estudio, complementado, entre otros, los análisis aquí presentados.

En criminología algunas de las aplicaciones más extendidas del aprendizaje automático son la *Entiti Extraction*, en la que se extrae información valiosa de datos de texto estructurados y no estructurados para identificar a individuos o empresas sospechosas. También la aplicación de técnicas *Cluster* para la identificación de potenciales delincuentes por su asociación con determinados delitos. Otras técnicas de clasificación para categorizar individuos sospechosos es el *Social Network Analysis* para conocer las

[36] *IEEE*, es la mayor asociación internacional sin ánimo de lucro formada por profesionales expertos en *nuevas tecnologías* dedicada a la estandarización y el desarrollo en áreas técnicas.

relaciones entre los individuos de las organizaciones criminales. También han sido aplicadas en la detección de patrones de delincuencia (Hassani *et al.*, 2016; Nath, 2006).

Sin embargo, la literatura disponible para la detección de patrones de delincuencia en el ámbito del blanqueo de capitales se reduce considerablemente. Se han utilizado técnicas de aprendizaje automático para la detección de patrones anómalos en bases de datos de transacciones financieras e implementación de soluciones basadas en el conocimiento para detectar patrones de blanqueo en bases de datos de entidades financieras (Khac y Kechadi, 2010; Cao y Do, 2012; Gao, 2009; Senator *et al.*, 1995). También se han implementado para la detección de precios anormales en la comercialización de ciertos productos (*Abnormal Trade Price*) (Zdanowicz, 2004). Así como para la detección de anomalías mediante técnicas *Knowledge Discovery Database* (KDD) en un enfoque basado en ensamble de modelos (Fan *et al.*, 2004). También se ha empleado el análisis de Redes Neuronales para identificar operativas sospechosas en la contabilidad de las empresas (Vikram *et al.*, 2004) y Árboles de Decisión para la evaluación del riesgo en SARs (Wang y Yang 2007). Dado que la creación de redes o estructuras empresariales es característico en el crimen organizado, los enfoques de análisis de redes sociales (*Social Network Analysis*) también se han utilizado contra el blanqueo de capitales (Zhang *et al.*, 2003). Con ello, la mayoría de trabajos desarrollan las técnicas de aprendizaje automático para procesar la información que se deriva de los informes de operaciones sospechosas (SAR) recibida en las unidades de inteligencia financiera (FIUs) (Yang y Wei, 2010).

En este trabajo se aplican las técnicas de aprendizaje automático para la detección de patrones en casos judiciales independientes. Se proponen dos nuevos acercamientos a la detección de indicios de blanqueo de capitales, por un lado, la aplicación de las técnicas de aprendizaje automático a bases de datos de la contabilidad interna de empresas investigadas por blanqueo de capitales, y por otro, ofrecer información a las autoridades investigadoras sobre cómo se organiza la red de blanqueo para orientar la investigación judicial hacia aquellas personas jurídicas o físicas que describan potenciales patrones de fraude.

Por lo que esta Tesis Doctoral representa uno de los primeros trabajos en los que se han aplicado las técnicas de aprendizaje automático para la detección de patrones de fraude en información obtenida en procesos judiciales de blanqueo de capitales. En España, y

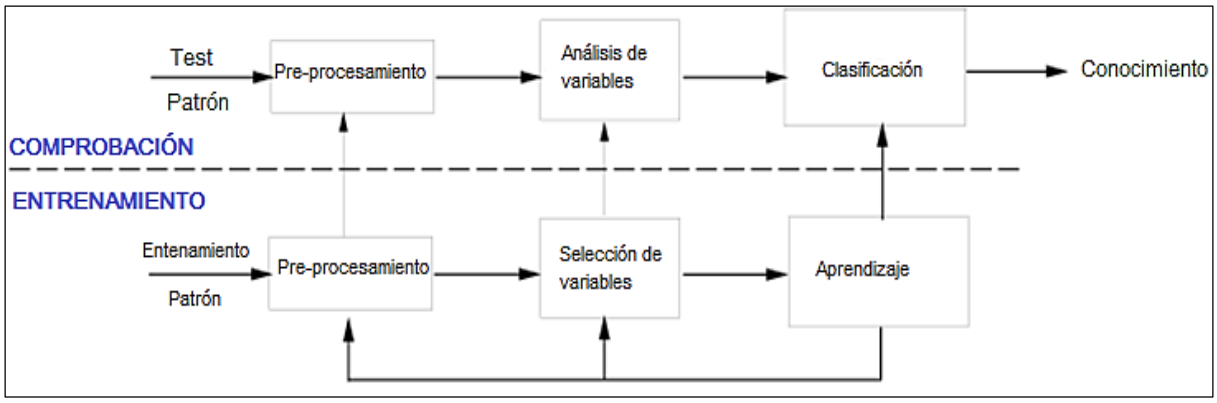
hasta donde la autora conoce, este trabajo de Tesis Doctoral supone la primera aproximación de las técnicas de detección de patrones basadas en metodologías de aprendizaje automático con aplicación directa en un caso judicial real de blanqueo de capitales.

En este Capítulo se describen las cuatro metodologías de aprendizaje automáticas aplicadas para la detección de patrones y la justificación de su elección en este trabajo, (1) la Regresión Logística Ridge, (2) el Árbol de Decisión, (3) la Red Neuronal Artificial y (4) el Bosque Aleatorio. Una vez descritos los modelos, se aborda, también, la Ley de Benford, en su aplicación propuesta por Nigrini (1996), como medida de detección de datos anómalos para combinarlos con las 4 metodologías anteriormente descritas (segundo enfoque propuesto en este trabajo). Por último, se trata el problema de los datos desequilibrados y posibles soluciones mediante la técnica SMOTE y Matriz de Costes.

III.1.- Aprendizaje automático supervisado para la detección de patrones

En el contexto de la detección de patrones, y desde la perspectiva del aprendizaje automático, un patrón es representado por un conjunto de m características, o atributos, vistos como un vector de entidad m-dimensional. Los conceptos de la teoría de la decisión estadística se utilizan para establecer los límites de decisión entre las clases de patrón. Concretamente en este trabajo, el sistema de reconocimiento se opera siguiendo un proceso de aprendizaje automático supervisado en dos fases: la fase de entrenamiento (aprendizaje) y la fase de clasificación (pruebas) (Jain *et al.*, 2000). Véase la Figura III.1.

Figura III.1
Modelo estadístico de reconocimiento de patrones



Fuente: Adaptado de Jain *et al.* (2000).

El objetivo del pre-procesamiento de datos supone convertir los datos disponibles en información que contribuya a definir una representación compacta del patrón que se persigue. Se elimina el ruido y se transforman las variables para incorporarlas en los modelos de detección de patrones.

En la fase de entrenamiento, relativa al módulo de extracción/selección de variables, se realiza la selección de predictores y se incorporan a los modelos de aprendizaje automático propuestos. La línea de retroalimentación permite al investigador optimizar las estrategias de pre-procesamiento y extracción/selección de características intercambiado los predictores del modelo.

En la fase de clasificación, los resultados obtenidos con la muestra entrenada son contrastados con los datos de entrenamiento para comprobar la calidad de la predicción del modelo (matriz de confusión), así como ofrecer información de los resultados obtenidos para complementar la toma de decisiones. Al emplear técnicas de aprendizaje dirigido, la información *a priori* ofrecida por los investigadores policiales es la base de la información incorporada a los distintos modelos. Si se obtiene un modelo que describa con suficiente precisión el comportamiento de las empresas que están cometiendo blanqueo de capitales, se deberá comprobar la capacidad predictiva de este modelo antes de tomarlo como válido. Como el número de empresas clasificadas previamente como fraudulentas es pequeño (datos desequilibrados), se realizará un análisis de sensibilidad de los resultados a los conjuntos de entrenamiento y comprobación.

El objetivo principal del aprendizaje automático es la extracción de la información contenida de un conjunto de datos con el fin de adquirir conocimiento que permita tomar decisiones sobre nuevos conjuntos de datos. Formalmente, se define como: Un sistema que aprende de la experiencia E con respecto a un conjunto de tareas T y una medida de rendimiento P, si su rendimiento en T, medido según P, mejora con la experiencia E (Mitchell, 2006).

Específicamente, en los modelos de aprendizaje automático la función de la predicción o clasificación es decidir la clase y' de un individuo x' basado en un conjunto de datos $D = (x_1, y_1), ..., (x_n, y_n)$ de individuos x_i de clase y_i.

Los x_i son usualmente *vectores m-dimensionales llamados* covariables o variables independientes (en el lenguaje estadístico) o predictores (en el contexto del aprendizaje automático).

La complejidad de los casos reales provoca que la mayoría de veces no exista relación funcional $y = f(x)$ entre y y x. Por tanto, la relación entre y y x normalmente tiene que ser descrita como una función de probabilidad $P(x, y)$, donde se asume que el conjunto de datos D contiene muestras independientes de P. Finalmente se elegirá la clase y que maximice la distribución $P(y \mid x)$. Existen dos tipos diferentes de

clasificación: el primero considera sólo una distinción dicotómica donde y solo toma los valores 0 o 1. El segundo intenta modelar la $P(y \mid x)$ lo que produce no sólo una etiqueta de clase para un elemento de datos, sino también una probabilidad de pertenecer a la clase.

En este trabajo se aborda la clasificación desde una perspectiva dicotómica, donde las clases y a las que pueden pertenecer los individuos son 0 o 1 (fraudulentas o no fraudulentas). Aunque los enfoques utilizados también permiten asignar probabilidades.

Se emplea un enfoque de aprendizaje automático supervisado para resolver el problema de la detección de patrones específicos en un caso de blanqueo de capitales en el que los datos de la empresa analizada provienen de la contabilidad interna de la empresa núcleo de un entramado empresarial de más de 600 empresas. El objetivo es clasificar a las empresas en legales o ilegales (fraudulentas/no fraudulentas) en función de las anomalías detectadas en la contabilidad de la empresa núcleo y con la ayuda del aprendizaje automático supervisado; se buscará conocer cuál es el comportamiento de las empresas defraudadoras, que así han sido etiquetadas por los expertos, en la información disponible *a priori*.

III.2.- Metodología aplicada para la detección de patrones de blanqueo de capitales

III.2.1.-Regresión Logística Ridge

La Regresión Logística representa un procedimiento clásico ampliamente utilizado para modelar la relación entre una variable dicotómica y una o más características (McCullagh y Nelder, 1989).

El análisis de Regresión Logística forma parte de los Modelos Lineales Generalizados que utilizan como función de enlace la función *logit*. Esta es una herramienta estadística ampliamente utilizada para estimar la probabilidad de que un individuo pertenezca a una clase $P(y \mid x)$ (Hosmer *et al.*, 2013). El objetivo de estos modelos es cuantificar (predecir) la probabilidad de permanencia a una categoría en función de unos predictores (predictores de tipo cualitativo y de tipo cuantitativo) (McCullagh y Nelder, 1989).

Al ser una de las técnicas más reconocidas como modelo de predicción, su aplicación en todas las áreas de conocimiento para abordar diferentes problemas de clasificación y predicción es abundante. La Regresión Logística se utiliza, por ejemplo, en el área económica (Cameron, 1988; Kleinman y Norton, 2009) o para la prevención de delincuencia (Bennell y Canter, 2002; Davies *et al.*, 1997).

III.2.1.1.- Aprendizaje del modelo de Regresión Logística

La aplicación de la Regresión Logística en el aprendizaje automático supone incorporar un predictor lineal para estimar las probabilidades posteriores que tiene una muestra de pertenecer, en general, a clases *K*. En este caso de estudio se propone un problema binario de clasificación.

En términos generales, la función logísticia o sigmoide, derivada de la *S-Shaperd*, del proceso de Regresión Logística se expresa como: $\alpha(x) = \frac{1}{1+e^{-x}}$

Calculada la inversa de la función sigmoide del modelo propuesto se obtienen los ratios de riesgo o *Log-odds*.

Entonces siguiendo la técnica *dummy*, el modelo es construido con K-1 ratios de riesgo o transformaciones logarítmicas (*log-odds*). De este modo, la función *logit* (o *log-odds*) es la razón logarítmica de las probabilidades de dos clases.

Una vez se tienen definidos los ratios de riesgo, el sistema de aprendizaje automático utiliza un método inspirado en los mínimos cuadrados en forma de iteraciones para ajustar los pesos de la función de máxima verosimilitud, este método es conocido como Mínimos Cuadrados Iterativos Reponderados (IRLS-*Iterative Reweighted Least Squares*) (Holland y Welsch, 1977). A pesar de ello no existe una solución cerrada, y las metodologías de aprendizaje automático permiten abordar un mismo problema mediante la combinación de diferentes técnicas.

Al minimizar los errores de clasificación en los datos de entrenamiento, se pueden estimar los diversos parámetros, β_k , del predictor lineal. El proceso de aprendizaje es iterativo hasta que se obtiene la mejor estimación.

Dado que en la Regresión Logística los parámetros β_k son una medida directa de la importancia relativa de los predictores, estos modelos permiten al investigador/diseñador identificar a las variables más relevantes (Pigeon *et al*., 2000).

En nuestro caso de estudio es aplicada la Regresión Logística Ridge (Le Cessie y van Houwelingen, 1992) que combina la regresión múltiple con métodos de regularización Ridge para minimizar el error del clasificador.

III.2.1.2.- Método de regularización Ridge

El estimador Ridge se engloba dentro de los métodos penalización cuadrática (Le Cessie y van Houwelingen, 1992). Esta técnica fue propuesta originalmente en los años setenta como un método para solventar el problema de colinealidad en un modelo lineal estimado por mínimos cuadrados, aún en el contexto $p < n$ (Hoerl y Kennard, 1970). En nuestro caso de estudio, dado que las variables no presentaban problemas de colinealidad, el estimador Ridge es aplicado para reducir el error medio del clasificador, es decir, maximizar la verosimilitud del clasificador.

El estimador Ridge regulariza los coeficientes para obtener el modelo de Regresión Logística con el menor error posible, de forma que una vez ajustados los parámetros se emplea el modelo más preciso como clasificador. En el contexto de la Regresión Logística el estimador Ridge quedaría definido como sigue:

$$\max_{\beta_0,\beta}\left\{\sum_{i=1}^{N}\left[y_i(\beta_0+\beta^T x_i)-\log(1+e^{\beta_0+\beta^T x_i})\right]-\lambda\sum_{j=1}^{p}\beta_j^2\right\}$$

Siendo $\lambda > 0$ el parámetro de contracción.

Este método tiende a reducir los coeficientes de regresión al incluir el término de penalización en la función objetivo: cuanto mayor sea λ existirá mayor penalización, y por tanto, mayor contracción de los coeficientes.

Una de las limitaciones de este método es que contrae todos los coeficientes hacia cero sin alcanzar la nulidad de ninguno de ellos. Por lo que al ser aplicado este método de penalización no se consigue una selección de variables, lo que resulta un inconveniente en aquellas aplicaciones que poseen un elevado número *p* de variables explicativas o predictores. Para eludir este inconveniente también es utilizada la regresión LASSO (Tibshibirani, 1996 y 2011) que, del mismo modo que el estimador Ridge, es una técnica de regresión regularizada. La técnica LASSO produce estimaciones nulas para algunos de los coeficientes, por lo que realiza una selección de variables en forma continua debido a la norma L1. LASSO reduce la variabilidad de las estimaciones por la reducción de los

coeficientes y al mismo tiempo produce modelos interpretables por la reducción de algunos coeficientes a cero. En concreto, adaptado a una Regresión Logística el estimador LASSO se obtiene resolviendo el problema de optimización:

$$\max_{\beta_0,\beta}\left\{\sum_{i=1}^{N}\left[\, y_i(\beta_0+\beta^T x_i)-\log(1+e^{\beta_0+\beta^T x_i})\,\right]-\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right\}$$

Siendo $\lambda \geq 0$ el parámetro de contracción.

El método LASSO al realizar una selección de variables ofrece cierta ventaja sobre la regresión Ridge, ya que produce modelos más simples y más interpretables que implica un único subconjunto de los predictores. Sin embargo, no hay un método predominante sobre el otro. El estimador Ridge obtiene mejores resultados cuando la respuesta es función de muchos factores predictivos, todos ellos con coeficientes de aproximadamente el mismo tamaño.

Finalmente, se opta por emplear el método de regularización Ridge con el objetivo de minimizar el error cuadrático global del clasificador propuesto.

III.2.2.- Árbol de Decisión

Los Árboles de Decisión representan otra técnica clásica de aprendizaje automático. Esta técnica puede ser vista como un diagrama de flujo con nodos de decisión que pueden interpretarse como reglas (Quinlan, 1986). El Árbol busca la combinación de cortes que hacen máxima la reducción de la incertidumbre en los nodos, dos de las medidas más empleadas son la Entropía de Shannon (Wang y Suen, 1984) y el Índice de impureza de Gini (Fayyad y Irani, 1992).

La flexibilidad de la estructura del Árbol permite incorporar valores desconocidos para las variables predictoras en los individuos en la fase de construcción del árbol y en la fase de predicción/clasificación. Además, en los Árboles de Decisión de clasificación, estos también permiten establecer una probabilidad *a priori* de las clases o ponderar las observaciones usando una variable *ad-hoc* (Brodley y Utgoff, 1995).

La estructura de la condición y la ramificación del Árbol son óptimas en la clasificación de individuos (Darasay, 1991). Dada su facilidad de interpretación, su agilidad de computación y la eficiente gestión de información innecesaria o ruido que realiza, los Árboles de Decisión han sido utilizados para tratar temas muy heterogéneos. Como uno de los métodos más populares de aprendizaje automático esta técnica se ha utilizado, por ejemplo, en la predicción de delincuencia (Pal y Pal, 2001; Li *et al.*, 2003; Jin *et al*., 2003; Nasridinov *et al.*, 2013; Sahin *et al.*, 2013) o en la detección de operaciones sospechosas (SARs) en casos de blanqueo de capitales (Wang y Yang, 2007).

Uno de los mayores retos que presentan las técnicas de aprendizaje automático en el área forense es la complejidad de los algoritmos que contienen. En este sentido, el Árbol de Decisión, presenta una ventaja respecto de las técnicas más vanguardistas en tanto que permite una interpretación intuitiva de sus resultados en forma de reglas (patrones) (Setiono *et al.*, 2000).

Por su estructura, el Árbol de Decisión particiona los datos recursivamente hasta que se cumple alguna condición, como la minimización de la entropía o la clasificación de todas las instancias. Debido a este proceso iterativo, el árbol generado tendrá una gran precisión en la clasificación de los datos de entrenamiento, pero muy poca precisión para clasificar

las instancias de los datos de comprobación (Wang y Suen, 1984; Maszczyk y Duch, 2008). Es decir, estará particularizando ejemplos en lugar de buscar un modelo generalista.

Este problema exige un procedimiento de poda *a posteriori* en la construcción del Árbol. El procedimiento de post-poda (que es donde realmente tiene efecto el enfoque específico del aprendizaje automático) consiste en medir el error estimado de cada nodo, de modo que si el error estimado para un nodo es menor que el error estimado para sus subnodos, entonces los subnodos se eliminan.

En la literatura se han desarrollado diferentes algoritmos o tipos de Árboles de Decisión con la filosofía machine learning. Algunos de los más importantes son: CART (Breiman *et al.*, 1984), ID3 (Quinlan, 1983) o C4.5 (Hunt *et al.*, 1966) (Rutkowski *et al.*, 2014).

El algoritmo ID3, por ejemplo, produce Árboles no binarios basados en la función de entropía de Shannon. Una versión actualizada del algoritmo ID3, basado en la función de ganancia de información es el algoritmo C4.5. En el algoritmo C4.5 se propone una función adicional que toma valores altos para atributos con dominios grandes, de este modo la relación de la eficiencia de la información es utilizada en los criterios de partición. En el tipo de Árbol CART se construyen árboles binarios y las particiones se forman en base al Índice de impureza de Gini (Rutkowski *et al.*, 2014).

A continuación se describe el Árbol de Decisión C4.5 (Quinlan, 1993 y 1996), al ser el modelo aplicado para detectar los patrones de fraude en este estudio.

III.2.2.1.- Especificación del modelo del Árbol de Decisión C4.5

El algoritmo C4.5 fue desarrollado para crear Árboles de Decisión bajo el criterio de "divide y vencerás" (Hunt, 1966). El siguiente algoritmo C4.5 genera un Árbol de Decisión de un conjunto de clasificación *D* como sigue:

Si D satisface un criterio de detención, donde D es una *hoja* asociada con la clase más frecuente. Una de las razones para detenerse en D es que contiene sólo los individuos de esta clase, aunque también pueden formularse otros criterios (Quinlan, 1996).

Algunos test T con resultados mutuamente excluyentes, $T_1;\ T_2;\ :::;\ T_k$, se utilizan para dividir D en subconjuntos $D_1;\ D_2;\ :::;\ D_k$, donde D_i contiene aquellos casos que tienen resultado T_i . El árbol para D se corresponde con el test T como su *raíz*, con un subárbol para cada resultado T_i que se construye aplicando el mismo procedimiento recursivamente a los casos en D_i .

En base a la iteración de estos criterios, cualquier test T que produzca una partición no trivial de D conducirá eventualmente a subconjuntos de clase única como los anteriores, siempre y cuando no existan casos con valores de atributos idénticos que pertenezcan a clases diferentes. Para alcanzar una partición eficiente del Árbol se examina una familia de pruebas posibles y se selecciona una de ellas para maximizar la eficiencia de la selección.

Entonces, para encontrar el umbral que maximiza el criterio de división, los casos en D se clasifican en sus valores del atributo A para dar valores ordenados $v_1,\ v_2,\ :::,\ v_N$.

El criterio de partición utilizado en este modelo es la *ganancia de información.*

$$Info(D) = -\sum_{(i=1)}^{C} p(D,i)\log_2(p(D,i))$$

De este modo, se determina la relación de ganancia de cada prueba posible y, entre las que tienen al menos la ganancia media, se selecciona la división con la relación de ganancia máxima.

En algunas situaciones, cada prueba posible divide D en subconjuntos que tienen la misma distribución de clase. Por lo que todas las pruebas tienen entonces una ganancia cero, y C4.5 utiliza esto como un criterio de detención adicional.

La estrategia de partición recursiva es eficiente en Árboles que son consistentes con los datos de entrenamiento. En la aplicación práctica de estos modelos, los datos suelen contener datos innecesarios o ruido, ello deriva en que los valores de los atributos se registren incorrectamente y los casos se clasifiquen erróneamente.

Dado que el ruido puede conducir a estructuras de Árboles demasiado complejas, este sistema efectúa un proceso *a posteriori* de poda del Árbol inicial, identificando los subnodos que contribuyen poco a la precisión predictiva (Quinlan, 1996).

El proceso de post-poda se fundamenta en la minimización del error del clasificador. Para ello se compara la cantidad de error que se reduce en cada poda y se decide en consecuencia en función de un "factor de confianza"; reduciendo el factor de confianza se reduce la cantidad de nodos eliminados. La post-poda se implementa en el algoritmo de forma inducida y en su proceso elimina los nodos estadísticamente insignificantes.

En el algoritmo C4.5 el proceso de post-poda supone evaluar el error de decisión en cada nodo y propagar este error en el resto del Árbol. En cada nodo, el algoritmo compara (1) el error ponderado de cada subnodo generado con (2) el error de clasificación erróneo si los subnodos fueran eliminados. Este procedimiento debe ser equilibrado ya que la reducción del árbol de decisión podría conducir a penalizaciones relevantes en la precisión del clasificador.

Como los clasificadores han sido implementados mediante el software libre WEKA, el Árbol de Decisión utilizado ha sido el clasificador J48, que combina el algoritmo C4.5 con un parámetro de confianza de 0,25.

III.2.3.- Redes Neuronales Artificiales

Desde los años 50 se han sucedido numerosas colaboraciones para avanzar los modelos de inteligencia artificial orientados a identificar patrones tratando de simular el funcionamiento de identificación del cerebro humano. Primero, se desarrollaron las técnicas *perceptron* (Minsky, 1954) y, posteriormente, estas dieron un gran salto cualitativo cuando, a mediados de los años 80, se desarrollaran los modelos de Redes Neuronales Artificiales (Hopfield, 1982) basadas en el modelo ADALINE (*ADAptative Linear Elements*) (Hopfield, 1982), que constituyó la primera Red Neuronal Artificial (Redes Neuronales en adelante) (Ripley, 2007).

Las Redes Neuronales están inspiradas en el cerebro humano, e intentan reproducir los aspectos esenciales de una neurona real a la hora de diseñar una neurona "artificial". Son modelos matemáticos estructurados en base al funcionamiento de las redes neuronales biológicas del cerebro humano (sistema nervioso), de forma que las unidades de procesamiento de una Red Neuronal son las neuronas artificiales, o nodos (Hilera y Martínez, 1995).

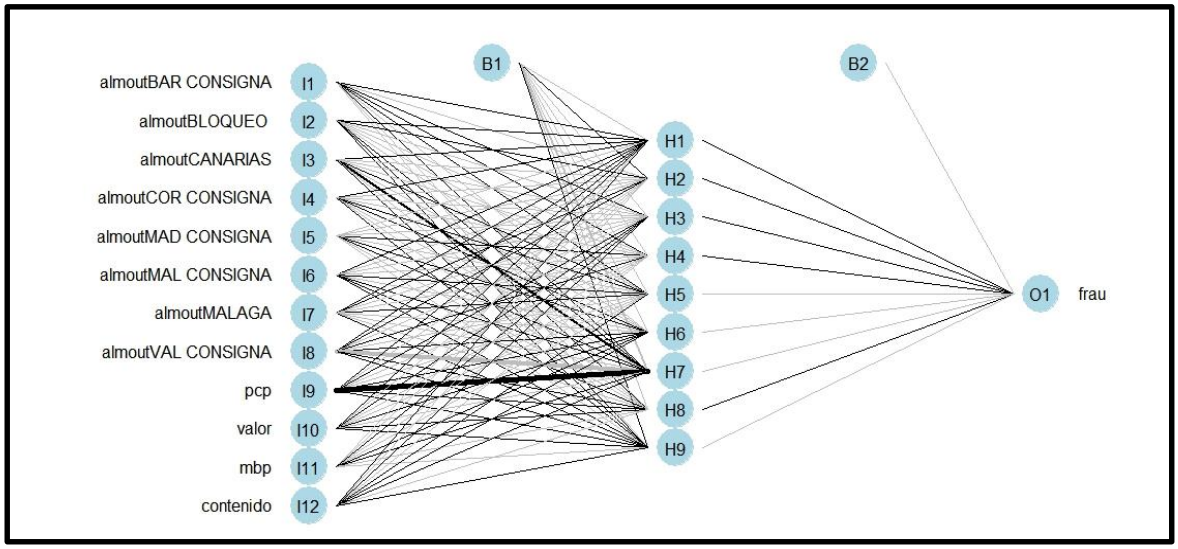
Dado que esta metodología ha alcanzado gran importancia en la inteligencia artificial, las herramientas estadísticas basadas en Redes Neuronales Artificiales para la detección de patrones de comportamiento se han aplicado con excelentes resultados en diversos campos, como en el área económica: (Caridad y Ceular, 2001; Muñiz y Alvarez, 1997; Olmedo *et al.*, 2007) o en la identificación de operaciones sospechosas de blanqueo de capitales, y en su extensión de fraude financiero (Lv *et al.*, 2008; Bolton y Hand, 2002; Chen e*t al.*, 2004; Ngai *et al.*, 2011).

Una de sus características diferenciales, y por la que destaca frente a las dos metodologías anteriores, es que pueden aprender de la experiencia a través de la generalización de casos (Bishop, 2006). Las Redes Neuronales Artificiales se constituyen en una técnica de procesamiento masivo y paralelo de la información que emula las características esenciales de la estructura neuronal del cerebro biológico.

Una Red Neuronal se caracteriza por cuatro elementos básicos: topología, mecanismo de aprendizaje, tipo de asociación realizada entre la información de entrada y salida, y la forma de representación de estas informaciones (Sosa, 2011).

Como se puede apreciar en la siguiente figura (Figura III.2), las neuronas se distribuyen en la red formando capas de un número determinado de elementos básicos. Es decir, existe una capa de entrada que recibe directamente la información proveniente de las fuentes externas de la red (I_i), capas ocultas que son internas a la red y no tienen contacto directo con el exterior (desde cero niveles hasta un número elevado) (H_i), pudiendo estar interconectadas de distintas maneras, lo que determina junto a su número las distintas topologías, y una capa de salida que transfiere la información de la red hacia el exterior (O_1).

Figura III.2
Esquema de la estructura de red ajustada aplicada en el primer enfoque propuesto.



Fuente: Badal y García (2016).

La topología de las Redes Neuronales es la forma de organización de las neuronas en la red formando capas o agrupaciones de neuronas más o menos alejadas de la entrada y la salida de la red. Por lo tanto, los parámetros fundamentales de la red serán: (1) el número de capas, (2) el número de neuronas por capa, (3) el grado de conectividad y (4) el tipo de conexiones ente neuronas (Bishop, 2006).

Se dispone fundamentalmente de dos tipos de redes en función del paradigma de aprendizaje: supervisado y no supervisado (Ripley, 1996). En el aprendizaje supervisado, que es el que se utiliza en el presente trabajo, a la red se le proporciona la respuesta correcta para cada una de las instancias de entrenamiento. De esta forma se pueden ajustar los pesos con la finalidad de aproximar la respuesta de la red a la proporcionada por los datos de la muestra. En el aprendizaje no supervisado se exploran patrones o correlaciones en los datos de entrada de la red, con la finalidad de poderlos clasificar (Ripley, 1996).

Los principales modelos de Redes Neuronales Artificiales son dos: (1) el Modelo perceptrón y (2) el Modelo Retro-propagación (Huerta *et al*., 2009).

1. Los Modelos de Redes Neuronales Perceptrón han sido ampliamente empleadas desde su descubrimiento (Rosemblatt, 1958) ya que poseen una alta capacidad para aprender a reconocer patrones sencillos (Demuth et al., 2014). Estos modelos se caracterizan por emplear perceptrones en la estructura de Red Neuronal. Un perceptrón es capaz de decidir cuándo una entrada presentada a la red pertenece a una de las dos clases que es capaz de reconocer. Un perceptrón está formado por varias neuronas lineales para recibir las entradas a la red y una neurona de salida.

El perceptrón (Liou et al., 2013) es un discriminador terciario que traza su entrada x (un vector binario) a un único valor de salida $f(x)$ (un solo valor binario) a través de una matriz que utiliza para representar la red:

$$f(x) = \begin{cases} 1 & Si \quad w \cdot x - u \\ 0 & \text{Resto} \end{cases}$$

Donde w es un vector de pesos reales y $w \cdot x$ es el producto escalar, y u es el "umbral" que define el grado de inhibición de la neurona.
La única neurona de salida del perceptrón realiza la suma ponderada de las entradas, resta el umbral y pasa el resultado a una función de transferencia de tipo escalón.

La regla de decisión es ponderar +1 si el patrón presentado pertenece a la clase A y –1 si el patrón pertenece a la clase B. De este modo los modelos solo son capaces de discriminar patrones sencillos y linealmente separables.

Esta característica obliga a las redes a generar sólo dos capas para la resolución de problemas en los cuáles el conjunto de puntos (correspondientes a los valores de entrada) sean separables geométricamente. Sin embargo, es posible resolver correctamente este problema usando una red de perceptrones o una red multiperceptrón.

2. El método *Back-Propagation network* (Rumelhart *et al.*, 1986), o método de propagación del error hacia atrás o retropropagación, define un algoritmo para que una Red Neuronal sea capaz de aprender la asociación que existe entre sus patrones de entrada y las clases correspondientes, pudiendo ser aplicada en modelos de redes multicapa. Esta metodología de Red Neuronal representa un algoritmo de ensamble de modelos (*ensemble*) en la que se combina el modelo de red con algoritmos para incrementar la aleatoriedad de los datos de entrenamiento (Kolen y Pollack, 1991).

El modelo se basa en la representación interna del conocimiento que es capaz de organizar en la capa o capas intermedias alcanzando la correspondencia entre la entrada y la salida de la red. El principal beneficio de la red de retropropagación es la capacidad de autoadaptar los pesos de las neuronas de las capas intermedias para aprender la relación que existe entre un conjunto de patrones dados como ejemplo y sus salidas correspondientes. Además, en un proceso posterior es capaz de utilizar esa misma relación a nuevos vectores de entrada con ruido o incompletos, dando una salida activa si la nueva entrada presenta características similares a las presentadas durante el aprendizaje (Huerta *et al*., 2009).

A pesar de su potencial eficiencia en la detección de patrones, la complejidad de los algoritmos e interacciones que emplea conlleva una gran dificultad de descripción de los mismos, siendo prácticamente imposible comprenderlos de forma llana. Los modelos de aprendizaje automático que presentan esta característica se les denominan modelos de "caja negra".

III.2.3.1.- Especificación del modelo de Red Neuronal *Back-Propagation Network*

Este análisis supone una primera aproximación a la implementación de modelos de Redes Neuronales al trabajo pericial para la detección de operaciones de fraude en casos judiciales de blanqueo. Por ello, se ha propuesto el empleo de una red de propagación de error hacia atrás.

En esta estructura de red hay tres elementos básicos: (1) las unidades de salida de la red (outputs, *Y*), (2) las unidades de entrada (inputs, *X*) y (3) las características derivadas de los inputs (*Z*, combinaciones lineales de *X*) (Hastie *et al.*, 2008).

Las unidades ocultas, Z, se obtienen como una combinación lineal de los inputs, transformada mediante una función de activación que se define como la función sigmoidea:

$$\sigma(\upsilon) = 1/(1+e^{-\upsilon})$$

donde:

$$\sigma(\upsilon) = [0,1] \text{, y } \upsilon =]-\infty, +\infty[$$

Para una clasificación binaria o en dos k clases, hay dos *K* unidades en la salida de la red, con la k-ésima unidad modelizando la probabilidad para la clase *k*. Hablamos, por tanto, de dos medidas objetivo Y_k, k = 1,2 codificadas como 0,1. La estructura de la red se representa mediante las tres expresiones siguientes:

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X), m = 1,...,M,$$

$$T_K = \beta_{0k} + \beta k_k^T Z, k = 1,2$$

$$f_k(X) = g_k(T), k = 1,2$$

donde: $Z = (Z_1, Z_2, ..., Z_M)$ y $T = (T_1, T_2)$.

Los parámetros del modelo (pesos o ponderaciones), inicialmente desconocidos, se ajustan utilizando los errores deviance (en el caso de redes de clasificación), definidos como:

$$R(\theta) = -\sum_{i=1}^{N} \sum_{k=i}^{K} y_{ik} \log fk(k_i)$$

Además, pueden incorporarse unidades de sesgo tanto en los nodos intermedios como en la función de salida que, pensadas como un input adicional, capturarían los interceptos α_{0m} y β_{0k}. El conjunto completo de pesos Θ se denota como:

$$\{\alpha_{0m}, \alpha_m; m = 1,2,...,M\} M(p+1) pesos,$$

$$\{\beta_{0k}, \beta_k; m = 1,2,...,M\} K(M+1) pesos,$$

Finalmente, la función de salida $g_k(T)$, permite la transformación final de los vectores de salida T, utilizando la función de transformación softmax (en el caso concreto) definida como:

$$g_k(T) = \frac{e^{T_k}}{\sum_{l=1}^{K} e^{T_L}}$$

Por tanto, el modelo de Red Neuronal es un modelo no lineal multilogit que utiliza una transformación de los inputs (X), mediante pesos (Θ) fijados a través de la minimización de los errores (deviance, $R(\Theta)$) mediante un procedimiento de actualización back-propagation (Ripley, 1996).

En el caso concreto que se analiza en este trabajo, existe una único nodo en la salida codificada 0 (no es posible concluir que la operación es fraudulenta) y 1 (la operación es identificada como fraudulenta), representada en el extremo derecho de la Figura III.2 como "O1" (Beck, 2015).

Las llamadas características derivadas (Z_m, m = 1,..,9) son los nodos H1 a H9, que se crean a partir de combinaciones lineales de las variables consideradas (inputs: I1 a I100, X_p, p = 1,...,100). En la estructura representada en la Figura III.2 puede también apreciarse la inclusión de variables que pretenden capturar el sesgo en cada uno de los nueve nodos de la capa.

III.2.4.- Bosque Aleatorio

El concepto de Bosque Aleatorio, desarrollado por Breiman (2001), nació como una variante de la metodología *Bagging* o agregación de *Bootstrap* (Dietterich, 2000b), que utiliza los Árboles de Decisión como clasificadores base (Breiman, 1996). Los Bosques Aleatorios son modelos que se encuentran enmarcados en los algoritmos de aprendizaje automático conocidos como ensamble de modelos, los cuales combinan diferentes técnicas de clasificación y técnicas de regresión para mejorar la estabilidad y la eficiencia de predicción del modelo.

Los Bosques Aleatorios son una metodología de clasificación y predicción que compite directamente con las metodologías más utilizadas en la detección de patrones, como son las Redes Neuronales o los Árboles de Decisión (Cutler *et al.*, 2007; Biau y Scornet, 2016). Las ventajas de los Bosques Aleatorios en comparación con otras metodologías de clasificación incluyen (Breiman, 1984; Ripley, 1996; Hastie, 2001; Cutler *et al.*, 2007): (1) precisión de clasificación muy alta, (2) un nuevo método para determinar la importancia de cada variable, (3) capacidad para modelar interacciones complejas entre las variables predictoras, (4) flexibilidad para realizar varios tipos de análisis de datos estadísticos, incluyendo regresión, clasificación, análisis de supervivencia y aprendizaje sin supervisión, y (5) disponibilidad de un algoritmo para imputar valores perdidos.

Dado que presenta ciertas ventajas sobre otras metodologías de aprendizaje automático, la aplicación de esta técnica se ha incrementado notoriamente en la última década (Bhattacharyya *et al.*, 2011; Bhattacharya, 2016; Lopez-Rojas y Axelsson, 2012; Sudjianto *et al.*, 2010).

Un Bosque Aleatorio es un clasificador consistente en una colección de clasificadores de Árboles de Decisión que son generados por un vector aleatorio distribuido idéntica e independientemente, y donde, cada árbol emite un voto (Breiman, 1996).

Cada árbol se construye utilizando una muestra bootstrap (muestreo aleatorio con reemplazamiento) diferente del conjunto de datos original de entrenamiento. Para clasificar un nuevo objeto, se le da el vector que lo describe a cada árbol, los cuales hacen

su clasificación independientemente. Los Árboles se construyen sin realizar ningún tipo de poda, dejando que alcance la mayor altura posible (Breiman, 2001).

Las instancias son clasificadas con la clase que obtiene un mayor número de votos de los árboles que conforman el ensamblado. Los resultados de los Bosques Aleatorios son difíciles de interpretar ya que son el resultado de la agregación de muchos Árboles de Decisión (modelos de caja negra).

Breiman (2001) expone que el error de los Bosques Aleatorios depende de dos factores fundamentalmente: (1) la correlación entre los árboles del ensamblado y (2) la efectividad de cada árbol individual. Al aplicar la metodología de *bagging* aumenta la estabilidad del árbol de decisión, que unido a la selección aleatoria de variables, incrementa la robustez sobre la presencia de variables redundantes, haciéndolo muy adecuado en conjuntos de datos con muchas variables (Breiman, 2001).

III.2.4.1.- Especificación del modelo de Bosque Aleatorio de clasificación supervisada

A pesar de que los Árboles de Decisión son intrínsecamente capaces de trabajar con problemas no binarios (Díaz-Uriarte y de Andrés, 2006), en este apartado consideramos únicamente el modelo de Bosque Aleatorio de clasificación binaria (Biau y Scornet, 2016), modelo propuesto en este trabajo.

En este contexto, tomando un conjunto de entrenamiento, $D_n = ((X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n))$ de variables aleatorias independientes (X, Y). La respuesta aleatoria, o de clasificación, Y, toma valores entre {0,1}.

Por lo que, mediante un conjunto X, se persigue estimar el valor de Y. Donde un clasificador, o regla de clasificación, m_n es una función de medida de *Borel* de X y D_n que tratan de estimar la clase Y de X en el conjunto D_n (Devroye *et al.*, 1996).

Atendiendo a ello, se dice que el clasificador m_n es consistente si su probabilidad condicionada de error:

$$L(m_n) = \mathrm{P}\left[m_n(X) \neq Y\right] \xrightarrow[n\to\infty]{} L^*$$

Donde L^* es el error de clasificación de Bayes desconocido:

$$m^*(x) = \begin{cases} 1 & Si \mathrm{P}\left[Y=1 \middle| X=x\right] > \mathrm{P}\left[Y=0 \middle| X=x\right] \\ 0 & \mathrm{Resto} \end{cases}$$

En el contexto de clasificación, el clasificador de Bosque Aleatorio se obtiene a través de un voto mayoritario entre los árboles de clasificación, es decir:

$$m_{M,n}(x;\theta_1,\theta_2,..,\theta_M,D_n) = \begin{cases} 1 & Si \frac{1}{M}\sum_{j=1}^{M} m_n(x;\theta_j;D_n) > 1/2 \\ 0 & \mathrm{Resto} \end{cases}$$

Si una hoja representa la región A, entonces un clasificador de árbol al azar toma la forma simple:

$$m_n(x;\theta_j,D_n) = \begin{cases} 1 & Si \sum_{i\in D_n(\theta)}^{*} 1x_i \in A, Y_i = 1 > \sum_{i\in D_n(\theta)}^{*} 1x_i \in A, Y_i = 0 \\ 0 & \mathrm{Resto} \end{cases}$$

Donde $\overset{*}{D}_n(\theta)$ contiene los puntos de datos seleccionados en la etapa de remuestreo. Es decir, en cada hoja se toma una mayoría de votos sobre todos los (X_i, Y_i) para los cuales X_i está en la misma región.

De este modo el algoritmo 1 se puede adaptar fácilmente para hacer la clasificación modificando el criterio de división CART para el ajuste binario. Dado que el criterio de división CART se fundamenta en la medida de impureza de Gini.

Para clasificar un conjunto de datos que caen en la celda A, se utiliza la regla que asigna un punto, uniformemente seleccionado de $\{ X_i \in A : (X_i, Y_i) \in D_n \}$, para etiquetarlo con valor l con probabilidad $p_{l,n}(A)$, para $j \in \{0,1\}$. La probabilidad estimada de que el elemento realmente sea de la clase l es $p_{l,n}(A)$.

Por lo tanto, el error estimado bajo el índice de impureza de Gini será $2_{p_{0,n}}(A)_{p_{1,n}}(A)$. En los Bosques Aleatorios de clasificación cada Árbol usa un voto mayoritario local. Por lo que de forma general, se suele establecer sólo un ejemplo en cada celda por debajo del cual la celda no está divida y $\sqrt{p}$ como número de direcciones posibles de división en cada nodo de cada Árbol (Liaw y Wiener, 2002).

III.3.- La Ley de Benford

Esta Ley fue descubierta por el astrónomo y matemático Simon Newcomb en 1881, aunque no fue puesta en valor hasta 57 años después cuando el físico Frank Benford la volvió a redescubrir. La Ley de Benford afirma que algunos conjuntos de datos numéricos tienen una distribución no uniforme de los diferentes dígitos, es decir, existe un patrón de comportamiento para cada uno de los dígitos y que sigue una ley logarítmica concreta (Benford, 1938).

En 1972, el economista norte americano Hal Varian propuso utilizar la Ley de Benford como una herramienta de diagnóstico para resultados de modelos proyectivos, en particular para pronosticar irregularidades en las auditorias que precisen inspecciones en mayor profundidad.

"(...) mientras que el análisis de la Ley de Benford por sí mismo podría no ser una manera infalible de detectar el fraude, puede ser una herramienta realmente útil para ayudar a identificar irregularidades, y por lo tanto, debe ayudar a los auditores en su búsqueda para detectar fraudes en la información contable" (Durtschi *et al.*, 2009, pág. 21).

Así, según si los datos económicos siguen la Ley de Benford, se dispondría de una herramienta para realizar comparaciones y poder detectar conjuntos de datos manipulados. Los datos que no siguen la Ley de Benford podrían haber sido manipulados y los investigadores tendrían un elemento de evidencia empírica de esa posible alteración (Pimbley, 2014). Consecuentemente, se podría utilizar para la detección del blanqueo de capitales o la evasión fiscal (Nigrini, 1992), ejemplos donde se produce una manipulación o alteración de los datos originales

"C*omo las acciones humanas no son aleatorias, los números falseados son poco proclives a seguir la Ley de Benford* (…)*"* (Nigrini, 1999).

El incumplimiento de la Ley de Benford es únicamente una evidencia que mostraría que los valores pueden estar manipulados, sin embargo, no es una muestra de que existe un delito en sí mismo. Esta ley no es una ley universal y habrá muchos conjuntos de datos

que no tienen por qué cumplir con ella (Nigrini, 1996). Sería una evidencia de la presencia de irregularidades en la contabilidad o las transacciones de determinadas empresas. Si los datos están manipulados, algo se debe esconder detrás de esa maniobra, y resultaría útil que se investigará el porqué de este comportamiento anómalo de los datos (Durtschi *et al.*, 2004).

Como sugieren Torres *et al.* (2007), es muy difícil que algunos datos manipulados o falsificados cumplan la Ley de Benford. De hecho, la mayoría de los datos contables falsificados no presentan una prevalencia del número 1 como el primer dígito significativo según la Ley de Benford, y en su lugar el número 5 y el número 6 son los dígitos prevalentes. Parece que psicológicamente, un falsificador trata de esconder datos falsos usando números a mitad de camino en la escala. La Ley de Benford ya es una herramienta real incluida en los programas informáticos de contabilidad y auditoría como otro mecanismo de análisis (Torres *et al.*, 2007; Bologna y Lindquist, 1995; Nigrini, 1996; Nigrini, 2012; Thomas, 1989; York, 2000).

En el área económico-contable, el profesor Nigrini (1992) demostró que podía servir para detectar fraudes en las declaraciones de renta y otros documentos contables (Nigrini, 1992).

Quick y Wolz (2003) examinan los datos de ingresos y el balance de situación de diversas compañías alemanas para los años 1994-1998. Sus resultados muestran que el primero y el segundo dígito en la mayoría de los casos (en un análisis año a año y también para todo el período) siguen aproximadamente la Ley de Benford.

Por otro lado, Günnel y Tödter (2009) demuestran que los controles para la manipulación de los datos deben centrarse en el primer dígito. Consideran que la Ley de Benford es una herramienta simple, objetiva y efectiva para detectar anomalías en grandes muestras que precisan una inspección más detallada. Sin embargo, Ramos (2006) afirma que la mejor parte del análisis es la de los tres primeros dígitos en donde realmente se obtiene un electro-cardiograma del archivo y se puede ver en detalle que ocurre en cada punto y cuáles son las operaciones con potencial fraude.

Alali y Romero (2013) analizan la información financiera de más de diez años de datos contables de empresas y concluyen que existe un error significativo en el ajuste de la Ley de Benford en parte del activo circulante de las empresas analizadas, esto es, en bienes de equipo, propiedades, cuentas por cobrar, lo que conlleva a que durante el periodo analizado existió una sobreestimación del activo.

Como se ha evidenciado, el empleo de la Ley de Benford en el ámbito de la contabilidad es amplio, lo que sugiere su capacidad para la detección de anomalías en los datos contables. De acuerdo a esa premisa, en este estudio se han propuesto diferentes medidas de ajuste de la muestra a la Ley de Benford como posibles indicadores para la detección de los patrones que esconden las operaciones fraudulentas, y así combinar la Ley de Benford con las metodologías de aprendizaje automáticas propuestas para detectar los patrones de blanqueo de las empresas fraudulentas. Seguidamente se expone como funciona esta Ley.

III.3.1.- Descripción de la Ley de Benford.

Empíricamente, Benford (1938) encontró que muchos conjuntos de datos y secuencias matemáticas no tienen una distribución uniforme del primer dígito como se podría esperar, y sin embargo tienen una función de probabilidad sesgada.

Un conjunto de números satisface la Ley de Benford si el primer dígito d_1,

$d_1 \in \{1,..., 9\}$, ocurre con una probabilidad como sigue:

$$P(d_1) = \log_{10}\left(1 + \frac{1}{d_1}\right)$$

Asimismo, la fórmula para la función de probabilidad del segundo dígito d_2,

$d_2 \in \{0,..., 9\}$:

$$P(d_2) = \sum_{k=1}^{9} \log_{10}\left(1 + \frac{1}{10 \cdot k + d_2}\right)$$

Con ello, las propiedades más importantes de la Ley de Benford son la invarianza en escala y la invarianza en base.

Sea X una variable aleatoria, distribuida aleatoriamente, entonces se podría esperar una distribución uniforme para el valor del primer dígito d_1=1, 2,…, 9. Sin embargo, una serie de variables muestra una distribución diferente para el primer dígito y, de acuerdo con las demostraciones de Pinkham (1961) y Hill (1995b), cumplirían con las dos propiedades indicadas:

1. *Invarianza en escala*. Se ha observado empíricamente que al efectuar cambios de escala en aquellas variables que se ajustan a la ley logarítmica también la nueva variable transformada se ajusta bien a esta ley. Si se cambian las unidades de medida la Ley de Benford se sigue cumpliendo, es decir, no depende del sistema de medición. En términos económicos, la moneda en que se mida la variable objeto de estudio es independiente para el resultado que se puede obtener.

2. *Invarianza en base*. La ley logarítmica es independiente de la base que se utilice, y es igualmente válida en base 10, en base binaria, o en cualquier otra base. Hill (1995) demostró que la distribución logarítmica es la única distribución continua que es invariante en base y que la invarianza en escala implica invarianza en base, aunque no viceversa.

Tabla III.1
Probabilidades de la Ley de Benford de 1º y 2 º dígitos

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Primer Dígito	-	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6
Segundo Dígito	12.0	11.4	10.9	10.4	10.0	9.7	9.3	9.0	8.8	8.5

Fuente: Elaboración Propia.

Además, la Ley de Benford es más robusta de lo que se puede imaginar. No todas las series numéricas siguen una distribución de Benford, sin embargo, si se seleccionan varias distribuciones de forma aleatoria, y las muestras tomadas de cada una de estas distribuciones son aleatorias, entonces la frecuencia de los dígitos del conjunto de datos

combinado convergerá a una distribución de Benford aunque las distribuciones separadas se desvíen de la Ley de Benford (Nigrini, 1996).

III.3.2.- El Estadístico Z

El Estadístico Z (Nigrini, 1992) se emplea para medir el ajuste a la Ley de Benford a cada uno de los primeros y segundos dígitos de las operaciones contables correspondientes a cada empresa. Se determina el grado de ajuste de los primero y segundos dígitos del importe de todas las operaciones de cada empresa.

Como las proporciones de Benford no representan una verdadera distribución, si no que se espera que se cumpla en el límite, los contrastes tradicionales como el contraste Chi-cuadrado, el de Kolmogorov-Smirnov o el contraste de Kuiper son demasiado rígidos para evaluar la bondad de ajuste (Tam y Gaines, 2007). Sin embargo, el Estadístico Z permite comprobar la conformidad de un conjunto de datos a la Ley de Benford, y la fórmula para su cálculo en cada dígito es la siguiente:

$$Z_i = \frac{n_{oi} - n_{Ti} - \left(\frac{1}{2N}\right)}{\sqrt{\frac{n_{Ti}(1-n_{Ti})}{N}}}$$

donde:

n_{oi} : Valor observado en la muestra.

n_{Ti} : Valor esperado derivado de Ley de Benford.

El término (1/2N) es un término de corrección de continuidad y solo se utiliza cuando es menor que el primer término del numerador.

Con el Estadístico Z_i se evalúa la proporción de los dígitos de forma separada, determinando qué dígitos difieren de la distribución de Benford. Esto implica que para el primer dígito haya nueve comparaciones, y que no se pueda tomar el nivel de

significación del 5% para comparar los *p-valores*. El proceso de reducción del nivel de significación se basa en la desigualdad de Bonferroni (Hogg *et al.*, 2005).

Cada p-valor es comparado con $\alpha/9 = 0{,}05/9 = 0{,}0056$; obteniéndose una probabilidad aproximada conjunta de rechazo de 0,05. Si P (IZI > 2.77) = 0,0056 cualquier Estadístico Z que sea mayor en valor absoluto que 2,77 implica el rechazo de la hipótesis nula.

En el caso de estudio, el Estadístico Z, del mismo modo que ocurre con los contrastes Chi-cuadrado y otros contrastes basados en los p-valores, rechaza la hipótesis nula cuando analizamos el ajuste a la ley del conjunto de todas las empresas (Amiram *et al.* 2015), esto se debe a la gran cuantía de datos que se analizan de forma global, en nuestro caso 285.774 operaciones comerciales de 643 empresas proveedoras. Por ello en este trabajo también se propone un test empírico basado en la simulación que no es sensible al tamaño muestral que se ha denominado OverBenford Test.

III.3.3.- Contraste de ajuste a la Ley de Benford propuesto: OverBenford Test

El OverBenford Test está fundamentado en la generación de 100 simulaciones de una distribución de Benford del mismo tamaño que cada una de las empresas, para posteriormente, realizar el contraste Chi-cuadrado de cada simulación. Este test imita los enfoques sugeridos para la bondad de ajuste de las distribuciones continuas implementadas en Pavía (2015).

Del resultado se obtiene una medida alternativa para cada empresa que no se ve influenciada por el tamaño muestral. El contraste propuesto es el siguiente:

$$T = \sum_{i=1}^{m} \frac{(n_{oi} - n_{Ti})^2}{n_{Ti}}$$

donde:

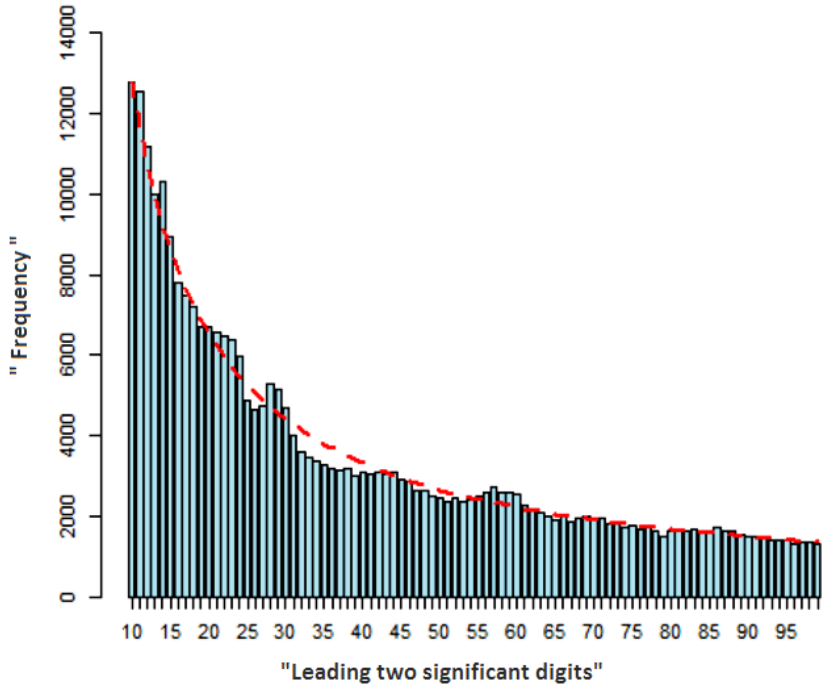
n_{oi} es el valor observado en la base de datos disponible.

n_{Ti} es el valor sintético generado procedente de una distribución de Benford del mismo tamaño que la muestra.

Si comparamos la distribución de toda la muestra con la Ley de Benford podemos comprobar visualmente como se ajusta de forma casi perfecta, como se indica en el siguiente Gráfico III.3.

Gráfico III.3

Análisis global de la muestra respecto de la Ley de Benford de 1º y 2º Dígito.



Fuente: Elaboración Propia.

Tabla III.2

Análisis global del ajuste a la Ley de Benford de primer dígito de la muestra.

Mean	*Var*	*Ex.Kurtosis*	*Skewness*
0.496	0.085	-1.224	0.026

Fuente: Elaboración Propia

Sin embargo, cuando se realizan los diferentes test referenciados en el apartado anterior (Test Chi-cuadrado y Estadístico Z) se observa que estos rechazan la hipótesis nula, véase la siguiente tabla.

Tabla III.3
Test de ajuste de la muestra a la Ley de Benford de primer dígito.

	X-Squared Test	***Z-Test***	***OverBenford Test***
P-Value	2,2e-16	2,3e-16	0,2303

Fuente: Elaboración Propia

En la Tabla III.3, se comprueba que los p-valores de los test tradicionales no servirían como medida para cuantificar el ajuste a la Ley de Benford en muestras con notable número de registros. Sin embargo, el OverBenford Test tiene la capacidad de medir el ajuste a esta Ley evitando la sensibilidad al tamaño muestral.

III.4.- Técnicas de balanceo de datos.

Un conjunto de datos está desequilibrado si las categorías de clasificación no están representadas aproximadamente igual en la base de datos. Durante los últimos años, las técnicas de aprendizaje automático se han aplicado para resolver problemas reales, los cuales en muchos casos se caracterizan por trabajar con datos desequilibrados (He y García, 2009; Chawla, 2005). Por ejemplo, en el proceso judicial que nos ocupa la muestra está claramente desequilibrada ya que del total de las 643 empresas que forman parte del entramado de blanqueo sólo se tiene certeza de que 26 son fraudulentas, es decir, las operaciones que realizan no se ajustan a la legalidad. Por tanto, sólo se dispone de información *a priori* facilitada por los investigadores policiales de que el 4% de las empresas son fraudulentas.

Los conjuntos de datos desequilibrados son bastante habituales en la literatura científica, y normalmente, la categoría que es más relevante para el análisis es la que tiene un porcentaje bajo de instancias (Kotsiantis, *et. al*, 2006; Raskutti y Kowalczykl, 2004; Wu y Chang, 2003; Yan *et al.*, 2003).

La cuestión fundamental con el problema de aprendizaje automático empleando datos desequilibrados es que estos pueden comprometer significativamente el rendimiento de la mayoría de los algoritmos de aprendizaje. Los algoritmos estándar suponen o esperan distribuciones equilibradas de las clases. Por lo tanto, cuando se aplican a conjuntos de datos desequilibrados complejos no representan las características distributivas de los datos y, en consecuencia, proporcionan predicciones sesgadas hacia la clase mayoritaria (He y García, 2009; Chawla, 2005). Sin embargo, son numerosos los trabajos que abordan el problema de datos desequilibrados en el contexto del aprendizaje automático.

Además de la cuestión de la distribución entre clases, otro problema que surge debido a la escasez de datos de la clase minoritaria es la distribución de datos dentro de cada clase (Japkowicz, 2001; Chawla, 2005), este problema también afecta directamente a los algoritmos de Árboles de Decisión. Por tanto, las bases de datos desequilibradas presentan los siguientes inconvenientes: la falta de información en los datos de entrenamiento, la superposición entre las clases, el impacto del ruido, la importancia de las instancias límite

para llevar a cabo una buena discriminación entre las clases, y las diferencias en la distribución del conjunto de entrenamiento y los datos de comprobación, también conocido como *dataset shift* en inglés (Kotsiantis *et al.*, 2006; Chawla, 2005; López *et al.*, 2013)

Teniendo en cuenta estas limitaciones en la aplicación de las técnicas de aprendizaje automático en bases de datos desequilibradas es aconsejable la aplicación de métodos que mejoren su distribución. Estos métodos se pueden categorizar en tres grandes grupos, siguiendo el trabajo López *et al.* (2013):

1. *Balancear el conjunto de entrenamiento*. El tratamiento de los datos desequilibrados se puede realizar mediante el submuestreo de la clase mayoritaria, aplicados en conjuntos de datos grandes y trabajando sobre la clase mayoritaria reduciendo el número de instancias de esta clase para balancearla con la clase minoritaria. Dado que con este método se descartan buena parte de las instancias de la clase mayoritaria, se pierde una información que podría ser relevante en el conjunto de entrenamiento.

Otro método de balanceo del conjunto de entrenamiento es el sobremuestreo de la clase minoritaria, que en oposición al submuestreo, trabaja con la clase minoritaria la cual es incrementada hasta ser equitativamente balanceada respecto a la clase mayoritaria. En este caso no se pierde información pero se incrementa el conjunto de entrenamiento mediante copiar y pegar observaciones de la clase minoritaria, lo que podría implicar problemas de sobreajuste del modelo.

Estas dos técnicas son fáciles de aplicar pero ambas presentan problemas para el rendimiento de los algoritmos, por lo que para balancear una clase con un enfoque más sofisticado se deberán utilizar alguna de las siguientes metodologías. En el presente trabajo, dada la escasa presencia de las empresas fraudulentas, es aconsejable utilizar otros métodos.

Otra de las metodologías de balanceo de datos es la generación de datos sintéticos en la clase minoritaria, esta técnica es una de las más utilizadas, concretamente la técnica *Synthetic Minority Oversampling Technique* (SMOTE) (Chawla *et al*. 2002), que como

se expondrá en el siguiente apartado presenta ventajas respecto de las otras dos técnicas. Otras técnicas de generación de datos sintéticos que se podrían enmarcar dentro de esta metodología son: la técnica *Adaptive Synthetic Sampling* (ADASYN) (He *et al.*, 2008), muestreo con técnicas de limpieza de datos (SMOTE+Tornek) (Batista *et al.*, 2004), aplicar técnicas de muestreo combinadas con metodologías Boosting (SMOTEBoost) (Chawla *et al.*, 2003b), también combinadas con técnicas Bootsrap como el algoritmo ROSE (Lunardon *et al.*, 2014) o aplicar técnicas *Cluster* (Cluster-Based Sampling Method) (Jo y Japkowicz, 2004).

2. *Modificar el algoritmo*. Este procedimiento está orientado hacia la adaptación de los métodos de aprendizaje automático estándar para estar en sintonía con las cuestiones de desequilibrio de clase, destacan el ajuste en el umbral de precisión o modificar el algoritmo para hacerlo más sensible a la clase minoritaria. También se están desarrollando las técnicas *Kernel-Based Methods* (Hong *et al.*, 2007). Sin embargo, estas técnicas requieren un alto conocimiento del problema y suponen la modificación del algoritmo siendo este un proceso complejo y menos automático que el balanceo de datos o la aplicación de la Matriz de Costes (Zadrozny y Elkan, 2001).

3. *Aprendizaje sensible a los costes*. Este tipo de soluciones incorporan enfoques en el nivel de datos o a nivel algorítmico, en ocasiones también es aplicado en ambos niveles combinados. Esta técnica se fundamenta en asignar un "coste" o "penalización" a los falsos positivos y a los falsos negativos, a través de la denominada Matriz de Costes. Se asignarán "costes" o "penalizaciones" para la clasificación errónea de los ejemplos de la clase positiva con respecto a la clase negativa y, por lo tanto, el algoritmo se orienta para tratar de minimizar los errores de mayor coste/penalización, destaca el empleo de la Matriz de Costes (Zadrozny *et al.*, 2003). De esta forma, el algoritmo lo que minimiza son los costes en lugar de los errores, y los costes estarán balanceados entre las dos categorías.

Como parece que los métodos de generación de datos sintéticos y los métodos de aprendizaje sensibles a los costes dominan a las técnicas de modificación de algoritmo (He y García, 2009), en este trabajo se utilizará un método de generación sintética de datos, concretamente la técnica SMOTE, y otro método de aprendizaje sensible a costes

para estudiar la mejora que se produce en relación con un análisis sin ninguna transformación de los datos originales.

En los siguientes apartados se especifica la técnica SMOTE y la técnica de Matriz de Costes, para ello, se establecen las siguientes notaciones.

Considerando un conjunto de datos de entrenamiento S dado con m ejemplos (es decir, $|S| = m$), entonces se define: $S = \{x_i, y_i\}, i = 1,...,m$, donde $x_i \in X$ es una instancia en el espacio *n-dimensional* de características $X = \{f_1, f_2,.., f_n\}$, y $y_i \in Y = \{1,...,C\}$ es una etiqueta de identidad de clase asociada con la instancia a la instancia x_i.

En particular, $C = 2$ representa el problema de clasificación de dos clases que es el que nos ocupa. Además, si se definen dos subconjuntos $S_{\min} \subset S$ y $S_{maj} \subset S$, donde $S_{\min}$ es el conjunto de ejemplos de clase minoritaria en S, y S_{maj} es el conjunto de ejemplos de clase mayoritaria en S, de modo que $S_{\min} \cap S_{maj} = \{\Phi\}$ y $S_{\min} \cup S_{maj} = \{S\}$. De este modo, cualquier conjunto generado a partir de procedimientos de muestreo en S son etiquetados E, con los subconjuntos disjuntos $E_{\min}$ y E_{maj} que representan las muestras minoritarias y mayoritarias de E, respectivamente, cuando se aplican (He y García, 2009).

III.4.1.- *Synthetic Minority Oversampling Technique* (SMOTE).

La técnica *Synthetic Minority Oversampling Technique* o SMOTE es un método potente que ha demostrado una gran capacidad de éxito en múltiples aplicaciones (Chawla *et al*. 2002). SMOTE proporciona información nueva relativa a la clase minoritaria además de equilibrar ambas clases.

Con esta técnica se consigue balancear un conjunto de datos generando datos artificiales, se crea un conjunto aleatorio de observaciones de la clase minoritaria para cambiar el sesgo de aprendizaje del clasificador hacia la clase minoritaria.

Los datos sintéticos se generan tomando un número determinado de valores de la clase minoritaria que están más cercanos a cada uno de los valores (en este caso, se seleccionan las 5 empresas fraudulentas más próximas a cada una de las empresas fraudulentas por las principales características de la muestra). Cada valor se une a cada una de las empresas más próximas y se selecciona aleatoriamente un punto de este segmento. El punto seleccionado será el nuevo dato sintético que tendrá similitud con los datos de la clase minoritaria pero será diferente, ya que sus características serán una combinación lineal (aleatoria) de los datos originales.

Así, para crear una muestra sintética, el algoritmo selecciona aleatoriamente uno de los *K-vecinos* más cercanos, luego multiplica la diferencia del vector de características correspondiente con un número aleatorio entre $[0,1]$, y finalmente añade este vector a x_i

$$x_{new} = x_i + (\hat{x}_i - x_i)\delta \ ,$$

donde $x_i \in S_{\min}$ es el caso minoritario considerado, $\hat{x}_i$ es uno de los *K-vecinos* más cercanos para $x_i : \hat{x}_i \in S_{\min}$, y $\delta \in [0,1]$ es un número aleatorio. Por lo tanto, la instancia sintética resultante de la expresión anterior es un punto a lo largo del segmento de la línea que une x_i y el *K-vecino* más cercano aleatoriamente seleccionado (He y García, 2009).

III.4.2.- Matriz de Costes

A diferencia de la técnica anterior, la Matriz de Costes (*Cost-Sensitive Learning*) no crea distribuciones de datos balanceadas, sino que busca equilibrar el aprendizaje mediante la aplicación de una Matriz de Costes que describe el coste de la clasificación errónea frente a la correcta (Elkan, 2001; Ting, 2002).

Esta técnica utiliza el coste asociado con la clasificación errónea de las observaciones mediante la aplicación de los pesos específicos de la clase en función de la pérdida (pesos más pequeños para las instancias de la clase mayoritaria y pesos más grandes para las de

la clase minoritaria). Los pesos se pueden configurar para que sean inversamente proporcionales a la fracción de las instancias de la clase correspondiente. En el caso de clasificación binaria el peso de una clase se puede ajustar mediante remuestreo para mejorar el poder predictivo del modelo (Chawla, 2005).

La Matriz de Costes es similar a la matriz de confusión. El objetivo es penalizar los errores (falsos positivos y falsos negativos) frente a los aciertos (verdaderos negativos y verdaderos positivos). Se le asignará un mayor coste a los errores de la clase minoritaria que a los errores de la clase mayoritaria, y no se aplicarán penalizaciones para los aciertos.

Siguiendo el escenario de clasificación expuesto anteriormente (He y García, 2009), se puede definir $C(Min, Maj)$ como el coste de clasificar erróneamente un ejemplo de clase mayoritaria como un ejemplo de clase minoritaria y $C(Maj, Min)$ como el coste del caso contrario.

Normalmente, no hay un coste para la clasificación correcta de ninguna clase y el coste de clasificar incorrectamente los ejemplos de las minorías es más alto que el caso contrario, es decir, $C(Maj, Min) > C(Min, Maj)$. El objetivo del aprendizaje sensible al coste es desarrollar hipótesis que minimicen el coste total en el conjunto de datos de entrenamiento, que es generalmente el riesgo condicionado de Bayes[37].

Al introducir diferentes costes los errores se ponderan en función de sus costes relativos y, por tanto, el objetivo que persigue esta técnica no es tanto minimizar errores sino minimizar costes.

En el caso que nos ocupa, el coste de señalar una empresa que no es fraudulenta como tal, generará unos costes de investigación (costes de personal y de realización de la investigación policial), y seguramente algunos problemas a la empresa. Sin embargo, los costes de no señalar a una empresa fraudulenta serán más elevados, en primer lugar, por los impuestos y sanciones que se dejan de recaudar y, en segundo lugar, por los daños causados por las actividades ilegales que han generado.

[37] *Bayes Conditional Risk.*

PARTE 2

DESCRIPCIÓN DEL PROCESO JUDICIAL Y RESULTADOS OBTENIDOS

1.- Exposición del proceso judicial

1.1.- Descripción de la estructura defraudadora

En este apartado se ofrece, resumidamente, la descripción del entramado empresarial investigado por delito de blanqueo de capitales en el proceso judicial objeto de estudio.

El supuesto entramado defraudador se integra por un elevado número de empresas proveedoras, estructuradas en clanes y organizadas jerárquicamente, coordinadas todas ellas por una única empresa núcleo (no participada directamente en ninguno de los clanes). La empresa núcleo estaría situada por encima del conjunto de empresas proveedoras del entramado defraudador, y habría sido la encargada de comprar ciertos bienes a sus más de 640 proveedores para exportarlos a dos empresas cliente situadas fuera de España. Dicha empresa a su vez formaba parte de un grupo empresarial internacional.

En este sentido, los posibles indicios de irregularidades del entramado defraudador que fueron observabas *a priori* por las autoridades policiales incluían:

1- Las ingentes cantidades de dinero que estaban circulando entre las empresas proveedoras y la empresa núcleo.

2- Los altos precios que ofertaba la empresa núcleo y que alcanzaban cifras muy por encima de los que ofrecían las empresas competidoras.

3- La utilización de canales de distribución opacos para las autoridades financieras y fiscales que dificultaba la identificación de los inversores y, por tanto, el conocimiento del origen de los fondos invertidos.

La disponibilidad de ingentes cantidades de dinero y los elevados precios que ofrecía la empresa núcleo propició que en pocos años esta empresa llegara a comercializar ella misma más de la mitad del total de exportaciones de un determinado producto en todo el territorio español.

Para hacer frente al gran volumen de pagos, las investigaciones policiales realizadas apuntan a que la empresa núcleo habría dispuesto de tres vías de financiación: (1) ingresos procedentes de la actividad realizada en España por la comercialización de ciertos productos, (2) la disposición de créditos comerciales por parte de las empresas clientes extranjeras y (3) la inyección en cuentas nacionales de dinero procedente de otras cuentas de su grupo empresarial en el extranjero, siendo esta tercera vía la más representativa en cuantía. De hecho, de acuerdo con las pesquisas realizadas, las transferencias de dinero presentaban una enorme regularidad y ofrecían a la empresa núcleo una línea de financiación con la que abarcar a proveedores de todo el territorio nacional y copar el sector de actividad en el que trabajó. Una desorbitada cantidad de dinero, procedente de las cuentas de su grupo en el extranjero, habría sido de ese modo introducido en el sistema financiero español.

Al objeto de poder determinar el origen y licitud del dinero procedente del extranjero y poder determinar la posible existencia, respecto de la fuente de financiación de la empresa núcleo, de un circuito de blanqueo de capitales, la autoridad policial cursó solicitud de Comisión Rogatoria a diversos países. Las informaciones recibidas apuntan a que en relación a dicho capital no habría podido ser acreditada su procedencia.

1.2.- Fases del delito de blanqueo de capitales por los bienes comercializados

Por parte de los investigadores policiales, no solo se dudaba del origen legal de los capitales con los que se financió la empresa núcleo, sino que también fue objeto de investigación la procedencia (u origen) de los bienes que se comercializaron. Adicionalmente, mediante la organización del entramado defraudador y la confección de facturación falsa o simulada entre los componentes de dicho entramado, esta organización supuestamente también habría "maquillado" el origen de bienes procedentes de actividades ilícitas o de naturaleza criminal. Se diferenciarían tres fases en el ciclo de blanqueo:

1. *Colocación*. Los grupos de proveedores defraudadores (clanes) habrían justificado el gran volumen de bienes comercializados bajo la cobertura (facturación falseada) de compras efectuadas a otras empresas del entramado fraudulento para dar origen, aspecto y cobertura de veracidad y legalidad a su procedencia, presumiblemente de origen ilícito. Se emplearían estrategias complicadas para alejar los bienes de su origen, cruzando compras y ventas entre empresas del mismo grupo.

2. *Estratificación.* Ocultado el origen real de los bienes, se cerraban y abrían empresas con idéntico objeto social y distintos administradores (aunque dirigidas por los mismos responsables de los clanes) que le habrían servido a la empresa núcleo para eludir el control de Hacienda y obtener la formalidad que esta le exigía, lo que le habría permitido seguir adquiriendo bienes ilícitos.

3. *Integración*: Una vez eliminada toda relación de los bienes comercializados con su origen y procedencia, estos productos eran introducidos en el sistema económico legal mediante la infraestructura perteneciente a la empresa núcleo. La empresa núcleo poseía infraestructura física y de personal que le permitía la reintegración de estos bienes al sistema económico con apariencia legal. Esta empresa se habría encargado de crear la documentación (facturación falsa) necesaria para aparentar, de cara a las autoridades, un ciclo de comercialización de los bienes "transparente". Confeccionadas las facturas a nombre de los proveedores del entramado defraudador, los bienes serían declarados con apariencia de ser adquiridos en el ejercicio de su actividad y eran integrados al circuito legal ya "blanqueados". De este modo, las empresas del entramado podrían disfrutar de los ingresos obtenidos en sus actividades ilegales.

1.3.- Indicios de blanqueo de capitales

La empresa núcleo, en el desarrollo de la actividad económica que realizó en el territorio español durante los años investigados, habría sido presuntamente la máxima responsable del entramado defraudador investigado. Y, por tanto, habría sido conocedora necesaria

de que su actividad era utilizada como instrumento para la defraudación fiscal y el blanqueo de capitales. De acuerdo con las pesquisas realizas, existirían evidencias suficientes para reconocerle, entre otros, los siguientes indicios de blanqueo de capitales:

1. Delito antecedente: el delito financiero.

En el ámbito del Impuesto sobre el Valor Añadido, la venta de los bienes que comercializó la empresa núcleo estaba sujeta al IVA. En este sentido, las compras documentadas con facturas emitidas por sociedades no declarantes (proveedores) dan lugar a la deducción del IVA soportado por el adquirente (empresa núcleo) que no ha sido declarado a la Hacienda Pública. Además, en los casos en que se destina a la exportación, ello da lugar a importantes cuantías de IVA a devolver.

En cuanto a otro posible indicio de fraude se encuentra el presunto fraude fiscal derivado del Impuesto sobre Sociedades. La cuota a ingresar a la Hacienda Pública surge de la aplicación de un tipo impositivo, del 30% en general, a la base imponible declarada. Dada la complejidad del entramado se intuía una irregularidad de las cifras declaradas a la hacienda pública referentes a los "aprovisionamientos" o el coste de los bienes comercializados.

Además, en el supuesto de que los bienes comercializados por el entramado empresarial fuesen de procedencia lícita, el delito financiero hubiese sido articulado siguiendo el llamado "fraude de carrusel de IVA" como sigue:

Una empresa proveedora del entramado habría comprado bienes a empresas radicadas fuera del territorio nacional. Al ser una adquisición intracomunitaria está exenta de IVA. A continuación, esta empresa (llamada "trucha" en el argot) habría vendido los bienes a la empresa núcleo, sin previamente haber ingresado la cuota correspondiente de IVA. Pasado un corto periodo de tiempo la empresa proveedora habría quedado disuelta.

A su vez, la empresa núcleo que compra dichos bienes paga a su proveedor ("trucha") con IVA soportado y exporta dichos productos a una empresa cliente radicada fuera del territorio español. Por último, la empresa núcleo habría reclamado a la Agencia Tributaria

el IVA que supuestamente habría soportado y pagado, dinero que en ningún caso hubiera sido ingresado a la Agencia Tributaria. El fraude se consuma cuando la empresa núcleo que ha comprado el bien en el territorio español hubiese reclamado la devolución del IVA por haberlo soportado.

Las denominadas empresas "truchas" desaparecen rápidamente, o quienes las dirigen las disuelven y constituyen otras, tratando de evitar la trazabilidad de los capitales. Suelen ser sociedades recién constituidas, sin apenas capital ni personal laboral, que carecen de una estructura empresarial real y siempre suelen ser administradas por testaferros o personas insolventes económicamente.

Esta práctica fraudulenta supone, no solo que la Agencia Tributaria no cobrará el IVA de la primera operación, sino que también deberá pagar la devolución solicitada por la empresa núcleo.

Para llevar a cabo este tipo de delito financiero es precisa una estructura coordinada de empresas necesarias (proveedoras y clientes), que bajo la dirección de la empresa núcleo, hubiesen falseado facturas, falseado transacciones y creado empresas ficticias (sociedades recién constituidas, sin apenas capital ni personal laboral, que carecían de una estructura empresarial real), resultando muy complicado seguirles la pista y exigirles responsabilidades.

Adicionalmente al fraude sobre el IVA, las posibles estrategias de defraudación a la Hacienda Pública serían:

La utilización de facturas irregulares para ocultar tanto el origen como la cuantía real del bien adquirido. La ocultación del verdadero suministrador/proveedor que impida contrastar con terceros el importe consignado en las facturas, e incluso, la realidad de la transacción.

La utilización de facturas irregulares que habrían sido utilizadas para justificar ante las entidades bancarias la relación de movimientos de capitales con alguna actividad comercial, en realidad inexistente. Es decir, para documentar como actividad empresarial

el blanqueo de capital. Dado que las propias sociedades proveedoras son las que elaboran sus facturas de compra, los importes reflejados obedecerán a los intereses de las mismas, entre los que puede encontrarse minimizar su beneficio declarado y, con ello, los impuestos a pagar.

Además, como el margen de beneficio que obtienen las empresas del entramado (clanes) y los importes declarados son irrisorios, el margen bruto de beneficio que estos declaran tener sobre el importe de sus adquisiciones es ligeramente superior al 0% de media. Teniendo en cuenta que a dicha cuantía habría que deducir los gastos inherentes a la actividad (personal, alquileres, suministro eléctrico, teléfono) incluso algunos de los proveedores habrían declarado tener pérdidas.

2. Blanqueo de capitales respecto a los bienes comercializados.

En los registros extraídos de la contabilidad interna de la empresa núcleo quedaron registradas las toneladas de producto que fueron comercializadas por dicha empresa. Del total de adquisiciones contabilizadas durante el periodo de investigación, y hasta el día de la intervención policial, ninguno de los clanes o empresas proveedoras investigada habría dispuesto de infraestructura ni medios suficientes para poder llevar a cabo la comercialización de tal cantidad de producto por su cuenta, siendo, por tanto, necesaria la presencia de la empresa núcleo para concluir el circuito comercial.

Además, las empresas proveedoras definidas como fraudulentas representaron un porcentaje alto del volumen total facturado. Estas empresas no habrían estado declarando a Hacienda los productos comercializados en algunos casos o habrían liquidado únicamente parte. En cualquier caso no habrían acreditado que ese producto adquirido en su origen tuviera una procedencia legal. Habrían sido documentadas las adquisiciones con empresas inexistentes o instrumentales, lo que presupondría el origen ilícito de la mayor parte de los bienes comercializados por el entramado defraudador.

Con ello, la empresa núcleo habría estado desoyendo la obligatoriedad que tiene impuesta por imperativo legal como “sujeto obligado” garante de la procedencia de los bienes que comercializó.

3. Blanqueo de capitales en relación al delito de falsedad documental

La facturación "maquillada" habría servido de soporte para documentar los bienes receptados (bienes de procedencia desconocida), siendo confeccionadas en nombre de empresas ficticias o irregulares, lo que claramente supondría un delito de falsedad documental.

4. Blanqueo de capitales en relación al delito de receptación

El conocimiento de que parte de los bienes comercializados que compraba a los proveedores eran de procedencia ilícita se pone de manifiesto en el ciclo de blanqueo que llevó a cabo la empresa núcleo. Esta empresa supuestamente habría sido la encargada de coordinar todas las empresas participantes en este complejo entramado defraudador para confeccionar la documentación ficticia necesaria que le hubiese permitido la integración de los bienes y capitales ilícitos en el sistema económico legal.

Además, a los cuatro indicios de blanqueo de capitales tratados anteriormente, habría que añadir un delito de blanqueo de capitales por financiación ilegal. Con ello, el desconocimiento del origen de las ingentes cuantías de dinero con las que se financió esta empresa pone de manifiesto la relevancia de este proceso judicial. Estaríamos frente a una de las redes de blanqueo de capitales más poderosas económicamente a nivel internacional.

2.- Descripción de la muestra

Una vez obtenida la información a analizar, y después de llevar a cabo el pre-procesamiento de datos[38], se obtiene la base de datos disponible para detectar los patrones de blanqueo de capitales que describen las empresas proveedoras fraudulentas (fraude demostrado).

Se emplearán técnicas de aprendizaje automático para clasificar el resto de empresas proveedoras (fraude sospechoso), de las cuales se desconoce si son fraudulentas o no. Con ello, se persigue orientar la investigación judicial hacia aquellas empresas que tengan mayor probabilidad de haber cometido fraude.

Se dispone de una gran base de datos[39] que contiene casi 300.000 operaciones comerciales realizadas por un entramado de alrededor 650 empresas (la empresa núcleo, los casi 650 proveedores y dos empresas cliente).

Los datos disponibles corresponden a las operaciones de compra de un bien registrados en la contabilidad interna de la empresa núcleo a más de 600 de sus proveedores, y su posterior exportación/comercialización al extranjero.

Se tiene certeza de que 26 de las empresas proveedoras son fraudulentas, es decir, las operaciones que realizan no se ajustan a la legalidad. Por tanto, sólo se dispone de información *a priori* de que aproximadamente el 4% de las empresas son fraudulentas, pero se desconoce si el resto de empresas los son o no.

Del total de las 300.000 operaciones registradas sólo se tiene acceso a la identificación de un pequeño número de operaciones realizadas por empresas fraudulentas (18,99% del total).

Es conviene concretar las siguientes definiciones:

[38] Nótese que la selección de variables en cada uno de los enfoques propuestos se especifica en los apartados correspondientes del Capítulo IV.

[39] La muestra que aquí se describe presenta datos encriptados y codificados para proteger el anonimato de las personas y empresas implicadas en este proceso judicial. Actualmente, el proceso judicial está en fase de investigación por lo que se encuentra bajo el secreto de sumario.

1. *Fraude demostrado*: se considera fraude demostrado a aquellas operaciones realizadas por empresas que hayan podido ser verificadas por la policía como empresas fraudulentas.

2. *Fraude sospechoso*: operaciones realizadas por el resto de empresas no categorizadas.

Por tanto, *a priori* no se tiene conocimiento de qué artículos, personas, lugares e incluso trabajadores están implicados en la trama de blanqueo de capitales. Asimismo, se dispone de las siguientes doce variables disponibles que caracterizan a cada operación registrada[40]:

1. *El artículo*: variable discreta que recoge los 42 tipos de artículos que se comercializaron.

2. *El almacén de la mercancía*: variable discreta que indica el lugar desde donde se realizó la compra de producto al proveedor, 12 localizaciones.

3. *El/la administrativo/a*: persona que gestionó la operación en el departamento de administración, 43 administrativos/as.

4. *El importe total pagado* al proveedor por la adquisición del producto, variable continua; "Importe de la Operación".

5. *La cantidad de material comercializada* en la operación, variable continua; "Cuantía".

6. *El margen bruto de beneficio de la operación*, importe total ingresado por la venta del material al cliente menos el importe total pagado al proveedor, variable continua.

7. *Margen de descuento aplicado en la operación*, variable continua.

[40] Aunque en el curso de la investigación como forense contable se tuvo acceso a información contable adicional del entramado investigado, en esta Tesis Doctoral se muestran únicamente las variables que no comprometen el total anonimato de los intervinientes.

8. *La fecha* de registro de las operaciones.

9. *El ID del proveedor*.

10. *El precio de cotización del producto en la fecha de compra*.

11. *El precio de cotización del producto en la fecha de venta*.

12. *El precio de venta esperado*, valor de venta esperado del producto.

A fin de ilustrar sobre la muestra disponible, a continuación se ofrece un ejercicio descriptivo con el recuento de los registros de las operaciones en función de las variables más representativas[41]. En la Tabla 2.1 se recoge el recuento de operaciones según artículo. Se muestran únicamente los 18 códigos de artículo más frecuentes, que representan el 94,27% de las operaciones. Se ordenan según el número de operaciones de fraude demostrado.

Tabla 2.1
Variable "Código de Artículo"

Código Artículo	Número Operaciones "Fraude Demostrado"	Número de Operaciones	% "Fraude Demostrado"
ART1	16.950	139.877	12,12
ART2	10.258	38.240	26,83
ART3	2.665	6.066	43,93
ART4	2.220	6.714	33,07
ART5	2.070	4.880	42,42
ART6	1.286	2.349	54,75
ART7	1.071	23.462	4,56
ART8	876	2.755	31,80
ART9	597	4.307	13,86
ART10	586	2.248	26,07
ART11	476	6.526	7,29
ART12	467	1.081	43,20
ART13	392	10.227	3,83
ART14	368	3.473	10,60
ART15	364	1.970	18,48
ART16	348	1.686	20,64
ART17	314	1.009	31,12
ART18	297	834	35,61

[41] Para atender a la legalidad del proceso judicial del que proceden los datos no se ofrece una descripción más exhaustiva de la muestra.

Fuente: elaboración propia.

En la Tabla 2.2 se muestra el recuento de operaciones para los 10 almacenes en los que se realizan las operaciones y en la Tabla 2.3 los 19 administrativos que más operaciones gestionaron, representando el 91,18% de las operaciones totales.

Tabla 2.2
Variable "Almacén"

Consigna	Operaciones "Fraude Demostrado"	Número de Operaciones	% "Fraude Demostrado"
ALM1	40.369	217.985	18,52
ALM2	1.284	6.936	18,51
ALM3	929	9.619	9,66
ALM4	912	21.966	4,15
ALM5	230	1.036	22,20
ALM6	79	79	100,00
ALM7	55	181	30,39
ALM8	0	15.241	0
ALM9	0	881	0
ALM10	0	334	0

Fuente: elaboración propia.

Tabla 2.3
Variable "Administrativos"

Administrativo	Operaciones "Fraude Demostrado"	Número de Operaciones	% "Fraude Demostrado"
PEX1	9.829	39.191	25,08
PEX2	7.246	20.494	35,36
PEX3	6.255	37.412	16,72
PEX4	3.903	14.308	27,28
PEX5	2.605	4.609	56,52
PEX6	2.042	12.501	16,33
PEX7	2.005	8.414	23,83
PEX8	1.917	24.947	7,68
PEX9	1.872	9.386	19,94
PEX10	925	4.530	20,42
PEX11	904	4.497	20,10
PEX12	804	2.565	31,35
PEX13	648	1.817	35,66
PEX14	620	10.768	5,76
PEX15	448	1.694	26,45
PEX16	429	12.538	3,42
PEX17	330	698	47,28
PEX18	282	1.463	19,28
PEX19	206	38.239	0,54

Fuente: elaboración propia.

Finalmente, en la Tabla 2.4 se recogen los descriptivos de las variables continuas disponibles. En concreto, se dispone del Importe total de cada operación de comercialización del bien, la Cantidad de material que contiene cada compra y los Márgenes de descuento y de beneficio.

Tabla 2.4
Variables continuas: Importe total de compra, Cantidad de material comprado, Margen de descuento y Margen bruto de beneficio.

	Mín.	**P25**	**Mediana**	**Media**	**P75**	**Máx.**	**Nº NA´s**
Importe total							
Fraude	-0,18	1,03	4,10	36,62	37,11	2.799,15	733,86
Resto	-0,18	1,74	12,91	68,81	68,00	1.818,30	153,30
Cantidad de Material							
Fraude	0	13	78	201	260	18.560	6.517
Resto	0	51	166	277	377	9.980	2.211
Margen de descuento							
Fraude	0	0	1,00	1,00	1,00	1,02	230
Resto	0	0	1,00	0,95	1,00	1,80	79
Margen Bruto de Beneficio							
Fraude	-25,53	0,01	0,03	0,66	0,04	40,60	5.612
Resto	-17,39	0,01	0,03	0,36	0,04	28,64	663

Fuente: elaboración propia.

Como se observa, *a priori* no se puede determinar que patrones se esconden detrás de los registros de las operaciones.

CAPÍTULO IV

RESULTADOS OBTENIDOS

CAPÍTULO IV: RESULTADOS OBTENIDOS

IV.1.- Enfoque 1. Detección del patrón de blanqueo mediante Redes Neuronales de clasificación[42]

Este análisis supone una primera aproximación a la implementación de modelos de Redes Neuronales al trabajo pericial para la detección de operaciones de fraude. Por ello se empleará, basadas en las técnicas de aprendizaje automático, la estructura de Red Neuronal propuesta por Hastie *et al.* (2008): la Red Neuronal de propagación hacia atrás (*Back-Propagation Network).*

Se emplea la Red Neuronal para identificar los patrones de blanqueo de capitales de las empresas proveedoras que así han sido identificadas previamente (fraude demostrado), clasificando al resto de empresas proveedoras como fraudulentas o no fraudulentas (fraude sospechoso).

Dado que se desconoce qué operaciones registradas por las empresas fraudulentas son efectivamente irregulares, lo que se pretende en este enfoque es identificar un patrón en las operaciones de estas empresas y comprobar su similitud con el resto de empresas sospechosas. El objetivo es detectar empresas fraudulentas a través de las operaciones que realizan.

Se incorporan al modelo como predictores las variables que la empresa núcleo registró en su contabilidad de las operaciones que comercializó con cada uno de los proveedores. El objetivo es conseguir un modelo eficiente que no presente tasas de error altas.

[42] Artículo completo en el *ANEXO III.*-Publicaciones. "*Detección de fraude financiero mediante Redes Neuronales de clasificación en un caso real español* ".

IV.1.- Selección de variables.

En este primer enfoque se pretende incorporar la máxima información disponible en el modelo, es decir, el mayor número de predictores posible. Sin embargo, debido a los problemas computacionales que cuantiosa base de datos conlleva, se realizó una selección de variables mediante la técnica de selección hacia delante, siguiendo una metodología *Wrapper* (Kohavi y John, 1995).

Del total de las 12 variables disponibles, finalmente se introdujeron 7 como predictores en el modelo de Red Neuronal de clasificación: "el Tipo de Artículo", "el Almacén", "el/la Administrativo/a", "el Importe de la operación", "el Margen bruto de beneficio de la operación", "la Cantidad de material" y "el Margen de descuento de cada una de las operaciones".

IV.1.2.- Estrategia de muestreo para el conjunto de entrenamiento.

El proceso de ajuste se inicia testando las limitaciones de los recursos informáticos disponibles. Es decir, teniendo en cuenta la magnitud de la base de datos disponible y la estructura de red que se desea aplicar es necesario establecer una estrategia para encontrar un compromiso entre estructura de la red y cantidad de datos a incluir en el conjunto de entrenamiento, que no comprometa la finalidad del análisis. La estrategia seguida es la siguiente:

1. El conjunto de entrenamiento será tal que el desequilibrio en los datos sea comparable a la muestra de operaciones de la policía, en el que la proporción de operaciones de empresas fraudulentas respecto al resto se sitúa en torno al 19%. Es decir, siendo A_m el número de operaciones de empresas fraudulentas de la muestra, B_m el número de resto de operaciones, Ace el número de operaciones fraude del conjunto de entrenamiento y Bce el número de resto de operaciones del conjunto de entrenamiento, se cumple que:

$$\frac{Am}{Bm} \approx \frac{Ace}{Bce} \approx 19\%$$

2. Si bien lo habitual es reservar el 20% de los datos para el conjunto comprobación y utilizar el 80% en el de entrenamiento, en este caso la proporción será la opuesta (80% comprobación, 20% entrenamiento) ya que se prioriza el número de nodos de la capa oculta de la red y la utilización de todas las variables disponibles.

Esta decisión se fundamenta en tres motivos: por una parte, se desea maximizar el aprovechamiento de la estructura de la Red Neuronal; por otra, el objetivo de incluir el número mayor de variables de entrada; finalmente, utilizar menos datos asegura que los resultados obtenidos, en cualquier caso, tiendan a estar por debajo de los que cabría esperar con mayor información (estrategia conservadora adecuada al enfoque de "primera aproximación" que se desea aportar con este trabajo).

Así, el ajuste se realiza manteniendo todos los inputs de entrada de la red (101), 9 nodos en la capa oculta de la red y un número de operaciones de entrenamiento por debajo del habitual (50.000 operaciones, el 20% del total de operaciones disponibles).

3. Además, siguiendo una estrategia conservadora, en el siguiente apartado se presentan los resultados del modelo de red especificado sin aplicar corrección alguna.

La finalidad de esa estrategia es obtener un resultado a partir del cual se puedan aplicar mejoras sucesivas (el balanceado de los datos se aborda en apartados sucesivos) y esperar un mayor ajuste del modelo conforme pudieran superarse las limitaciones informáticas.

IV.1.3.- Resultados del modelo ajustado con datos desequilibrados (Raw Data)

Utilizando el entorno de programación de R (R Core Team, 2015) se ha empleado la paquete de R *"nnet"* (Venables y Ripley, 2002) de la librería del mismo nombre, ajustando sus parámetros conforme a las especificaciones descritas y se evalúa el ajuste del modelo mediante el enfoque tradicional: construyendo una matriz de confusión (Tabla IV.1.1), donde se confrontan las operaciones bien y mal clasificadas basadas en el conjunto de comprobación. En ella se especifican los recuentos utilizados para el cálculo de las tasas con las que evaluar los resultados (Tabla IV.1.2), considerando todas las operaciones de la muestra disponible mediante los siguientes índices: la Tasa Global de Aciertos (*Correctly Classified*), la Tasa de Operaciones de empresas fraudulentas Bien Clasificadas o verdaderos positivos (TP Rate-*True Positive Rate*) y la Tasa de operaciones de empresas fraudulentas Mal Clasificadas o falsos negativos (FN Rate-*False Negative Rate*).

Formalmente sería:

$$CorrectlyClasified = \frac{\left(P_{11} + P_{22}\right)}{\left(P_{11} + P_{12} + P_{21} + P_{22}\right)} 100; TPRate = \frac{\left(P_{22}\right)}{\left(P_{21} + P_{22}\right)} 100; FNRate = \frac{\left(P_{21}\right)}{\left(P_{21} + P_{22}\right)} 100$$

Tabla IV.1.1
Matriz de confusión (1).

		Clasificación según el modelo	
		Resto (*No*)	Fraude (*Yes*)
Clasificación Policial	Resto (*No*)	P_{11}	P_{12}
	Fraude (*Yes*)	P_{21}	P_{22}

Fuente: elaboración propia.

Tabla IV.1.2
Tasas obtenidas a partir de la matriz de confusión

Correctly Classified	72,16%
Incorrectly Classified	27,83%
TP Rate	17,63%
FN Rate	82,36%

Nota: TP Rate = Tasa de Verdaderos Positivos, operaciones de empresas fraudulentas bien clasificadas; FN Rate = Tasa de falsos negativos, operaciones de empresas fraudulentas mal clasificadas; Operaciones bien clasificadas (*Correctly Calssified*). Fuente: Elaboración Propia.

De la Tabla IV.1.2 se desprende que la tasa de ajuste global del modelo es reseñable, con un 72,16% de operaciones bien clasificadas. Esta tasa desciende al 17,63% cuando se trata de operaciones de fraude bien clasificadas. Por su parte, la tasa de falsos negativos es elevada (superior al 80%) aunque hay que interpretarla con cuidado, ya que incluye en su recuento operaciones de empreasas fraudulentas que en en realidad no son irregulares. Asimismo, las operaciones que no han sido detectadas como fraude no dejan de ser sospechosas.

Teniendo en cuenta que no se han utilizado técnicas para compensar el desequilibrio inicial de la base de datos, ni para considerar el coste de los errores de clasificación, el resultado es, al menos, prometedor.

IV.1.4.- Sensibilidad a cambios en el conjunto de entrenamiento

En este apartado se realiza un experimento para evaluar la sensibilidad del modelo a cambios en el conjunto de entrenamiento con dos objetivos. En primer lugar, dado que se ha seguido una estrategia de muestreo por la que se limita la cantidad de información en el entrenamiento de la red al 20% de los datos disponibles es conveniente obtener una medida de la influencia de esta decisión en el ajuste. Esta medida, además, es una magnitud de las posibilidades de mejora del ajuste que no se limita sólo a superar las restricciones de potencia de cálculo informático, sino que puede afrontarse aplicando técnicas de balanceado de los datos mediante remuestreo (tal como se aborda en el apartado siguiente).

En segundo lugar, es un hecho conocido que los pesos que el modelo asigna a los nodos de la red son altamente sensibles a cambios en el conjunto de datos de entrenamiento debido al procedimiento de actualización de gradiente descendente del diseño *Back-propagation*.

La estrategia seguida es la siguiente: se establecen 100 conjuntos de entrenamiento del mismo tamaño que el anterior (50.000 operaciones diferentes cada uno) y con el mismo ratio Ace/Bce (19%). Se ajustan 100 modelos de red con la misma estructura anterior, fijando los mismos pesos iniciales y utilizando todo el conjunto de inputs disponibles (nuevamente, se emplea la función del paquete de R nnet). Para cada modelo se han obtenido los ratios antes definidos a partir de las correspondientes matrices de confusión y sus conjuntos de comprobación.

La Tabla IV.1.3 recoge el valor promedio y la desviación típica del ajuste para los 100 modelos. La variabilidad alrededor de la media (11,2% para los verdaderos positivos y 88,98% para los falsos positivos) se sitúa por debajo de los 4 puntos porcentuales.

Tabla IV.1.3
Ajuste de los 100 modelos de red

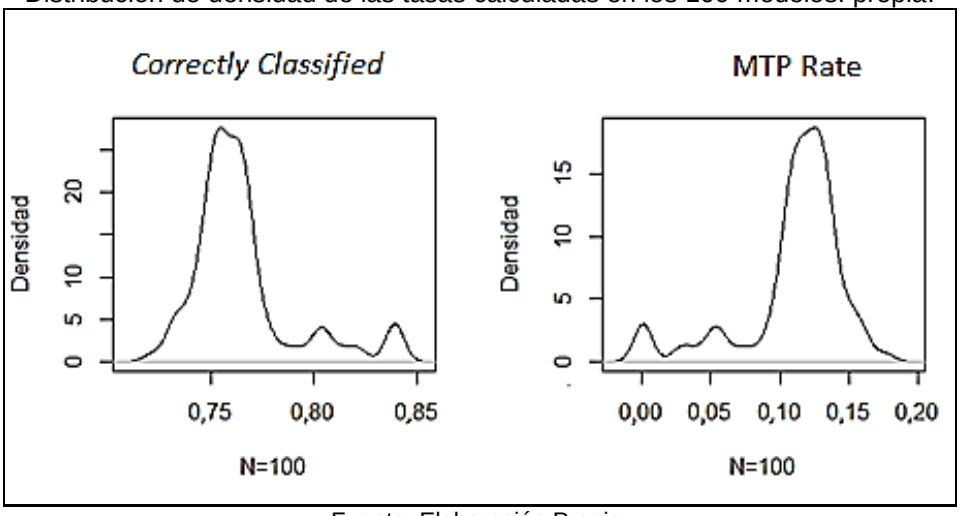
	Media	Desviación Típica	Coeficiente de asimetría
Correctly Classified	76,50%	2,5113%	1,549
Incorrectly Classified	23,50%		
MTP Rate	11,02%	3,697%	1,547
MFN Rate	88,98%		

MTP Rate = Tasa Media de operaciones de empresas fraudulentas Bien Clasificadas.
MFN Rate= Tasa Media de operaciones de empresas fraudulentas Mal Clasificadas.

$$MTPRate = \sum_{i=1}^{100} \frac{\left(P_{22}^{i}\right)}{\left(P_{21}^{i} + P_{22}^{i}\right)}; MFNRate = \sum_{i=1}^{100} \frac{\left(P_{21}^{i}\right)}{\left(P_{21}^{i} + P_{22}^{i}\right)}.$$

Fuente: Elaboración Propia.

Figura IV.1.1
Distribución de densidad de las tasas calculadas en los 100 modelos. propia.



Fuente: Elaboración Propia.

En el experimento realizado para detectar la sensibilidad del ajuste a cambios en el conjunto de entrenamiento, se observa que la distribución de verdaderos positivos es asimétrica (Figura IV.1.1) debido a un grupo de 14 modelos cuya proporción cae por debajo del 7,5% sin los que se obtiene que la tasa de verdaderos positivos $TPRate \sim N(12,3\%,1,7\%)$. La asimetría indicaría que la exclusión/inclusión de casos en el conjunto de entrenamiento no es neutral: la estrategia de selección aleatoria deja margen para la mejora.

IV.1.5.- Resultados del modelo mediante balanceado del conjunto de entrenamiento (SMOTE)

En el caso que nos ocupa cabe destacar que la proporción de operaciones de empresas fraudulentas en la muestra es muy inferior respecto del resto de operaciones y que la estrategia de selección del conjunto de entrenamiento no es neutral.

Las técnicas de muestreo permiten compensar la falta de información de la clase minoritaria (operaciones de empresas fraudulentas) en el conjunto de entrenamiento, a la vez que implican una forma alternativa para considerar en el proceso de aprendizaje los costes asociados a los errores de clasificación. Una combinación de sobremuestreo de la clase minoritaria e inframuestreo de la mayoritaria puede ser la estrategia con mejores resultados (Japkowicz, 2000).

Al aplicar la técnica de SMOTE[43] al mismo modelo del Apartado IV.1.3, a partir de su matriz de confusión (Tabla IV.1.4) se obtienen los resultados recogidos en la Tabla IV.1.5.

Tabla IV.1.4
Matriz de confusión al aplicar SMOTE.

		Clasificación según el modelo	
		Resto (*No*)	Fraude (*Yes*)
Clasificación Policial	Resto (*No*)	41,72%	42,33%
	Fraude (*Yes*)	4,88%	11,05%

Fuente: elaboración propia.

Tabla IV.1.5
Tasas obtenidas a partir de la matriz de confusión al aplicar SMOTE.

Correctly Classified	52,77%
Incorrectly Classified	47,22%
TP Rate	69,36%
FN Rate	30,63%

Nota: TP Rate = Tasa de Verdaderos Positivos, operaciones de empresas fraudulentas bien clasificadas; FN Rate = Tasa de falsos negativos, operaciones de empresas fraudulentas mal clasificadas; Operaciones bien clasificadas (*Correctly Calssified*). Fuente: Elaboración Propia.

[43] Se utiliza la función SMOTE del paquete de R "DMwR" (Torgo, 2010).

La mejora obtenida es reseñable por varios motivos. En primer lugar, destaca el descenso de la tasa de falsos negativos (FP Rate), de 82,36% del modelo con un conjunto de entrenamiento desequilibrado (Tabla IV.1.2) a 30,63% (Tabla IV.1.5). En segundo lugar, la proporción de operaciones de empresas fraudulentas bien clasificadas (TP Rate) asciende del 17,66% del modelo inicial al 69,36% en el modelo con SMOTE.

En la Tabla IV.1.5 se observa también un claro descenso de las operaciones bien clasificadas de forma global (*Correctly Classified*) (52,77%) en este nuevo modelo en relación a los resultados de la Tabla IV.1.2 (72,16%). Sin embargo, tal como se discute en los párrafos anteriores lo relevante es acertar bien las operaciones de empresas fraudulentas, dado que las operaciones del resto de empresas en realidad pudieran serlo.

Por tanto, el uso de estos modelos en la detección de casos reales de fraude financiero, junto con la implementación de técnicas de muestreo como SMOTE que mejoren el desequilibrio de los datos, puede servir de herramienta a los investigadores para conocer patrones de comportamiento en casos de blanqueo de capitales y ofrecer mayor información para su detección, así como también orientar a las autoridades a aquellas empresas cuyos comportamientos aparentan ser fraudulentos.

2.- Enfoque 2. Aplicación de la Ley de Benford y técnicas de aprendizaje automático para la detección de patrones de blanqueo de capitales[44].

En este segundo enfoque se propone una aproximación a la detección de patrones más ambiciosa que la anterior. Se combina la Ley de Benford, como herramienta para caracterizar los registros contables de las operaciones comerciales entre la empresa núcleo y las empresas proveedoras, con cuatro modelos de clasificación: Regresión Logística Ridge (LG), Red Neuronal (NN), Árbol de Decisión C4.5 (DT) y Bosque Aleatorio (RF), para identificar otros proveedores potencialmente fraudulentos (caracterizados mediante las operaciones que realizaron).

A diferencia del enfoque anterior, en esta propuesta lo que se incorpora al modelo no son las variables que definen las operaciones que realizó cada empresa proveedora, sino que se pretende clasificar a las empresas sospechosas en función de su ajuste a la Ley de Benford.

En el modelo son incorporados como predictores el número de operaciones de cada empresa (variable "Frequency") y los p-valores de los ajustes de los primeros dos dígitos de la Ley de Benford de las distribuciones de la variable "Importe de la Operación". Como variable respuesta la consideración por parte de la policía de cada operación como perteneciente a empresa proveedora fraudulenta o no perteneciente (variable "Invest*"*).

Dada la predisposición de las empresas que cometen fraude financiero y blanqueo de capitales a la generación del máximo número de operaciones posibles con el objetivo de ocultar entre ellas su estrategia de fraude (Dionysios, 2011), se ha optado por incluir "el recuento del número de operaciones de cada empresa" (variable "Frequency") que, como se observará más adelante, es una de las que mayor correlación presenta con la variable respuesta ("Invest").

[44] Artículo completo en el *ANEXO IV.*-Publicaciones*:* "*Combining Benford's Law and Machine Learning to detect Money Laundering. An actual Spanish Court case".*

El ajuste a la Ley de Benford de cada empresa proveedora se proyecta en un espacio 20-dimensional de p-valores obtenidos del Z-Test de la Ley de Benford y del OverBenford Test (Apartado III.3.2), donde P1 a P9 denota los p-valores asociados a los primeros dígitos, S0 a S9 los vinculados a los segundos dígitos y V1.y al OverBenford Test.

IV.2.1.- Selección de variables

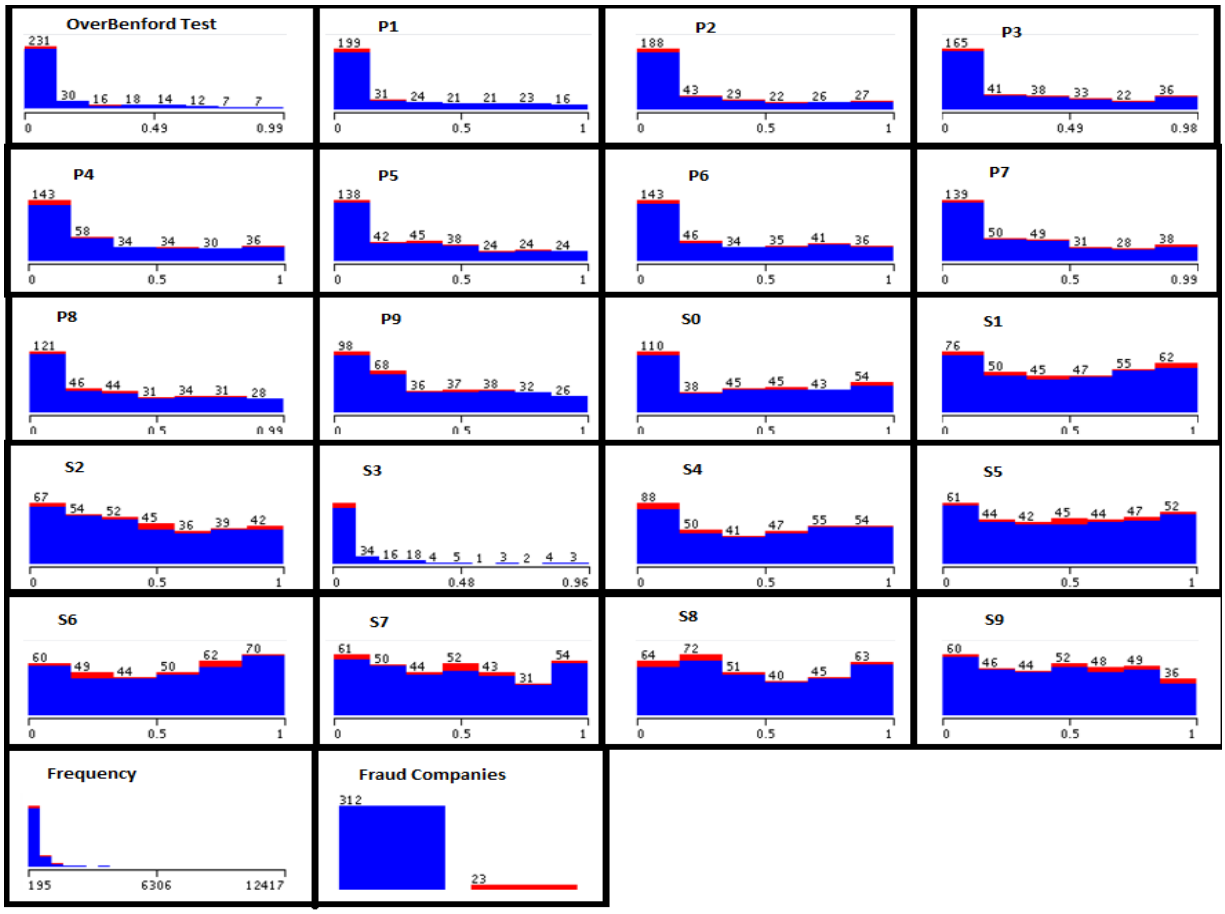
Se dispone para cada una de las empresas 21 variables independientes: la medida de ajuste obtenida mediante el OverBenford Test (V1.y), los 9 (P1 a P9) y los 10 (S0 a S9) p-valores del Estadístico Z de cada uno de los primeros y segumdos dígitos y la frecuencia de operaciones de las empresas analizadas.

Con ello, el objetivo es analizar si existe algún tipo de patrón de comportamiento que diferencie a las empresas que cometen fraude de las que no lo hacen.

La selección de variables se efectúa realizando una selección previa entre el conjunto de predictores potenciales utilizando el algoritmo *Ranker Search Method*[45] (Frank, Hall, y Witten, 2016) del software libre Weka; el cual se basa en las correlaciones entre predictores y la variable respuesta dentro del grupo de métodos integrados de selección de variables.

[45] Eibe Frank, Mark A. Hall, and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Figura IV.2.1
Distribución de Variables



Fuente: Elaboración propia.

La Figura IV.2.2 ofrece los predictores del modelo ordenados por grado de relación con la variable respuesta. Del total de predictores considerados inicialmente, se incluyen finalmente en los modelos predictivos aquellos que superan el valor 0.05. En total se seleccionan 11 predictores.

Figura IV.2.2
Ránking de correlación de predictores.

Frequency	F7	S4	S9	S8	OverBenford	S3	F3	F9	S2	F5
0.3157	0.1392	0.1348	0.1281	0.106	0.1011	0.0911	0.0713	0.0678	0.0658	0.0573
S0	F2	F8	F6	S7	S5	S6	F1	F4	S1	
0.0424	0.0389	0.036	0.0308	0.0274	0.0226	0.0198	0.0189	0.0166	0.0137	

Fuente: Elaboración Propia.

Además, como consecuencia del alto número de operaciones realizadas por las empresas proveedoras fraudulentas (la media de operaciones de una empresa fraudulenta es de 2.042,09 operaciones, mientras que la del resto de empresas es de 635,45 operaciones) también se reduce el número de empresas incluidas en el análisis, de modo que el número de operaciones de cada empresa (variable "frecuencia") mínimo para que una empresa se incorpore en el análisis es de 195.

De esta forma, de la inmensa base de datos disponible, finalmente se incluye en la muestra 335 empresas proveedoras, de las cuales los expertos han identificado como fraudulentas 23. Sólo un 6,87% de las instancias pertenecen a la clase minoritaria, estando claramente ante un conjunto con datos desequilibrados, de nuevo, será oportuno aplicar estrategias de balanceado a la muestra.

Tras la selección de predictores y de empresas, la base de datos disponible queda compuesta por 335 empresas (245.227 operaciones), 11 predictores y una variable respuesta. Sobre estos datos es sobre los que se ha aplicado la metodología propuesta.

A continuación se exponen los resultados obtenidos con los cuatro modelos, los cuales se podrían diferenciar en dos niveles:

(1) Evaluación de los modelos en relación con los diferentes tratamientos efectuados en los datos desequilibrados, aplicación del método SMOTE y Matriz de Costes.

(2) Análisis de la sensibilidad de la predicción para la transformación de los datos originales mediante SMOTE.

En el primer nivel, y dadas las pocas empresas que se han identificado como fraudulentas, se ha optado por utilizar validación cruzada para determinar el grado de precisión de la clasificación de los diferentes modelos propuestos.

En el segundo caso, se dividirá el conjunto inicial de datos en dos subconjuntos, uno de entrenamiento 70% y otro de comprobación (30%). La selección se realiza de forma aleatoria, por lo que se ha repetido 10 veces la selección, creando 10 conjuntos diferentes de entrenamiento y otros 10 conjuntos de comprobación. Sobre los conjuntos de

entrenamiento se ha efectuado una transformación de los datos mediante el algoritmo SMOTE para balancear las categorías de la variable objetivo.

IV.2.2.- Evaluación de los modelos

En este apartado se muestran los resultados obtenidos por los modelos de clasificación siguiendo diferentes estrategias de tratamiento de datos desequilibrados:

(1) Modelizar los datos sin aplicar ningún tipo de transformación (Raw Data).

(2) Aplicar aprendizaje sensible a los costes (Matriz de Costes)

(3) Generar datos sintéticos mediante el algoritmo SMOTE para la balancear la variable dependiente.

Los resultados se muestran en una tabla como sigue:

Tabla IV.2.1
Tabla de resultados

	LG		***DT***		***NN***		***RF***	
	NO	***YES***	***NO***	***YES***	***NO***	***YES***	***NO***	***YES***
NO	P^{LG}_{11}	P^{LG}_{12}	P^{DT}_{11}	P^{DT}_{12}	P^{NN}_{11}	P^{NN}_{12}	P^{RF}_{11}	P^{RF}_{12}
YES	P^{LG}_{21}	P^{LG}_{22}	P^{DT}_{21}	P^{DT}_{22}	P^{NN}_{21}	P^{NN}_{22}	P^{RF}_{21}	P^{RF}_{22}
Correctly Classified								
Incorrectly Classified								
TN Rate (No)								
TP Rate (Yes)								
FN Rate (Yes)								
FP Rate (No)								

Fuente: Elaboración Propia.

Donde las cuatro primeras filas corresponden a los resultados obtenidos con cada uno de los cuatro modelos, Regresión Logística Ridge (LG), Árbol de Decisión (DT), Red Neuronal (NN) y Bosque Aleatorio (RF), presentados en forma de matriz de confusión (Véase Tabla IV.1.1).

Las siguientes filas de la tabla se corresponden con las tasas de aciertos y errores de clasificación de cada modelo según las siguientes definiciones:

Correctly Classified: Tasa Global de Aciertos; $P_{11}+ P_{22}/ (P_{22}+ P_{21} +P_{11} P_{12})$

Incorrectly Classified: Tasa Global de Errores; $P_{21}+ P_{12}/ (P_{22}+ P_{21} +P_{11} P_{12})$

TN Rate (No): Tasa de Verdaderos Negativos*;* $P_{11}/ (P_{11}+ P_{12})$

TP Rate (Yes): Tasa de Verdaderos Positivos*;* $P_{22}/ (P_{21}+ P_{22})$

FN Rate (Yes): Tasa de Falsos Negativos*;* $P_{21}/ (P_{21}+ P_{22})$

FP Rate (No): Tasa de Falsos Positivos*;* $P_{12}/ (P_{11}+ P_{12})$

IV.2.2.1.- Resultados con datos desequilibrados (Raw Data)

Una vez efectuada la clasificación de las operaciones de los cuatro modelos, y sin realizar ninguna transformación sobre los datos, se obtienen los siguientes resultados mediante validación cruzada:

Tabla IV.2.2
Matriz de confusión de los modelos sin transformación

	LG		***DT***		***NN***		***RF***	
	NO	***YES***	***NO***	***YES***	***NO***	***YES***	***NO***	***YES***
NO	311	1	302	10	301	11	312	0
YES	20	3	17	6	15	8	19	4
Correctly Classified	93,73%		91,94%		92,24%		94,33%	
Incorrectly Classified	6,27%		8,06%		7,76%		5,67%	
TN Rate (No)	99,68%		96,79%		96,47%		100,00%	
TP Rate (Yes)	13,04%		26,09%		34,78%		17,39%	
FN Rate (Yes)	86,96%		73,91%		65,22%		82,61%	
FP Rate (No)	0,32%		3,21%		3,53%		0,00%	

Fuente: Elaboración Propia.

La capacidad predictiva de los modelos es muy alta, pero se presentan unas tasas de verdaderos positivos muy bajas (TP Rate), entre el 13,04% obtenido por la Regresión Logística y el 34,78% obtenido en la Red Neuronal. Al tener unos datos tan desequilibrados, los algoritmos, debido a las medidas de precisión utilizadas, tienden a favorecer la clasificación en la categoría dominante, identificando muy pocas empresas fraudulentas.

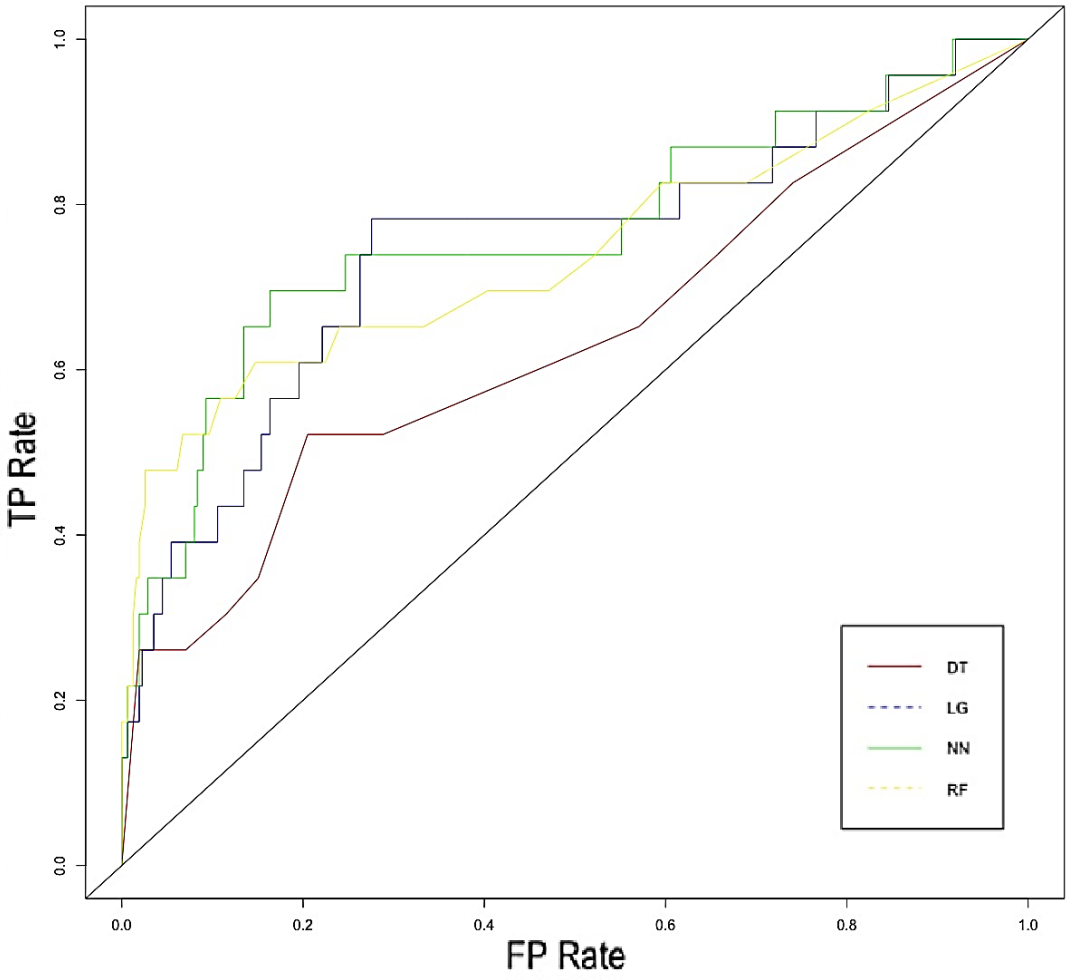
Estos resultados pueden ser mostrados gráficamente empleando la Curva ROC (*Receiver Operating Characteristic*). La Curva ROC, o Curva de Rendimiento Diagnóstico, es una técnica estándar utilizada para conocer el rendimiento de un modelo de clasificación entre las tasa de positivos correctamente clasificados (TP Rate) y la tasa de negativos incorrectamente clasificados como positivos (FP Rate) (Swets, 1988).

En una curva ROC, el eje X representa el porcentaje de falsos positivos o Tasa de Especificidad (FP Rate) y el eje Y representa el porcentaje de verdaderos positivos o Tasa de Sensibilidad (TP Rate). El punto ideal en la curva ROC sería 1, es decir, todos los ejemplos positivos se clasifican correctamente y no se clasifican erróneamente los ejemplos negativos como positivos.

Esta Curva es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación. Así, el área bajo la curva (AUC- *Area Under the Curve*) es una medida de rendimiento de los modelos de clasificación ampliamente aceptada (Bradley, 1997; He y García, 2009).

De forma visual, los resultados obtenidos en los clasificadores con datos desequilibrados (Raw Data) son los que se muestran en el siguiente Gráfico IV.2.1.

Gráfico IV.2.1
Curva ROC con datos desequilibrados (Raw Data)



Fuente: Elaboración Propia.

Con datos desequilibrados, el área bajo la curva (AUC) que define el Árbol de Decisión (Línea Roja) es la más pequeña y, por tanto, representaría al clasificador menos eficiente. En general, los cuatro clasificadores muestran una curva irregular que se aleja del punto óptimo conforme avanza el eje X. Estos resultados muestran la baja tasa de verdaderos positivos (TP Rate), entre el 13,04% que alcanza la Regresión Logística (LG) y el 34,78% que obtiene la Red Neuronal (NN), y la baja Tasa de Falsos Positivos (FP Rate), que se sitúa entre el 0% (RF) y el 3,53% (NN).

IV.2.2.2.- Aplicación de la Matriz de Costes

Se parte del supuesto que los costes de una incorrecta clasificación son diferentes. Los falsos positivos únicamente tendrían el coste de realizar la investigación correspondiente hasta determinar su clasificación errónea, sin embargo, los falsos negativos supondrían un coste mucho mayor. La Matriz de Costes permitirá balancear la variable objetivo sin necesidad de realizar ningún tipo de transformación en los datos. Se ha aplicado una Matriz de Costes con pesos el valor inverso del desequilibrio de las clases.

Los resultados que se han obtenido son los siguientes:

Tabla IV.2.3
Matriz de confusión de los modelos con Matriz de Costes

	LG		***DT***		***NN***		***RF***	
	NO	***YES***	***NO***	***YES***	***NO***	***YES***	***NO***	***YES***
NO	234	78	290	22	285	27	309	3
YES	10	13	16	7	15	8	17	6
Correctly Classified	73,73%		88,66%		87,46%		94,03%	
Incorrectly Classified	26,27%		11,34%		12,54%		5,97%	
TN Rate (No)	75,00%		92,95%		91,35%		99,04%	
TP Rate (Yes)	56,52%		30,43%		34,78%		26,09%	
FN Rate (Yes)	43,48%		69,57%		65,22%		73,91%	
FP Rate (No)	25,00%		7,05%		8,65%		0,96%	

Fuente: Elaboración Propia.

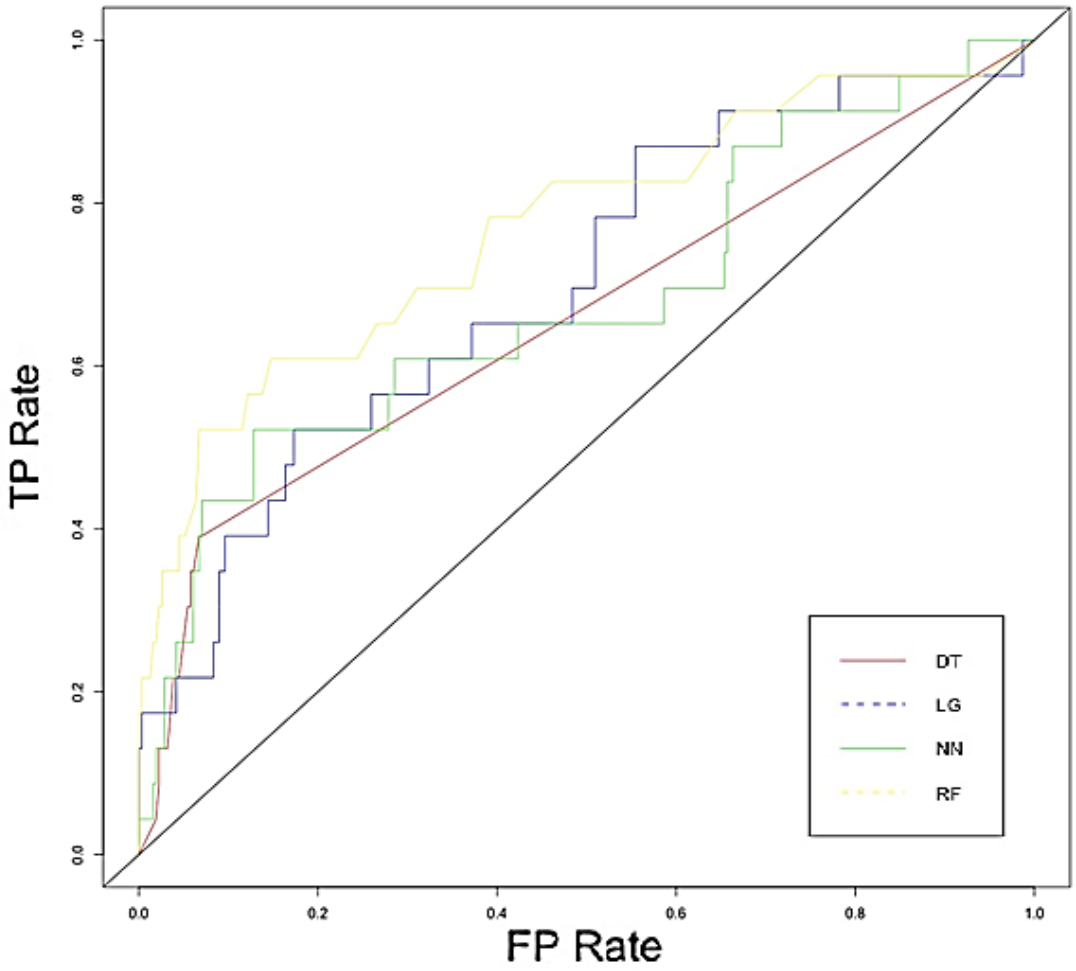
Si se comparan los resultados con los del apartado anterior, inmediatamente se identifica una bajada importante de la precisión del modelo. El Bosque Aleatorio es el único que mantiene un 94% de instancias correctamente clasificadas, reduciéndose en el resto de los casos hasta el 73,73% del Árbol de Decisión.

No obstante, la tasa de verdaderos positivos (TP Rate) se ha mejorado sustancialmente. Esta mejora se ha debido a que el algoritmo identifica correctamente más empresas como posibles defraudadoras por la inclusión de la Matriz de Costes.

Esta metodología ha incrementado la detección de verdaderos positivos a cambio de elevar considerablemente los falsos positivos. En el único caso que esto no ocurre es en el Bosque Aleatorio, que es el que menos verdaderos positivos identifica aunque también menos positivos.

Como se observa en el Gráfico IV.2.2, el área bajo la curva (AUC) que define el Bosque Aleatorio (Línea amarilla) es la mayor y, por tanto, representa al clasificador más eficiente. En general, los cuatro clasificadores mejoran respecto al apartado anterior (Raw Data), la tasa de verdaderos positivos. Sin embargo, también aumentan la tasa de falsos positivos, por lo que las curvas que definen se encuentran todavía alejadas del punto óptimo.

Gráfico IV.2.2
Curva ROC con Matriz de Costes



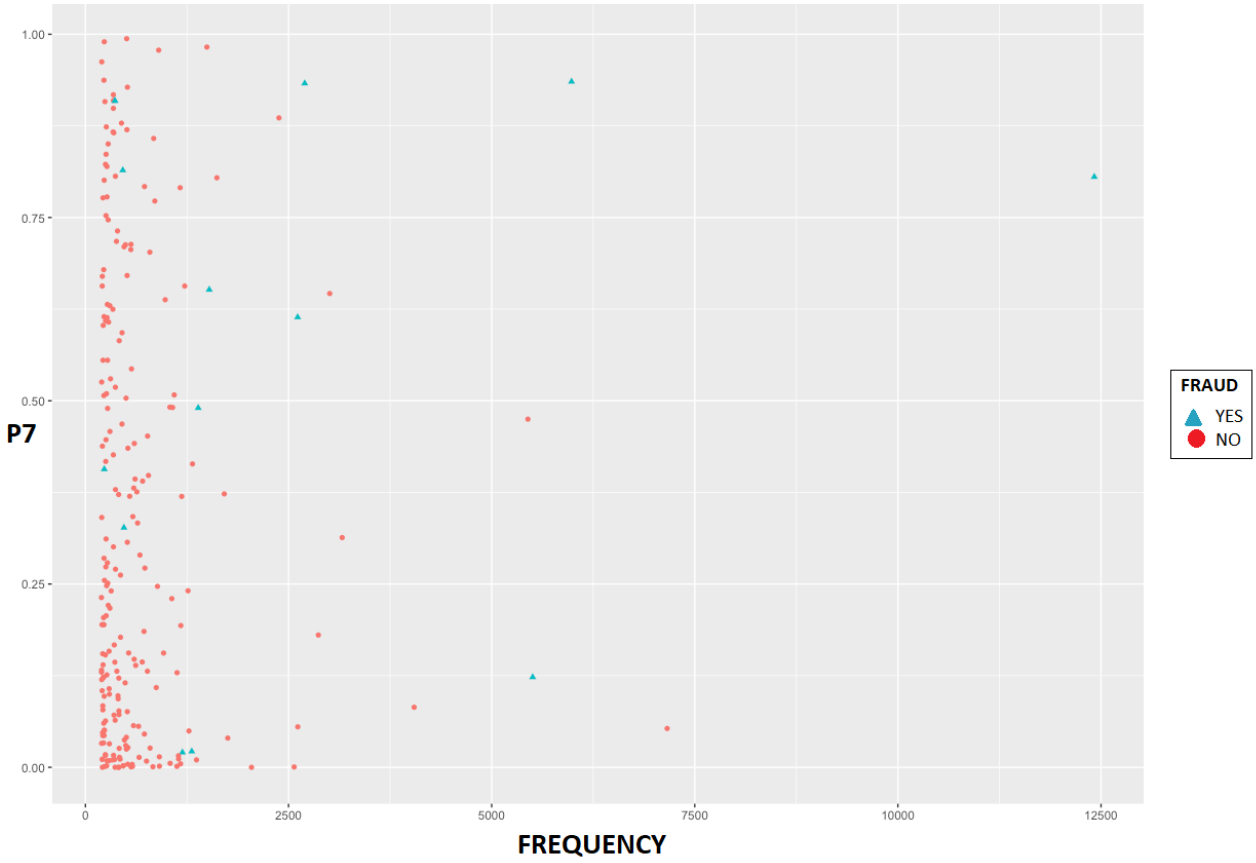
Fuente: Elaboración Propia.

IV.2.2.3.- Resultados mediante la aplicación de SMOTE

Una de las transformaciones de los datos más utilizadas para el balanceo de las categorías es el algoritmo SMOTE. Partiendo de los anteriores datos, donde había 312 empresas sospechosas y 23 no legales, con esta técnica se han generado datos sintéticos de empresas fraudulentas hasta un total de 299. De esta forma se tiene un 51,06% de "*No*" y un 49,94% de "*Yes*". Sobre este nuevo conjunto de entrenamiento se han creado los nuevos modelos con las técnicas ya expuestas.

Los siguientes Gráficos IV.2.3 y IV.2.4 muestran la correlación entre las variables de mayor relación con la variable objetivo para cada una de las empresas sospechosas (Fraud "*No*") y de las empresas fraudulentas (Fraud "*Yes*") antes y después de aplicar el algoritmo SMOTE.

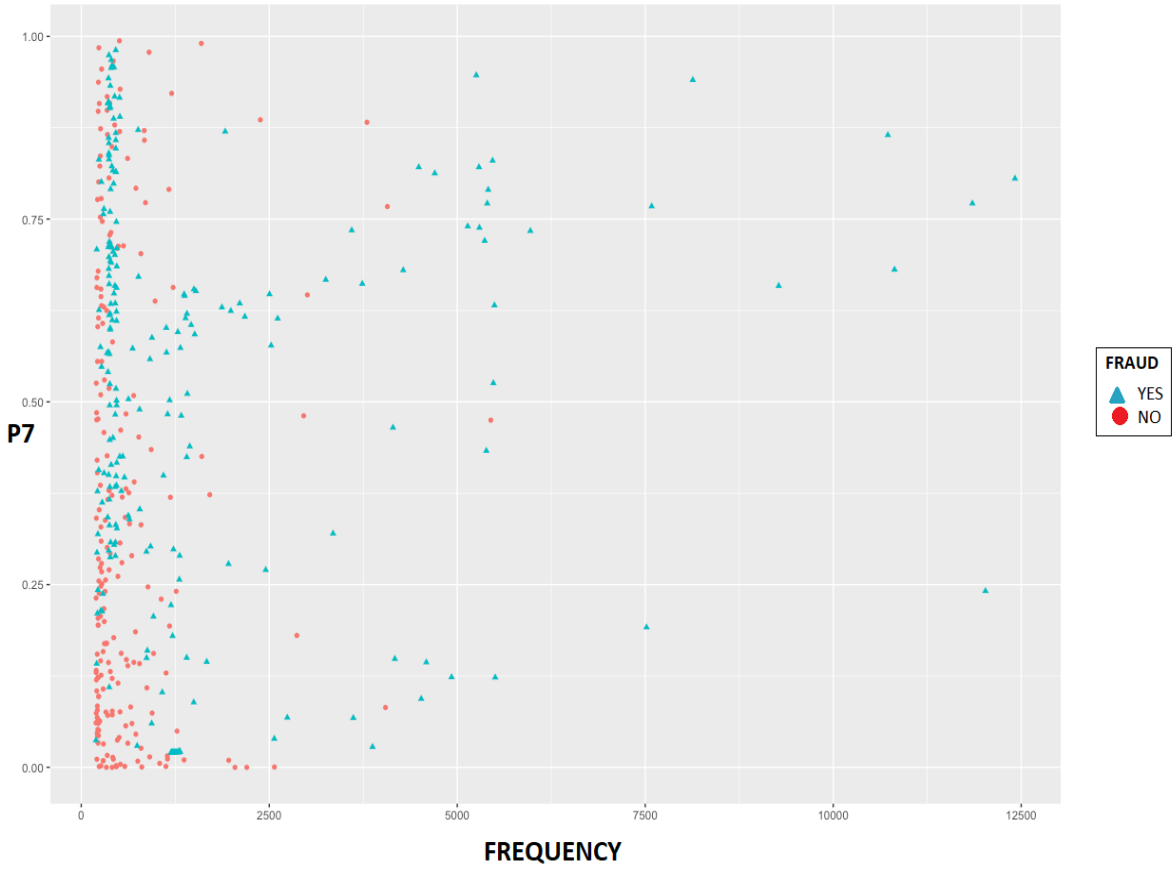
Gráfico IV.2.3
Correlación entre la variable "P7" y la variable "Frequency" con datos desequilibrados



Fuente: Elaboración Propia.

Teniendo en cuenta la distribución de los datos, se puede apreciar un claro predominio de empresas sospechosas ("*No*"). Llama la atención la escasa correlación entre las variables "P7" y "Frequency", distribuyéndose los datos de forma vertical muy próximos al eje Y.

Gráfico IV.2.4
Correlación entre la variable “P7” y la variable “Frequency” al aplicar SMOTE



Fuente: Elaboración Propia

Una vez originados los datos sintéticos mediante el algoritmo SMOTE, destaca la existencia de un predominio mucho más equitativo de la clase minoritaria. Sin embargo, a pesar de modificar la distribución de los datos la correlación entre ambas variables no se ve altamente modificada, se aprecia cómo el algoritmo trata de mantener la estructura inicial de los datos.

Los resultados obtenidos se muestran a continuación:

Tabla IV.2.4
Matriz de confusión de los modelos con SMOTE

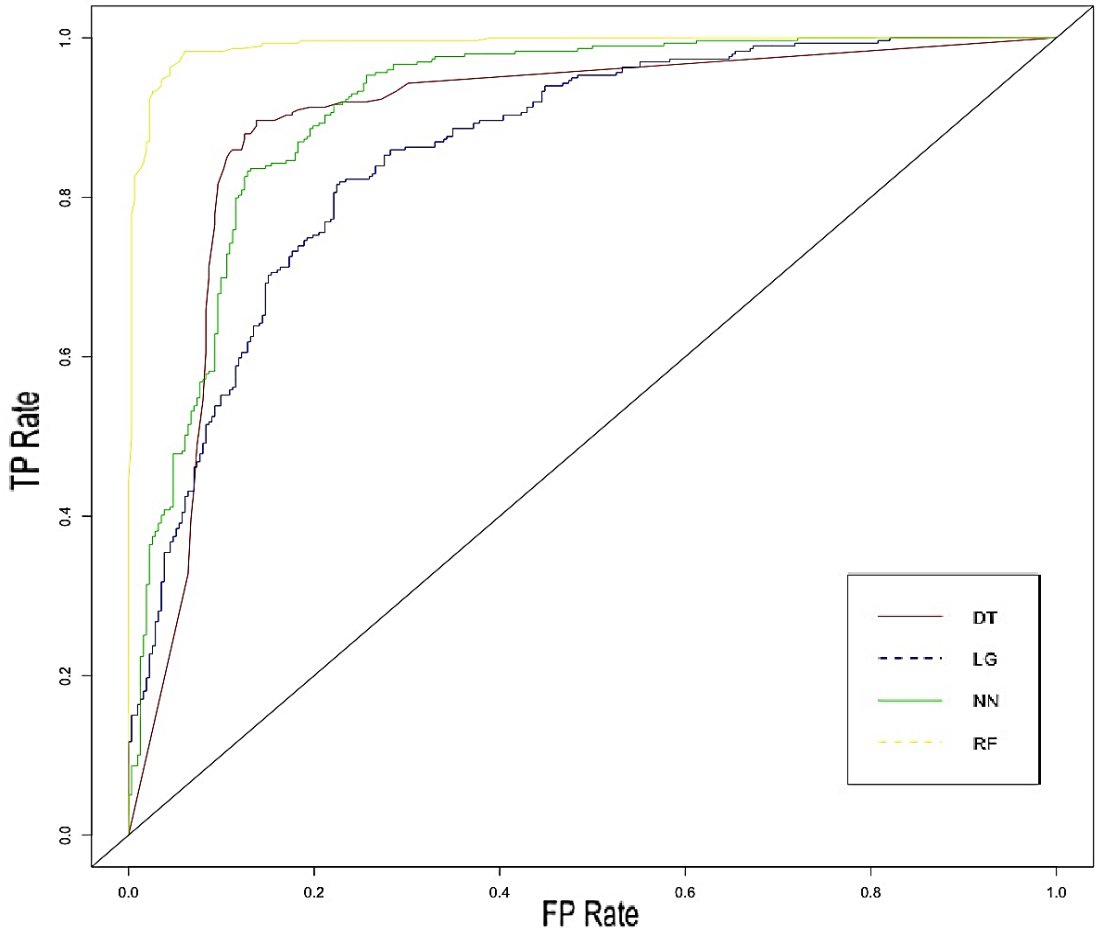
	LG		***DT***		***NN***		***RF***	
	NO	***YES***	***NO***	***YES***	***NO***	***YES***	***NO***	***YES***
NO	239	73	269	43	252	60	300	12
YES	53	246	31	268	38	261	15	284
Correctly Classified	79,38%		87,89%		83,96%		95,58%	
Incorrectly Classified	20,62%		12,11%		16,04%		4,42%	
TN Rate (No)	76,60%		86,22%		80,77%		96,15%	
TP Rate (Yes)	82,27%		89,63%		87,29%		94,98%	
FN Rate (Yes)	17,73%		10,37%		12,71%		5,02%	
FP Rate (No)	23,40%		13,78%		19,23%		3,85%	

Fuente: Elaboración Propia.

Los resultados no mejoran sustancialmente la capacidad predictiva de los modelos en comparación con la utilización de la Matriz de Costes. No obstante, la tasa de verdaderos positivos ha mejorado en todos los casos. Con los datos originales se obtenían tasas de verdaderos positivos (TP Rate) entre el 13,04% de la Regresión Logística y el 34,78% de la Red Neuronal. A su vez, con la Matriz de Costes se mejoraron los resultados hasta alcanzar entre el 26,09% de acierto del Bosque Aleatorio y el 56,52% de la Regresión Logística (que es la que más mejora). Finalmente, con la generación de los datos SMOTE, las tasas de verdaderos positivos oscilan entre el 82,27% de la Regresión Logística y el 94,98% del Bosque Aleatorio.

La capacidad para identificar empresas ilegales de esta tercera metodología es muy superior a las dos anteriores, obteniéndose en el caso del Bosque Aleatorio un resultado muy satisfactorio. De 611 instancias, únicamente clasifica de forma incorrecta 27 (4,42%), de las cuales 15 son falsos negativos y 12 falsos positivos.

Gráfico IV.2.5
Curva ROC con SMOTE



Fuente: Elaboración Propia.

Comparativamente con los dos gráficos anteriores (Gráfico IV.2.1 y Gráfico IV.2.2), se puede apreciar en el Gráfico IV.2.5 como al introducir la técnica SMOTE las Curvas ROC que definen los clasificadores mejoran sustancialmente, aproximándose exitosamente en todos los casos al punto óptimo. El Bosque Aleatorio (Línea Amarilla) es el clasificador que define una mayor área bajo la curva (0,989) y, por tanto, es el clasificador más eficiente.

IV.2.2.4.- Medidas para la evaluación de los modelos

La evaluación de los modelos y las técnicas de balanceado se han realizado utilizando las medidas del área ROC, el estadístico Kappa y el RMSE (Root Mean Squared Error).

Tabla IV.2.5
Medidas de precisión de los modelos

	Sin Transformación			Matriz de Costes			SMOTE		
	ROC	***Kappa***	***RMSE***	***ROC***	***Kappa***	***RMSE***	***ROC***	***Kappa***	***RMSE***
LG	0,747	0,2061	0,236	0,711	0,3243	0,4227	0,844	0,5675	0,4012
DT	0,635	0,2664	0,2702	0,615	0,2086	0,332	0,894	0,7348	0,3499
NN	0,765	0,34	0,2578	0,63	0,2104	0,3306	0,926	0,7252	0,3392
RF	0,74	0,2817	0,2268	0,773	0,3499	0,2415	0,989	0,9116	0,2088

Fuente: Elaboración Propia.

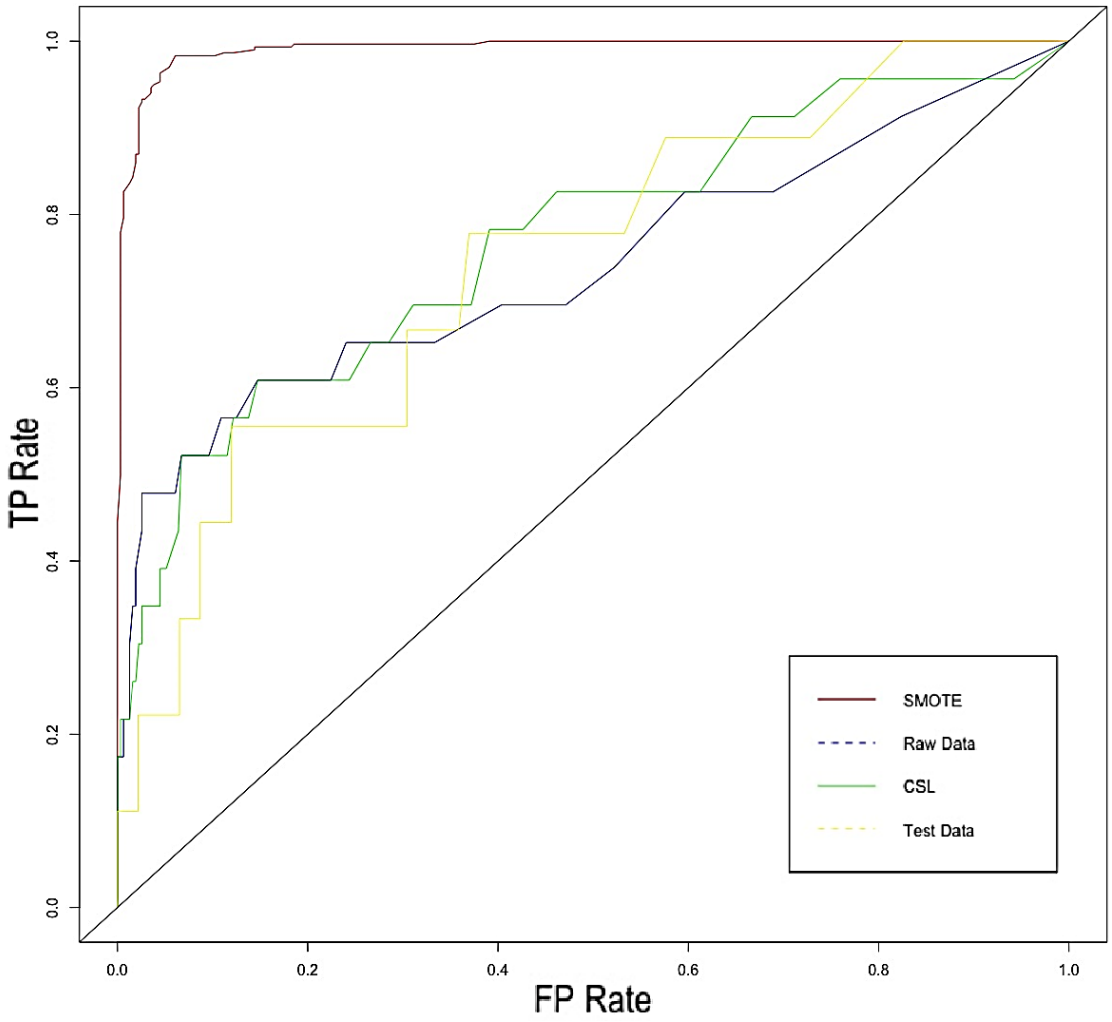
De la tabla anterior (Tabla IV.2.5) se deduce que los mejores resultados se obtienen con el algoritmo SMOTE frente a los datos sin transformar y la aplicación de la Matriz de Costes.

En cuanto a la técnica de clasificación, el mejor modelo es el Bosque Aleatorio con un área ROC de 0,989 y un estadístico Kappa de 0,9116, en ambos casos muy próximos a 1, y el menor RMSE de todos los modelos aplicados.

Si observamos los resultados obtenidos en los cuatro clasificadores para cada uno de los 3 escenarios analizados, se aprecia como la técnica de equilibrio SMOTE es la que más incrementa la precisión de los modelos.

Tomando como ejemplo el modelo de Bosque Aleatorio, se puede comprobar como las curvas ROC han ido mejorando en función de la metodología utilizada sobre los datos. A continuación, se muestra el gráfico con la curva ROC correspondientes al Bosque Aleatorio aplicado sobre los datos iniciales (Raw Data), los datos con matriz de costes (CSL) y los datos generados con SMOTE (SMOTE).

Gráfico IV.2.6
Curva ROC para los resultados del Bosque Aleatorio (RF)



Fuente: Elaboración Propia.

Las curvas ROC de las diferentes metodologías son muy similares, excepto la generada por los datos con SMOTE (Línea Roja) que claramente presenta una mejora sustancial sobre el resto. En el resto de modelos (Regresión Logística Ridge, Árbol de Decisión y Red Neuronal) sucede una situación parecida, mejorado considerablemente la capacidad de los modelos al emplear la técnica SMOTE.

IV.2.3.- Análisis de Sensibilidad

Dados los excelentes resultados obtenidos con la técnica de balanceado SMOTE, a continuación se efectúa un análisis de sensibilidad aplicando la transformación SMOTE a los datos incluidos en el análisis.

Sobre el conjunto de datos aleatoriamente se crean dos subconjuntos, uno con el 70% de los datos (entrenamiento) y otro con el restante 30% (comprobación). Esta operación se repite 10 veces para analizar si afecta la distribución de los datos entre los dos conjuntos a los resultados de la clasificación y de esta forma poder calibrar correctamente los modelos.

Tabla IV.2.6
Distribución de las muestras en las 10 repeticiones

	Conjunto Entrenamiento					*Conjunto Comprobación*				
	no	***yes***	***No***	***yes***	***Total***	***no***	***yes***	***no***	***yes***	***total***
1	219	15	93,59%	6,41%	234	93	8	92,08%	7,92%	101
2	220	14	94,02%	5,98%	234	92	9	91,09%	8,91%	101
3	216	18	92,31%	7,69%	234	96	5	95,05%	4,95%	101
4	216	18	92,31%	7,69%	234	96	5	95,05%	4,95%	101
5	217	17	92,74%	7,26%	234	95	6	94,06%	5,94%	101
6	217	17	92,74%	7,26%	234	95	6	94,06%	5,94%	101
7	215	19	91,88%	8,12%	234	97	4	96,04%	3,96%	101
8	219	15	93,59%	6,41%	234	93	8	92,08%	7,92%	101
9	219	15	93,59%	6,41%	234	93	8	92,08%	7,92%	101
10	219	15	93,59%	6,41%	234	93	8	92,08%	7,92%	101
	2177	**163**	**93,03%**	**6,97%**	**2340**	**943**	**67**	**93,37%**	**6,63%**	**1010**

Fuente: Elaboración propia.

Sobre el conjunto de entrenamiento (70%) se aplica el algoritmo SMOTE para balancear las categorías. Posteriormente se genera un modelo de clasificación con las metodologías

de aprendizaje automático seleccionadas (Regresión Logística Ridge (LG), Redes Neuronales (NN), Árboles de Decisión C4.5 (DT) y Bosque Aleatorio (RF)) y se obtiene la precisión o exactitud del mismo por validación cruzada.

Con el modelo obtenido, se predicen los resultados del conjunto de entrenamiento y se comparan con los resultados reales, obteniéndose la capacidad predictiva de los diferentes modelos.

El conjunto inicial de datos tiene 335 empresas con 312 en la categoría negativa (93,13%) y 23 empresas fraudulentas (6,87%), sin embargo, al efectuarse la selección aleatoria de cada uno de los conjuntos, se puede comprobar que las particiones no se alejan de estos porcentajes. Como se pretende analizar la sensibilidad de los modelos a los datos se ha optado por la selección aleatoria de todos los datos frente a una selección de datos estratificada en función de su categoría.

Tomando la primera división, el conjunto de entrenamiento presenta 234 instancias (70%), de las que 219 corresponden a empresas sospechosas y 15 a empresas fraudulentas. El conjunto de comprobación tendrá 101 instancias (30%) con una distribución de 93 "*No*" y 8 "*Yes*".

El número de empresas fraudulentas oscila en el conjunto de entrenamiento entre un mínimo de 14 y un máximo de 19, mientras que en el conjunto de comprobación será a la inversa, oscilando entre un máximo de 9 y un mínimo de 4.

En el conjunto de entrenamiento se generan datos sintéticos con el algoritmo SMOTE para balancear la categoría minoritaria de la variable objetivo. A continuación se modeliza con las cuatro técnicas indicadas en los apartados anteriores, y se calcula la precisión de los modelos mediante validación cruzada.

IV.2.3.1.- Análisis de los resultados

Los resultados para la primera división son los siguientes:

Tabla IV.2.7
Matriz de confusión de los modelos con SMOTE (conjunto de entrenamiento)

	LG		***DT***		***NN***		***RF***	
	NO	***YES***	***NO***	***YES***	***NO***	***YES***	***NO***	***YES***
NO	165	54	176	43	187	32	196	23
YES	47	186	26	207	20	213	13	220
Correctly Classified	77,65%		84,73%		88,50%		92,04%	
Incorrectly Classified	22,35%		15,27%		11,50%		7,96%	
TP Rate (No)	75,34%		80,37%		85,39%		89,50%	
TP Rate (Yes)	79,83%		88,84%		91,42%		94,42%	
FP Rate (Yes)	20,17%		11,16%		8,58%		5,58%	
FP Rate (No)	24,66%		19,63%		14,61%		10,50%	

Fuente: Elaboración Propia.

Aunque se ha reducido el número de instancias para la modelización, los resultados son muy parecidos a los obtenidos con el conjunto total de datos. Como se observa en la Tabla IV.2.7, tomando la media de las diez divisiones, los resultados obtenidos son prácticamente igual a los resultados obtenidos con el algoritmo SMOTE en el apartado anterior. El número de instancias en el conjunto de entrenamiento no afecta para la precisión general de los modelos, ya que la capacidad explicativa se mantiene, a pesar de reducir el conjunto.

Con los modelos obtenidos en el conjunto de entrenamiento balanceado con SMOTE se predicen los resultados de la variable dependiente con el conjunto de confirmación, y se comprueban con los datos reales.

Para la primera división del conjunto de comprobación se obtienen los siguientes resultados que se presentan en la Tabla IV.2.8:

Tabla IV.2.8
Matriz de confusión de los modelos (conjunto de comprobación)

	LG		***DT***		***NN***		***RF***	
	NO	***YES***	***NO***	***YES***	***NO***	***YES***	***NO***	***YES***
NO	70	23	79	14	77	16	89	4
YES	3	5	3	5	4	4	5	3
Correctly Classified	74,26%		83,17%		80,20%		91,09%	
Incorrectly Classified	25,74%		16,83%		19,80%		8,91%	
TP Rate (No)	75,27%		84,95%		82,80%		95,70%	
TP Rate (Yes)	62,50%		62,50%		50,00%		37,50%	
FP Rate (Yes)	37,50%		37,50%		50,00%		62,50%	
FP Rate (No)	24,73%		15,05%		17,20%		4,30%	

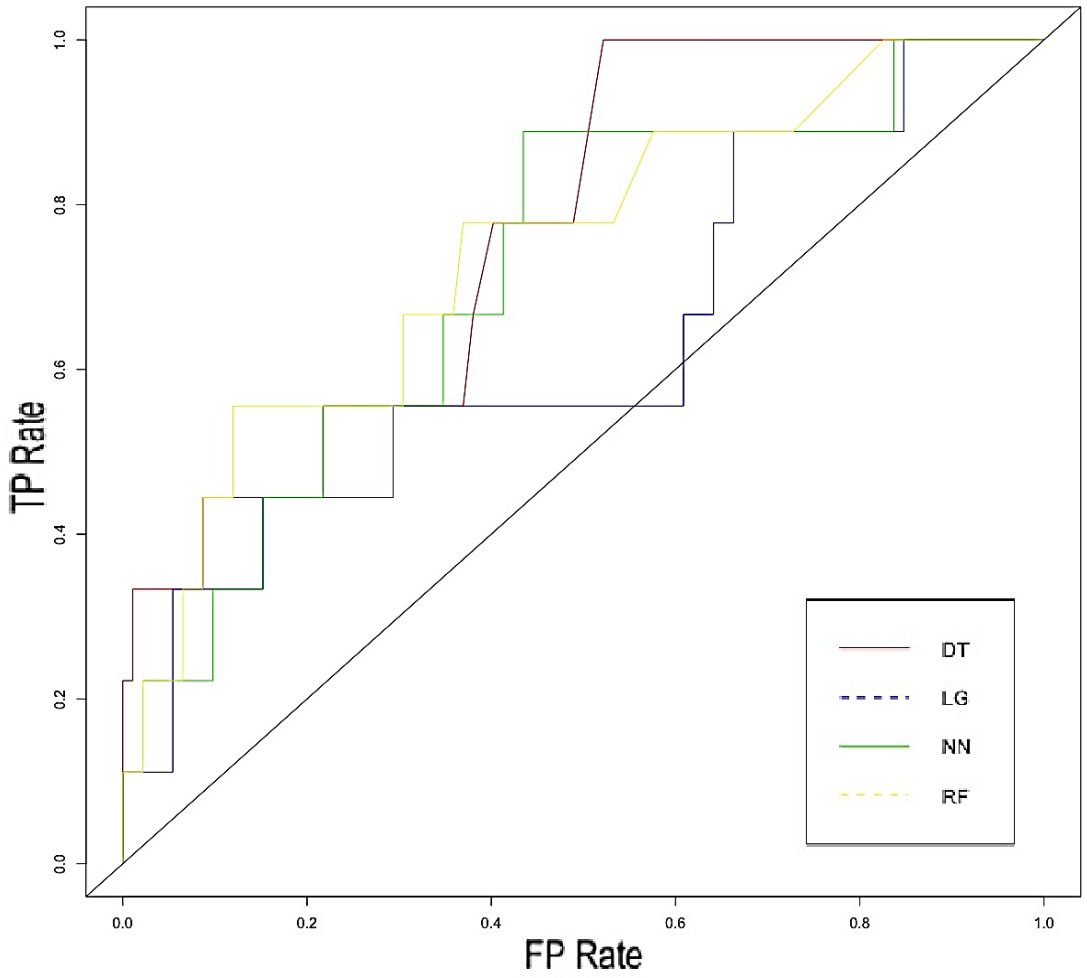
Fuente: Elaboración Propia.

La capacidad predictiva de todos los modelos se ha visto reducida, sobre todo en la tasa de verdaderos positivos (TP Rate), siendo muy llamativa la pérdida que se produce en el Bosque Aleatorio. Se ha pasado de obtener los mejores resultados con el conjunto de entrenamiento a los peores con el conjunto de confirmación.

En promedio de las 10 divisiones, la Red Neuronal y la Regresión Logística son las técnicas que mejor predicen sobre el conjunto de comprobación. Sin embargo, el Bosque Aleatorio era el que mejor explicaba en el conjunto de entrenamiento. En el presente caso, el Bosque Aleatorio está sobreestimando, motivo por el cual tendría la menor capacidad predictiva de los cuatro al no generalizar suficientemente.

La reducción de capacidad predictiva, especialmente en el caso del Bosque Aleatorio, se puede apreciar en el siguiente gráfico (Gráfico IV.2.7), donde visualmente se muestran los resultados obtenidos por los clasificadores sobre el conjunto de comprobación.

Gráfico IV.2.7
Curva ROC con SMOTE sobre el conjunto de comprobación



Fuente: Elaboración Propia.

El Gráfico IV.2.5, que mostraba los resultados con datos equilibrados con la técnica SMOTE sobre el conjunto de entrenamiento, presentaba unas Curvas ROC mucho más cercanas al punto óptimo, definiendo unas tasas más favorables que en el Gráfico IV.2.6. Sin embargo, en el Gráfico anterior (Gráfico IV.2.7) llama la atención la reducción de precisión del Bosque Aleatorio (línea amarilla en ambos gráficos) que, al igual que el resto de clasificadores, en este caso define un área bajo la curva más reducida y más irregular.

Asimismo, cabría destacar los resultados obtenidos en el conjunto de comprobación por los modelos de la tercera división, y que son los siguientes:

Tabla IV.2.9
Matriz de confusión de los modelos (conjunto de comprobación nº 3)

	LG		***DT***		***NN***		***RF***	
	NO	***YES***	***NO***	***YES***	***NO***	***YES***	***NO***	***YES***
NO	72	24	82	14	82	14	90	6
YES	1	4	1	4	0	5	0	5
Correctly Classified	75,25%		85,15%		86,14%		94,06%	
Incorrectly Classified	24,75%		14,85%		13,86%		5,94%	
TP Rate (No)	75,00%		85,42%		85,42%		93,75%	
TP Rate (Yes)	80,00%		80,00%		100,00%		100,00%	
FP Rate (No)	20,00%		20,00%		0,00%		0,00%	
FP Rate (Yes)	25,00%		14,58%		14,58%		6,25%	

Fuente: Elaboración Propia.

En esta división, la capacidad predictiva de la Red Neuronal y del Bosque Aleatorio sobre la categoría "*Yes*" es del 100%, presentado los mejores resultados de todos los modelos probados.

IV.2.3.2.- Análisis de las medidas de evaluación

Las medidas de precisión muestran el mismo comportamiento que las matrices de confusión. Los resultados para los datos de entrenamiento son de gran calidad como se observa en la Tabla IV.2.10, obteniéndose para el Bosque Aleatorio una media del área ROC del 0,9873 con un rango de 0,019. El estadístico Kappa, también obtiene los mejores resultados el Bosque Aleatorio, con un valor muy elevado (0,89024) y rango pequeño (0,1103). La metodología que minimiza los errores vuelve a ser el Bosque Aleatorio con una media del RMSE de 0,22126 y un rango de 0,0688.

Tabla IV.2.10
Medidas de precisión de los modelos en el conjunto de entrenamiento

	ROC			***Kappa***			***RMSE***		
	Mean	***Min***	***Max***	***Mean***	***Min***	***Max***	***Mean***	***Min***	***Max***
LG	0,8581	0,825	0,906	0,55947	0,496	0,7023	0,3875	0,3391	0,4123
DT	0,8901	0,871	0,918	0,76576	0,6802	0,8422	0,32817	0,2784	0,3783
NN	0,9253	0,873	0,959	0,73131	0,611	0,8128	0,34025	0,2801	0,4148
RF	**0,9873**	**0,977**	**0,996**	**0,89024**	**0,8403**	**0,9376**	**0,22161**	**0,1875**	**0,2563**

Fuente: Elaboración Propia.

Al comparar las medidas de precisión del conjunto de entrenamiento con el conjunto de comprobación que se recogen en la Tabla IV.2.11, se muestra una disminución considerable, mayor en unos métodos que en otros.

Tomando como referencia la media del área ROC, la mayor perdida la sufre el Bosque Aleatorio que tiene una diferencia de 0,2331 entre el conjunto de entrenamiento y el de comprobación. La menor diferencia se observa en la Regresión Logística, que era la que obtenía peores resultados en el conjunto de entrenamiento.

Tabla IV.2.11
Medidas de precisión de los modelos en el conjunto de comprobación

	ROC			***Kappa***			***RMSE***		
	Mean	***Min***	***Max***	***Mean***	***Min***	***Max***	***Mean***	***Min***	***Max***
LG	0,7787	**0,708**	0,881	0,14947	0,0555	0,2914	0,40051	0,3493	0,4661
DT	0,657	0,558	0,796	0,1859	**0,0901**	0,293	0,39203	0,3226	0,4372
NN	0,7361	0,508	0,948	0,22728	0,0249	0,3671	0,39697	0,3397	0,4651
RF	**0,7886**	0,62	**1**	**0,26919**	0,0804	**0,5976**	**0,27975**	**0,2388**	**0,3224**

Fuente: Elaboración Propia.

El Bosque Aleatorio es el que tiene una pérdida mayor del estadístico Kappa, siendo la diferencia de 0,62105. Finalmente, cabe destacar que el incremento en la media del RSME es muy similar para el Árbol de Decisión, la Red Neuronal y el Bosque Aleatorio, siendo muy pequeño para la Regresión Logística, sólo un incremento de 0,0102.

Los rangos (diferencia entre el máximo y el mínimo) de las medidas de precisión de los conjuntos de entrenamiento son estrechos, por lo que los modelos son bastante estables y no dependen del conjunto de datos de entrenamiento seleccionado. Sin embargo, al tomar los rangos de las medidas de precisión en el conjunto de comprobación, estos son mucho más amplios, dependiendo la capacidad predictiva del modelo del conjunto de datos seleccionados.

Este resultado sugeriría la existencia de un número mayor de empresas fraudulentas no investigadas por las autoridades. Cuando una empresa es investigada y finalmente se le fija una etiqueta, se tiene la seguridad que es legal o ilegal. No obstante, como se conoce, hay muchas empresas que no presentan inicialmente ningún indicio, o sobre las que no hay ninguna sospecha, que podrían ser ilegales y que no han sido identificadas como tales.

Finalmente, los resultados de este estudio han permitido identificar un número de empresas que presentaban un patrón de comportamiento similar al de las empresas ya identificadas como fraudulentas. Esta información ha sido presentada a las autoridades en el transcurso del proceso judicial a fin de ofrecer información adicional sobre el conjunto de empresas sospechosas, destacando el comportamiento de algunas de ellas. Debido al estado del caso, no se ha obtenido confirmación policial sobre si ciertas empresas detectadas como falsos positivos (FP Rate) son realmente empresas fraudulentas, aunque se trataría de conocer esta situación en el futuro.

CONCLUSIONS

International Monetary Fund, 2017, p.1.

"*Money laundering and the financing of terrorism are financial crimes with economic effects. They can threaten the stability of a country's financial sector or its external stability more generally. Effective regimes to combat these threats are essential to protect the integrity of markets and of the global financial framework as they help prevent financial abuses. Action against money laundering and terrorist financing thus responds not only to a moral imperative but also to an economic need*."

CONCLUSIONS

The analysis of money laundering evolution (Section I.6) shows a rise in this serious crime in Spain over the last few decades. New technologies, electronic transfers and online payment services are excellent alternatives for “cyber laundering methodologies” that complicate the traceability of illicit capital (Souto, 2013). In this context, researchers need automatic and effective detection tools to ensure an efficient judicial system as well as an effective system of prevention (FATF, 2012).

This study provides criminal investigators with an objective way of identifying companies that show a fraudulent pattern within a business organization suspected of money laundering. In this way, the proposed methodology helps to focus available investigation resources on those companies that present a greater number of suspicious operations. It is hoped that the proposed machine learning models will be as good as the "police eye" and provide a new way of detecting a greater number of cases of fraud. As such, the implemented models aim to correctly obtain the largest possible number of fraudulent operations, whilst minimizing the proportion of fraudulent misclassification (False Positive Rate).

In the judicial process under study, a proportion of suppliers have been identified as fraudulent (proven fraud). However, it is not clear whether the other suppliers who were involved in this complex plot of money laundering were fraudulent or not (suspected fraud). Detecting potential defrauding companies through their similarity to those already investigated is one of the objectives of this work. For this, a process of analysis of measurable information is developed that favours the interpretation of the results and that is useful during the judicial process. The expert-assisted investigation begins with a complete pre-processing of data that gives depth and robustness to the models. Once the database has been debugged and the available variables checked, the information is entered into the classification models. Finally, the classifiers are subjected to a sensitivity analysis that allows the evaluation of the results.

Even in "real-life" problems where one depends on the criterion of an expert to determine the tag of the target variable, one can never be sure of the classification algorithm to be

used (Bishop, 2006). The first approach proposes to use Neural Network models as a working tool in this expert-assisted activity, from which good results were obtained. In this framework, it was possible to incorporate all the information (variables) available in the estimation of the model. The way with which the adjustment was achieved was besides noteworthy.

In the adjusted model with unbalanced training data, the adjustment rate was seen to improve, with 72.16% of operations being correctly classified. This rate drops to 17.63% when dealing with correctly classified fraudulent transactions (TP Rate). It should be noted that the false negative rate (FN Rate), above 80%, must be interpreted with care, since possible non-fraudulent operations are included in the count.

In addition, the sampling strategy followed for the selection of the training data set is not neutral, as evidenced by the experiment performed with 100 different training sets to detect adjustment sensitivity to changes in the training set. The distribution of true positives is seen to be asymmetric. The asymmetry would indicate that the exclusion/inclusion of cases in the training set is not neutral, suggesting the random selection strategy leaves room for improvement.

The previous experiment and the methodology of adjustment of the network structure used, as well as the unknown impact associated with classification errors, led to the inclusion of an improvement strategy with balanced training data using the SMOTE technique. Despite the decrease in the overall success rate (*Correctly Classified*), the detection capacity of fraudulent operations (TP Rate) was seen to improve to almost 70%, which practically reaches the overall success rate of the unbalanced model.

Having achieved a notable adjustment, significant opportunities for improvement were highlighted through a review of the case selection strategy for the training set. These include implementing other network structures or other classification methodologies, and even incorporating other balancing systems.

In the second approach, different classification structures are tried: Ridge Logistic Regression (LG), Neural Networks (NN), Decision Trees C4.5 (DT) and Random Forests (RF) which, combined with Benford's Law, are used for the detection of money

laundering patterns. The Cost Matrix and the SMOTE technique are incorporated as data balancing techniques.

The results obtained in this approach again show that the SMOTE algorithm obtains better results on the true positives than the unbalanced sample, and higher than with the application of the cost matrix. However, the overall accuracy of the model is very similar, so the proportion of false positives increases with the SMOTE methodology.

In a similar way, applying the cost matrix for data balancing identifies less positives overall. Consequently, companies indicated as potential fraudsters would be less, and the companies investigated *a posteriori* by the police authority would diminish. Selecting a large number of companies to investigate would have two negative points. The first would be the increase of the costs of the investigation: more investigated companies will require more time, more personnel and more resources. The second drawback would be the disruption caused to legal companies that would have to endure an investigation and its consequences.

Overall, the Random Forest showed the best results with the SMOTE transformation, obtaining 96.15% of true negatives (TN Rate) and 94.98% of true positives (TP Rate). The classification capacity of this methodology is undoubtedly very high.

The advantages offered by the new ensemble model methodologies, particularly the random forest methodology, over the independent classification models are especially beneficial, since (1) they mainly improve the accuracy obtained and reduce the bias of the classifier and (2) they apportion robustness to the models.

However, the predictive capacity of the models with the SMOTE transformation is seen to reduce when a dataset is used for training and another for testing. While the results obtained by cross-validation for the total data do not differ substantially from the results obtained with only 70% of the data, the results of the prediction with the test set are lower.

Nevertheless, the results obtained in the 10 replicates are quite stable, so the models proposed in the second approach are not very sensitive to the division of the data. Both the descriptive capacity and the predictive capacity of the model have narrow ranges, which would confirm that the models are stable.

It should be noted that the speed with which machine learning techniques evolve and the wide variety of machine learning techniques available offer researchers and forensic accountants different algorithms that could increase the performance and efficiency of the classification models proposed here. This could include experimentation with other more complex network structures or other learning methods (Ripley, 2007; Bishop, 2006), and even their integration into a Boosting strategy (Freund and Schapire, 1996) together with the SMOTE technique (SMOTE-Boost de Chawla et al., 2003b), or other model assembly techniques (Wang and Shao, 2009).

An optimal solution would be to use different methods and algorithms to evaluate additional approaches. This would provide the police authorities with as much information as possible to properly analyse the results from applying the machine learning techniques. For example, data without transformation classifies true negatives very well, and this could also be exploited if ensemble model methods are used for the final classification.

Therefore, the results obtained here open a wide range of possibilities for the improvement of expert-assisted research in the detection of financial fraud and money laundering, using these types of predictive tools for whatever is required, through the use of sensitivity and performance analysis of machine learning techniques to search for patterns of fraud that can be described *a priori*. This reinforces the role of the forensic accountant in this type of investigation, since one of the main problems faced by money laundering investigators is the inability to translate the vast amount of information derived from complex business structures into behaviour patterns of clearly fraudulent individuals.

From the list of companies not categorized by the Judicial Police as fraudulent in the previous investigation (suspicious fraud), applied expert systems have successfully offered information about certain supplier companies that presented behavioural patterns similar to those categorized as fraudulent with the information available *a priori*. Unfortunately, the current state of the judicial process (still sub judice) does not allow a rigorous evaluation *a posteriori* of the indications that were offered to the police authorities.

Detection of patterns of fraudulent behaviour by these types of expert systems is not a crime in itself, however, but it is a further indication that complements the information available to criminal investigators. The difficulty of describing the fraud patterns detected by the machine learning techniques makes it impossible to provide a detailed and coherent explanation of them, and this proof does not represent an indicator sufficiently consolidated to be presented as evidence in the judicial process.

Based on the experience of this research, which is framed in a current judicial process and limited to a pioneering collaboration with the Money Laundering Squad of the Spanish Judicial Police, it´s obtained an outcome that would indicate the viability of this type of model as a tool in the planning and prioritization of police investigation tasks.

In summary, this research has evidenced that machine learning techniques represent a new tool that is able to effectively and objectively orientate researchers to those companies that are more likely to be launderers of money.

However, as it is the first time that machine learning techniques have been used in a real case for the detection of fraud and money laundering, and considering the time constraints for its implementation, many further issues could be studied and analysed for future research, as mentioned below.

New learning methodologies such as Deep Learning or Extreme Gradient Boosting (XGBoost algorithm) could be used. In both cases, the computation time increases considerably compared to the methodologies used in this PhD Dissertation. However, the results obtained in other fields encourage their use for these tasks, and would suggest that their use and application in the financial sector could be more significant.

Consideration should also be given to the inclusion of additional information on the targeted companies. The accounting information of companies is very important for detecting certain irregularities but it is true that, in most cases, the accounts will have already been manipulated what might limit the ability of the forensic accountant to get information. Considering accounting model to detect accounting manipulation as well as the inclusion of social networking information, public records, financial market prices, etc. would provide data that would improve the reliability of the models used.

Moreover, through business transactions between companies and through Graph Theory, connections between companies that work together in the laundering process within a fraudulent business group could be identified. The Graph Theory would make it clear what the relationship was between companies that *a priori* did not have direct commercial connection and would facilitate the traceability of the capital and the identification of new potentially fraudulent companies.

In addition, the knowledge acquired in this research would reinforce the use of machine learning techniques in order to analyse business information (transactions, business organization, billing...) from institutions such as the Tax Agency, financial institutions or Financial Intelligence Units. In any case, future lines of research should be developed for its application as a tool for the analysis of text of reports of suspicious operations (SARs) with the objective of coding them according to their level of risk. This would expedite a process that currently consumes excessive human and time resources and would facilitate the classification of suspicious companies or business groups, helping to prioritize available resources.

Furthermore, another future line of research could be in its application for the detection of money laundering patterns in financial markets with an on-line platform. The existence of virtual currencies reveals new mechanisms of laundering that require the use of automatic techniques for its prevention, detection and investigation.

The use of automatic tools, such as those proposed in this PhD Dissertation which improve the efficiency of the international methods used in the fight against this crime (AML/CFT), would increase opportunities to combat money laundering and reduce the sophisticated systems by which terrorism is financed (FATF, 2012; UNODC, 2008).

REFERENCIAS BIBLIOGRÁFICAS

REFERENCIAS BIBLIOGRÁFICAS

Aguinis, H.; Gottfredson, R. y Joo, H. (2013). Best-Practice Recommendations for Defining Identifying and Handling Outliers. *Organizational Research Methods*, 16(2): 270–301.

Alali, F. A. y Romero, S. (2013). Benford's Law: Analysing a decade of financial data. *Journal of Emerging Technologies in Accounting, 10(1)*, 1-39. http://dx.doi.org/10.2308/jeta-50749.

Albus, J. E., Anderson, R. H., Brayer, J. M., DeMori, R. Feng, H. Y., Horowitz, S. L. y Vamos, T. (2012). *Syntactic Pattern Recognition, Applications*. (Vol. 14). Springer Science and Business Media.

Alhosani, W. (2016). *Anti-Money Laundering. A Comparative and Critical Analysis of the UK and UAE's Financial Intelligence Units*. Springer.

Ali, K. M. y Pazzani, M. J. (1996). Error reduction through learning multiple descriptions. *Machine Learning*, 24 (3), 173-202.

Amiram, D., Bozanic, Z., y Rouen, E. (2015). Financial statement errors: evidence from the distributional properties of financial statement numbers. *Review of Accounting Studies*, *20*(4), 1540-1593.

Anscombe, F. J. (1960). Rejection of outliers. *Technometrics*, *2*(2), 123-146.

Badal-Valero, E. y García-Cárceles, B. (2016). Detección de fraude financiero mediante redes neuronales de clasificación en un caso real español. *Estudios de Economía Aplicada*, *34* (3), 693-709.

Barnett, V. y Lewis, T. (1994). *Outliers in Statistical Data* (3ed.). New York: Wiley.

Bartlett, Brent L. (2002). The negative effects of money laundering on economic development. *Platypus Magazine*, December, 18-23.

Barone, R. y Masciandaro, D. (2011). Organized crime, money laundering and legal economy: theory and simulations. *European Journal of Law and Economics,* 32, 115-142.

Batista, G. E., Prati, R. C. y Monard, M. C. (2004). A study of the behaviour of several methods for balancing machine learning training data. *ACM Sigkdd Explorations Newsletter*, *6*(1), 20-29.

Bauer, E. y Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*, *36*(1), 105-139.

Becker, G. S. (1968). *The Economic Approach to Human Behavior*. Chicago: University Chicago Press.

Beck, M.W. (2015). *Neural Net Tools: Visualization and Analysis Tools for Neural Networks.* Version 1.4.0. Disponible en: http://cran.r-project.org/web/packages/NeuralNetTools/ [27/07/2016].

Benesty, J., Chen, J., Huang, Y. y Cohen, I. (2009). *Pearson Correlation Coefficient. In Noise Reduction in Speech Processing*. Springer, Berlin Heidelberg, 1-4.

Benford, F. (1938). The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 551-572.

Bennell, C. y Canter, D. V. (2002). Linking commercial burglaries by modus operandi: Tests using regression and ROC analysis. *Science and Justice*, *42*(3), 153-164.

Bernasconi, P. (1995). La criminalité organise et d'affaires internationales. *Changes in Society, Crime and Criminal Justice in Europe*. Volume II: international organized and corporate crime, eds. Cyrille Fisnaut, Johan Goethals, Tony Peters and Lode Walgrave. The Hague: Kluwer Law International.

Biau, G. y Scornet, E. (2016). A random forest guided tour. *Test*, *25* (2), 197-227.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Bologna, G. y R. Lindquist. (1995). *Fraud Auditing and Forensic Accounting: New Tools and Techniques*. 2nd Ed. New York, NY: John Wiley & Sons.

Bolton, R. J. y Hand, D. J. (2002). Statistical fraud detection: A review. *Statistical Science*, 235-249.

Brantingham P. J. y Brantingham P. L. (1981) *Environmental Criminology*. Prospect Height, IL: Waveland Press.

Breiman, L., Friedman, J., Stone, C. J. y Olshen, R. A. (1984). *Classification and Regression Trees*. CRC press.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123-140.

Breiman, L. (2000). Randomizing outputs to increase prediction accuracy. *Machine Learning*, *40*(3), 229-242.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1): 5–32.

Brodley, C. E. y Utgoff, P. E. (1995). Multivariate decision trees. *Machine Learning*, *19*(1), 45-77.

Camdessus, M. (1998).Money laundering: the importance of international countermeasures, address and the plenary meeting of the Financial Action Task Force. *Money Laundering*, Paris. Accessed April 10, 2011.http://www.imf.org./external/np/speeches/1998/021098.htm.

Cameron, T. A. (1988). A new paradigm for valuing non-market goods using referendum data: maximum likelihood estimation by censored logistic regression. *Journal of Environmental Economics and Management*, *15*(3), 355-379.

Cao, D. K. y Do, P. (2012). Applying data mining in money laundering detection for the vietnamese banking industry. *Asian Conference on Intelligent Information and Database Systems,* (207-216). Springer, Berlin Heidelberg.

Caridad, J. M. y Ceular, N. (2001). Un análisis del mercado de la vivienda a través de redes neuronales artificiales. *Estudios de Economía Aplicada*, (18), pp. 67-81.

Carpio Delgado, J. (2011). La posesión y utilización como nuevas conductas en el delito de blanqueo de capitales. *Revista General del Derecho Penal,* 15: 1-28.

CCBE. (2014). *A Lawyer's Guide to Detecting and Preventing Money Laundering*.

Chapman, F. y Scheffer, J., (2000). An analysis of the Missing Data Methodology for Different Types of Data. *Unpublished Master's Thesis*, Massey University.

Chawla, N., Hall, L., K.W., B. y Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chawla, N. V. (2003a). C4. 5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. *Proceedings of the ICML* (Vol. 3).

Chawla, N. V., Lazarevic, A., Hall, L. O. y Bowyer, K. W. (2003b). SMOTEBoost: Improving prediction of the minority class in boosting. *European Conference on Principles of Data Mining and Knowledge Discovery,* (pp 107-119). Springer, Berlin Heidelberg.

Chawla, N. V., Japkowicz, N. y Kokz, A., editors (2004). *SIGKDD Special Issue on Learning from Imbalanced Datasets*.

Chawla, N. V. (2005). Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook* (pp. 853-867). Springer US.

Chen, H., Chung, W., Xu, J. J., Wang, G., Qin, Y. y Chau, M. (2004). Crime data mining: a general framework and some examples. *Computer*, 37(4), 50-56.

Chen, C. H. (Ed.). (2015). *Handbook of Pattern Recognition and Computer Vision*. World Scientific.

Clarke, R. V. (1980) Situational crime prevention: theory and practice. *British Journal of Criminology*, 20, 136-147.

Cordero, I. B. (2002). *El delito de blanqueo de capitales*. 2ª Edición. *Navarra*: Editora Aranzadi.

Cordero, I. B. (2009) Eficacia del Sistema de prevención del blanqueo de capitales–Estudio del cumplimiento normativo (*compliance*) desde una perspectiva criminológica. *Eguzkilore* 23, 117-138

Chong, A. y López-De-Silanes, F. (2007). *Money Laundering and Its Regulation*. Washington DC: Inter-American Development Bank.

Chong, A., y Lopez-De-Silanes, F. (2015). Money Laundering and Its Regulation. *Economics & Politics*, *27*(1), 78-123.

Council of Bars and Law Societies of Europe (*CCBE*), International Bar Association y American Bar Association. (2014). *A Lawyer's Guide to Detecting and Preventing Money Laundering.*

Crumbley, D. L., Heitger, L. E., y Smith, G. S. (2005). *Forensic and Investigative Accounting* (Vol. 4025). CCH Incorporated.

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J. y Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, *88*(11), 2783-2792.

Dahbur, K. y Muscarello, T. (2003): Classification system for serial criminal patterns. *Artificial Intelligence and Law,* 11(4), 251–269.

Dayton, C. M. (1992). Logistic regression analysis. *Stat*, 474-574.

Darasay, B. V. (1991). *Nearest Neighbor Pattern Classification Techniques*.

Davies, A., Wittebrood, K. y Jackson, J. L. (1997). Predicting the criminal antecedents of a stranger rapist from his offence behaviour. *Science y Justice*, *37*(3), 161-170.

Devijver, P. A. y Kittler, J. (1982). *Pattern Recognition: A Statistical Approach*. Prentice hall.

Devroye, L. Györfi, L. y G. Lugosi. (1996). *A Probabilistic Theory of Pattern Recognition* Springer, New York.

De la Fuente, S. (2011). Regresión Logística. *España: Facultad de Ciencias Económicas y Empresariales, UNAM-Universidad Autónoma de Madrid*.

Del Cid, J. M. (2007): *Blanqueo internacional de capitales. Cómo detectarlo y prevenirlo*. El País, 20 de marzo.

Demetis, D. S. (2010).*Technology and anti-money laundering: A systems theory and risk-based approach*. Edward Elgar Publishing.

Demetis, D. S. (2011). *Unfolding Dimensions of an Anti-Money Laundering/Counter-Terrorist Financing Complex*. System. LexisNexis

Demuth, H. B., Beale, M. H., De Jess, O., y Hagan, M. T. (2014). *Neural Network Design*. Martin Hagan.

D'Souza, J. (2011).*Terrorist financing, money laundering, and tax evasion: Examining the performance of financial intelligence units*. CRC Press.

Díaz-Uriarte R. y Alvarez de Andrés S. (2006).Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:1–13.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157.

Dietterich, T., Margineantu, D., Provost, F. y Turney, P. Editors (2003). *Proceedings of the ICML'2000.* Workshop on cost-sensitive learning.

Dixon, W. J. (1950). Analysis of extreme values. *The Annals of Mathematical Statistics*, *21*(4), 488-506.

Duda, R. O., Hart, P. E. y Stork, D. G. (1973). *Pattern classification* (Vol. 2). New York: Wiley.

Durtschi, C., Hillison, W. y Pacini, C. (2004). The effective use of Benford's law to assist in detecting fraud in accounting data. *Journal of Forensic Accounting*, *5*(1), 17-34.

Dutta, S. K. (2013). *Statistical techniques for forensic accounting: understanding the theory and application of data analysis*. FT Press.

Duyne, P. C. (2003). Money laundering, fears and facts. *Criminal Finances and Organizing Crime in Europe*, eds. Petrus C. van Duyne, Klaus von Lampe and James L Newell (Ed.), Nijmegen: Wolf Legal Publishers.

Frank, E, Hall, M. A. y Ian H. Witten, I. H. (2016). The WEKA Workbench. Online Appendix for *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, Fourth Edition.

Elkan, C. (2001). The foundations of cost-sensitive learning. *International Joint Conference on Artificial Intelligence* (Vol. 17, No. 1, pp. 973-978). Lawrence Erlbaum Associates Ltd.

Fayyad, U. M. y Irani, K. B. (1992). The attribute selection problem in decision tree generation. *AAAI* (pp. 104-110).

Farrell, G., Tilley, N. y Tseloni, A. (2014). Why the crime drop? In Tonry M (ed.). Why crime rates fall and why they don't. *Crime and Justice: A Review of Research*. Chicago: University of Chicago Press.

FATF (2008). Typologies report on Proliferation Financing. Paris: FATF/OECD. http://www.fatf-gafi.org/media/fatf/documents/reports/Typologies%20Report%20on%20Proliferation%20Financing.pdf.

FAFT (2003): Técnicas complejas de lavado de dinero. *Informe Sobre Tipologías*.

FATF (2012). The FATF *Recommendations: International Standards on Combating Money Laundering and the Financing of Terrorism and Proliferation*. Paris: FATF/OECD.

FATF (2013). *Methodology: For Assessing Technical Compliance with the FATF Recommendations and the Effectiveness of AML/CFT Systems*. FATF/OECD.

Fernández, A. (2012). Financial channels of money laundering in Spain. *British Journal of Criminology, 52, 908-931.*

Ferwerda, J. (2009). The economics of crime and money laundering: does ant-money laundering policy reduce crime?. *Review of Law and Economics*, 5(2): 903-929.

Ferwerda J, Kattenberg M, Chang H-H, *et al*. (2013). Gravity models of trade-based money laundering. *Applied Economics* 45(22): 3170-3182.

Fan, J. y Li, R. (2001). Variable selection via noncave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.

Fay, R. E. (1991). *A design-based perspective on missing data variance*. In Proc. Seventh Annual Res. Conf., Washington, D.C.: U.S. Bureau of the Census, 429-440.

Fondo Monetario Internacional (FMI). (2014). *Review of the fund's strategy on Anti-Money Laundering and Combating the Financing of Terrorism*. https://www.imf.org/external/np/pp/eng/2014/022014a.pdf.

Fondo Monetario Internacional (FMI). (2017). *The IMF and the Fight Against Money Laundering and the Financing of Terrorism*. https://www.imf.org/About/Factsheets/Sheets/2016/08/01/16/31/Fight-Against-Money-Laundering-the-Financing-of-Terrorism?pdf=1

Filzmoser P. y Gschwandtner M. (2017). *mvoutlier: Multivariate Outlier Detection Based on Robust Methods*. R package version 2.0.8.

Forget, L. y Hočevar, V. Š. (2004). *Financial Intelligence Units: An Overview*. International Monetary Fund.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, *3*, 1289-1305.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, *7*(3), 397-416.

Fukunaga, K. (2013). *Introduction to Statistical Pattern Recognition*. Academic press.

FAFT (2003). Técnicas complejas de lavado de dinero. *Informe Sobre Tipologías*.

Galán, M.A. (2016). *Asesoramiento Fiscal y Blanqueo de Capitales*. Mitos y realidades del Blanqueo de Capitales. Fundación de Estudios Bursátiles y Financieros, Valencia.

Gao, Z. y Ye, M. (2007). A framework for data mining-based anti-money laundering research. *Journal of Money Laundering Control*, *10*(2), 170-179.

Gao, Z. (2009). Application of cluster-based local outlier factor algorithm in anti-money laundering. *Management and Service Science, 2009. MASS'09,* (pp. 1-4). IEEE.

García, S., Luengo, J. y Herrera, F. (2015). *Data Processing in Data Mining*. Springer International, Switzerland.

Garrett G. y Hadley W. (2011). Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.

Giles, D. E. (2007). Benford's law and naturally occurring prices in certain eBay auctions. *Applied Economics Letters, 14(3)*, 157-161.

Gray, J. y Shenoy, P. (2000). Rules of thumb in data engineering. *Data Engineering, 2000. Proceedings. 16th International Conference* (pp. 3-10). IEEE.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Annals of Mathematical Statistics*.

Guiora, A. N. y Field, B. J. (2007). Using and abusing the financial markets: money laundering as the Aschilles' heel of terrorism. University of Pennsylvania. *Journal of International Economic Law*, 29(1), 59-104.

Günnel, S. y Tödter, K. H. (2009). Does Benford's Law hold in economic research and forecasting?. *Empirica*, *36(3)*, 273-292

Hadley W. (2017). *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.6.1. http://CRAN.R-project.org/package=tidyr.

Hansen, L., y Salamon, P. (1990). Neural network ensembles. IEEE Trans. *Pattern Analysis and Machine*. Intel l., 12, 993-1001.

Hassani H., Huang X., Silva E. S. y Ghodosi (2016) A Review of data mining applications in crime. *Statistical Analysis and Data Mining*.

Hastie T., Tibshirani R. y Friedman J. (2001). *The Elements of Statistical Learning*. New York: Springer.

Hastie, T., Tibshirani, R. y Friedan, J. (2008). The Elements of Statistical Learning*. Data Mining, Inference, and Prediction* (2nd ed.). Standfor: Springer. (pp. 392-396).

Hastie, T., Tibshirani, R. y Friedman, J. (2009). The Elements of Statistical Learning*. Data Mining, Inference, and Prediction*, Second Edition, Springer: New York.

Hawkins, D.M. (1980). *Identification of Outliers*. London: Chapman y Hall.

He H., Bai Y., Garcia E. A. y Li S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *Neural Networks, 2008. IJCNN 2008 (IEEE World Congress on Computational Intelligence).* IEEE International Joint Conference (pp. 1322-1328). IEEE.

He, H. y Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, *21*(9), 1263-1284.

Hill, T. (1995). The Significant-Digit Phenomenon*. The American Mathematical Monthly, 102(4)*, 322-327.

Hill, T. (1995b). A statistical derivation of the significant-digit law. *Statistical Science*, 10, 354-363.

Hilera, J. y Martínez, V. (1995). *Redes Neuronales Artificiales: Fundamentos, Modelos y Aplicaciones*. Madrid: Addison-Wesley Iberoamericana. RA-MA. 390 p.

Hirschey, J. K. (2014). Symbiotic Relationships: Pragmatic Acceptance of Data Scraping. *Berkeley Tech.* LJ, 29, 897.

Hoaglin, D. C., Iglewicz, B., y Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, *81*(396), 991-999.

Hoerl, A. E. y Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

Hogg R., J. W. McKean J. y Craig A. (2005), *Introduction to Mathematical Statistics*, 6th ed., Pearson Prentice Hall, Upper Saddle River, New Jersey.

Holland, P. W. y Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics-theory and Methods*, *6*(9), 813-827.

Holmes, G., Donkin, A., y Witten, I. H. (1994). Weka: A machine learning workbench. *Intelligent Information Systems, 1994.* Proceedings of the 1994 Second Australian and New Zealand Conference (pp. 357-361). IEEE.

Hong, X., Chen, S. y Harris, C. J. (2007). A kernel-based two-class classifier for imbalanced data sets. *IEEE Transactions on neural networks*, *18*(1), 28-41.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, *79*(8), 2554-2558.

Hosmer Jr, D. W., Lemeshow, S. y Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Huerta, H. V., Vásquez, A. C., Dueñas, A. M. H., Loayza, L. A. y Naupari, P. J. R. (2009). Reconocimiento de patrones mediante redes neuronales artificiales. *Revista de Investigación de Sistemas e Informática*, *6*(2), 17-26.

Hunt, E. B., Marin, J. y Stone, P. J. (1966). *Experiments in Induction*. New York: Academic Press.

Izenman, A. J. (2013). Linear discriminant analysis. *Modern multivariate statistical techniques* (pp. 237-280). Springer New York.

Jackson, P. (1998). *Introduction To Expert Systems* (3 edición), Addison Wesley, p. 2, ISBN 978-0-201-87686-4

Jain, A. K., Duin, R. P. W. y Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(1), 4-37.

Japkowicz, N. (2000). The Class Imbalance Problem: Significance and Strategies*. Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000*): Special Track on Inductive Learning, Las Vegas, Nevada.

Japkowicz, N. (2000b). Learning from Imbalanced Data sets: A Comparison of Various Strategies. *Proceedings of the AAA1'2000 Workshop on Learning from Imbalanced Data Sets*, Austin, TX.

Japkowicz, N. (2001). Supervised versus unsupervised binary-learning by feedforward neural networks. *Machine Learning*, 42(1/2):97-122.

Jiménez, C. (2009). *El blanqueo de capitales*. PhD dissertation, Universidad Rey Juan Carlos.

Jin, G., Qian, J., Qian, J. y Huang, W. (2003). A Forecasting Model of Crime-risk Using Data-mining Based on Decision-tree. *Computer Engineering*, *9*, 070.

Jo, T. y Japkowicz, N. (2004). Class imbalances versus small disjuncts. *ACM Sigkdd Explorations Newsletter*, *6*(1), 40-49.

Judd, C. M. y McClelland, G. H. (1989). *Data analysis: A model comparison approach.* San Diego, CA: Harcourt Brace Jovanovich.

Judge, G. y Schechter, L. (2009). Detecting problems in survey data using Benford's Law. *Journal of Human Resources*, *44(1)*, 1-24

Khac, N. L. y Kechadi, M. (2010). *Application of Data Mining for Anti-Money Laundering Detection: A Case Study*. IEEE International Conference on Data Mining Workshops.

Kleinman, L. C. y Norton, E. C. (2009). What's the risk? A simple approach for estimating adjusted risk measures from nonlinear models including logistic regression. *Health Services Research*, *44*(1), 288-302.

Kohavi, R. y Kunz, C. (1997). Option decision trees with majority votes. *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 161–169). San Francisco, CA: Morgan Kaufman.

Kohavi, R. y John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 273-324.

Kolen, J. F. y Pollack, J. B. (1991). Back propagation is sensitive to initial conditions. *Advances in Neural Information Processing Systems*, Vol. 3, pp. 860-867 San Francisco, CA. Morgan Kaufmann.

Kotsiantis, S., Kanellopoulos, D. y Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1), 25-36

Kwok, S. W. y Carter, C. (1990). Multiple decision trees. In Schachter, R. D., Levitt, T. S., Kannal, L. N., y Lemmer, J. F. (Eds.), *Uncertainty in Artificial Intelligence 4*, pp. 327-335. Elsevier Science, Amsterdam.

Lee, M. L., Lu, H., Ling, T. W. y Ko, Y. T. (1999). Cleansing data for mining and warehousing. *International Conference on Database and Expert Systems Applications* (pp. 751-760). Springer Berlin Heidelberg.

Le Cessie, S. y Van Houwekingen, J. C. (1992). Ridge estimators in logistic regression, *Applied Statistics, 41*, 191-201.

Levi, M. (2002) Money Laundering and Its Regulation. *The ANNALS of the American Academy of Political and Social Science*, 582(V), 181-194.

Levi, M. y Reuter, P. (2006). Money laundering. *Crime and Justice*, *34*(1), 289-375.

Levi, M. (2015). Money for crime and money from crime: financing crime and laundering crime proceeds. *European Journal on Criminal Policy and Research*, *21*(2), 275-297.

Li, X. B., Sweigart, J. R., Teng, J. T., Donohue, J. M., Thombs, L. A. y Wang, S. M. (2003). Multivariate decision trees using linear discriminants and tabu search. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *33*(2), 194-205.

Liaw, A. y Wiener, M. (2002) Classification and regression by randomForest. *R News*, 2, 18–22.

Lin-Tao, A. Ji, N. y Zhang, J.-L. (2008). A RBF neural network model for anty-money laundering. *International Conference on Wavalet Analysis and Pattern Recognition*, 209-215.

Liou, D.-R., Liou, J.-W. y Liou, C.-Y. (2013). *Learning Behaviors of Perceptron*. iConcept Press. ISBN 978-1-477554-73-9.

Little, R. J. y Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.

Lopez-Rojas, E. A. y Axelsson, S. (2012). Multi agent based simulation (mabs) of financial transactions for anti-money laundering (aml). *Nordic Conference on Secure IT Systems*. Blekinge Institute of Technology.

López, M. J. C. (2015). Blanqueo de capitales: técnicas de blanqueo y relación con el sistema tributario. *Cuadernos de Formación. Colaboración*, *4*(15).

López, V., Fernández, A., García, S., Palade, V., y Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113-141.

Lohr, S. (2012). The age of big data. *New York Times*, *11*.

Lukasz K. (2011). *outliers: Tests for outliers. R package version 0.14*. http://CRAN.R-project.org/package=outliers.

Lunardon, N., Menardi, G. y Torelli, N. (2014). ROSE: a package for binary imbalanced learning. *R Journal*, *6*(1), 79-89.

Lv, L. T., Ji, N. y Zhang, J. L. (2008). A RBF neural network model for anti-money laundering. *Wavelet Analysis and Pattern Recognition, 2008. ICWAPR'08. International Conference* (Vol. 1, pp. 209-215). IEEE.

Mackrell, N. (1996) Economic consequences of money laundering. InGraycar A and Gravosky PN (eds). *Money Laundering in the 21st Century: Risks and Countermeasures*. Camberra: Australian Institute of Criminology.

Mahalanobis, P.C. (1936). *On the generalised distance in statistics*. *Proceedings of the National Institute of Science of India*, 12, 49-55.

Mallada, C. (2012). *Fiscalidad y Blanqueo de Capitales*. Tesis Doctoral. Universidad de Oviedo, 38-42.

Malm A. y Bichler G. (2013) Using friends for money: The positional importance of money-launderers in organized crime. *Trends in Organized Crime*, *16(4), 365-381*.

Masciandaro D (1999) Money laundering: the economics of regulation. *European Journal of Law and Economics*, 7, 225-240.

Maszczyk, T. y Duch, W. (2008). Comparison of Shannon, Renyi and Tsallis entropy used in decision trees. *International Conference on Artificial Intelligence and Soft Computing* (pp. 643-651). Springer Berlin Heidelberg.

McCullagh, P. y Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edn. Chapman and Hall: London.

Minsky, M. (1954). *Neural Nets and the Brain-Model Problem*. Unpublished doctoral dissertation, Princeton University, NJ.

Mitchell, T. (2006). *The Discipline of Machine Learning*. Technical Report CMU ML-06 108.

Moore, G. E. (1995). Lithography and the future of Moore's law. *SPIE's 1995 Symposium on Microlithography* (pp. 2-17). International Society for Optics and Photonics.

Müller, H. y Freytag, J. C. (2005). *Problems, methods, and challenges in comprehensive data cleansing*. Professoren des Inst. Für Informatik.

Muñiz, P. y J. A. Alvarez (1997). Comportamiento del Mercado: Hipótesis Alternativas. *Revista de Bolsas y Mercados Españoles*, 60, 29-33.

Muñoz, J. F. (2009). *Métodos de imputación para el tratamiento de datos faltantes: aplicación mediante R/Splus*. Universidad de Granada.

Murthy, S. K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery*, *2*(4), 345-389.

Nagy, G. (1968). State of the art in pattern recognition. *Proceedings of the IEEE*, *56*(5), 836-863.

Nasridinov, A., Ihm, S. Y. y Park, Y. H. (2013). A decision tree-based classification model for crime prediction. *Information Technology Convergence* (pp. 531-538). Springer Netherlands.

Nath, S. V. (2006). Crime pattern detection using data mining. *Proceedings of the International Conference on Web Intelligence and Intelligent Agent Technology Workshops*, Hong Kong, 41–44.

Newcomb, S. (1881) Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics, 4*, 39-40.

Ngai, E., Hu, Y., Wong y Chen, Y. y Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification frame work and an academic review of literature. *Decision Support Systems*, 50 (3), 559-569.

Nigrini, M. J. (1992). *The detection of income escape through an analysis of digital distributions*. PhD Tesis University of Cincinatti.

Nigrini, M. J. (1996). A taxpayer compliance application of Benford's Law. *The Journal of the American Taxation Association, 18(1)*, 72-91.

Nigrini, M. J y Mittermaier, L. J. (1997). The Use of Benford's Law as an Aid in Analytical Procedures. *Auditing: A Journal of Practice y Theory, 16(2)*, 52-67.

Nigrini, M. J. y Miller, S. J. (2009). Data diagnostics using second-order tests of Benford's law. *Auditing: A Journal of Practice y Theory, 28(2)*, 305-324. http://dx.doi.org/10.2308/aud.2009.28.2.305

Nigrini, M. (2011). *Forensic analytics: methods and techniques for forensic accounting investigations* (Vol. 558). John Wiley & Sons.

Nigrini, M. (2012). *Benford's Law: Applications for forensic accounting, auditing, and fraud detection* (Vol. 586). John Wiley & Sons.

Nikulin, M. S. (1973) Chi-square test for normality. *International Vilnius Conference on Probability Theory and Mathematical Statistics*.

Olmedo, E., Velasco, F. y Valderas, J. M. (2007). *Caracterización no lineal y predicción no paramétrica en el IBEX35*. *Estudios de Economía Aplicada*, 25(3).

Osborne, J. W. y Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research y Evaluation*, *9*(6), 1-12.

Owojori, A. A. y Asaolu, T. O. (2009). The role of forensic accounting in solving the vexed problem of corporate world. *European Journal of Scientific Research*, *29*(2), 183-187.

Pal, S. K. y Pal, A. (2001). *Pattern recognition: from classical to modern approaches*. World Scientific.

Pavía, J. M. (2015). Testing Goodness-of-Fit with the Kernel Density Estimator: GoFKernel. *Journal of Statistical Software, 66(1)*, 1-27.

Pearsall, B., (2010). Predictive policing: The future of law enforcement?. *National Institute of Justice Journal*, 266.

Petrucelli, J. (2012). *Detecting Fraud in Organizations: Techniques, Tools, and Resources*. Washington DC: John Wiley & Sons, Inc.

Pigeon, S., Druyts, P. y Verlinde, P. (2000). Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. *Digital Signal Processing*, *10*(1-3), 237-248.

Pimbley, J. M. (2014). Benford's law and the risk of financial fraud. *Risk Professional (May)*, 1-7.

Pinkham, R. S. (1961). On the distribution of first significant digits. *The Annals of Mathematical Statistics*, *32*(4), 1223-1230

Quinlan, J., 1983.Learning Efficient Classification Procedures and Their Application to Chess end Games, Morgan Kaufman, San Francisco, CA.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81-106.

Quinlan, J., (1993). C4.5: *Programs for Machine Learning*. Morgan Kaufman, San Francisco, CA.

Quinlan, J. R. (1996). Improved use of continuous attributes in C4. 5. *Journal of Artificial Intelligence Research*, *4*, 77-90.

Quick, R. y Wolz, M. (2003). Benford's Law in deutschen Rechnungslegungsdaten. *Betriebswirtschaftliche Forschung und Praxis, 2*, 208–224.

Quirk, P. (1996*). Macroeconomics Implications of Money Laundering*, Working Paper, FIM.

R Core Team (2015). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org/

Rahm, E. y Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Eng.* Bull., *23*(4), 3-13.

Rahn, R. W. (2001). *The Case Against Federalizing Airport Security*. Cato Institute. Available from: <http://www.cato.org/pub_display.php?pub_id3865>

Ramos, D. (2006). Fraude: un nuevo enfoque para combatirlo. *Auditoría Pública, 38*, 99-104.

Rao, J.N.K. y Shao, J. (1992). Jackknife Variance Estimation With Survey Data Under Hot-Deck Imputation. *Biometrika*, 79 811-822.

Raskutti, B. y Kowalczyk, A. (2004). Extreme re-balancing for SVMs: a case study. *ACM Sigkdd Explorations Newsletter*, *6*(1), 60-69.

Rasmussen, J. L. (1988). Evaluating outlier identification tests: Mahalanobis D squared and Comrey Dk. *Multivariate Behavioral Research*, *23*(2), 189-202.

Reaven, G. M. y Miller, R. G. (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia*,*16*(1), 17-24.

Ripley, B. D. (2007). *Pattern Recognition and Neural Networks*. Cambridge university press.

Rivero, G. (2011). Análisis de datos incompletos en Ciencias Sociales*. Publicaciones del CIS, Cuadernos Metodológicos nº 46.*

Rorabacher, D.B. (1991) *Statistical Treatment for Rejection of Deviant Values: Critical Values of Dixon Q Parameter and Related Subrange Ratios at the 95 percent Confidence Level*. Anal. Chem., 63 (2), 139–146.

Rubin, D. (1976). Inference and missing data*. Biometrika*, 63: 581-592

Rutkowski, L., Jaworski, M., Pietruczuk, L. y Duda, P. (2014). The CART decision tree for mining data streams. *Information Sciences*, *266*, 1-15.

Sahin, Y., Bulkan, S. y Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications, 40(15)*, 5916-5923.

Samuel, A. L. (1967). Some studies in machine learning using the game of checkers. In recent progress. *IBM Journal of Research and Development*, *11*(6), 601-617.

Schapire, R. E. (1997). Using output codes to boost multiclass learning problems. In Proceedings of the Fourteenth *International Conference on Machine Learning*, pp. 313{321 San Francisco, CA. Morgan Kaufmann.

Shapire, Y., Freund, P. Bartlett y Lee W. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26 (5):1651–1686, 1998. 18.

Sharma, A. y Panigrahi, P. K. (2013). *A review of financial accounting fraud detection based on data mining techniques*. arXiv preprint arXiv:1309.3944.

Scheffer, J. (2002). *Dealing with Missing Data*. Res. Lett. Inf. Math. Sci, 3, 153-160.

Schürmann, J. (1996). *Pattern classification: a unified view of statistical and neural approaches* (pp. I-XVII). New York: Wiley.

Schwager, S. J. y Margolin, B. H. (1982). Detection of multivariate outliers. *The Annals of Statistics, 10*, 943-954.

Seagrave, S. (1995). *Lords of the rim: The invisible empire of the overseas Chinese*. Putnam Adult.

Senator, T. E., Goldberg, H. G., Wooton, J., Cottini, M. A., Khan, A. U., Klinger, C. D. y Wong, R. W. (1995). Financial Crimes Enforcement Network AI System (FAIS) Identifying Potential Money Laundering from Reports of Large Cash Transactions. *AI Magazine*, *16*(4), 21.

SEPBLAC (2008): *Informe sobre Tipologías de Blanqueo de Capitales*. http://www.sepblac.es/espanol/informes_y_publicaciones/informe_sobre_tipologias.pdf

SEPBLAC (2013): *Memoria Anual*. http://www.sepblac.es/espanol/informes_y_publicaciones/memoria2013.pdf

Setiono, R., Leow, W. K. y Thong, J. Y. (2000). Opening the neural network black box: an algorithm for extracting rules from function approximating artificial neural networks. *Proceedings of the twenty first international conference on Information systems* (pp. 176-186). Association for Information Systems.

Singleton T. W., Bologna G. J., Lindquist R.J. y Singleton A. J. (2006), *Fraud Auditing and Forensic Accounting* (Third Edition), John Wiley & Sons, Inc, New Jersey.

Schneider F. (2005). Shadow economies around the world: what do we really know?, *European Journal of Political Economy* 21(3): 598-642.

Sremack, J. (2015). *Big Data Forensics–Learning Hadoop Investigations*. Packt Publishing Ltd.

Sokal, R. R. y F Rohlf, J., (1995). *Biometry*. New York: WH Freeman and Company.

Sosa Sierra, M. D. C. (2011). Inteligencia artificial en la gestión financiera empresarial. *Revista Científica Pensamiento y Gestión, 4(23)*, 153-186.

Soudijn MR. J. (2012) Removing excuses in money laundering. *Trends in Organized Crime*, 15(2-3), 146-163.

Soudijn MR. J. (2014) Using strangers for money: a discussion on money-launderers in organized crime. *Trends in Organized Crime*, 17(3), 1-19.

Sudjianto, A., Nair, S., Yuan, M., Zhang, A., Kern, D. y Cela-Díaz, F. (2010). Statistical methods for fighting financial crimes. *Technometrics*, *52*(1), 5-19.

Svetnik, V. A. Liaw, C. Tong, J. C. Culberson, R.P. Sheridan, y B.P. Feuston (2003). Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, 43:1947–1958.

Tam, W. K. y Gaines, B. J. (2007). Breaking the (Benford) law: Statistical fraud detection in campaign finance. *The American Statistician*, *61(3)*, 218-223

Tang, J. y Yin, J. (2005). Developing an intelligent data discriminating system of anti-money laundering based on SVM. *Machine Learning and Cybernetics, 2005. Proceedings of 2005 International Conference IEEE,* 6, 3453-3457.

Thomas, J. K. (1989). Unusual patterns in reported earnings. *The Accounting Review*. 54 (4):773-787.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Society*, Ser. B, 58, 267-288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso. A retrospective, *Journal of the Royal Statistical Society*: Series B (Methodological), 73(3): pages. 273-282.

Ting, K. M. (2002). An instance-weighting method to induce cost-sensitive trees. *IEEE Transactions on Knowledge and Data Engineering*, *14*(3), 659-665.

Torgo, L. (2010) *Data Mining using R: learning with case studies*, CRC Press (ISBN: 9781439810187).

Torres, J., Fernandez, S., Gamero, A. y Sola, A. (2007). How do numbers begin? (The first digit law). *European Journal of Physics*, *28(3)*, 17-25.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology*, *49*(11), 1225-1231.

Tumer, K. y Ghosh, J. (1996). Error correlation and error reduction in ensemble classifiers. *Connection Science*, 8 (3{4), 385{404.

Unger, B. (2007). *The scale and impact of money laundering*. Cheltenham, UK: Edward Elgar.

Unger, B. (2009). Money laundering. A newly emerging topic on the international agenda. *Review of Law and Economics*, 5(2), 809-819.

Unger, B. y Joras, F. (2011). *Money laundering in the real state sector: suspicious property*. Cheltenham, UK: Edward Elgar.

Unger, B. (2013). Can money laundering decrease. *Public Finance Review,* 41(5), 658-676.

Unger, B. y Hertog, J. (2012). Water always finds it way: identifying new forms of money laundering. *Crime Law and Social Change*,57 287-304.

United Nations Office on Drugs and Crime (UNODC). (2006). *The Integrity and Accountability of the Police: Criminal Justice Assessment Toolkit*.

United Nations Office on Drugs and Crime (UNODC). (2011). *Estimating illicit financial flows resulting from drug trafficking and other transnational organized crimes*. UNODC Research Report, October 2011.

U.S. Congress, Office of Technology Assessment. (1995). *Information Technologies for Control of Money Laundering*. Washington, DC: U.S: U.S. Government Printing Office. pp. 55-72.

Varian, H. R. (1972). Benfords law. *American Statistician*, *26(3)*, 65-66.

Varese, F. (2011). *Mafias on the move: How organized crime conquers new territories*. Princeton University Press.

Vaithilingam, S. y Mahendhiran N. (2009). Mapping global money laundering trends: lessons from the space setters. *Research in International Business and Finance*, 23, 18-30.

Venables, W. N. y Ripley, B. (2002). *Modern Applied Statistics with S*. 4th Edition. New York: Springer.

Verhage A (2009) Compliance and AML in Belgium: a booming sector with growing pains. *Journal of Money Laundering Control*, 12, 113-133.

Vikram, A., Chennuru, S., Rao, H. R. y Upadhyaya, S. (2004). A solution architecture for financial institutions to handle illegal activities: a neural networks approach. *System Sciences, 2004. Proceedings of the 37th Annual Hawaii International Conference IEE,* 181-190.

Wainer, H. (1976). Robust statistics: A survey and some prescriptions. *Journal of Educational Statistics*, *1*(4), 285-312.

Walker, J. (1999). How big is global money laundering?. *Journal of Money Laundering Control*, 3(1), 25-37.

Walker, J. y Unger, B. (2009). Measuring global money laundering: "The Walker Gravity Model". *Review of Law and Economics*, 5, 821-853.

Wang, Q. R. y Suen, C. Y. (1984). Analysis and design of a decision tree based on entropy reduction and its application to large character set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (4), 406-417.

Wang, S. N. y Yang, J. G. (2007). A money laundering risk evaluation method based on decision tree. *Machine Learning and Cybernetics, 2007 International Conference* (Vol. 1, pp. 283-286). IEEE.

Wang, S. y Yao, X. (2009). Diversity analysis on imbalanced data sets by using ensemble models. *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE SymposiuM IEEE,* 324-331.

Wang, T., Rudin, C., Wagner, D. y Sevieri, R. (2013). Learning to detect patterns of crime. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 515-530). Springer: Berlin Heidelberg.

West, J. y Bhattacharya, M. (2016). Intelligent financial fraud detection: a comprehensive review. *Computers y Security*, *57*, 47-66.

Wickham, H. (2015). *stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.1.0*. https://CRAN.R-project.org/package=stringr

Wickham, H, y Chang, W. (2016). *devtools: Tools to Make Developing R Packages Easier. R package version 1.12.0*. https://CRAN.R-project.org/package=devtools

Wright, J., Ma, Y., Mairal, J., Sapiro, G., Huang, T. S. y Yan, S. (2010). Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE*, *98*(6), 1031-1044.

Woda, K. (2006). Money laundering techniques with electronic payment systems. *Information and Security International Journal*, 18, 27–47.

Wu, G. y Chang, E. (2003). Class-Boundary Alignment for Imbalanced Dataset Learning. ICML 2003. *Workshop on Learning from Imbalanced Data Sets II*, Washington, DC.

Yadav, S. y Yadav, S. (2013). Forensic accounting: A new dynamic approach to investigate fraud cases. *EXCEL International Journal of Multidisciplinary Management Studies*, *3*(7), 1-9.

Yan, R., Liu, Y., Jin, R. y Hauptmann, A. (2003). On predicting rare classes with SVM ensembles in scene classification. *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03). 2003. IEEE International Conference,* 3, 3-21).

York, D. (2000). Auditing technique: Benford's law. *Accountancy*, 1283, 126.

Yang, S. y Wei, L. (2010). Detecting money laundering using filtering techniques: a multiple-criteria index. *Journal of Economic Policy Reform*, *13*(2), 159-178.

Zadrozny, B. y Elkan, C. (2001). Learning and making decisions when costs and probabilities are both unknown. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* , 204-213. ACM.

Zadrozny, B., Langford, J. y Abe, N. (2003). Cost-sensitive learning by cost-proportionate example weighting. *Data Mining, 2003. ICDM 2003. Third IEEE International Conference,* 435-442.

Zdanowicz, J. (2009). Trade-based money laundering and terrorist financing. *Review of Law and Economics,* 5(2), 855-878.

Zeng, Z., Pantic, M., Roisman, G. I. y Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1), 39-58.

Zhang, Z. M., Salerno, J. J. y Yu, P. S. (2003). Applying data mining in investigating money laundering crimes. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 747-752). ACM.

Zimmerman, D. W. (1995). Increasing the power of nonparametric tests by detecting and downweighting outliers. *Journal of Experimental Education,* 64(1), 71-78.

Zimmerman, D. W. (1998). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *Journal of Experimental Education,* 67(1), 55-68.

Zou, H. y Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society:* Series B (Statistical Methodology), *67*(2), 301-320.

ANEXOS

ANEXO I. Marco normativo y legislación aplicable

La legislación vigente que se aplica en los casos procesales que sean objeto de investigación por blanqueo de capitales se escala en dos niveles, por una parte, la normativa de la Unión Europea y la normativa española, y por otra, en las recomendaciones efectuadas por el Grupo de Acción Financiera Internacional (F.A.T.F.-G.A.F.I.).

1. Normativa de la Unión Europea

(i) Directiva 91/308/CEE, del Consejo de la Comunidad Económica Europea, de 10 de junio de 1991, relativa a la prevención de la utilización del sistema financiero para el blanqueo de capitales.

(ii) Directiva 2005/60/CE, del Parlamento Europeo y del Consejo, de 26 de octubre de 2005, relativa a la prevención de la utilización del sistema financiero para el blanqueo de capitales y para la financiación del terrorismo.

(iii) Directiva 2006/70/CE, de la Comisión, de 1 de agosto de 2006, por la que se establecen disposiciones de aplicación de la Directiva 2005/60/CE del Parlamento Europeo y del Consejo en lo relativo a la definición de «personas del medio político» y los criterios técnicos aplicables en los procedimientos simplificados de diligencia debida con respecto al cliente así como en lo que atañe a la exención por razones de actividad financiera ocasional o muy limitada.

(iv) Directiva 2015/849/CE[46], del Parlamento Europeo y del Consejo, de 20 de mayo de 2015, relativa a la prevención de la utilización del sistema financiero para el blanqueo de capitales o la financiación del terrorismo.

[46] A la aplicación de esta directiva se modifica el Reglamento (UE) n° 648/2012 del Parlamento Europeo y del Consejo, y se derogan la Directiva 2005/60/CE del Parlamento Europeo y del Consejo y la Directiva 2006/70/CE de la Comisión.

2. *Normativa española*

Régimen jurídico-administrativo

(i) Ley 10/2010, de 28 de abril, de prevención del blanqueo de capitales y de la financiación del terrorismo.

(ii) Real Decreto 304/2014, de 5 de mayo, por el que se aprueba el Reglamento de la Ley 10/2010, de 28 de abril, de prevención del blanqueo de capitales y de la financiación del terrorismo.

(iii) Orden EHA/2444/2007, de 31 de julio, por la que se desarrolla el Reglamento de la Ley 19/1993 de 28 de diciembre, en relación con el informe de experto externo sobre los procedimientos y órganos de control interno y comunicación establecidos para prevenir el blanqueo de capitales.

(iv) Recomendaciones de la Comisión de Blanqueo de Capitales e Infracciones Monetarias *(SEPBLAC)*, de 4 de Abril de 2013, sobre las medidas de control interno para la prevención del blanqueo de capitales y la financiación del terrorismo.

Régimen jurídico-penal

Artículos 301 y 304 del Código Penal (Ley orgánica 10/1995), del Capítulo XIV de la receptación y el blanqueo de capitales.

Artículos 305-310 del Código Penal (Ley Orgánica 10/1995), del Título XIV de los delitos contra la Hacienda Pública y contra la Seguridad Social.

3. Recomendaciones del Grupo de Acción Financiera Internacional (*F.A.T.F.-G.A.F.I.*), *FATF* en adelante.

Las recomendaciones del Grupo de Acción Financiera Internacional (*FATF*) establecen un marco amplio y coherente de medidas que los países deben aplicar en sus reglamentos

para combatir el blanqueo de capitales y la financiación del terrorismo, y también la financiación de la proliferación de armas de destrucción en masa.

Los países cuentan con diversos marcos legales, administrativos y operativos y con diferentes sistemas financieros, por lo que no todos pueden adoptar medidas idénticas para contrarrestar estas amenazas. Por consiguiente, las recomendaciones del *FATF* establecen una norma internacional que los países deben aplicar mediante medidas adaptadas a sus circunstancias particulares.

Las 49 recomendaciones del *FATF*, 40 recomendaciones relativas al blanqueo de capitales y 9 adicionales referentes a la financiación del terrorismo, establecen las medidas esenciales que deben adoptar los países para: (1) identificar los riesgos y desarrollar políticas y coordinación interna; (2) perseguir el blanqueo de capitales, la financiación del terrorismo y la financiación de la proliferación; (3) aplicar medidas preventivas para el sector financiero y otros sectores designados; (4) establecer los poderes y responsabilidades de las autoridades competentes (por ejemplo, las autoridades de investigación, de aplicación de la ley y de supervisión) y otras medidas institucionales; (5) mejorar la transparencia y la disponibilidad de información sobre la propiedad beneficiaria de personas jurídicas y arreglos; y (6) facilitar la cooperación internacional. Revisado en (FATF, 2012).

ANEXO II. Publicaciones. "*Money laundering trend in Spain: Offences and arrests over 15 years"*

"Money laundering trend in Spain: Offences and arrests over 15 years"

Autores: Badal-Valero, Elena; Benavent-Corai, José; Pérez-Poveda, Eugenio

En revisión, Journal on Criminal Policy and Research

Springer

MONEY LAUNDERING TREND IN SPAIN:

OFFENCES AND ARRESTS OVER 15 YEARS

Elena BADAL-VALERO[47], José BENAVENT-CORAI[48], Eugenio PEREZ-POVEDA[48]

[47] Financial Crime and Money Laundering Research Group, Universitat de València, Spain.
[48] Laundering Money Group – Drug and Organized Criminality Unit, Provincial Judicial Police Brigade – Headquarter of the Comunidad Valenciana, Spanish National Police Force, Spain.

Abstract

Money laundering affects the integrity of financial systems. Global efforts have been made to set up the international standards and enact national Anti-Money Laundering (AML) regulations. Nevertheless, their efficiency remains unknown. Whereas some authors support a legal framework would reduce money laundering rates; others support the opposite and even expect an increase.From the period 1998-2012; when looking specifically at criminality rates in Spain, we focused on the two main pillars of the AML: prevention and enforcement. Thus broaching the question if these two areas have decreased money laundering offences and arrests. Secondly, we examined if the AML regime has decreased criminality rates of predicate offences and arrests.

Hence, our results have shown that the money laundering increase was partially explained by the AML framework, and priority programs of police/enforcement agencies. However, the decrease of predicate offences and arrests was not explained by the money laundering dynamic, nor by preventative efforts of legislation. These results suggest that Spanish law; derived from international standards and European Union mandates over twenty years, fails to combat money laundering. Failures in law implementation and underestimation of money laundering rates would explain limited efficiency of the AML regime in Spain. Therefore adopting a criminological approach to more available statistics will help measure the effectiveness of the nation-state to legal adequacy while regulating money laundering.

Keywords: money laundering, long-term, asymmetric eigenvector maps, predicate offences, variance partitioning.

Acknowledgments. The authors want to thank Professor B.J. Biringer of the Florida State University for his English revision.

Funding. This work was supported by funds of the Action Plan of the National Police Force against economic crimes and money laundering of the General Directorate of the Police of the Spanish Ministry of Interior and also by the Council on Culture, Education and Sport of the Community of Valencia, Spain.

Introduction

Professional Career Offenders and Criminal organizations (*sensu lato*) looking for stability or growing have the need to inject funds from criminal activity into legal economic system, not raising suspicions of their origin (Unger, 2007). The most common methods are to disguise the source, to change the form of the economic profit or to hide the funds previously obtained through the predicate offences (revisited in Unger, 2007; UNODC, 2011). These *modus operandi* are known as Money Laundering (ML), and they have been employed since the famous gangster Al Capone used launderettes to turn cashes profits from illegal alcohol trafficking (Duyne, 2003; Unger, 2009).

Despite money laundering being the last step forwarded for criminals before reaching economic profit, it still represents an opportunity for police forces; a starting point for a successful investigation able to identify authors and their illicit funds (UNODC, 2011). Given that without economic profit many crimes make no sense, prints detected in money tracing should be the focus of the police and the Achilles' heel of any criminal organization (Bernasconi, 1995; Guiora and Feeld, 2007). Studies dealing with the framework of money laundering are a good way to terminate transnational organizations and protect the integrity of the international financial system (Unger, 2007; UNODC, 2011; IMF, 2014).

Although there is no a universally accepted methodology to estimate the volume of the money laundered every year in the world (Unger, 2007; Barone and Masciandaro, 2011; UNODC, 2011), a consensus of different "guesstimates" place it in the range between 2% to 5% of the Global Gross Domestic Product (Camdessus, 1998; Unger, 2007; UNODC, 2011; Unger, 2013). This represents approximately 2 trillion US Dollars per year (Camdesus, 1998; Walker and Unger, 2009; UNODC, 2011). Unfortunately, reports have

quantified that only 1% of the benefits from crime are intercepted when they are laundered through the financial system (UNODC, 2011).

Money laundering has multiple negative effects operating direct and indirectly at different levels: socioeconomic, spatial and temporal (Bartlett, 2002; Unger, 2007; UNODC, 2011). It stimulates unfair business competition, illegal money capital outflows, political and police corruption and social disaffection with the institutions (including the courts of justice). These negative effects, which branch out to real estate, financial and public sectors, do not remain constrained on a national scale. They expand through the international market system to other economies (Unger, 2007). Hence, money laundering is receiving close review in the agendas of the main international organizations (Unger, 2009). The impact of money laundering in a country is related to both its degree of technological development and the strength of its legal framework, tax system and financial institutions (Chong and López-De-Silanes, 2007; Vaithilingam and Nair, 2009).

Over the last two decades, there has been a strong, unified global effort in developing an Anti-Money Laundering (AML) regime (Unger and Hertog, 2012). This is reflected in the Financial Action Task Force (FATF) that has raised forty anti-money laundering recommendations and nine anti-terrorist financing recommendations (FATF, 2012) and the European Union regulations (91/308/CEE; 2005/60/EC; 2006/70/EC; 2015/849/CE). Both organizations (TATF and EU) have set the international standards that their member countries have to implement in their national laws.

From a criminology perspective, AML approach considers delinquents as "*homo economicus*" whose decision making depends on the previous balance of the costs and benefits of committing a crime (Becker, 1968; Wilson and Herrnstein, 1985; Van

Overtveldt, 2007) and sets prevention and enforcement as its two fundamentals pillars (Reuter and Truman, 2004). On the one hand, prevention aims to deter launderers through the development of administrative laws that facilitate to trace the money. These laws define customer due diligence (CDD) and regulate and supervise the obligation of financial institutions and professionals to report to authorities suspicious activities. On the other hand, enforcement aims to punish launderers through the development of penal laws that define the range of predicate offences, investigation, punishment and confiscation (Reuter and Truman, 2004).

It is expected that the increase of money laundering regulation on both global and domestic levels (*i.e.* within each country) will force the observed ML trend to decrease (Chong and López-De-Silanes, 2007; Ferwerda, 2009). However, there are authors who disagree and suggest that these laws will have no effects (Camdessus, 1998; Rahn, 2001). They argue that through investments in more complex illegal activities, money laundering could increase predicate offences that themselves would promote an increase of money laundered (Masciandaro, 1999; Unger, 2007). Thereupon, it is also expected that money laundering would increase as a consequence of positive feedback from the promoted crimes with economic incomings (Unger, 2007). In other words, the dilemma about the effectiveness of a strong regulation and about its impact on ML dynamics remains unresolved today. The lack of empirical data and of economic models supporting both hypotheses only permits to detect the shadow of money laundering (Levi and Reuter, 2006; Unger and Hertog, 2012; Unger, 2013).

Papers and reports dealing with the trend in money laundering have principally focused on theoretical and empirical estimates of the volume of money laundered in the world, an issue that remains a source of controversy (Unger, 2007; 2013). Furthermore, these

studies are scarce and were adopted only under economic approaches which present difficulties in testing the presence of a conclusive pattern (Unger 2013). Hence, a more empirical and promising way of research would be to use the rates of money laundering offences and arrests derived from criminality data, as they originate when they are real evidence of a crime. Nevertheless, upon our knowledge, no study dealing with police records on money laundering fighting has been yet published, as this information remains reserved in many countries, even for research purposes. Thus, the aim of this paper is to adopt another point of view and to test the money laundering trend through criminality records over several years. Indeed, according to Levi and Reuter (2006) and Unger and Hertog (2012), predicate offences that previously generated the illicit economic profits can be used for monitoring ML dynamic.

Specifically in Spain, money laundering estimates differ considerably and range between €36 million per year, as calculated from 367 cases judged (Fernández Steinko, 2012), and $56US billion per year, as calculated from a theoretical model (Walker, 1999). Unger (2007), following Walker (1999), estimated an increasing index of attractiveness to money launderers that ranged from 0 to 55.4 for Luxembourg, which showed the highest value. In Unger's index, Spain scored in the range 9-10 and ranked 51 out of 228 countries (Unger, 2007). More recently, Vaithilingam and Nair (2009) mapped money laundering pervasiveness for 88 countries and Spain ranked in a middle position. Spain is an emerging attractor for organized crime investments due to its geographical position, the presence of multiple different criminal groups, its tourism industry, and an important real estate market (Palomo et al., 2015).

Since 1990, Spain is a member of the Financial Action Task Force (FATF) fulfilling their recommendations on money laundering regulations (Chong and López-De-Silanes, 2007;

FATF, 2010). The anti-money laundering regime in Spain reflects international standards and European Union mandates, applying both prevention and enforcement approaches. Spain is therefore a good benchmark for studying the evolution of money laundering over the last 15 years and the effectiveness of AML regime implemented by the European initiatives.

In short, the aim of this paper is twofold. Firstly, we focus on the AML's two pillars of prevention and enforcement, in relation to the trend of money laundering offences and arrests in Spain over the period 1998-2012. Secondly, given that the final aim of AML policies is to reduce source crimes, we also study whether the AML initiatives was decreased criminality rates of predicate offences and arrests. The study as a whole intends to show how criminality rates reflect failures in the application of AML laws in Spain.

Material and Methods

Criminality Data sets

The data of criminality (offences and arrests) for the 1998-2012 period were extracted from the annual statistical reports published every year by the Ministry of Interior of the Spanish Government, which can be freely downloaded from its institutional website: http://www.interior.gob.es/. Data includes records from the Spanish Security Forces, mainly composed by both the National Police Force and the Civil Guard. The data sets also include the statistics collected by other three police forces created over the last twenty years: Mossos d'Esquadra in Catalonia, Ertzaintza in the Bask Country and Policía Foral in Navarre, which are progressively replacing the National Police Force and Civil Guard activity in their respective territories. Although the records from the regional police forces are still scarce compared to those provide by both the National Police Force and the Civil Guard, they have been accounted when available. Arrest data from 2011 and 2012 were excluded from the analysis because for these years reports only present accused persons by security forces, without detailing if they were previously arrested or not.

Predicate offences and arrested persons accounted in crimes with economic incomings are classified in the Spanish Penal Code as crimes against patrimony and Public Administration, also drug trafficking. Statistics in Spanish reports dealing with crimes against patrimony include: robbery, theft, fraud, counterfeiting, and damage. However, this last offence was excluded because damages do not generate economic incomings. Corruption crimes include: bribery, influence peddling, embezzlement of public funds, disobedience of penal law, and prevarication. Despite the fact that tax fraud and evasion could produce high economic benefits that need to be laundered (Unger, 2007;2013;

Unger and Hertog, 2012; Walker and Unger, 2009), our data did not include fraud to public institutions, such as tax fraud or capital evasion. This information is not available in detail for all the years covering the period 1998-2012. During that period, tax fraud and capital evasion has been mainly investigated by the Spanish Tax Agency (AEAT) and offences have been communicated directly to judges, without arresting the accused and without reporting them to the security forces. Using sentences data (n = 367), Fernández Steinko (2012) published a study showing that for the 1985-2010 period, €898.5millions were laundered in Spain, mainly by drug trafficking (64.5%) and corruption (30.2%). Thus predicate offences and arrested persons selected for the present study represented a realistic sample.

To better understand offence and arrest figures, they are expressed and analysed in population relative terms. Demographic data are obtained from the Spanish Statistical Office (http://www.ine.es/). Rates of Money Laundering Offences (R_{MLO}), Rates of Money Laundering Arrests (R_{MLA}), Rates of Predicate Offences (R_{PO}), and Arrested Persons (R_{PA}) are expressed by millions of persons.

Anti-Money laundering (AML) regime in Spain

The Anti-Money Laundering regime in Spain is composed of both the enforcement and prevention pillars. They have suffered several modifications during the period of study, following the standards of European Union regulations. Since 1995, the Spanish Penal Code has progressively reformed (LO 10/1995; LO 15/2003; LO 5/2010; revised in Carpio Delgado, 2011) and extended the definition of predicate offences as those crimes penalized with almost five years of prison (1996-2004), any crime included in the penal

code (2005-2010) and any activity suspicious of being criminal (since 2011). Despite the changes, punishment has remained constant with incarceration up to six months to six years.

The prevention approach has been developed through administrative laws, which have continuously been modified with the aim of limiting the maximum amounts of cash payments and of identifying actual owners of money and properties (revised in Blanco Cordero, 2009). The development of compliance measures has increased the obligation of financial institutions and professionals to report to authorities suspicious activities. The consequences of this evolution can be quantified with the statistics of the Spanish Finance Intelligence Unit (FIU), also known as SEPBLAC (Spanish acronym of "Servicio Ejecutivo de la Comisión de Prevención del Blanqueo de Capitales e Infracciones Monetarias"), which records both the number of suspicious transaction reports (STR) received (available 2000-2012) and the census of regulated institutions obligated to report to the Spanish FIU and to apply the CDD measures (available 2001-2012). In both time series, data for 2009 was not available and was estimated by averaging values from 2008 and 2010. All these statistics were obtained from the annual reports of the Spanish Finance Intelligence Unit (http://www.sepblac.es/espanol/home_esp.htm).

Hypotheses testing

Money laundering trend. Time series of both money laundering offences and arrests has been fitted to linear models applying least square regression (Sokal and Rohlf, 1995). Fits has been tested through the Monte Carlo methodology on the sum of the squared residuals and the significance quantified permuting values of the response variables, keeping in the same order the values of predictor variables. This methodology allows for relaxation of both normality assumption and sample size (Gotelli and Ellison, 2004).

Effects of money laundering enforcement. The successive reforms of the *anti-money laundering* laws has resulted in three waves of enforcement regimes, thus it was tested through visual analysis of money laundering time series on the three periods with different definitions of predicate offences: crimes penalized with almost 5 years of prison (1998-2004), all crimes included in the penal code (2005-2010) and all types of activity suspicious of being criminal (since 2011).

Effect of money laundering prevention– was tested applying linear regressions with both money laundering offences, and arrests as response variables with the following predictor variables: the number of suspicious transaction reports received in the Spanish FIU and the census of regulated institutions.

Efficiency of AML framework on criminality rates. The volume of predicate offences and arrested persons could be used for measuring the efficiency of the AML, but its realistic

measures are difficult to estimate (Levi and Reuter, 2006). Nevertheless, some criticism emerges because not all the potential predicate offences lead to money laundering procedures. Thus, a noise effect could mask the detection of the variations of target predicate offences and induce erroneous conclusions like the absence of performance.

The main objective of the AML regime is to reduce criminality rates related to professional criminals, organized crime, and terrorism while protecting the general public. They act as a keystone enhancing the dynamic of related offences or insecurity sensitivity. For example, the activity of little dealers, that deal but do not launder money due to low benefits, are directly dependent of criminal organizations that traffic drugs into the country and provide material. Moreover, vehicle theft is related to jewellery robberies or phone shops. Finally, there are urban zones managed by criminal organizations, with high rates of drug dealing, trafficking, and forced prostitution that attract other offences like robberies, thefts or violent crimes. Even opportunistic are committed by those who are not professional criminals.

Therefore, the performance of the AML framework on criminality rates was tested by applying linear regressions with predicated offences as a response variable and money laundering offences as a predictor. Also, efficiency was tested through the linear regression of the rate of predicate arrests as a response variable with money laundering arrests as the predictor. Similar analyses were carried out with the number of suspicious transaction reports received by the Spanish FIU, and the census of regulated institutions as predictors; in order to evaluate the effects of prevention efforts on criminality rates.

Statistical methods

Criminality records sampled over time are not independent among them because criminality rates of year *n* could affect rates of year *n*+1. Next, classical linear regression testing should not be applied because the independence assumption among data is not met (Sokal and Rohlf, 1995). Also the relation between a response variable and its predictors could be highly correlated because they share a linear trend, and not because of variations in the predictor explaining variation in the response variable (Legendre and Legendre, 2012). It is preferable to adopt a time series approach, and test the relation between criminality rates and its predictors by applying a multiple linear regression analysis that includes a temporal vector as a covariable (Blanchet et al., 2011; Legendre and Legendre, 2012). Secondly, the pure effect of predictors could be quantified with variance partition procedures excluding the trend effect (Legendre and Legendre, 2012).Then, in order to test the effect of prevention effort on money laundering, and the efficiency of the AML regime on predicate crimes, we applied multiple linear regression that included a trend covariable which was calculated through Asymmetric Eigenvector Maps Methodology (Blanchet et al. 2011, Legendre and Legendre, 2012). The pure effect of each predictor on criminality rates was quantified with the adjusted explained variance, once the variance explained by the trend was excluded. The significance was quantified permuting residuals of reduced regression model (Legendre and Legendre, 2012).

Statistical Software

The trend vector was estimated with Asymmetric Eigenvectors Maps analyses carried out with the "aem.time" function of the "AEM" package (Borcard et al., 2011; Blanchet and Legendre, 2012). Multiple linear regressions, variance partitioning and significance estimation were done with the "varpart" and "rda" functions of the "vegan" package (Borcard et al., 2011; Oksanen et al., 2013). All the above packages were implemented in R 2.15.1 statistical software (R Development Core Team, 2012).

Results

Money laundering trend

Criminality rates of money laundering in Spain have increased over 15 years. During the 1998-2012 period, money laundering offences in Spain (R_{MLO}) increased nearly 244% whereas arrests (R_{MLA}) increased 431% (see figure 1a,b). R_{MLO} time series was significantly explained by the linear model ($R^2 = 0.78$, Pval< 0.01) and increased every year 0.2 units per million of inhabitants. Also, the linear model explained significantly nearly 70 % of the variance of R_{MLA} ($R^2 = 0.71$, Pval< 0.01) which increased every year 0.3 units per million of inhabitants.

Effects of money laundering enforcement

Figure 1a, representing the increasing dynamic of money laundering offences, shows two differentiated periods. First, from 1998 to 2004, presents rates of money laundering offence (MLO) ranging between 1 and 2 offences per million of inhabitants whereas the second, between 2005 and 2012, present rates ranging between 2 and 4 offences per million of habitants. The first period corresponds to the application of the original penal code LO 10/1995, which defined predicate offences as crimes penalized with almost 5 years of prison. The second period corresponds to the application of the law LO 15/2003 which extends the range of predicate offences to all crimes included in the penal code. Then, the extension of the range of predicate offences would explain the observed increases in money laundering offences.

Nevertheless figure 1b, which represents the dynamic of money laundering arrests, presents a different pattern. On one hand, peaks in years 2005 and 2010 could reflect security forces efforts, for they match with the enhancement of priority programs of the Spanish Government against organized crime through the creation of specialized units and gubernatorial campaigns (Abel Souto, 2013a).Yet, decreases in the years 2003-2004 and 2007could be explained because money laundering investigations needed time to be achieved.

Effects of money laundering prevention

To illustrate, between 2000 and 2012, the number of regulated institutions increased following a linear model with a slope of 1,550 institutions per year (R^2 = 0.97, Pval< 0.001, figure 1c). Local decrease during 2004 and 2005 is the result of the merger or consolidation of companies, and the exclusion of real estate companies with variable capital from the census of regulate institutions as an outcome of law RD 54/2005 (SEPBLAC, 2005; 2006). The census of regulated institutions explained the rate of money laundering arrests (R^2_{adj} = 0.16, Pval = 0.052) once the effect of the trend was excluded, but not the rate of money laundering offences (R^2_{adj} = 0.00, Pval = 0.643). Therefore, efforts in prevention through the increase of institutions obligate the application of customer due diligences when explaining the increase of money laundering arrests.

Again, between 2001 and 2012, the number of suspicious transaction reports (STRs) increased linearly (R^2 = 0.87, Pval< 0.001, figure 1d) with a rate of 65,220.00 per year. The highest increase, between 2006 and 2007, was due to the implementation of two

administrative laws (EHA/1439/2006 and EHA/2619/2006) in regulating the declaration of payments, and the obligation of both money exchange and external transfer institutions to report suspicious activities (SEPBLAC, 2007; 2008). Once the effect of this trend was taken into account, the number of STRs did not explain money laundering offences (R^2_{adj} = 0.01, Pval = 0.304), neither money laundering arrests (R^2_{adj} = 0.07, Pval = 0.149). As a result, the communications of STRs had no effect on money laundering rates.

Efficiency of AML framework on criminality rates

In contrast to money laundering dynamics, predicate offences (R_{PO}) and arrested persons (R_{PA}) decreased nearly 30 % and 16 % respectively (see figure 1e,f). During this period of study, R_{PO} decreased significantly (R^2 = 0.58, Pval< 0.01) every year 436.7 units per million of inhabitants and R_{PA} decreased significantly (R^2 = 0.47, Pval< 0.01) every year 40.8 units per million of habitants.

Once trend effect was excluded, both money laundering offences (R^2_{adj} = 0.01, Pval = 0.294) and arrests (R^2_{adj} = 0.00, Pval = 0.950) did not explain the dynamic of predicate offences. Also the rate of persons arrested for predicate offences was neither explained by the rate of money laundering offences (R^2_{adj} = 0.00, Pval = 0.622) nor money laundering arrests (R^2_{adj} = 0.00, Pval = 0.730). Therefore, decreasing criminality rates of predicate crimes are not explained by money laundering increase.

Meanwhile, once trend effect was excluded, the number of regulated institutions did not explain neither predicate offences (R^2_{adj} = 0.04, Pval = 0.215) nor predicate arrests (R^2_{adj} = 0.00, Pval = 0.701). Even more, the number of suspicious transactions reports did not explained neither predicate offences (R^2_{adj} = 0.09, Pval = 0.108) nor predicate arrests

(R^2_{adj} = 0.06, Pval = 0.174).Therefore, decrease in the predicate rate does not result from prevention efforts of the Spanish Anti-Money Laundering regime.

Discussion

Money laundering increase is partly explained by AML framework and priority programs of police/enforcement agencies

First, the European Union has made a considerable effort in developing anti-money laundering initiatives (91/308/CEE; 2005/60/EC; 2006/70/EC; 2015/849/CE), however the effectiveness of its implementation in member countries remains unknown. Our results have shown that over the last 15 years, rates of money laundering offences and arrests have increased in Spain. This increase was partly explained as the result of tightening the laws by following two strategies. The enforcement approach has extended the range of predicate offences by modifying the Spanish Penal Code. By the same token, the prevention approach has developed an anti-money laundering (AML) framework modifying administrative laws to reduce cash flow, and identify real owners of money by obligating an increasing number of private institutions to report suspicious behavior. A similar pattern has been observed in Canada between 2001 and 2006, where criminality rates of money laundering have also increased, partially explained as the result of the implementation of a new legislation which increased the power of the police and courts to prosecute this crime (Brennan and Vaillancourt, 2011).Therefore, both studies on criminality rates support conclusions from economic studies: strengthen the AML regime at both global and local levels (*i.e.* within each country) increases money laundering trend (Schneider, 2005; Zdanowicz, 2009; Ferwerdaet al., 2013;revised in Unger, 2013); on the

contrary to authors that have suggested a decrease (Chong and López-De-Silanes, 2007; Ferwerda, 2009) or no effect (Camdessus, 1998; Rahn, 2001).

However, money laundering rates stabilized in Canada after 2006 whereas in Spain it continues to have a bubble effect. This phenomenon would be the result of an increased effort of the police/law enforcement agencies to link predicate offences with money laundering, promoted by the interest of both (i) the Spanish Government in satisfying international agreements through the implementation of priority programs against organized crime, and (ii) the media through the impact of corruption scandals published (revised in Abel Souto, 2013a,b). Moreover, modifications of the Spanish Penal Code would condition the efforts of the Spanish Police to fight this crime as Unger (2007) suggested for the Dutch Police.

In addition, the rise in money laundering rates could also be explained by the indirect effects of the AML framework that cannot be measured with quantitative variables. An increase in the regulation of the traditional banking sector forces criminals to look for less controlled new ways of investments, or to switch to outside the financial sector by adopting alternative and modern forms of laundering (Woda, 2006; FernándezSteinko, 2012; Unger and Hertog, 2012). These steps go beyond the Spanish regulation. In fact, new technology such as internet, electronic transfers, and payment services with mobile phones are excellent alternatives of "cyber laundering methodologies" that have not been considered by AML legislation (Abel Souto, 2013b). This migration to new economic sectors would explain why results from some economic studies suggest that money laundering would decrease as a result of the global effort in developing anti-money laundering regulation and an improved disclosure (Chong and López-De-Silanes, 2007;Chong and Lopez-De-Silanes, 2015). Additionally, the complexity of money

laundering procedures have also increased and forced launderers to adopt a high degree of specialization, and division of roles resulting in the participation of more criminals (Unger, 2007; UNODC, 2011; Fernández Steinko, 2012). Hence, the observed rise in money laundering criminality rates would be explained because the implementation of regulations would have promoted an arms race enhancing launderers to overcome legal difficulties through diversity and innovation of methodologies (Unger and Hertog 2012; Levi, 2015). Whereas it was expected that *Situational Crime Prevention* (Clarke, 1980) would play a role in the extinction of money laundering, criminals have evolved by adapting to new economic environments. Therefore, flexibility of money laundering should be taken into account for developing new laws.

Finally, the increase of suspicious transaction reports (STRs) in Spain was not related to money laundering rates as it has been observed before in Belgium (Verhage, 2009). Both similar patterns would support suggestions that the impact of the reporting system on money laundering repression is limited, and it would be overestimated (Blanco Cordero, 2009; Verhage, 2009). Moreover, the lack of relationship would also results from a noise effect if a large amount of transactions were erroneously considered as suspicious. Longer time series of data and more statistics on financial information units are needed to reach a definitive conclusion. In contrast, the increase of the number of regulated institutions obligated to adopt customer due diligence measures, partially explained the rise of money laundering rates. In fact, the measures adopted focus on the recompilation of a large amount of information about clients that can be used by enforcement authorities' requirements. Thus facilitation the identification of offenders and shorter criminal investigations, and increasing money laundering rates.

Anti-Money laundering regime in Spain fails in deterring predicate offences

Some authors have suggested that money laundering could increase criminality rates because criminal behaviour could be encouraged, and illicit legitimized funds could finance new crimes maintaining criminal activity and expanding criminal organizations (Levi, 2002; Mackrell, 1996; Unger, 2007). This feedback hypothesis was supported by Masciandaro (1999) and Unger (2007) with a theoretic economical model based on a closed economy whose results suggested that money laundering would multiply criminality. Otherwise, our results from criminality data do not support this hypothesis because both predicate offences, and arrested persons presented a decreasing trend during the 1998-2012 period.

Besides, the true target off AML regime is to reduce the volume of predicate offences (Levi and Reuter, 2006); however our results show that the AML regime did not explain the decrease in criminality rates. For that reason, both prevention and enforcement efforts have not been effective. Failures in law implementation and underestimation of money laundering rates would explain limited efficiency of the AML regime in Spain.

First, the AML strategy has focused on the "follow the money" rule for lighting the traceability of illicit benefits obtained from crime and identify the real owners of money (revised in Levi and Reuter, 2006). Due to money launderers floating between illicit markets and legal economy, international standards and European Union decrees have promoted the development of a booming business (Verhage, 2009; Levi 2015). Private institutions and professionals have to invest in compliance to reach the AML goals of both public institutions and international cooperation agreements. Nevertheless, for a few thousand euros AML barriers can easily be overcome by criminals with the creation of

shell companies, networks and figureheads where bank accounts are controlled through the internet.

Therefore, maximum prison punishment for a launderer is up to six years in Spain, and cheap alternatives for overcoming legal barriers seem to be very light in the theory of balance of rational choice theory (Becker, 1968; Wilson and Herrnstein, 1985) as compared to the amount of euros expected to be gained from illicit activities.

Next, criminality rates could also not be affected by the money laundering dynamic due to the fact that criminal organizations commit crimes in their location and launder benefits in different countries through different kind of investments (Varese, 2011). Notwithstanding, the need of letters rogatory and international collaborations would deter enforcement authorities to continue their investigation to follow money. The long response time and the idiosyncrasy of the country targeted would be the main barriers. In addition, even if the Spanish Penal Code also penalized money laundering crimes when predicate offences are originated in other countries, the great majority of money laundering investigations of Spanish polices forces start with predicate offences originated in Spain. The Spanish real estate sector is a powerful attractor for investments of foreign criminal organizations (Palomo et al., 2015), and it is extremely difficult to start investigations on money laundering when predicate offences are unknown or committed in other countries.

Moreover, time is a fundamental dimension for explaining offence dynamics (Brantingham and Brantingham, 1981), however there is a lack of studies dealing with response time of crime rates to law implementation, even less studies dealing with the AML regime. Both penal and administrative laws for combating money laundering in

Spain have continuously been introduced or modified (Abel Souto 2013b). Then, the lack of relationship between law efficiency and criminality rates could be explained because the AML framework is in the incipient stage, and it has not spent the time necessary for fully being implemented. Moreover, because money laundering has effects at different time scales (Unger, 2007), the period of study would reflect the initial phase when criminals have not yet learned the lesson and are caught. A short term view would be expected increases in money laundering; whereas criminality rates remain unaffected because the message has not been transmitted. On the other hand in the future a decrease would be expected in money laundering offences and arrests; either because the AML regulation has an effect or because criminals have improved their procedures. In other words, the response of criminality rates to law implementation could present sinusoidal (i.e. periodic) patterns different to the classical linear trend. Whether different temporal scales can also explain the money laundering trend remains to be seen in futures studies.

Finally, law inefficiency would result from an underestimation of the census of launderers and a fully loaded capacity of enforcement authorities would also explain the observed pattern: increasing rates of money laundering while illustrating a decrease in predicate crimes. Unfortunately, there is not realistic data on specialized personnel dedicated to combating money laundering in Spain; in order to estimate the reactive capacity of the AML regime.

Crime decrease since the early 1990s has been observed before in other countries of Europe, North America, and Australasia (Farrell et al., 2014). This phenomenon could result from a complex interaction of controlling factors, yet one seems to predominate. The security hypothesis considers technological advances that have promoted security measures for preventing keystone crimes, which determine dynamic of interrelated

offences (Farrell et al., 2014). Thus, predicated offences would have decreased responding to an increase of security rather than the control of money laundering criminality.

Applying criminology for understanding money laundering

Analyses of time series are an important step toward understanding and forecasting money laundering (Unger, 2013). Like other variables used to estimate the volume of laundered money as suspicious, or unusual transactions reported to financial intelligence units (Unger and Rawlings,2008; Barone and Masciandaro, 2011; UNODC,2011), monitoring criminality data could be used for estimating the size of money launderers' population through capture-mark-recapture methodologies employed in biological sciences to estimate wild animal populations (Lindberg, 2012). Since arrested persons should be presumed innocent until proven guilty during judgment (United Nations, 1948), criminality data is not the last step in the AML combat. Upon our knowledge in Spain, a tracing program has not been implemented neither in other countries, nor in international institutions. This is necessary from the initial laundering detection (*e.g.* police investigations, FIUs or tax agencies reports, etc …) to its end when criminals are sentenced with a penalty and illicit funds are confiscated. The analysis of this data under a criminalistic approach would considerably increase the knowledge about predicate offences, seizure volumes, the network of launderers, and methodologies used (Unger, 2009). The implementation of a tracing program would be necessary to improve our understanding of the complex relationship between law efficiency and money laundering rates, today limited by the low availability of criminality data.

Conclusions

In the final analysis, the present study shows that prevention and enforcement efforts of the Anti-Money Laundering regime have partially explained the money laundering increase in Spain. However, none of them explained the predicate offence decrease. These results suggest that Spanish Law, derived from international standards and European Union mandates over twenty years, would fail to combat money laundering. Adopting a criminological approach based on more available statistics will help to measure the effectiveness of nation-state to law adequacy and regulate money laundering.

References

Abel Souto M. 2013a. Volumen mundial del blanqueo de dinero, evolución del delito en España y jurisprudencia reciente sobre las últimas modificaciones del código penal. *Revista General de Derecho Penal* 20: 1-53.

Abel Souto M. 2013b. Money laundering, new technologies, FATF and Spanish penal reform. *Journal of Money Laundering Control* 16: 266-284.

Barone R. and Masciandaro D. 2011. Organized crime, money laundering and legal economy: theory and simulations. *European Journal of Law and Economics* 32(1): 115-142.

Becker G.S. 1968. Crime and punishment: an economic approach. *Journal of Political Economy* 76: 169-217

Bernasconi P. 1995. La criminalité organisée et d'affaires internationales. In Fisnaut C, Goethals J, Peters T., et al. (eds.) *Changes in society, crime and criminal justice in Europe (Volume II : International Organized and Corporate Crime)*. Norwell: Kluwer Law International.

Blanchet FG., Legendre P. 2012. *AEM: tools to construct Asymmetric Eigenvector Maps (AEM) spatial variables*. R package version 0.4-1/r104. Available from: <http://R-Forge.R-project.org/projects/sedar/>

Blanchet FG., Legendre P., Maranger R., et al. 2011. Modelling the effect of directional spatial ecological processes at different scales. *Oecologia* 166(2): 357-368.

Blanco Cordero I. 2009. Eficacia del Sistema de prevención del blanqueo de capitales – Estudio del cumplimiento normativo (compliance) desde una perspectiva criminológica. *Eguzkilore* 23, 117-138

Borcard D., Gillet F. and Legendre P. 2011. *Numerical ecology with R*. New York: Springer Verlag.

Brantingham P. J. and Brantingham P. L. 1981. *Environmental Criminology*. Prospect Height, IL: Wavel and Press.

Brennan S. and Vaillancourt R. 2011. Money laundering in Canada, 2009. *Juristat Bulletin. Canadian Centre for Justice Statistics* 85-005-x.

Camdessus M. 1998. Money Laundering: the importance of international countermeasures, address and the plenary meeting of the Financial Action Task Force. Available from: <http://www.imf.org./external/np/speeches/1998/021098.htm>

Carpio Delgado J. 2011. La posesión y utilización como nuevas conductas en el delito de blanqueo de capitales. *Revista General del Derecho Penal* 15: 1-28.

Chong A. and López-De-Silanes F. 2007. *Money laundering and its regulation*. Inter-American Development Bank. Working Paper 493.

Clarke R.V. 1980. Situational crime prevention: theory and practice. *British Journal of Criminology* 20: 136-147.

Farrell G., Tilley N. and Tseloni A. 2014. Why the crime drop? In Tonry M (ed.) *Why crime rates fall and why they don't (volume 43: Crime and Justice: A Review of Research)*. Chicago: University of Chicago Press.

FATF, Financial Action Task Force. 2010. Mutual evaluation fourth follow-up report: anti-money laundering and combating the financing of terrorism, Spain. Available from: http://www.fatf-gafi.org/media/fatf/documents/reports/mer/FoR%20Spain.pdf

FATF, Financial Action Task Force. 2012. The FATF recommendations: international standards on combating money laundering and the financing of terrorism and proliferation. Available from: <http://www.fatf-gafi.org/media/fatf/documents/recommendations/pdfs/FATF_Recommendations.pdf>

Fernández Steinko A. 2012. Financial channels of money laundering in Spain. *The British Journal of Criminology* 52(5): 908-931.

Ferwerda J. 2009. The economics of crime and money laundering: does anti-money laundering policy reduce crime. *Review of Law and Economics* 5(2): 903-929.

Ferwerda J., Kattenberg M., Chang H-H., et al. 2013. Gravity models of trade-based money laundering. *Applied Economics* 45(22): 3170-3182.

Gotelli NJ and Ellison A. M. 2004. *A Primer of Ecological Statistics.* Sunderland: Sinauer Associates Inc.

Guiora A. N. and Field B. J. 2007. Using and abusing the financial markets: money laundering as the Aschilles' heel of terrorism. *University of Pennsylvania Journal of International Economic Law* 29(1): 59-104.

IMF, International Monetary Fund. 2014. *Review of the fund's strategy on anti-money laundering and combating the financing of terrorism*. International Monetary Fund Policy Papers. Available from: <http://www.imf.org/external/np/pp/eng/2014/022014a.pdf>

Legendre P. and Legendre L. 2012. *Numerical Ecology*. Amsterdam: Elsevier BV.

Levi M. 2002. Money Laundering and Its Regulation. *The ANNALS of the American Academy of Political and Social Science* 582(V): 181-194.

Levi M. 2015. Money for crime and money from crime: financing crime and laundering crime proceeds. *European Journal on Criminal Policy and Research* 21: 275-297.

Levi M. and Reuter P. 2006. Money laundering. *Crime and Justice* 34:289-375.

Lindberg M. S. 2012. A review of designs for capture-mark-recapture studies in discrete time. *Journal of Ornithology* 152(2): 355-370.

Mackrell N. 1996. Economic consequences of money laundering. In Graycar A and Gravosky PN (eds) *Money laundering in the 21st century: risks and countermeasures*. Camberra: Australian Institute of Criminology.

Masciandaro D. 1999. Money laundering: the economics of regulation. *European Journal of Law and Economics* 7: 225-240.

Oksanen J., Blanchet F. G., Kindt R., et al. 2013. *Vegan: Community Ecology Package*. R package version 2.0. Available from: <http://CRAN.R-project.org/package=vegan>

Palomo J., Marquez J., Laguna P., et al. 2015. From illegal markets to legitimate businesses: the portfolio of organised crime in Spain. In Savona EU and Riccardi M (eds.) *From illegal markets to legitimate businesses: the portfolio of organised crime in Europe*. Trento: Transcrime – Università degli Studi di Trento.

Rahn R. W. 2001. *The Case Against Federalizing Airport Security*. Cato Institute. Available from: <http://www.cato.org/pub_display.php?pub_id3865>

Reuter P. and Truman E. 2004. *Chasing dirty money: the fight against money laundering.* Washington, DC: Institute for International Economics.

R Development Core Team. 2012. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Available from: <http://www.R-project.org>

Schneider F. 2005. Shadow economies around the world: what do we really know? *European Journal of Political Economy* 21(3): 598-642.

SEPBLAC, Servicio Ejecutivo de la Comisión de Prevención del Blanqueo de Capitales e Infracciones Monetarias. 2005. Memoria anual 2004. Madrid: SEPBLAC.

SEPBLAC, Servicio Ejecutivo de la Comisión de Prevención del Blanqueo de Capitales e Infracciones Monetarias. 2006. Memoria anual 2005. Madrid: SEPBLAC.

SEPBLAC, Servicio Ejecutivo de la Comisión de Prevención del Blanqueo de Capitales e Infracciones Monetarias. 2007. Memoria anual 2006. Madrid: SEPBLAC.

SEPBLAC, Servicio Ejecutivo de la Comisión de Prevención del Blanqueo de Capitales e Infracciones Monetarias. 2008. Memoria anual 2007. Madrid: SEPBLAC.

Sokal R. R. and Rohlf F. J. 1995. *Biometry*. San Francisco: WH Freeman and Company.

Unger B. 2007. *The scale and impact of money laundering*. Cheltenham: Edward Elgar.

Unger B. 2009. Money laundering – a newly emerging topic on the international agenda. *Review of Law and Economics* 5(2): 809-819.

Unger B. 2013. Can money laundering decrease. *Public Finance Review* 41(5): 658-676.

Unger B. and Ferwerda J. 2011. *Money laundering in the real state sector: suspicious property*. Cheltenham: Edward Elgar.

Unger B. and Hertog J. den 2012. Water always finds it way: identifying new forms of money laundering. *Crime Law and Social Change* 57(3): 287-304.

Unger B. and Rawlings G. 2008. Competing for criminal money. *Global Business and Economic Review* 10: 331-352.

United Nations. 1948. *The Universal Declaration of Human Rights*. The United Nations. Available from: <http://www.un.org/en/documents/udhr/>

UNODC, United Nations Office on Drugs and Crime (2011). *Estimating illicit financial flows resulting from drug trafficking and other transnational organized crimes*. UNODC Research Report. Available from: <http://www.unodc.org/documents/data-and-analysis/Studies/Illicit_financial_flows_2011_web.pdf>

Vaithilingam S. and Nair M. 2009. Mapping global money laundering trends: lessons from the space setters. *Research in International Business and Finance* 23(1): 18-30.

van Duyne P. C. 2003. Money laundering, fears and facts. In van Duyne PC, von Lampe K and Newell JL (eds) *Criminal finances and organizing crime in Europe*. Nijmegen: Wolf Legal Publishers.

van Overtvel dt J. 2007. *The Chicago School: how the University of Chicago assembled the thinkers who revolutionized economics and business*. Chicago, IL: Agate Publishing.

Verhage A. 2009. Compliance and AML in Belgium: a booming sector with growing pains. *Journal of Money Laundering Control* 12:113-133.

Walker J. 1999. How big is global money laundering? *Journal of Money Laundering Control* 3(1): 25-37.

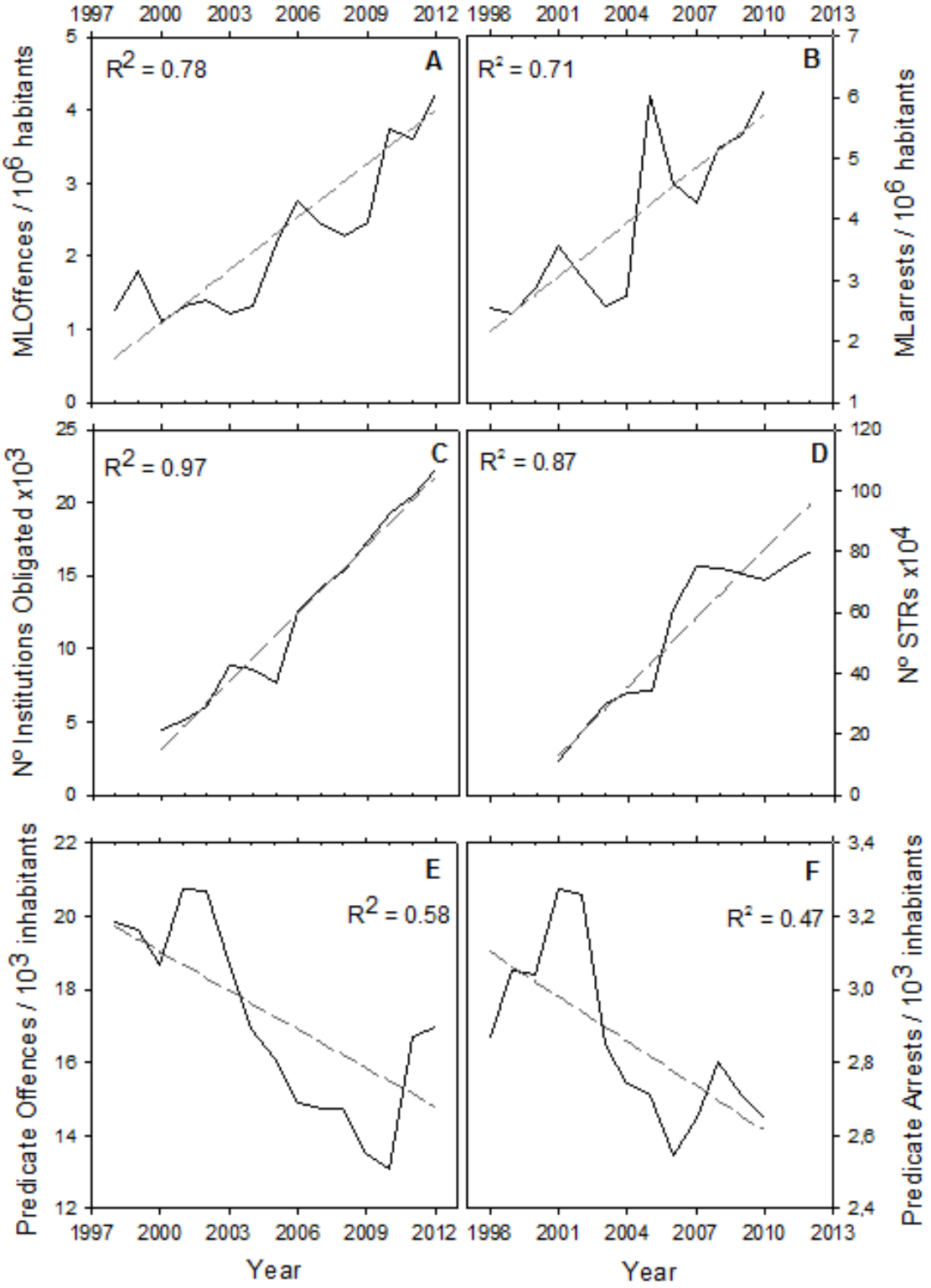
Walker J. and Unger B. 2009. Measuring global money laundering: the Walker gravity model. *Review of Law and Economics* 5(2): 821-853.

Wilson J. Q. and Herrstein R. J. 1985. *Crime and Human nature. The definite study of the causes of crime.* New York: The Free Press.

Woda K. 2006. Money laundering techniques with electronic payment systems. *Information and Security International Journal* 18: 27–47.

Zdanowicz J. 2009. Trade-based money laundering and terrorist financing. *Review of Law and Economics* 5(2): 855-878.

Figure 1. Upper panel: (A) time series of the rates of money laundering (ML) offences, and (B) arrests, in Spain during 1998-2012 period. Middle panel: (C) census of institutions obligate to report suspicious activities and apply customer due diligence measures, and (D) number of suspicious transactions reports (STRs). Lower panel: (E) time series of the rate of predicate offences, and (F) arrests. Discontinuous line represents the fit to a linear model with its explained variance (R^2).


A
R2 = 0.78
B
R² = 0.71
C
R2 = 0.97
D
R² = 0.87
E
R2 = 0.58
F
R² = 0.47
MLOffences / 10^6 habitants
MLarrests / 10^6 habitants
Nº Institutions Obligated x10^3
Nº STRs x10^4
Predicate Offences / 10^3 inhabitants
Predicate Arrests / 10^3 inhabitants
Year
Year

ANEXO III. Publicaciones. "*Detección de fraude financiero mediante redes neuronales de clasificación en un caso real español*"

"Detección de fraude financiero mediante redes neuronales de clasificación en un caso real español"

Autores: Badal-Valero, Elena; García-Cárceles, Belén

Estudios de Economía Aplicada, Vol. 34, núm. 3, 2016, pp. 693-709

Asociación Internacional de Economía Aplicada

Valladolid, España

ESTUDIOS DE ECONOMÍA APLICADA VOL.34, 2016. PÁGS.693-709

Detección de fraude financiero mediante redes neuronales de clasificación en un caso real español

ELENA BADAL-VALERO
Departamento de Economía Aplicada, UNIVERSIDAD DE VALENCIA, ESPAÑA
e-mail: Elena.Badal@uv.es

BELÉN GRACÍA-CÁRCELES
Departamento de Análisis económico, UNIVERSIDAD DE VALENCIA, ESPAÑA
e-mail: Belen.Garcia-Carceles@uv.es

RESUMEN

Este análisis supone una primera aproximación a la implementación de modelos de redes neuronales al trabajo pericial para la detección de operaciones de fraude. Los datos analizados provienen de un caso real de blanqueo de capitales en el que se está colaborando con la Policía Nacional Española. En ellos se cuenta con información de operaciones contables individuales entre las que se cuenta con una proporción de operaciones bien identificadas como fraudulentas con la que es posible entrenar un modelo de clasificación. En este trabajo, tras describir brevemente la metodología utilizada y la estrategia de ajuste se obtiene un modelo con una capacidad predictiva reseñable, incluso con datos de entrenamiento fuertemente desequilibrados. Además, al aplicar técnicas de balanceado de los datos de entrenamiento (SMOTE) se obtiene un resultado que indicaría la viabilidad de este tipo de modelos como herramienta en la planificación y priorización de las tareas de investigación policial, ya que uno de los principales problemas de los investigadores expertos en estos delitos financieros es la incapacidad para traducir la gran cantidad de información que se deriva de las empresas implicadas en patrones de compra de los individuos claramente fraudulentos.

Palabras clave: Redes Neuronales, data mining, fraude financiero, blanqueo de capitales.

Detecting financial fraud using neural network classification models in a real Spanish case

ABSTRACT

This paper explores the possibilities offered by statistical tools based on artificial neural networks for pattern recognition in export work for money-laundering detection. The data is provided by the Spanish Police Department and comes from a case in which is actually working at. Account information is provided, where some accounting entries are identified as fraud. Hence it is possible to use this information to train a classification model. In this analysis, after briefly describing methodology used and fitting strategy, it is presented a model with a promising predictive capacity, even with strongly unbalanced

training data set. After applying balancing technique to the training data (SMOTE) the result is remarkably improved which would indicate the viability of those models as tool for police experts planification, providing a way to reduce the use of expensive research resources.

Keywords: Artificial neural network, data mining, financial fraud, money laundering, forensic accounting.

Clasificación JEL: C450, D220, L51.

1. INTRODUCCIÓN

El blanqueo de capitales es un delito económico que ha evolucionado en el tiempo y que se ejecuta a distintos niveles y en diversas magnitudes. Las cuantías defraudadas oscilan desde el tradicional blanqueo de pequeñas cantidades de dinero proveniente del tráfico minorista y local de drogas, hasta las grandes cantidades (miles de millones de euros) de macro estructuras empresariales que han surgido en las últimas décadas y que operan a escala internacional (Khac y Kechadi, 2010).

A este problema se suma el hecho de que los actuales avances tecnológicos, así como los sistemas de comunicación, han puesto a disposición de los solventes delincuentes herramientas con las que crear estructuras grandes, complejas y coordinadas de empresas con las que evitar la detección del fraude económico (ingeniería financiera) (Petrucelli, 2012).

Por ello, el reto que afrontan los cuerpos de seguridad especializados en esta área es doble: por una parte deben ser capaces de identificar las grandes redes de blanqueo, y por otra, conseguir seleccionar, de la cantidad de información que se deriva de las empresas implicadas, aquella que permita identificar claramente los patrones de compra de los individuos que ocultan blanqueo. Sin embargo, con las técnicas tradicionales se hace muy complicado cumplir estos objetivos (Dutta, 2013).

La aplicación de herramientas estadísticas basadas en redes neuronales artificiales para la detección de patrones de comportamiento han sido aplicadas con buenos resultados en campos tan heterogéneos como el mercado inmobiliario (Caridad y Ceular, 2001), el área biomédica (Uberbacher y Mural, 1991) o en mercados financieros (Muñiz y Alvarez, 1997) (Olmedo y Velasco, 2007), por lo que en este estudio se explora las posibilidades que ofrecen estas herramientas estadísticas para el reconocimiento de patrones de fraude en operaciones individuales.

La base de datos corresponde a datos contables de una empresa matriz investigada por la Policía Nacional Española que contiene información proveniente de los registros efectuados en el transcurso de una investigación por delitos de blanqueo de capitales. Concretamente, contiene más de doce millones de datos extraídos de la contabilidad interna de la empresa núcleo de una estructura empresarial potencialmente defraudadora.

La autoridad policial desarrolla un papel fundamental en el proceso de análisis ya que son los responsables de detectar, investigar y extraer la información que se procesa en este estudio y, es de acuerdo a sus conclusiones que se consiguen identificar las operaciones que son fraudulentas. Dado que los recursos de los investigadores son limitados, en la mayoría de casos no es realista esperar que sea posible hacer un seguimiento exhaustivo de todas las empresas de la organización y por tanto se centran en las que consideran más delictivas[49]. Es por este motivo que la red se entrena con un número limitado de operaciones fraudulentas detectadas lo que permite el ajuste de un modelo de clasificación mediante la combinación de las características de las operaciones, de modo que al introducir nuevas puedan ser clasificadas como potencialmente fraudulentas.

Por el procedimiento de análisis aquí descrito se persigue ofrecer a los investigadores una forma objetiva de identificar, de entre las operaciones pendientes de investigar,

[49] En el caso objeto de estudio se tiene constancia de que la red de empresas la formaban más de 500 empresas entre proveedores y mayoristas.

aquellas que pudieran ser fraudulentas, ayudando a priorizar los recursos de investigación disponibles hacia las empresas que concentran mayor potencial de fraude.

El artículo se organiza como sigue: En el apartado segundo se presentan los conceptos de fraude demostrado y sospecha de fraude y se especifica el modelo de red neuronal artificial que va a utilizarse en todo el análisis. El tercer apartado recoge la estrategia de ajuste, su justificación y limitaciones. Es también en este apartado donde se explora la sensibilidad del modelo al conjunto de entrenamiento y las posibilidades de mejora al balancearlo utilizando SMOTE. El último apartado recoge brevemente las conclusiones y las posibilidades de investigación que se abren a la luz de los resultados obtenidos.

2. DESCRIPCIÓN DE LA METODOLOGÍA

Es de esperar que el modelo de Redes Neuronales sea tan bueno como el "ojo policial" y que aporte una forma de detectar un mayor número de operaciones de fraude. Por tanto, el objetivo del modelo ajustado es identificar correctamente el mayor número de operaciones fraudulentas posible, siendo importante minimizar la proporción de operaciones fraudulentas mal clasificadas (falsos positivos).

Las variables disponibles en la base de datos que caracterizan una operación:

1. El artículo: variable discreta que recoge los 42 tipos de artículos que se comercializaron.

2. El almacén de la mercancía: variable discreta que indica el lugar donde se realizó la compra de producto al proveedor, 12 localizaciones.

3. El/la Administrativo/a: persona que gestiona la operación en el departamento de administración. 43 administrativos/as.

4. El importe total pagado al proveedor por la adquisición del producto. Variable continua.

5. El margen bruto de beneficio de la operación, importe total ingresado por la venta del material al cliente menos el importe total pagado al proveedor.

Variable continua.

6. La cantidad de material contenida en el artículo. Variable continua.

7. Margen de descuento aplicado en la operación. Variable continua.

No se tiene conocimiento a priori de qué artículos, personas o lugares están envueltos en la trama de blanqueo de capitales, sólo se tiene acceso a la identificación de un pequeño número de operaciones fraudulentas (18,99% del total) que la policía ha podido demostrar como tal a través de sus procedimientos de investigación.

Es de reseñar que estos procedimientos se basan en la vigilancia de las personas involucradas en la trama, siendo posterior la detección de dichas operaciones a través de los movimientos contables. Por tanto, aquellas operaciones que no se han detectado como fraudulentas no dejan de ser sospechosas. Es por esto que conviene concretar las siguientes definiciones:

1. Fraude demostrado: consideraremos como fraude demostrado aquellas operaciones que hayan podido ser verificadas por la policía como tal.

2. Fraude sospechoso: el resto de operaciones.

2.1. El fraude demostrado

En primer lugar, se analizan las características mencionadas en las operaciones de fraude demostrado. Con estas descripciones no se pretende hacer juicios de valor "a priori" sobre qué variables deben tenerse en cuenta a la hora de identificar un fraude potencial en el modelo de red. Precisamente, una de las ventajas de este tipo de modelo es que pueden incorporar toda la información disponible en su estructura.

En la Tabla 1 se recoge el recuento de operaciones según artículo. Se muestra únicamente los 18 códigos de artículo más frecuentes, que representan el 94,27% de las operaciones. Se ordena según el número de operaciones de fraude demostrado.

Tabla 1
Variable "Código de Artículo".

Código Artículo	Operaciones "Fraude Demostrado"	Número de Operaciones	% "Fraude Demostrado"
ART1	16.950	139.877	12,12
ART2	10.258	38.240	26,83
ART3	2.665	6.066	43,93
ART4	2.220	6.714	33,07
ART5	2.070	4.880	42,42
ART6	1.286	2.349	54,75
ART7	1.071	23.462	4,56
ART8	876	2.755	31,80
ART9	597	4.307	13,86
ART10	586	2.248	26,07
ART11	476	6.526	7,29
ART12	467	1.081	43,20
ART13	392	10.227	3,83
ART14	368	3.473	10,60
ART15	364	1.970	18,48
ART16	348	1.686	20,64
ART17	314	1.009	31,12
ART18	297	834	35,61

Fuente: elaboración propia.

Por su parte, la Tabla 2 muestra el recuento de operaciones para todos los almacenes en los que se realizan las operaciones y la Tabla 3 los 20 administrativos que más operaciones gestionan, representando el 91,18% de las operaciones totales.

Tabla 2
Variable “Almacén”.

Consigna	Operaciones “Fraude Demostrado”	Número de Operaciones	% “Fraude Demostrado”
ALM1	40.369	217.985	18,52
ALM2	1.284	6.936	18,51
ALM3	929	9.619	9,66
ALM4	912	21.966	4,15
ALM5	230	1.036	22,20
ALM6	79	79	100,00
ALM7	55	181	30,39
ALM8	0	15.241	0
ALM9	0	881	0
ALM10	0	334	0

Fuente: elaboración propia.

Tabla 3
Variable “Administrativos”.

Administrativo	Operaciones “Fraude Demostrado”	Número de Operaciones	% “Fraude Demostrado”
PEX1	9.829	39.191	25,08
PEX2	7.246	20.494	35,36
PEX3	6.255	37.412	16,72
PEX4	3.903	14.308	27,28
PEX5	2.605	4.609	56,52
PEX6	2.042	12.501	16,33
PEX7	2.005	8.414	23,83
PEX8	1.917	24.947	7,68
PEX9	1.872	9.386	19,94
PEX10	925	4.530	20,42
PEX11	904	4.497	20,10
PEX12	804	2.565	31,35
PEX13	648	1.817	35,66
PEX14	620	10.768	5,76
PEX15	448	1.694	26,45
PEX16	429	12.538	3,42
PEX17	330	698	47,28
PEX18	282	1.463	19,28
PEX19	206	38.239	0,54

Fuente: elaboración propia.

Finalmente, en la Tabla 4 se recogen los descriptivos de las variables continuas disponibles, en concreto, se dispone del importe total de cada operación, la cantidad de material que contiene cada compra y los márgenes de descuento y de beneficio.

Tabla 4
Variables continuas: importe total, cantidad de material, margen de descuento y margen bruto de beneficio.

	Mín.	P25	Mediana	Media	P75	Máx.	NA´s
Importe total							
Fraude	-0,18	1,03	4,10	36,62	37,11	2.799,15	733,86
Resto	-0,18	1,74	12,91	68,81	68,00	1.818,30	153,30
Cantidad de Material							
Fraude	0	13	78	201	260	18.560	6.517
Resto	0	51	166	277	377	9.980	2.211
Margen de descuento							
Fraude	0	0	1,00	1,00	1,00	1,02	230
Resto	0	0	1,00	0,95	1,00	1,80	79
Margen Bruto de Beneficio							
Fraude	-25,53	0,01	0,03	0,66	0,04	40,60	5.612
Resto	-17,39	0,01	0,03	0,36	0,04	28,64	663

Fuente: elaboración propia.

Al realizar este sencillo ejercicio descriptivo los datos parecen determinar ciertos patrones de fraude (artículos y administrativos con mayor proporción de implicación en las operaciones de fraude, mayor margen bruto de beneficio …) sin embargo, la ingeniería financiera a la que pueden acceder estas empresas hace desconfiar de la fiabilidad de estas apariencias.

De hecho, la realidad muestra que dichos patrones no son tan determinantes como a priori pudiera parecer para que, finalmente, se pueda demostrar que una operación (y, sobre todo las personas implicadas en ellas) estén delinquiendo. De ahí que se busque una alternativa en la que pueda incorporarse la mayor cantidad de información disponible que, en modelos más sencillos e interpretables, no sea posible captar la complejidad de las relaciones que lleven a una adecuada detección de operaciones de fraude en base a sus características.

2.2. Especificación del modelo de red

Este análisis supone una primera aproximación a la implementación de modelos de redes neuronales al trabajo pericial para la detección de operaciones de fraude. Por ello se ha utilizado una estructura de red muy sencilla: la red de propagación hacia atrás (del inglés “Back-Propagation network”) o perceptron de una capa oculta (del inglés “Single Hidden Layer Perceptron”).

En esta estructura de red hay tres elementos: las unidades de salida de la red (outputs, *Y*), las unidades de entrada (inputs, *X*) y las características derivadas de los inputs (*Z*) (Hastie *et al.*, 2008).

Las unidades ocultas, Z, se obtienen como una combinación lineal de los inputs, transformada mediante una función de activación que se define como la función sigmoidea:

$$\sigma(\upsilon) = 1/(1+e^{-\upsilon})$$

donde: $\sigma(\upsilon) = [0,1]$,y $\upsilon =]-\infty, +\infty[$

Para una clasificación en k clases, hay *K* unidades en la salida de la red, con la k-ésima unidad modelizando la probabilidad para la clase *k*. Hablamos, por tanto de *k* medidas objetivo Y_k, $k = 1,\ldots, K$, codificadas como 0,1. La estructura de la red se representa mediante las tres expresiones siguientes:

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X),\ m = 1,...,M,$$
$$T_K = \beta_{0k} + \beta_k^T Z,\ k = 1,...,K,$$
$$f_k(X) = g_k(T)\ k = 1,...,K,$$

donde: $Z = (Z_1, Z_2, ..., Z_M)$, y $T = (T_1, T_2, ..., T_k)$.

Los parámetros del modelo (pesos o ponderaciones), inicialmente desconocidos, se ajustan utilizando los errores Deviance (en el caso de redes de clasificación), definidos como:

$$R(\theta) = -\sum_{i=1}^{N} \sum_{k=i}^{K} y_{ik} \log fk(k_i)$$

Además, pueden incorporarse unidades de sesgo tanto en los nodos intermedios como en la función de salida que, pensadas como un input adicional, capturarían los interceptos α_{0m} y β_{0k}. El conjunto completo de pesos Θ se denota como:

$$\{\alpha_{0m}, \alpha_m; m = 1,2,...,M\}\ M(p+1)\ pesos,$$
$$\{\beta_{0k}, \beta_k; m = 1,2,...,M\}\ K(M+1)\ pesos,$$

Finalmente, la función de salida $g_k(T)$, permite la transformación final de los vectores de salida T, utilizando la función de transformación softmax (en el caso concreto) definida como:

$$g_k(T) = \frac{e^{T_k}}{\sum_{l=1}^{K} e^{T_L}}$$

Por tanto, el modelo de red neuronal es un modelo no lineal multilogit que utiliza una transformación de los inputs (*X*), mediante pesos (Θ) fijados a través de la minimización de los errores (Deviance, $R(\Theta)$) mediante un procedimiento de actualización back-propagation (Ripley, 1996).

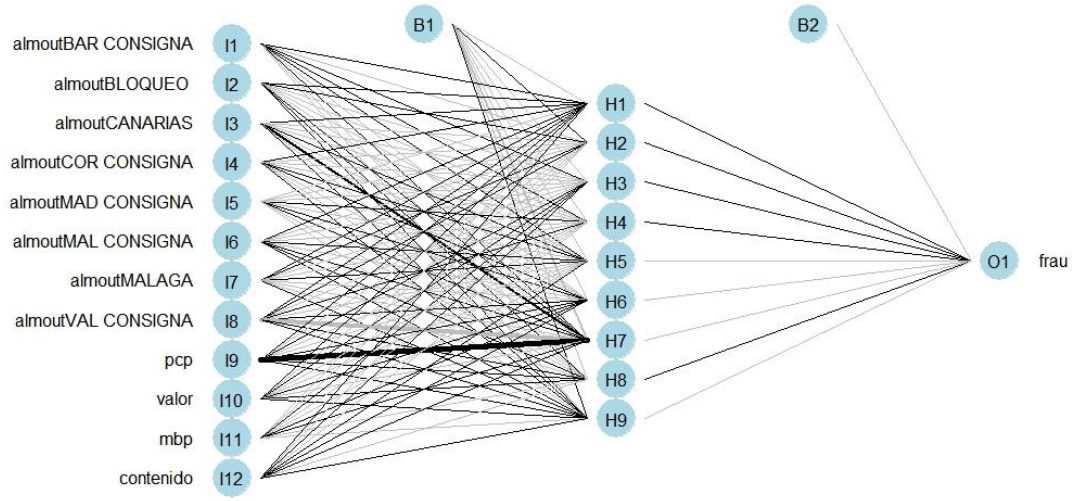


Figura 1: esquema de la estructura de red ajustada. Nota: sólo se representan las variables input I1 a I12 para mayor claridad. Si bien la estructura de la red en el caso de estudio considera 101 inputs. Fuente: elaboración propia.

En el caso concreto que se analiza en este trabajo, la estructura de la red se ha recogido en la Figura 1 (Beck, 2015). En ella existe una único nodo en la salida (Y_k=1) codificada 0 (no es posible concluir que la operación es fraudulenta) y 1 (la operación es identificada como fraudulenta), representada en el extremo derecho de la figura con O1.

Las llamadas características derivadas (Z_m, m = 1,..,9) son los nodos H1 a H9, que se crean a partir de combinaciones lineales de las variables consideradas (inputs: I1 a I100, X_p, p = 1,...,100). En la estructura representada en la Figura 1 puede también apreciarse la inclusión de variables que pretenden capturar el sesgo en cada uno de los nueve nodos de la capa oculta (el vector de pesos B1, α_{0m}) y en la función de salida (el peso B2, β_{0k}).

3. IDONEIDAD DE LA ESTRUCTURA DE RED "BACK PROPAGATION NETWORK" PARA LA DETECCIÓN DEL FRAUDE DEMOSTRADO

3.1. Estrategia de muestreo para el set de entrenamiento

El proceso de ajuste se inicia testando las limitaciones de los recursos informáticos disponibles. Es decir, teniendo en cuenta la magnitud de la base de datos disponible y la estructura de red que se desea aplicar es necesario establecer una estrategia para encontrar un compromiso entre estructura de la red y cantidad de datos a incluir en el set de entrenamiento, que no comprometa la finalidad del análisis.

La estrategia seguida es la siguiente:

1. El set de entrenamiento será tal que el desequilibrio en los datos sea comparable a la muestra de operaciones de la policía, en el que la proporción de operaciones fraudulentas respecto al resto se sitúa en torno al 19%. Es decir, siendo Am el número operaciones fraude de la muestra, Bm el número de resto de operaciones, Ace el

número operaciones fraude del conjunto de entrenamiento y Bce el número de resto de operaciones del conjunto de entrenamiento, se cumple que:

$$\frac{Am}{Bm} \approx \frac{Ace}{Bce} \approx 19\%$$

2. Si bien lo habitual es reservar el 20% de los datos para el conjunto comprobación y utilizar el 80% en el de entrenamiento, en este caso la proporción será la opuesta (80% comprobación, 20% entrenamiento) ya que se prioriza el número de nodos de la capa oculta de la red y la utilización de todas las variables disponibles.

Esta decisión se fundamenta en tres motivos: por una parte se desea maximizar el aprovechamiento de la estructura; por otra, si no se tomaran todas las variables de entrada se estarían tomando decisiones "a priori" en relación a la inclusión/exclusión de variables, lo que iría en contra de los objetivos establecidos en este trabajo; finalmente, utilizar menos datos asegura que los resultados obtenidos, en cualquier caso, tiendan a estar por debajo de los que cabría esperar con mayor información (estrategia conservadora adecuada al enfoque de "primera aproximación" que se desea aportar con este trabajo).

Así, el ajuste se realiza manteniendo todos los inputs de entrada de la red (101), 9 nodos en la capa oculta de la red y un número de operaciones de entrenamiento por debajo del habitual (50.000 operaciones, el 20% del total de operaciones disponibles).

3. Además, siguiendo una estrategia conservadora, en este apartado se presentan los resultados del modelo de red especificado sin aplicar corrección alguna, es decir, el ajuste se realiza sin utilizar técnicas que permitieran compensar alguno de los problemas que habitualmente se señalan en relación a este tipo de datos (remuestreo (Buhlmann y Yu, 2002), ensamblado (Meir y Rätsch, 2003), técnicas de sobremuestreo o inframuestreo (SMOTE-Boost de Chawla et al. 2003)).

La finalidad de esa estrategia es obtener un resultado a partir del cual se puedan aplicar mejoras sucesivas (el balanceado de los datos se aborda en el apartado siguiente) y esperar un mayor ajuste del modelo conforme pudieran superarse las limitaciones informáticas.

3.2. Punto de partida: modelo ajustado con datos desequilibrados

Utilizando el entorno de programación de R (R Core Team, 2015) se ha empleado el paquete de R nnet (Venables y Ripley, 2002) de la librería del mismo nombre, ajustando sus parámetros conforme a las especificaciones descritas y se evalúa el ajuste del modelo mediante el enfoque tradicional: construyendo una matriz de confusión (Tabla 5), donde se confrontan las operaciones bien y mal clasificadas basadas en el conjunto de comprobación. En ella se especifican los recuentos utilizados para el cálculo de las tasas con las que evaluar los resultados (Tabla 6) mediante la tasa de operaciones bien clasificadas (*TBC*), la de operaciones fraudulentas bien clasificadas (verdaderos positivos, *sensitivity* o *recall, TVP*) y c) operaciones fraudulentas mal clasificadas (falsos positivos, TFP).

Formalmente sería: $TBC = \frac{(P_{11} + P_{22})}{(P_{11} + P_{12} + P_{21} + P_{22})} 100; TVP = \frac{(P_{22})}{(P_{21} + P_{22})} 100; TFP = \frac{(P_{21})}{(P_{21} + P_{22})} 100$

Tabla 5
Matriz de confusión.

		Clasificación según el modelo	
		Resto (0)	Fraude (1)
Clasificación Policial	Resto (0)	P_{11}	P_{12}
	Fraude (1)	P_{21}	P_{22}

Fuente: elaboración propia.

Tabla 6
Tasas obtenidas a partir de la matriz de confusión.

Operaciones Bien Clasificadas	72,16%
Operaciones Mal Clasificadas	27,83%
TVP (*recall)*	17,63%
TFP	82,36%

Nota: TVP = Tasa de Verdaderos Positivos, operaciones fraudulentas bien clasificadas; TFP = Tasa de falsos positivos, operaciones fraudulentas mal clasificadas; Operaciones bien clasificadas (TBC). Fuente: elaboración propia.

La tasa de ajuste global del modelo es reseñable, con un 72,16% de operaciones bien clasificadas. Esta tasa desciende al 17,63% cuando se trata de operaciones de fraude bien clasificadas. Por su parte, la tasa de falsos positivos es elevada (superior al 80%) aunque hay que interpretarla con cuidado, ya que incluye en su recuento operaciones fraudulentas no detectadas por la policía pero que en realidad lo son. Es decir, como se ha explicado al inicio del punto 2, las operaciones que no han sido detectadas como fraude no dejan de ser "sospechosas".

Teniendo en cuenta que no se han utilizado técnicas para compensar el desequilibrio inicial de la base de datos, ni para considerar el coste de los errores de clasificación, el resultado es, por lo menos, prometedor.

3.3. Posibilidades de mejora: Sensibilidad a cambios en el conjunto de entrenamiento

En este apartado se realiza un experimento para evaluar la sensibilidad del modelo a cambios en el conjunto de entrenamiento con dos objetivos. En primer lugar, dado que se ha seguido una estrategia de muestreo por la que se limita la cantidad de información en el entrenamiento de la red al 20% de los datos disponibles (ver apartado 3.1) es conveniente obtener una medida de la influencia de esta decisión en el ajuste. Esta medida, además, es una magnitud de las posibilidades de mejora del ajuste que no se limita sólo a superar las restricciones de potencia de cálculo informático, sino que puede afrontarse aplicando técnicas de balanceado de los datos mediante remuestreo (tal como se aborda en el apartado siguiente). En segundo lugar, es un hecho conocido que los pesos que el modelo asigna a los nodos de la red son altamente sensibles a cambios en el conjunto de datos de entrenamiento debido al procedimiento de actualización de gradiente descendente del diseño *back-propagation*.

La estrategia seguida es la siguiente: se establecen 100 conjuntos de entrenamiento del mismo tamaño que el anterior (50.000 operaciones diferentes cada uno) y con el mismo ratio Ace/Bce (19%). Se ajustan 100 modelos de red con la misma estructura anterior, fijando los mismos pesos iniciales y utilizando todo el conjunto de inputs disponibles (nuevamente, se emplea la función del paquete de R nnet (*op. cit.*)). Para cada modelo se han obtenido los ratios antes definidos a partir de las correspondientes matrices de confusión y sus conjuntos de comprobación.

La Tabla 7 recoge el valor promedio y la desviación típica del ajuste para los 100 modelos. La variabilidad alrededor de la media (11,2% para los verdaderos positivos y 88,98% para los falsos positivos) se sitúa por debajo de los 4 puntos porcentuales.

Tabla 7
Ajuste de los 100 modelos de red.

	Media	Desviación Típica	Coeficiente de asimetría
Operaciones Bien Clasificadas Operaciones Mal Clasificadas	76,50% 23,50%	2,5113%	1,549
TMVP TMFP	11,02% 88,98%	3,697%	1,547

TMVP = Tasa Media de Operaciones Fraudulentas bien clasificadas.
TMFP= Tasa Media de Operaciones Fraudulentas mal clasificadas.

$$TMVP = \sum_{i=1}^{100} \frac{\left(P_{22}^{i}\right)}{\left(P_{21}^{i} + P_{22}^{i}\right)}; TMFP = \sum_{i=1}^{100} \frac{\left(P_{21}^{i}\right)}{\left(P_{21}^{i} + P_{22}^{i}\right)} .$$

Fuente: elaboración propia.

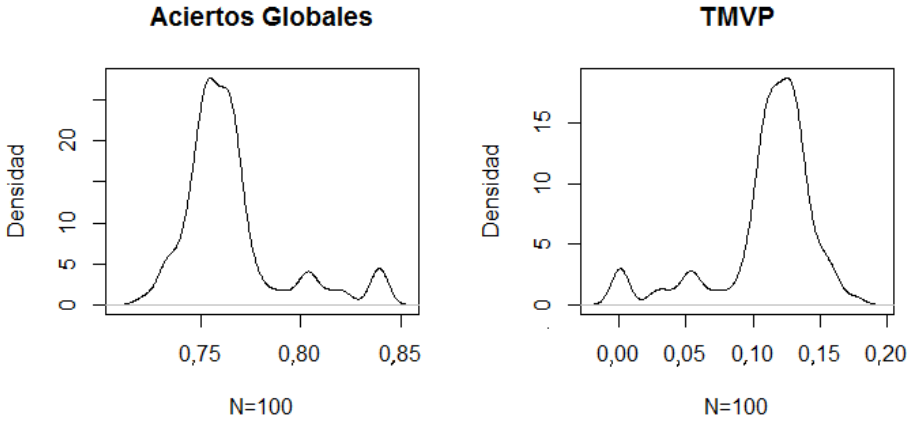


Figura 2: distribución de densidad de las tasas calculadas en los 100 modelos. Fuente: elaboración propia.

En el experimento realizado para detectar la sensibilidad del ajuste a cambios en el conjunto de entrenamiento, se observa que la distribución de verdaderos positivos es asimétrica (Figura 2) debido a un grupo de 14 modelos cuya proporción cae por debajo del 7,5% sin los que se obtiene que la tasa de verdaderos positivos $TVP \sim N(12,3\%,1,7\%)$. La asimetría indicaría que la exclusión/inclusión de casos en el conjunto de entrenamiento no es neutral: la estrategia de selección aleatoria deja margen para la mejora.

3.4. Ajuste mediante balanceado del conjunto de entrenamiento

En el caso que nos ocupa cabe destacar que: a) no se dispone de información acerca del coste asociado a clasificar erróneamente una operación fraudulenta lo que limita el uso de medidas sensibles al coste (matriz de costes, curvas de coste), b) la proporción de operaciones fraudulentas en la muestra es muy inferior respecto del resto de operaciones y c) la estrategia de selección del conjunto de entrenamiento no es neutral.

Las técnicas de muestreo permiten compensar la falta de información de la clase minoritaria (operaciones fraudulentas) en el conjunto de entrenamiento a la vez que

implican una forma alternativa para considerar en el proceso de aprendizaje los costes asociados a los errores de clasificación. Una combinación de sobremuestreo de la clase minoritaria e inframuestreo de la mayoritaria puede ser la estrategia con mejores resultados (Japkowicz, 2000). Ahora bien, si se aplica de forma aleatoria, el inframuestreo puede llevar a eliminar ejemplos importantes de la información de entrenamiento, cuestión que, como se ha visto en el apartado anterior, puede llevar a un rendimiento muy pobre del clasificador. Por su parte, el sobremuestreo aleatorio de la clase minoritaria puede llevar al sobreajuste y a la falta de generalidad, ya que crea regiones de decisión más pequeñas y específicas.

Una alternativa es generar sintéticamente nuevos ejemplos que amplíen la información para aquellos casos de la clase minoritaria difíciles de clasificar. SMOTE (de *Synthetic Minority Oversampling Technique*) es una técnica que proporciona información nueva relativa a la clase minoritaria además de infrarrepresentar la clase mayoritaria (Chawla *et al*. 2002). Con esta técnica se generan ejemplos sintéticos en el segmento que une un ejemplo de la clase minoritaria con sus *k* vecinos más próximos, dependiendo de la cantidad de ejemplos sintéticos que se requieran, se escogerán aleatoriamente vecinos de estos *k* vecinos. Para cada clase se considera el voto de los vecinos más próximos con lo que la región de decisión es más grande y menos específica (hay más puntos, está mejor descrita), a diferencia de las regiones más pequeñas y específicas que se crean con el muestreo aleatorio.

Al aplicar la técnica de SMOTE[50] al mismo modelo del apartado 3.2, a partir de su matriz de confusión (Tabla 8) se obtienen los resultados recogidos en la Tabla 9.

Tabla 8

Matriz de confusión al aplicar SMOTE.

		Clasificación según el modelo	
		Resto (0)	Fraude (1)
Clasificación Policial	Resto (0)	41,72%	42,33%
	Fraude (1)	4,88%	11,05%

Fuente: elaboración propia.

Tabla 9

Tasas obtenidas a partir de la matriz de confusión al aplicar SMOTE.

Operaciones Bien Clasificadas	52,77%
Operaciones Mal Clasificadas	47,22%
TP RATE (*recall*)	69,36%
FP RATE	30,63%

Nota: TP RATE = Tasa de Verdaderos Positivos, operaciones fraudulentas bien clasificadas; FP RATE = Tasa de falsos positivos, operaciones fraudulentas mal clasificadas; Operaciones bien clasificadas (TBC). Fuente: elaboración propia.

La mejora obtenida es reseñable por varios motivos. En primer lugar destaca el descenso de la tasa de falsos positivos de 82,36% del modelo con un conjunto de entrenamiento desequilibrado (Tabla 6) a 30,63%. En segundo lugar, la proporción de operaciones fraudulentas bien clasificadas asciende del 17,66% del modelo inicial al 69,36% en el modelo con SMOTE.

[50] Se utiliza la función SMOTE del paquete de R DMwR (Torgo, L. 2010).

La importancia de esta mejora no es baladí, dado que la diferencia de utilizar el modelo del apartado 3.2 como soporte a las tareas periciales frente al modelo con SMOTE, implica reducir considerablemente la probabilidad de malgastar recursos policiales en investigar empresas y personas detrás de operaciones que el modelo indica como fraudulentas y que en realidad no lo son.

En la tabla 8 se observa también un claro descenso de las Operaciones Bien Clasificadas de forma global (52,77%) en este nuevo modelo en relación a los resultados de la Tabla 6 (72,16%). Sin embargo, tal como se discute en los párrafos anteriores lo relevante es acertar bien las operaciones fraudulentas, dado que las operaciones no fraudulentas en realidad pudieran serlo.

Por tanto, el uso de estos modelos en la detección de casos reales de fraude financiero, junto con la implementación de técnicas de muestreo como SMOTE que mejoren el desequilibrio de los datos, puede servir de herramienta a los investigadores para conocer patrones de comportamiento en casos de blanqueo de capitales y ofrecer mayor información para su detección, así como también orientar a las autoridades a aquellas empresas cuyos comportamientos aparentan ser fraudulentos.

4. CONCLUSIONES

Los resultados obtenidos en la exploración de los modelos de red como herramienta de trabajo en la actividad pericial han sido notables. Por un lado, ha sido posible incorporar toda la información (variables) disponibles en la estimación del modelo. Por otro lado, la rapidez con la que se ha obtenido el ajuste ha sido destacable.

En el modelo ajustado con datos de entrenamiento sin balancear se observa que la tasa de ajuste es reseñable, con un 72,16% de operaciones bien clasificadas. Esta tasa desciende al 17,63% cuando se trata de operaciones de fraude bien clasificadas. Por su parte, la tasa de falsos positivos (superior al 80%) hay que interpretarla con cuidado, ya que incluye en su recuento operaciones fraudulentas no detectadas por la policía pero que en realidad lo son.

La estrategia de muestreo seguida para la selección del conjunto de datos de entrenamiento no es neutral, tal como se constata con el experimento realizado con 100 conjuntos de entrenamiento distintos para detectar la sensibilidad del ajuste a cambios en el conjunto de entrenamiento. Se observa que la distribución de verdaderos positivos es asimétrica debido a un grupo de 14 modelos cuya proporción cae por debajo del 7,5% sin los que se obtiene que la tasa de verdaderos positivos $TVP \sim N(12,3\%,1,7\%)$. La asimetría indicaría que la exclusión/inclusión de casos en el conjunto de entrenamiento no es neutral: la estrategia de selección aleatoria deja margen para la mejora.

Del experimento anterior y de la propia metodología de ajuste propia de la estructura de red utilizada, así como el desconocimiento de los costes asociados a los errores de clasificación llevan a aplicar una estrategia de mejora con datos de entrenamiento balanceados utilizando la técnica SMOTE. Se observa cómo, a pesar de descender la tasa de aciertos globales, la capacidad de detección de operaciones fraudulentas mejora hasta alcanzar casi un 70%, que prácticamente alcanza la tasa de aciertos globales del modelo sin balancear.

Así, habiéndose conseguido un ajuste reseñable se han detectado importantes oportunidades de mejora a través de la revisión de la estrategia de selección de casos para el conjunto de entrenamiento, como por ejemplo, la exploración de otras estructuras de

red y otros métodos de aprendizaje. Además, también se podría combinar el modelo con diferentes estrategias de balanceado como la de tipo "*Boosting*" (Freund y Schapire, 1996) que junto con la técnica SMOTE (SMOTE Boost de Chawala *et al.*, 2003) permitiría mejorar la identificación de las operaciones fraudulentas.

Finalmente, los resultados aquí obtenidos abren un amplio abanico de posibilidades a la mejora del trabajo pericial en la detección del fraude financiero y el blanqueo de capitales utilizando este tipo de herramientas predictivas para lo que sería deseable, mediante el uso de análisis de sensibilidad, la búsqueda de patrones de fraude que puedan describirse "a priori".

REFERENCIAS BIBLIOGRÁFICAS

Beck, M.W. (2015). *"NeuralNetTools: Visualization and Analysis Tools for Neural Networks".* Version 1.4.0. Disponible en: http://cran.r-project.org/web/packages/NeuralNetTools/ [27/07/2016].

Büchlmann, P., Yu B. (2002). "Analyzing Bagging".The Annals of Statistics. Vol. 30, No. 4, pp. 927-961.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). "*Smote: Synthetic minority over-sampling technique*". Journal of Artificial Intelligence Research, pp. 321-357.

Caridad, J. M., & Ceular, N.. (2001). "*Un análisis del mercado de la vivienda a través de redes neuronales artificiales"*. Estudios de economía aplicada, (18), pp. 67-81.

Chawla, N. V., Lazarevic, A., Hall, L. O., and Bowyer, K. W. (2003b). *"Smoteboost: Improving Prediction of the Minority Class in Boosting"*. In Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases, Vol.16, pp. 107-1 19, Dubrovnik, Croatia.

Duncan, L.T., y CRAN Team.(2016). Package 'RCurl'. Disponible en: https://cran.r-project.org/web/packages/RCurl/index.html [27/07/2016].

Dutta, S. (2013). *Statistical Techniques for Forensic Accounting*. Upper Saddle River (NJ): FT Press.

Goh, A.T. (1995). "Back-propagation neural networks for modelling complex systems". *Artificial Intelligence in Engineering*, Vol.9, nº3, pp. 143-151.

Hastie, T., Tibshirani, R., y Friedman, J. (2008). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction* (2nd ed.). Standfor: Springer. (pp. 392-396).

Heidarinia, N., Harounabadi, A., y Sadeghzadeh, M. (2014). *"An intelligent Anti-Money Laundering Method for Detecting Risky Users in the Banking Systems"*. International Journal of Computer Applications. No.22, pp. 35-39.

Japkowicz, N. (2000), *"The Class Imbalance Problem: Significance and Strategies".* In Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI'2000): Special Track on Inductive Learning, Las Vegas, Nevada.

Khac, N. L., y Kechadi, M. (2010). *"Application of Data Mining for Anti-Money Laundering Detection: A Case Study"*. IEEE International Conference on Data Mining Workshops.

Lin-Tao, Ji, N., y Zhang, J.-L. (2008). *"A RBF neural network model for anty-money laundering"*. International Conference on Wavalet Analysis and Pattern Recognition, pp. 209-215.

Meir, R., y Rästch, G. (2003). "*An introduction to boosting and leveraging*". Lecture Notes in Computer Science, pp 118-183.

Muñiz, P. y J. A. Alvarez (1997). "*Comportamiento del Mercado: Hipótesis alternativas*". Revista de Bolsas y Mercados Españoles, Vol.60, pp 29-33.

Ngai, E., Hu, Y., Wong, Y..,Chen, Y., y Sun, X. (2011). *"The application of data mining techniques in financial fraud detection: A classification frame work and an academic review of literature"*. Decision Support Systems, Vol.50, nº3, pp. 559-569.

Olden, D. (2005). "*Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks".* Ecological Modelling, nº 154, pp. 135-150.

Olmedo, E., Velasco, F., & Valderas, J. M. (2007). "*Caracterización no lineal y predicción no paramétrica en el IBEX35*". Estudios de Economía Aplicada, 25(3).

Petrucelli, J. (2012). *Detecting Fraud in Organizations: Techniques, Tools, and Resources*. Washington DC: John Wiley & Sons, Inc.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org/

Ripley. (1996). *Pattern Recognition and Neural Networks*. Cambridge University:Press.

Shmield, R. y Ames M. (2013). *"Next generation detection engine for fraud and compliance"*. SAS Global Forum, pp. 1-6.

Torgo, L. (2010) *Data Mining using R: learning with case studies*, CRC Press (ISBN: 9781439810187).

Uberbacher, E. C., & Mural, R. J. (1991). "*Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach*". Proceedings of the National Academy of Sciences, 88(24), pp.11261-11265.

U.S. Congress, Office of Technology Assessment. (1995). *"Information Technologies for Control of Money Laundering"*. Washington, DC: U.S: U.S. Government Printing Office. pp. 55-72.

Venables , W.N. y Ripley, B. (2002). *Modern Applied Statistics with S*. 4th Edition. New York: Springer.

Wickham, H. (2015). stringr: Simple, Consistent Wrappers for Common String Operations. R package version 1.1.0. https://CRAN.R-project.org/package=stringr

Wickham, H, y Chang, W. (2016). devtools: Tools to Make Developing R Packages Easier. R package version 1.12.0. https://CRAN.R-project.org/package=devtools

ANEXO IV. Publicaciones."*Combining Benford's Law and Machine Learning to detect Money Laundering. An actual Spanish Court case*"

"Combining Benford's Law and Machine Learning to detect Money Laundering. An actual Spanish Court case"

Badal-Valero, Elena; Pavía Miralles, José Manuel; Álvarez Jareño, José Antonio

En revisión, Forensic Science Internacional

Elsevier

Combining Benford's Law and Machine Learning to detect Money Laundering. An actual Spanish Court case.

ELENA BADAL-VALERO
Department of Applied Economics. University of Valencia
Avenida de los Naranjos, s/n, 46022 Valencia

JOSÉ A. ALVAREZ-JAREÑO
Department of Applied Economics. University of Valencia
Avenida de los Naranjos, s/n, 46022 Valencia

JOSE M. PAVÍA (corresponding author)
Department of Applied Economics. University of Valencia
Avenida de los Naranjos, s/n, 46022 Valencia
e-mail: pavia@uv.es

Abstract

This paper is based on the analysis of the database of operations from a macro-case on money laundering orchestrated between a core company and a group of its suppliers, 26 of which had already been identified by the police as fraudulent companies. In the face of a well-founded suspicion that more companies have perpetrated criminal acts and in order to make better use of what are very limited police resources, we aim to construct a tool to detect money laundering criminals by combining Benford's Law and machine learning algorithms. After mapping each supplier's set of accounting data into a 21-dimensional space using Benford's Law, we apply machine learning algorithms to flag up additional companies that could merit further scrutiny.

Key Words: Benford's Law, fraud, money laundering, neural networks, random forests, ridge logistic regression.

Highlights

A new approach for detecting money laundering perpetrators is proposed.

Benford's Law is used to map sets of accounting data into a 21-dimensional space.

Further potential criminals are found modelling accounting patterns by machine learning.

Acknowledgements

The authors wish to thank M. Hodkinson for translation of the paper into English. This work has been supported by the Spanish Ministry of Economics and Competitiveness under grant CSO2013-43054-R.

1. INTRODUCTION

Practically on a daily basis, newspapers as well as radio and television news programs report on the occurrence of some or other economic crime: tax fraud, money laundering, corruption, embezzlement of public funds, etc. These are referred to as white collar crimes, crimes which call for more intelligence than brute force. Consequently, the tools for their detection and prosecution also have to be more sophisticated. In 1972, the American economist Hal Varian (1972) proposed the use of Benford's Law as a prospective diagnostic tool for highlighting sets of economic and financial operations that require more in-depth scrutiny.

The Benford Law was discovered by the astronomer and mathematician Simon Newcomb in 1881 (Newcomb, 1881), although its true value was not recognised until 57 years later when the physicist Frank Benford rediscovered it. Benford's Law affirms that the frequency distribution of leading digits in many real-life collections of numbers is not uniform. Benford's Law defines a biased distribution based on a logarithm law.

In the business and economics world, many data sets obey Benford's Law. Hence, if the economic data follow Benford´s Law naturally, its non-compliance could be indicating the possible presence of irregularities in accounting or business-to-business transactions. Benford's Law can be used as a tool to direct us to an economic crime of money laundering or tax evasion (Nigrini, 1992).

Failure to comply with Benford´s Law is only evidence that the values of a set of numbers can be manipulated. It does not itself identify a crime. Benford's Law is not a universal law, like the law of gravity, and there will be data sets that do not conform to it. However, if the data appear manipulated, something must be behind this, and it would therefore be appropriate to investigate the reason for this anomalous behaviour.

On this basis, and in the context of a real police investigation, we analyse a database composed of the operations carried out between a company suspected of money laundering (parent or core company) and a group of more than 600 suppliers, some of which had previously been identified by police authorities as fraudulent or cooperative. The aim is to find patterns of behaviour in this set of companies which would then enable the identification of other companies that might deserve a more detailed scrutiny.

We use Benford's Law as a tool to characterize the accounting records of business operations between the core company and the suppliers and we apply four classification models (logistic regression, neural networks, decision trees and random forests) to identify other potential fraudulent suppliers. In the models, we incorporate the knowledge provided by the police on which companies have already been identified as collaborators. The ultimate aim is to uncover the largest number of fraudulent companies possible and, at the same time, reduce the likelihood of wrongly targeting companies who are operating correctly. Through the use of this methodology, a group of companies have been identified that show a greater probability of fraudulent

operations. This enables the scarce resources of the police investigators to be used more efficiently by focusing more on these companies.

This paper has been completed in the context of a police investigation from a Spanish case of money laundering in which the authors have collaborated as forensic accountants. As far as we know, this work represents the first step towards the use of machine learning for the detection of financial fraud in Spanish judicial cases.

The rest of the paper is organized as follows. Section 2 briefly reviews the use of Benford's Law in the literature. Section 3 focuses on methodological issues. In this section, we introduce Benford's Law, we detail the statistical tests implemented, describe the machine learning methods used and, after drawing attention to the challenge that entails handling clearly imbalanced data sets, we present the strategies used to deal with this. The data and the treatments to which they have been subjected are presented in section 4. Section 5 shows the results obtained after applying the methods considered. A section of conclusions ends the paper.

1. A REVISION OF THE LITERATURE

Outside the area of economics, Benford´s Law has been applied to different fields of knowledge. In computing, Torres, Fernández, Gamero and Sola (2007) have verified that the size of the files stored in a personal computer follows Benford´s Law. This knowledge can help to develop more effective data storage procedures, to carry out maintenance, or as a tool for detecting viruses or errors. In mathematics, Luque and Lacasa (2009) have uncovered a statistical behaviour in the sequence of prime numbers and of zeros in the Riemann zeta function that coincides with the generalized Benford´s Law. In election forensics, Benford's Law has been extensively used as a tool for detecting election fraud (eg, Mebane, 2007; Mebane, Alvarez, Hall and Hyde, 2008; Pericchi and Torres, 2011), although its effectiveness in this area has been called into question by Deckert, Myagkov and Ordeshook (2011). Furthermore, Benford's Law has also been applied to study the length of rivers (Rauch *et Al*., 2011), to detect scientific fraud (Diekmann, 2007), to assess quality of survey data (Judge and Schechter, 2009) and to discover manipulation in self-reported toxic emissions data (de Marchi and Hamilton, 2006).

The use of Benford's Law in the economic area is more widespread. Nigrini (1992) suggests that it can be used to detect fraud in income tax returns and other accounting documents. This idea is reinforced in Nigrini (1994), who states that "*individuals, either through psychological habits or other constraints peculiar to the situation, will invent fraudulent numbers that will not adhere to the expected digital frequencies*"; and in Nigrini and Mittermaier (1997), who proposed extending its use as a regular tool in auditing. Currently, a line of development in accounting applies Benford's Law to detect fraud, or the "manufacture" of data, in accounting and financial documents (Durtschi, Hillison and Pacini, 2004).

Quick and Wolz (2003), Tam Cho and Gaines (2007) and Alali and Romero (2013) constitute other examples. Quick and Wolz (2003) examine data of income and from the balance sheets of various German companies for the years 1994 to 1998 and find that the series of numbers of the first and second digits are, in most cases, adjusted to Benford´s Law. They find these patterns both when the analysis is performed on an annual basis and also when the inspection is carried out for the whole period. Tam Cho and Gaines (2007) scrutinize financial transactions undertaken in the context of election campaign finance and point to pockets of data that merit more careful inspection. In a more recent study, Alali and Romero (2013) analyse the financial information corresponding to more than ten years of accounting data of a large sample of US public companies and find that the current assets (equipment, property, accounts receivable) do not conform to Benford's Law. This result leads them to conclude that during the period analysed there was an overestimation of the asset.

Günnel and Tödter (2009) consider that Benford's Law is a simple, objective, powerful, and effective tool for identifying anomalies in big samples of data that require a detailed inspection; a vision shared by many authors. There are fewer consensuses on how the knowledge provided by Benford's Law should be used however. Günnel and Tödter (2009) argue that controls over data manipulation should focus on the first digit. Whereas, Ramos (2006) states that the analysis should focus on the first three digits. According to Ramos, the analysis of the first three digits offers a realistic electrocardiogram of the set of numbers, allowing a detailed observation of what happens at each point and which are the potentially fraudulent operations.

As has been shown, the use of Benford´s Law in the field of accounting is prominent, having shown itself able to detect anomalies in accounting data. In accordance with this premise, in subsection 3.2 we propose different measures based on Benford's Law. These measures are used, along with other variables, as indicators for the detection of patterns that conceal fraudulent operations with the ultimate aim of directing law enforcement authorities to companies that are more likely to have engaged in fraudulent operations.

2. THEORY AND METHODS

The aim of the current paper is to classify a set of suppliers as legal or illegal based solely on the data available in the undisclosed accounting ledgers of a large company investigated for laundering huge amounts of money. This is carried out by analysing the monetary payments from commercial operations carried out between the suppliers and the core company. In this research, we rely on machine learning techniques to find out which patterns can identify, within a binary decision model, companies labelled by police experts as fraudulent. The aim is to determine those companies which can be classified as legal and those as illegal, or likely to be illegal.

A major problem in the dataset lies in the fact that companies classified as fraudulent are very few in relation to the total number of suppliers contained in the dataset. That is, the response

variable is very imbalanced. Therefore, we consider statistical techniques of balancing with the purpose of improving the predictive capacity of the models. This section details the methodological aspects related to the approaches and techniques used and their validation.

3.1. Benford's Law

Empirically, Benford (1938) realized that, contrary to what might be expected, the frequency distribution of leading digits in many real-life collections of numbers is not uniform. He discovered that the frequency distribution that arises in many natural sets of numerical data is biased towards small numbers. Benford's Law sets down that the probability of occurrence of each first digit d_1 (=1, 2,…, 9) in many sets of numbers responds to the following probability mass function:

$$f_1(d_1) = P(X_1 = d_1) = log_{10}\left(1 + \frac{1}{d_1}\right) \quad d_1 = 1,2,\ldots.,9$$

Its cumulative distribution function being:

$$F_1(d_1) = P(X_1 \leq d_1) = log_{10}(1 + d_1) \quad d_1 = 1,2,\ldots.,9$$

From the first-digit distribution, it is not difficult to derive the frequency distribution for the second digit (e.g., Hill, 1995). The second-digit distribution can be given as:

$$\mathrm{f_2(d_2) = P(X_2 = d_2) = \sum_{k=1}^{9} \log_{10}\left(1 + \frac{1}{10k + d_2}\right) \quad d_2 = 0,1,2,\ldots.,9}$$

Notably, if a set of numbers follows Benford´s Law, the percentages expected for the first and second digit are given in Table 1.

Table 1. Probability distributions of first- and second-digit Benford's Laws.

	0	1	2	3	4	5	6	7	8	9
First-digit	-	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6
Second-digit	12.0	11.4	10.9	10.4	10.0	9.7	9.3	9.0	8.8	8.5

Benford's Law has as its main properties invariance in scale and in base (Pinkham, 1961; Hill, 1995). The scale-invariant property implies that Benford's Law continues to be fulfilled even if the units of measurement are changed. That is, the level of fit of some data to Benford's Law is independent of the measurement system. In economic terms, the currency in which the variable is measured does not influence the result. The base-invariant property states that the logarithmic law remains independent of the base used. It is equally valid in base 10, in binary basis, or in any other base. Hill (1995) proves that Benford's Law is the unique continuous distribution base-invariant and that scale-invariance (a property impossible for continuous variables) entails base-invariance, the reverse being untrue.

Although Benford's Law is not universally applicable, it is more "robust" than one might guess. For instance, if we randomly select probability distributions and from each of them we take random samples, we see that the significant-digit frequencies of the combined set will converge to the Benford's distribution even if each particular distribution deviates from Benford's Law (Hill, 1998). This last result is interesting, because it is a deep-rooted fact in psychology that people cannot behave truly randomly, even when it is to their benefit to do so (Wagenaar, 1972). Hence, due to human biases, when people manufacture data manually, the data rarely fits Benford's Law.

3.2. Statistical Tests

A first step towards identifying fraudulent companies might be to calculate the frequencies of leading digits of the monetary amounts corresponding to each supplier and to study the degree of fit to Benford´s Law that the data present, using any of the classic tests of goodness-of-fit (χ^2, Kolmogorov-Smirnov, Kuiper). However, this approach would not be adequate. On the one hand, as is well known, goodness-of-fit tests tend to reject the null hypothesis as soon as the sample size grows: they have too much statistical power. On the other hand, as stated by Giles (2007) and Tam Cho and Gaines (2007), these tests can be excessively rigid for economic data. In real-life data sets, Benford's Law does not represent a true distribution but a distribution that we would expect to occur in the limit.

Under these conditions, we have opted for a more flexible alternative and we have mapped the distributions of monetary amounts of each supplier in a 20-dimensional space of p-values. This strategy allows us, on the one hand, to characterize the distribution of monetary amounts of each supplier and, on the other hand, to have a large battery of features that, along with other variables, can be included in a model of machine learning. Specifically, we have calculated the p-values of individual fit to Benford´s Law of the frequencies of each of the first and second digits and computed the p-value for a fit of each data set to the first-digit Benford's Law using a statistical test based on the distance χ^2, but more flexible.

To measure the fit of first and second digit of each supplier to Benford's Law, we have computed the Z_i statistic and its associated p-value in a two-tailed test under the hypothesis that Z_i follows a standard Normal distribution.

$$Z_i = \frac{|n_{Oi} - n_{Ti}| - \frac{1}{2N}}{\sqrt{\frac{n_{Ti}(1 - n_{Ti})}{N}}}$$

where n_{Oi} is the frequency of either first or second digits equal to i in the subsample corresponding to the supplier for which the Z measurement is being computed, $n_{Ti} = Nf(i)$ is the associated expected frequency under Benford's Law (with f being equal to f_1 or f_2, as appropriate) and N is the number of operations corresponding to the supplier company in the accounts book.

Moreover, we have used an empirical test based on simulation to calculate the degree of global fit of the data for each supplier to the first-digit Benford's Law. The test, which we call the OverBenford test, eliminates the sample size effect. We obtain the p-value for the OverBenford test by simulation. Firstly, we draw B samples from f_1 with the same size N as our actual sample. Secondly, we compute for each of these new samples the χ^2-distance to the expected distribution. Finally, we calculate the p-value as the proportion of times the sample B distances computed in (ii) exceeds the χ^2-distance obtained from the observed sample. This approach mimics the approaches suggested for goodness-of-fit of continuous distributions implemented in Pavía (2015).

3.3. Machine Learning Methods

Benford´s Law provides the basis for classifying companies as legal or fraudulent. In a first phase, the p-values corresponding to the OverBenford statistic and the different Z_i are calculated. These p-values serve to characterize the behaviour of each of the companies in their daily operations and are used to perform the classification. In addition to the p-values, we have also used as classificatory variable the number of operations. These variables constitute the predictors (features) that are included in the machine learning models of classification to be used in this research. The automatic learning methodologies used have been: ridge logistic regression (LG), neuronal networks (NN), C4.5 decision trees (DT) and random forests (RF). In what follows in this subsection we outline the different procedures.

3.3.1. Ridge Logistic Regression

Logistic regression models are classic procedures widely used to model the relationship between a dichotomous variable and one or more features (McCullagh and Nelder, 1989). Logistic regression models are used to either (i) quantify the importance of the relationship between each of the features and the binary response variable or, (ii) classify instances between two categories. We use ridge logistic regression (Le Cessie and van Houwekingen, 1992) for this second purpose. In ridge logistic regression, penalization is used to prevent overfitting occurring due to either collinearity among the predictors or high-dimensionality. The aim of shrinkage and penalization is to improve the predictive accuracy of the model.

3.3.2. Neural Networks

Neural networks take their inspiration from the human brain. Neural networks as learning methods have been used for over 50 years and were developed separately in statistics and artificial intelligence to mirror the way human brains solve problems (Hastie, Tibshirani and Friedman,

2009). Artificial neural networks use concepts borrowed from our understanding of how the human brain responds to stimuli from model arbitrary functions. The neural networks are made up of a set of simpler elements, which we call neurons, that are interconnected in parallel in hierarchical form and that interact like the neural systems. A neural network has four basic elements: the number of layers, the number of neurons per layer, the degree of connectivity and the type of connections between neurons. The central idea of a neural network system is to model the response variable as a nonlinear function of the features by processing linear combinations of the inputs as derived features. In this sense, they are referred to as black-box algorithms because they use a complex and obscure mechanism to transform the inputs into a response. Neural networks are prediction tools, difficult to interpret, that can be used to predict both categorical and continuous variables. In this research, the neural network is trained within a supervised learning paradigm with the aim of identifying suspicious suppliers.

3.3.3. C4.5 Decision Trees

Decision trees represent another classical technique of machine learning. Decision tree learners build a model in the form of a tree structure. The decision tree classifies the instances according to an objective based on the available features, which can be quantitative or qualitative. A decision tree can be seen as a flowchart with decision nodes that can be interpreted as rules. Decision tree algorithms partition the data recursively until some condition is met, such as minimization of entropy or classification of all instances. Due to this procedure, the tendency is to generate trees with many nodes and nodes with many leaves, which leads to over-adjustment or overtraining. The tree will have great precision in the classification of the training data, but very little precision in classifying the instances of the test data. This problem is solved with a pruning procedure a posteriori. In this research, we used the algorithm C4.5 developed by Quinlam (1993).

3.3.4. Random Forests

A random forest is an ensemble-based method that uses decision trees as building blocks to construct more powerful prediction models. Breiman (2001) develops random forests as an improvement of bagging by adding additional diversity to the decision tree models. In a random forest, each tree is built using a random subsample of the full feature set. Typically, in each split, the number of predictors considered is approximately equal to the square root of the total number of predictors. Compared to bagging, this has the effect of *decorrelating* the trees. After the ensemble of trees (the forest) is obtained, the model uses a vote to combine the trees' predictions. The instances are classified in the class that obtains the greatest number of votes from the trees that make up the forest. Random forests are viewed as one of the most popular machine learning algorithms due to their power, versatility, and ease of use.

3.4. Imbalanced Data Sets

The performance of machine learning algorithms is typically assessed using global predictive accuracy. However, this is not advisable when the consequences of the different kinds of errors vary markedly and/or when the data is imbalanced (Brown and Mues, 2012). This is a situation that often occurs in the real world, where the relevant category is usually the one that has a significantly lower percentage of instances (Kotsiantis, Kanellopoulos and Pintelas, 2006). According to López et al. (2013), the machine learning community has addressed the issue of class imbalance using three basic strategies. The first consists in balancing the training set by undersampling the majority class, oversampling the minority class or generating synthetic data in the minority class. The second focuses on modifying the algorithm through an adjustment in the precision threshold or making a change to make it more sensitive to the minority class. The third seeks to make the learning cost-sensitive to errors arising especially in the minority class.

Undersampling can be used with very large datasets and applied to the majority class, reducing the number of instances of this class to balance it with the minority class. Since this method discards most of the instances of the majority class, information that could be relevant in the training set is lost. Oversampling works with the minority class, which is increased to balance it with the majority class. In this case no information is lost, but the training set is increased by copying and pasting observations of the minority class, which could lead to other problems.

Although these two techniques are easy to apply, given the scant presence of fraudulent companies in our base it is advisable to use other methods. In this work, we use a synthetic data generation method and a cost-sensitive learning method as alternatives to not balancing the sample, and we study the improvement that occurs with respect to using, directly with the original data, the machine learning techniques mentioned above. Recent examples with these approximations can be found in Rivera and Xanthopoulos (2016) and Sahin, Bulkan and Duman (2013).

3.4.1. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE combines synthetic oversampling of the minority class with undersampling of the majority class as a tool to balance the sample (Chawla et al., 2002). This technique is based on a form of oversampling that provides new information relative to the minority class through the random generation of new instances in the minority class.

To generate the synthetic random set, SMOTE makes use of bootstrap and k-NN (k-Nearest Neighbors algorithm), combined. The minority class is oversampled by introducing synthetic examples by convex combining each minority class instance with a random sample drawn with replacement among its k minority class nearest neighbours in the feature space.

3.4.2. Cost Matrix (Cost Sensitive Learning)

This technique does not create distributions of balanced data, but seeks to balance learning by applying a cost matrix that accounts for the cost of an erroneous classification versus a correct one. This technique applies smaller costs (weights) to the instances of the majority class and larger costs to those of the minority class. The weights can be set to be inversely proportional to the fraction of instances of the corresponding class. In the case of binary classification the weight of a class can be adjusted by resampling to improve the predictive power of the model.

In fraud detection, identifying a company that has committed fraud as fraudulent or non-fraudulent to one that has not been fraudulent carries no cost. However, the cost associated with identifying a company that has committed fraud as negative (false negative) carries more costs than identifying a company that has not caused fraud as positive (false positive). The cost matrix is similar to the confusion matrix. The aim is to penalize the errors (false positives and false negatives) against the correct ones (true negatives and true positives). In our application, we do not penalize correct ones and we assign a higher cost to errors in the minority class than to errors in the majority class.

3.5. Assessing Predictive Accuracy

In order to evaluate the predictive capacity of the different models, we have applied the traditional method of splitting the initial set into two subsets, one of training (70% of the data) and one of test (30%). If the model selected with the training data has a good predictive ability it will be able to correctly classify the instances of the test set.

Since the number of original positives in the dependent variable is very low and the procedure of increasing the original data is revealed as the most advisable, we have chosen to repeat the procedure 10 times to assess the robustness of the conclusions. That is, 10 sets of training data with their corresponding sets of test data are randomly selected on the original data set. Cross-validation was used for the training sets to fit the models. The measures selected to evaluate the accuracy of the predictions are: (i) the area of the ROC curve, (ii) the Kappa statistic, and (iii) RMSE (Root Mean Squared Error).

These measures have been used to compare the different procedures (combination of automatic learning method and data balancing method) with respect to their predictive capacity in both training data (explanation) and in the test data (prediction). Some models may well explain the data with which they have been trained but may not be able to predict positives in the test set. Although a greater explanatory capacity than predictive capacity is to be expected, what is important, however, is the sensitivity of the model to the data of training and verification.

4. DATA

As in all research using real data, much of the work has been devoted to purification and treatment. The quality of any analysis relies heavily on the quality of the data used. This section describes the database, the treatment criteria implemented, the selection of variables and the characteristics of the learning groups and test group built.

4.1. The data set

The criminal case being analysed is one of the most voluminous cases of money laundering in Spanish history, both in terms of economic value and in the number of companies involved. There is a large database containing 285,774 commercial operations carried out by 643 suppliers with the core company under investigation. This study is based on the analysis of the amounts of commercial operations, represented by analysis of the variable "Amount of the Operation".

Of the total number of suppliers, we are only certain that 26 of them are fraudulent, that is, they carry out operations that do not comply with the law. These are money laundering operations linked to economic gains made through the commission of other highly profitable crimes. This *a priori* information is provided by the police authorities based on the investigation of the associated judicial process.

A priori information only provides certainty that 4% of the suppliers are fraudulent; the fraudulent or non-fraudulent status of other companies not being known. When the classic goodness-of-fit tests are applied, the hypothesis that the data relating to commercial operations conform to Benford's Law is massively rejected. This result can be misleading since, as discussed in section 3.2, the traditional tests do not serve as a measure to evaluate the fit in very large samples. In fact, if we compare graphically the fit of the whole sample to Benford´s Law of the first two leading digits (see Figure 1), we see that the data does seem to fit this distribution, despite the classic tests systematically reject the null hypothesis. This generalized rejection does not occur using the OverBenford test proposed in this paper. For instance, the p-value of testing the fit of the whole dataset to the first digit Benford's Law is 0.2303.

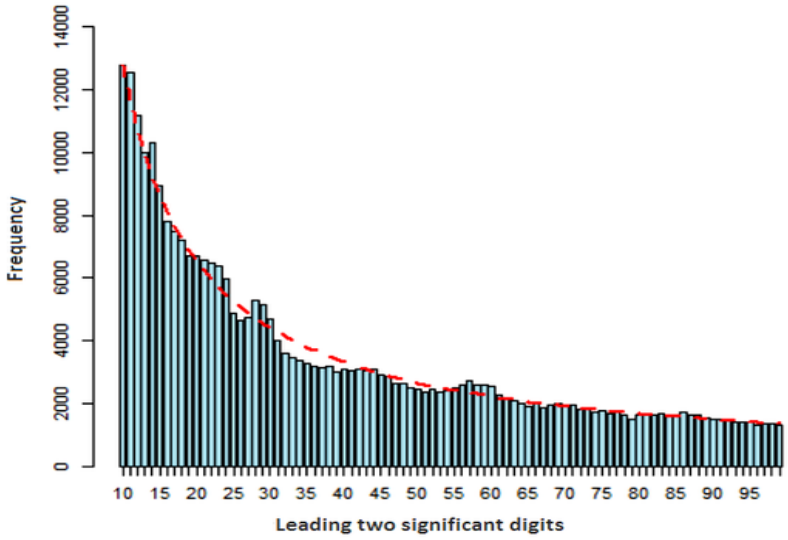


Figure 1. Fit of the data to Benford's first two leading digit frequency distribution.

In any case, to evaluate if a company should be included or not in the group of defrauding companies, using exclusively the annotations corresponding to its operations with the core company —i.e., without using other external information (such as wiretaps or criminal records)—, it seems reasonable to have a minimum amount of data about the company: a minimum number of operations. Hence, we have screened the initial database to consider exclusively those companies for which a minimum number of operations (variable *Frequency*) exist. Specifically, companies that have performed at least 195 operations have been incorporated into the automatic detection analysis. This decision has led to the analysis focusing on 335 companies, which account for a total of 245,227 operations with the core company.

Of the 335 companies studied, police experts have identified 23 of them as fraudulent. That is, 6.87% of the instances belong to the minority class. This means that we have an imbalanced data set, and it is therefore appropriate to apply strategies to the sample such as those described in section 3.4.

The average number of operations of a company identified as fraudulent is 2042.09 operations, while the average of other companies is 635.45 operations. This is consistent with what has been discussed in previous literature, which has revealed that companies that commit financial fraud and money laundering often show a certain predisposition to performing a large number of operations (Demetis, 2011). In this type of crime, criminals tend to perform as many operations as possible with the dual objective of hiding the fraud strategy and laundering as much money as possible. It has been decided, therefore, to include the frequency variable as predictor. As shown below, this variable has a high correlation with the fraud variable.

To sum up, for each of the companies analysed we have as predictors the number of operations registered (variable *Frequency*) and a set of 20 p-values —corresponding to Z-tests (where F1 to F9 denotes the p-values associated with the first digits and S0 to S9 those linked to the second digits) and to the OverBenford test—, and as response variable the police consideration of each company as fraudulent or non-fraudulent (variable *Invest*). A detail of the marginal distributions of such variables is given in Figure 2. The aim is to analyse if there is any type of behaviour pattern that distinguishes companies that commit fraud from those that do not in order to highlight other companies that might merit scrutiny by police and judicial authorities.

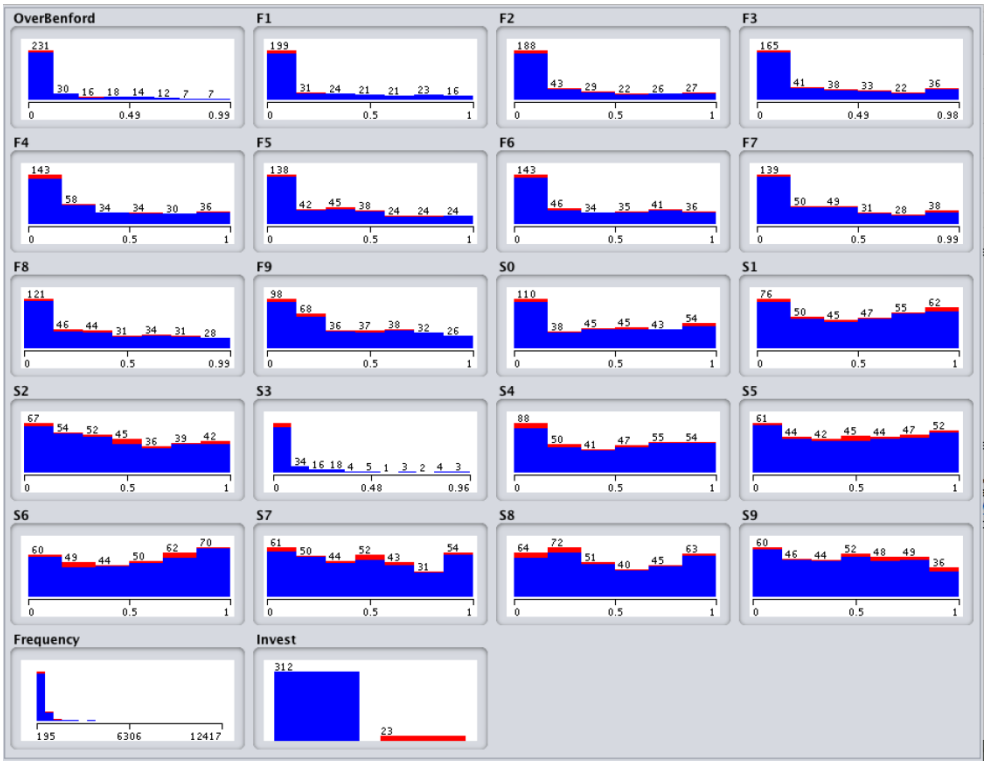


Figure 2. Marginal frequency distribution of predictors and response variable. Weka output.

4.2. Selecting predictors

Although the machine learning procedures considered are designed to avoid the dangers of a high number of predictors in terms of multicollinearity and overparameterization, it is sometimes better to make a previous selection of predictors (Seo and Choi, 2016). This is the case in this study, where we found that the models showed a lower predictive capacity without a previous selection of features. Hence, we have made a previous selection among the set of potential predictors using the Weka Ranker Search Method algorithm; which is based on correlations between predictors and response variable. Table 2 gives the predictors ordered by degree of

relation to the response variable. From the total of predictors initially considered, we have included in the models those that exceeded the value 0.05. In total, there are 11 predictors.

Table 2. Correlation ranking of the predictors. Output of Weka.

Frequency	F7	S4	S9	S8	OverBenford	S3	F3	F9	S2	F5
0.3157	0.1392	0.1348	0.1281	0.1060	0.1011	0.0911	0.0713	0.0678	0.0658	0.0573
S0	F2	F8	F6	S7	S5	S6	F1	F4	S1	
0.0424	0.0389	0.0360	0.0308	0.0274	0.0226	0.0198	0.0189	0.0166	0.0137	

4.3. Splitting the sample

Once the predictors have been selected, the base is composed of 335 companies, 11 predictors and a response variable. These data are those on which the methodologies described in section 3 have been applied.

Since the analysis (see section 5) has been structured in two parts, we have also used two different strategies to divide the sample. On the one hand, the initial fit of the models is done with the complete base, using the machine learning approach based on cross-validation that allows determination of the tuning parameters that for each algorithm offer a greater predictive capacity in the training set. Specifically, we have randomly divided the learning set into 10 folds, each with 10% of the data. On the other hand, in order to analyse the impact that using the SMOTE strategy has on the predictive quality of the different models, we have also used the classic strategy of dividing the data into two groups, one of training (70% of the instances) and the other of test (30% of the instances). This last process has been repeated ten times (applying in all cases 10-fold cross-validation to fit the models with the training sets) to discount the possibility that the solutions could be dependent on the actual partition carried out. In all cases, instead of using the option of a stratified selection of the instances, with the categories as strata, the groups have been constructed by extracting instances randomly from the total data set. This has caused some variability in the internal composition of the groups. For example, training samples composed of 70% of the cases vary between a minimum of 14 positive cases (5.98%) and a maximum of 19 (8.12%).

5. RESULTS

This section discusses the results. We have grouped the analyses into two subsections. The first subsection focuses on evaluating the explanatory/predictive capacity of the models and the impact of the different solutions implemented to deal with the challenge that entails working with

such imbalanced data. From this analysis, we deduce that the SMOTE strategy, based on the generation of synthetic instances of the minority class, is the one that produces the best results. Thus, a second subsection is included, in which we perform an analysis of the sensitivity of the predictive power of the models when the original data are balanced using the SMOTE strategy.

In the first block of analysis, the different models are trained using cross-validation, after dividing the sample into 10 sub-samples of equal size. In the second block of analysis, the initial division of the data set into two subsets (one of training, 70%, and another of test, 30%) was repeated 10 times, with the models again trained using cross-validation models on the training sets. On each of the training sets, we have applied the transformation of the data using the SMOTE algorithm to balance the target variable.

5.1. Assessing the models

The explanatory capacity of each of the four machine learning algorithms analysed has been evaluated under each of the three scenarios considered in terms of imbalanced data: (i) directly modelling the data, without considering the imbalance presented by the two categories of the response variable; (ii) using cost-sensitive learning (cost matrix); and (iii) applying a transformation to the data using the SMOTE algorithm to balance the dependent variable

5.1.1. Imbalanced dataset

The results of training the models using the complete database without performing any transformation on the data (i.e., without taking into account the imbalance presented by the two categories of the response variable) are presented in Table 3.

Table 3. Confusion matrix of the models. Imbalanced dataset.

	LG		DT		NN		RF	
	NO	YES	NO	YES	NO	YES	NO	YES
NO	311	1	302	10	301	11	312	0
YES	20	3	17	6	15	8	19	4
Correctly Classified	93.73%		91.94%		92.24%		94.33%	
Incorrectly Classified	6.27%		8.06%		7.76%		5.67%	
TN Rate (No)	99.68%		96.79%		96.47%		100.00%	
TP Rate (Yes)	13.04%		26.09%		34.78%		17.39%	
FN Rate (Yes)	86.96%		73.91%		65.22%		82.61%	
FP Rate (No)	0.32%		3.21%		3.53%		0.00%	

The explanatory ability of the models is very high, but they present very low true positive rates in the target category, ranging from 13.04% for the logistic regression model to 34.78% for the neural network approach. By having such imbalanced data, the algorithms tend to favour classification in the dominant category, identifying very few fraudulent companies.

5.1.2. Cost Matrix

In a cost-sensitive fit, we assume that the losses of an incorrect classification are asymmetric. The cost of checking the misclassification of the false positives would be minimal, while false negatives would entail a much higher cost, not only in terms of tax, but also in economic and security terms. The cost matrix allows the process to be balanced without having to perform any type of transformation on the data. Table 4 gives the results of training the models using the cost matrix.

On comparing the results of Tables 3 and 4, it is clear that there is a significant decrease in the overall accuracy of the models. The random forest model is the only algorithm that maintains 94% of instances correctly classified, while the rest of the models show a lower figure, with the logistic regression model showing a minimum of 73.73%. However, the rate of true positives improves substantially. The consequence of identifying a positive as a negative is that the algorithms correctly identify more companies as defrauders.

Table 4. Confusion matrix of the models. Cost Matrix.

	LR		DT		NN		RF	
	NO	YES	NO	YES	NO	YES	NO	YES
NO	234	78	290	22	285	27	309	3
YES	10	13	16	7	15	8	17	6
Correctly Classified	73.73%		88.66%		87.46%		94.03%	
Incorrectly Classified	26.27%		11.34%		12.54%		5.97%	
TN Rate (No)	75.00%		92.95%		91.35%		99.04%	
TP Rate (Yes)	56.52%		30.43%		34.78%		26.09%	
FN Rate (Yes)	43.48%		69.57%		65.22%		73.91%	
FP Rate (No)	25.00%		7.05%		8.65%		0.96%	

The cost-based approach has increased the detection of true positives at the expense of significantly raising false positives. The only case where this does not happen is in the random forest model, which shows the least true positives but also less false positives.

5.1.3. SMOTE balance

One of the transformations of the data most used for balancing the categories is the algorithm SMOTE. Based on the above data, where there were 312 legal and 23 illegal companies, new instances of fraudulent companies had been synthetically generated, up to a total of 299. This gives a 51.06% "no" and a 49.94% of "yes". The models were subsequently trained on this new set. Table 5 provides a summary of outcomes.

Table 5. Confusion matrix of the models. SMOTE.

	LR		DT		NN		RF	
	NO	YES	NO	YES	NO	YES	NO	YES
NO	239	73	269	43	252	60	300	12
YES	53	246	31	268	38	261	15	284
Correctly Classified	79.38%		87.89%		83.96%		95.58%	
Incorrectly Classified	20.62%		12.11%		16.04%		4.42%	
TN Rate (No)	76.60%		86.22%		80.77%		96.15%	
TP Rate (Yes)	82.27%		89.63%		87.29%		94.98%	
FN Rate (Yes)	17.73%		10.37%		12.71%		5.02%	
FP Rate (No)	23.40%		13.78%		19.23%		3.85%	

The SMOTE approach does not substantially improve the overall explanatory ability of the models compared to the use of the cost matrix. However, the true positive rate has improved in all cases. With the original data, true positive rates ranged between 13.04% for logistic regression and 34.78% for neural network. With the cost matrix, results improved, reaching 26.09% for random forest and 56.52% for logistic regression. With the SMOTE transformation, true positive rates range from 82.27% for the logistic regression model to 94.98% for random forest.

The capacity of this third balancing strategy to identify illegal companies is greatly superior to the two previous ones; obtaining highly satisfactory results in the case of random forest. Of 611 instances, it only incorrectly classified 27 (4.42%), of which 15 were false negatives and 12 false positives.

5.1.4. Measurements of precision

Finally, the measurements of the ROC area, the Kappa statistic, and RMSE (Root Mean Squared Error) are taken in order to evaluate the different procedures as well as the joint action of models and techniques of balancing, as shown in Table 6.

Table 6. Measurements of precision of the procedures.

	Imbalanced dataset			Cost Matrix			SMOTE		
	ROC	**Kappa**	**RMSE**	**ROC**	**Kappa**	**RMSE**	**ROC**	**Kappa**	**RMSE**
LG	0.747	0.2061	0.2360	0.711	0.3243	0.4227	0.844	0.5675	0.4012
DT	0.635	0.2664	0.2702	0.615	0.2086	0.3320	0.894	0.7348	0.3499
NN	0.765	0.3400	0.2578	0.630	0.2104	0.3306	0.926	0.7252	0.3392
RF	0.740	0.2817	0.2268	0.773	0.3499	0.2415	0.989	0.9116	0.2088

From Table 6 it follows that the best results are obtained when the sample is balanced using SMOTE. When the original data is used with or without cost matrix the fit worsens. As for the classification algorithm, the best results are obtained for random forest, with a ROC area of 0.989 and a Kappa statistic of 0.912 (in both cases very close to 1) and the lowest RMSE compared to all the other models applied.

5.2. Sensitivity Analysis

The results of true positives detected when balancing the training data with the SMOTE algorithm are the ones with the most promising outputs. Therefore, in this subsection, we verify the predictive ability of the models based on a set of data independent of the one used for training.

With that purpose, we have performed 10 random divisions of the initial set of data, generating 10 pairs of subsets with respectively 70% of the cases (training) and 30% of the instances (test), and have applied the same methods as before with SMOTE. The number of fraudulent companies in training sets ranges from a minimum of 14 to a maximum of 19, while in the test set it is the inverse, ranging from a maximum of 9 to a minimum of 4.

Due to the different compositions of the different training and test sets, the subsets are not comparable in terms of the number of fraudulent and non-fraudulent instances and so Tables 7 and 8 do not have "yes/no" matrices but only show the means for cases classified correctly and incorrectly. Table 7 focuses on training sets and Table 8 is dedicated to test sets.

Table 7. Average of confusion matrix summaries of the models (training set).

	LR	DT	NN	RF
Correctly Classified	78.06%	88.31%	86.60%	95.16%
Incorrectly Classified	21.94%	11.69%	13.40%	4.84%
TN Rate (No)	77.01%	86.55%	83.60%	93.87%
TP Rate (Yes)	79.09%	90.03%	89.54%	96.16%
FN Rate (Yes)	20.91%	9.97%	10.46%	3.84%
FP Rate (No)	22.99%	13.45%	16.40%	6.13%

Although we have fewer instances to train the models, comparing the results of Tables 5 and 7 we observe that the explanatory ability (training set) of the models is very similar. If we now evaluate the predictive ability (test set) of the different models (see Table 8), we see that it has been reduced, especially in the true positives; particularly striking is the loss that occurs in the random forest model. The random forest goes from registering the best results with the training set to the worst ones with the validation set. In the average of the 10 divisions, neural network and logistic regression are the techniques that best predict in the test set.

Table 8. Average of confusion matrix summaries of the models (test set).

	LR	DT	NN	RF
Correctly Classified	76.04%	83.66%	82.41%	90.40%
Incorrectly Classified	23.96%	16.34%	17.59%	9.60%
TN Rate (No)	77.41%	86.43%	84.13%	94.38%
TP Rate (Yes)	56.72%	44.78%	58.21%	34.33%
FN Rate (Yes)	43.28%	55.22%	41.79%	65.67%
FP Rate (No)	22.59%	13.57%	15.87%	5.62%

Precision measurements draw the same results as the matrices of confusion. The fit for the training data is significant, random forest obtaining an average of the ROC area of 0.9873 with a range of 0.019. The Kappa statistic also draws the best results in random forest, with a very high value (0.89024) and small range (0.1103). The methodology that minimizes errors is again random forest with an RMSE mean of 0.22126 and a range of 0.0688.

When comparing the precision measurements of the training set with the test set (see Table 9), a significant loss of precision is observed, albeit different for each method. Taking the mean of the ROC area as a reference, the greatest loss is for random forest with a difference of 0.2331 between the training set and the test set. The smallest difference was observed in logistic

regression, which was the one that obtained worse results in the training set. Random forest is the one that has a greater loss in the Kappa statistic, the difference being 0.62105. Finally, the increase in the mean of the RSME is very similar for decision tree, neural network and random forest, being very small for logistic regression with only a slight increase of 0.0102.

Table 9. Summary of measurements of precision over the test sample.

	ROC			Kappa			RMSE		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
LG	0.7787	**0.708**	0.881	0.14947	0.0555	0.2914	0.40051	0.3493	0.4661
DT	0.6570	0.558	0.796	0.18590	**0.0901**	0.2930	0.39203	0.3226	0.4372
NN	0.7361	0.508	0.948	0.22728	0.0249	0.3671	0.39697	0.3397	0.4651
RF	**0.7886**	0.620	**1.000**	**0.26919**	0.0804	**0.5976**	**0.27975**	**0.2388**	**0.3224**

As the ranges (the difference between the maximum and the minimum) of the precision measurements in the training sets are narrow, it seems the models are fairly stable and do not depend on the training data set selected. However, the ranges of precision measurements in the test set are much broader, depending on the predictive ability of the model in the set of data selected. This result can lead to question the robustness of the approaches. An alternative explanation is that the models are identifying (in the test sets) more actual fraudulent companies than initially identified, suggesting that as suspected even more companies may have perpetrated criminal acts. The implemented procedures provide a guide for law enforcement investigators to focus their inquiries and their resources on those companies that are systematically identified as fraudulent by a variety of methods.

6. CONCLUSIONS

Many real financial and economic datasets conform to Benford's Law, but this is not widely known. Hence, under the assumption that it is highly unlikely that the fit to the Benford distribution would be preserved when people fabricate data, Benford's Law has been used as a tool to detect accounting irregularities. In this work, we combine Benford's Law and machine learning algorithms as a tool to detect money laundering criminals in the context of a real Spanish court case.

To this end, we analyse the hidden accounting of a company investigated for money laundering and characterize, based on Benford´s Law, the operations carried out by each of its suppliers by projecting all of its operations in a space of 21 dimensions. The information provided by the police about the fraudulent status of a subgroup of companies is used to detect, using machine learning

models, other companies that could be potentially fraudulent. The proposed procedures have to consider the problem of working with highly imbalanced categories.

The results show that, in general, the strategy used to balance the sample has more weight than the machine learning algorithm. The SMOTE strategy achieves better results on the true positives than the cost matrix. However, the overall accuracy of the model is very similar, so the proportion of false positives increases with the SMOTE methodology. The cost matrix identifies fewer positives, and consequently, fewer companies would be investigated.

Selecting a large number of companies to investigate would have two negative points. The first would be an increase in investigation costs: more companies investigated means more time, more personnel and more resources. The second is the disruption caused to companies deemed legal, who have to endure an investigation and the intrusion that accompanies it. In our opinion, the best solution is to use different methods and algorithms to evaluate the different alternatives and to identify potential criminals, leaving the final decision to police experts as to which companies deserve more in depth scrutiny.

References

Alali, F. A., & Romero, S. (2013). Benford's Law: Analyzing a decade of financial data. *Journal of Emerging Technologies in Accounting, 10(1)*, 1-39.

Asllani, A., & Naco, M. (2014). Using Benford's Law for Fraud Detection in Accounting Practices. *Journal of Social Science Studies, 2(1)*, 129-143.

Beneish, M. D. (1999). The Detection of Earnings Manipulation. *Financial Analysts Journal, 55(5)*, 24-36.

Benford, F. (1938). The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society*, 78 (4), 551-572.

Breiman, L. (2001). Random Forests. *Machine Learning, 45 (1)*, 5-32.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39(3)*, 3446-3453.

Chawla, N., Bowyer, K.W., Hall, L., & Kegelmeyer, W. (2002). SMOTE: Synthetic Minority Oversampling Technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Deckert, J., Myagkov, M., & Ordeshook, P. C. (2011). Benford's Law and the detection of election fraud. *Political Analysis, 19(3)*, 245-268.

de Marchi, S., & Hamilton, J. T. (2006). Assessing the accuracy of self-reported data: an evaluation of the toxics release inventory. *Journal of Risk and Uncertainty*, *32(1)*, 57-76.

Demetis, S. D. (2011) Unfolding Dimensions of an Anti-Money Laundering/Counter-Terrorist Financing Complex System. *Lexis Nexis*, *Emerging Issues 6019*.

Diekmann, A. (2007). Not the First Digit! Using Benford's Law to Detect Fraudulent Scientific Data. *Journal of Applied Statistics*, *34(3)*, 321-329.

Durtschi, C., Hillison, W., & Pacini, C. (2004). The effective use of Benford's law to assist in detecting fraud in accounting data. *Journal of Forensic Accounting, 5(1)*, 17-34.

Giles, D. E. (2007). Benford's Law and naturally occurring prices in certain eBay auctions. *Applied Economics Letters, 14(3)*, 157-161.

Günnel, S., & Tödter, K. H. (2009). Does Benford's Law hold in economic research and forecasting? *Empirica*, *36(3)*, 273-292

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Second Edition, Springer: New York.

Hill, T. (1995). The Significant-Digit Phenomenon*. The American Mathematical Monthly, 102(4)*, 322-327

Hill, T. (1995b). A statistical derivation of the significant-digit law. *Statistical Science*, 10, 354-363.

Hill, T. (1998), The First Digit Phenomenon. *American Scientist, 86*, 358–363.

Judge, G., & Schechter, L. (2009). Detecting problems in survey data using Benford's Law. *Journal of Human Resources*, *44(1)*, 1-24

Kotsiantis, S., Kanellopoulos, D., & Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering, 30(1)*, 25-36

Le Cessie, S., & van Houwekingen, J. C. (1992). Ridge estimators in logistic regression, *Applied Statistics, 41*, 191-201.

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113-141.

Luque, B., & Lacasa, L. (2009). The first-digit frequencies of prime numbers and Riemann zeta zeros. *Proceedings of the Royal Society of London, 465*, 2197-2216.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edn. Chapman and Hall: London.

Mebane W. (2007). *Election forensics: Statistics, recounts and fraud*. Paper presented at the 2007 Annual Meeting of the Midwest Political Science Association. Chicago, IL, April 12–16

Mebane, W., Alvarez, R.M, Hall, T.E., & Hyde, S.D. (2008). Election forensics: The Second Digit Benford's Law Test and recent American presidential elections. In *Election fraud*. Washington, DC: Brookings.

Newcomb, S. (1881). Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics, 4*, 39-40.

Nigrini, M. J. (1992). *The detection of income escape through an analysis of digital distributions*. PhD Tesis University of Cincinatti.

Nigrini, M. J. (1994). Using digital frequency to detect fraud. *The White Paper*, April, 1-3.

Nigrini, M. J. (1996). A taxpayer compliance application of Benford's Law. *The Journal of the American Taxation Association, 18(1)*, 72-91.

Nigrini, M. J, & Mittermaier, L. J. (1997). The Use of Benford's Law as an Aid in Analytical Procedures. *Auditing: A Journal of Practice & Theory, 16(2)*, 52-67.

Nigrini, M. J., & Miller, S. J. (2009). Data diagnostics using second-order tests of Benford's Law. *Auditing: A Journal of Practice & Theory, 28(2)*, 305-324.

Pavía, J. M. (2015). Testing Goodness-of-Fit with the Kernel Density Estimator: GoFKernel. *Journal of Statistical Software, 66(1)*, 1-27.

Pericchi, L., & Torres, D. (2011). Quick Anomaly Detection by the Newcomb—Benford Law, with Applications to Electoral Processes Data from the USA, Puerto Rico and Venezuela. *Statistical Science, 26(4)*, 502-516.

Pinkham, R. S. (1961). On the distribution of first significant digits. *The Annals of Mathematical Statistics*, *32*(4), 1223-1230.

Quick, R., & Wolz, M. (2003). Benford's Law in deutschen Rechnungslegungsdaten. *Betriebswirtschaftliche Forschung und Praxis, 2*, 208–224.

Ramos, D. (2006). Fraude: un nuevo enfoque para combatirlo. *Auditoría Pública, 38*, 99-104.

Rivera, W. A., & Xanthopoulos, P. (2016) A priori synthetic oversampling methods for increasing classification sensitivity in imbalanced data sets. *Expert Systems with Applications*, 66, 124-135.

Sahin, Y., Bulkan, S., & Duman, E. (2013). A cost-sensitive decision tree approach for fraud detection. *Expert Systems with Applications, 40(15)*, 5916-5923.

Seo, J. H., & Choi, D. (2016). Feature Selection for Chargeback Fraud Detection based on Machine Learning Algorithms. *International Journal of Applied Engineering Research*, *11*(22), 10960-10966.

Tam, W. K., & Gaines, B. J. (2007). Breaking the (Benford) law: Statistical fraud detection in campaign finance. *The American Statistician*, *61(3)*, 218-223

Tödter, K. H. (2007). Das Benford-Gesetz und die Anfangsziffern von Aktienkursen. *WiSt-Wirtschaftswissenschaftliches Studium, 36(2)*, 93-98.

Torres, J., Fernández, S., Gamero, A., & Sola, A. (2007). How do numbers begin? (The first digit law). *European Journal of Physics*, *28(3)*, 17-25.

Varian, H. R. (1972). Benfords Law. *American Statistician*, *26(3)*, 65-66.

Wagenaar, W. A. (1972). Generation of random sequences by human subjects: A critical survey of the literature. *Psychological Bulletin, 77(1)*, 65-72.

ANEXO V. Introducción y Conclusiones (Español)

INTRODUCCIÓN

El blanqueo de capitales es un delito económico cuya operativa ha evolucionado en el tiempo y que se ejecuta a distintos niveles y en diversas magnitudes. Las cuantías defraudadas oscilan desde el tradicional blanqueo de pequeñas cantidades de dinero proveniente del tráfico minorista y local de drogas, hasta las grandes cantidades (miles de millones de euros) de macro-estructuras empresariales que han surgido en las últimas décadas y que operan a escala internacional (Khac y Kechadi, 2010).

Atendiendo a la alta repercusión socio-económica de este delito, se derivan directa e indirectamente (delitos de blanqueo de capitales y delitos antecedentes del blanqueo de capitales) múltiples efectos negativos a la sociedad en su conjunto, que van desde la víctima del delito que genera el beneficio económico hasta su efecto en la economía mundial en el corto y en el largo plazo (Bartlett, 2002; Quirk, 1996; UNODC, 2011; Unger, 2007). A nivel internacional los efectos negativos afectan a todos los niveles: en los sistemas económicos, en las instituciones financieras, en los organismos públicos y en las empresas (Unger, 2007).

La lucha contra los delitos financieros y el blanqueo de capitales se ha intensificado por parte de las administraciones públicas de todos los países desarrollados dada la relación directa de ambos delitos con la financiación del terrorismo y las armas de destrucción masiva (FATF, 2012). La amenaza que los delitos financieros y el blanqueo de capitales suponen para cualquier país obliga a disponer de todos los recursos necesarios, económicos y de inteligencia, para luchar contra estas actividades ilícitas y las que propicia.

El principal objetivo del régimen de lucha contra el blanqueo de capitales y la financiación del terrorismo es reducir las tasas de criminalidad relacionadas con la delincuencia profesional, el crimen organizado y el terrorismo, y a su vez proteger a la sociedad en su conjunto (CCBE, 2014).

Todos los agentes involucrados en una organización criminal, salvo pocas excepciones, realizan las actividades ilegales con el único objetivo de lucrarse (López, 2015). Entendiendo el blanqueo de capitales como el "Talón de Aquiles" de cualquier

organización criminal, combatir este delito evita que los delincuentes lleven a cabo las actividades ilegales (delitos antecedentes de blanqueo de capitales) e impide el disfrute de los capitales ilícitos (Alhosani, 2016).

El escenario mundial actual, marcado por la libre circulación, la globalización y la liberalización del comercio mundial, ha propiciado un entramado de estructuras financieras y de empresas que se favorecen entre ellas para integrar en el mercado financiero el capital procedente de sus actividades ilegales (Demetis, 2011; FATF, 2013).

En el ámbito europeo, las circunstancias económicas y políticas hicieron avanzar a la Unión Económica y Monetaria de la Unión Europea (UEE) para adaptarla a un mundo cambiante y cada vez más globalizado. En una primera fase, la UEE introdujo plena libertad para la circulación de capitales, prevista en la Directiva del Consejo 88/361/CEE. Posteriormente, en una segunda fase, con la firma del Tratado de Maastricht en 1992 se previó la prohibición de todas las restricciones a los movimientos de capitales y sobre los pagos, tanto entre Estados miembros como entre Estados miembros y terceros (Galán, 2016). Actualmente, las únicas excepciones se limitan fundamentalmente a movimientos de capitales relacionados con terceros países, artículos 64 y 65 del Tratado de Funcionamiento de la Unión Europea.

El nuevo orden en el comercio mundial, que emana de la globalización de la economía, hace muy difícil ejercer un control efectivo sobre el movimiento del dinero (UNODC, 2011). Con todo, tratar de poner mecanismos de obstrucción iría contra la liberalización del comercio mundial que defienden, además de la mayoría de empresas y gobiernos, instituciones como el Banco Mundial y la Organización Mundial del Comercio.

En este contexto, los esfuerzos para combatir estos delitos económicos se articulan en todos los niveles. A nivel internacional, los mecanismos de lucha contra el banqueo de capitales y la financiación del terrorismo se estructuran de forma coordinada entre instituciones internacionales, como el Grupo de Acción Financiera Internacional (FATF), el Fondo Monetario Internacional, las Naciones Unidas o la Unión Europea y los distintos países (Unidades Financieras de Inteligencia-FIUs) (Unger, 2009). Lo que supone la existencia de un régimen mundial coordinado de lucha contra el blanqueo de capitales y la financiación del terrorismo (FATF, 2012).

El Grupo de Acción Financiera Internacional ha desarrollado 40 recomendaciones contra el blanqueo de capitales, que junto con las 9 recomendaciones contra la financiación del terrorismo, establecen el marco básico para la detección, prevención y eliminación del blanqueo de capitales y de la financiación del terrorismo (FATF, 2012).

Estas recomendaciones constituyen una serie de principios que los países deben trasladar a sus reglamentos. La Unión Europea ha difundo estas recomendaciones a sus Estados miembros en una serie de Directivas (91/308/CEE, 2005/60/CE, 2006/70/CE, 2015/849/CE).

A nivel nacional, las Unidades de Inteligencia Financiera (FIUs) constituyen un componente importante de estas estrategias, en tanto que estas instituciones son las encargadas de recibir, analizar y transmitir las transacciones sospechosas de blanqueo a las autoridades competentes. Todas ellas coordinadas por el Grupo Egmont de Unidades de Inteligencia Financiera (Forget y Hočevar, 2004).

Concretamente en España, la Unidad Financiera de Inteligencia es la Comisión de Blanqueo de Capitales e Infracciones Monetarias (SEPBLAC), máximo responsable del desarrollo de la política preventiva y de lucha contra el blanqueo de capitales en nuestro país.

Sin embargo, a pesar de los esfuerzos de coordinación, actualmente ni tan siquiera existe una metodología universalmente aceptada para estimar la cuantía de dinero de procedencia ilícita que se integra en los sistemas financieros de todo el mundo, siendo, por tanto, desconocida la eficiencia del sistema internacional contra este tipo de delitos (AML/CFT) (Unger, 2007; Barone y Masciandaro, 2011). Un consenso de "*guesstimates"* diferentes lo sitúa entre el 2% y el 5% del Producto Interior Bruto Global (Camdessus, 1998; Unger, 2007; UNODC, 2011; Unger, 2013). En España, las estimaciones del dinero blanqueado difieren considerablemente y oscilan entre los 36 millones de euros (Fernández, 2012) y los 56 mil millones de dólares americanos por año (Walker, 1999).

En el contexto descrito en los párrafos anteriores es en el que se realiza la actual investigación. Como punto de partida, se ha llevado a cabo un estudio, junto con dos agentes de la Policía Judicial, Grupo de Blanqueo de Capitales y Grupo de Delincuencia

Económica, para analizar la evolución del banqueo de capitales en España, entre los años 1998 y 2012[51].

Los resultados obtenidos, presentados en el Apartado I.6, muestran como durante el período 1998-2012 la tasa del número de delitos de blanqueo de capitales en España aumentó casi un 244%, mientras que la tasa del número de arrestados por delito de blanqueo de capitales aumentó un 431%. Por otro lado, la tasa que recoge el número de delitos antecedentes o subyacentes del blanqueo de capitales disminuyó casi un 30%, y la tasa que representa el número de personas detenidas en delitos antecedentes o subyacentes también decreció un 16% en el periodo analizado.

El estudio indicaría que los esfuerzos en materia de prevención y legislativos del régimen de lucha contra el blanqueo de capitales estarían teniendo resultados positivos a corto plazo, sin embargo, se conocen factores que estarían limitando su eficiencia.

Actualmente, se observa un incremento del número de casos de blanqueo y de la sofisticación de los mismos, dificultando su identificación y consecuentemente incrementando los efectos negativos en los sistemas económicos internacionales y repercutiendo en el desarrollo mundial (Zhongfei *et al.*, 2003). Además, los actuales avances tecnológicos, así como los sistemas de comunicación, han puesto a disposición de los delincuentes herramientas con las que crear estructuras grandes, complejas y coordinadas de empresas a fin de evitar la detección de la trazabilidad de los capitales (ingeniería financiera e ingeniería numérica) (Petrucelli, 2012).

Por ello, los investigadores especializados en esta área se enfrentan a un doble reto: por una parte, deben ser capaces de identificar las grandes redes de blanqueo, y por otra, deben seleccionar, de la cantidad de información que se deriva de las empresas implicadas, aquella que permita identificar claramente los patrones de comportamiento de los individuos que ocultan blanqueo. Sin embargo, con las técnicas tradicionales se hace muy complicado cumplir estos objetivos (Sremack, 2015).

[51] Artículo completo en el *ANEXO II*. Publicaciones. "*Money laundering trend in Spain: offences and arrests over 15 years*".

Concretamente, los procesos judiciales de blanqueo de capitales requieren que los investigadores tengan conocimientos avanzados de análisis de datos, de economía y contabilidad y, también, conocimientos del manejo de los softwares más avanzados en gestión de datos y técnicas estadísticas (Dutta, 2013).

En este sentido, las aportaciones más destacables del forense contable en procesos judiciales de blanqueo de capitales son: (1) la detección de operativas sospechosas (patrones de fraude), (2) el análisis de los posibles delitos financieros y de fraude (trazabilidad de los capitales) y (3) el desarrollo de técnicas que ayuden a detectar comportamientos delictivos incipientes (Owojori y Asaolu, 2009; Dutta, 2013).

Las exigencias del trabajo que desempeña el forense contable, y particularmente en los procesos judiciales de blanqueo de capitales, requieren el empleo de técnicas vanguardistas para el análisis de la información y su rápida aplicación en las bases de datos disponibles. Estas técnicas se engloban en el concepto de aprendizaje automático, y ofrecen una amplia gama de posibilidades y de recursos que crece exponencialmente (Sremack, 2015).

Pese a la cuantiosa bibliografía sobre detección de patrones en numerosas áreas de conocimiento, son todavía insuficientes los trabajos que aplican técnicas de aprendizaje automático para la detección de patrones de delincuencia, especialmente cuando se trata de trabajos que aplican las vanguardistas metodologías de aprendizaje automático en información obtenida en procesos judiciales de blanqueo de capitales (Hassani *et al.*, 2016; Pearsall, 2010). Así, la mayoría de las investigaciones se centran en desarrollar técnicas de aprendizaje automático para procesar la información de operaciones sospechosas, también llamadas *Suspicius Activity Reports* (SARs), que reciben las distintas unidades de inteligencia financiera o en abordar el problema desde una perspectiva general (Nath, 2006; Zhang *et al.*, 2003; Bolton y Hand, 2002; Tang y Yin, 2005).

Este trabajo representa dos nuevos acercamientos a la detección de indicios de blanqueo de capitales: (1) la aplicación de técnicas de aprendizaje automático en bases de datos de la contabilidad interna de empresas para la detección de blanqueo de capitales, y (2) ofrecer información a las autoridades investigadoras sobre cómo se organiza la red de

blanqueo, con el objetivo de orientar la investigación judicial hacia aquellas personas jurídicas o físicas que describan patrones sospechosos.

De este modo, y en el contexto de un macro-caso judicial real sobre blanqueo de capitales en el que se ha colaborado como forense contable, se analiza la base de datos disponible compuesta por las operaciones realizadas entre una empresa núcleo y un conjunto de 643 empresas proveedoras, 26 de las cuales ya han sido identificadas *a priori* por la Policía Judicial como fraudulentas.

Ante la sospecha fundada de que otras empresas proveedoras del entramado hayan perpetrado hechos delictivos, y a fin de dirigir más eficientemente los escasos recursos policiales disponibles, se proponen las técnicas de aprendizaje automático, en dos enfoques diferenciados, para la detección de los patrones de fraude.

El primer enfoque propuesto (Apartado IV.1[52]) supone una aproximación a la implementación de modelos de Redes Neuronales al trabajo pericial para la detección de operaciones de fraude. Para ello se aplica, basadas en las técnicas de aprendizaje automático, la estructura de red propuesta por Hastie *et al.* (2008): la red de propagación hacia atrás (del inglés *Back-Propagation Network*).

En el segundo enfoque (Apartado IV.2[53]) se propone una aproximación a la detección de patrones más ambiciosa que la anterior, en la que se combina la Ley de Benford *(*Nigrini y Mittermaider, 1997), como herramienta para caracterizar los registros contables de las operaciones comerciales entre la empresa núcleo y las empresas proveedoras, con cuatro modelos de clasificación: Regresión Logística con regularización Ridge (*Ridge Logistic Regression*) (Le Cessie y van Houwelingen, 1992), Redes Neuronales Artificiales (*Neural Networks*) (Hastie *et al.*, 2008), Árbol de Decisión C4.5 (*Decision Trees C4.5*) (Quinlan, 1993 y 1996) y Bosque Aleatorio (*Random Forest*) (Breiman, 2001).

Las técnicas de aprendizaje automático se emplean, utilizando la información *a priori* facilitada por la Policía Judicial sobre las empresas proveedoras que son fraudulentas

[52] Artículo completo en el *ANEXO III*.- Publicaciones. "*Detección de fraude financiero mediante Redes Neuronales de clasificación en un caso real español* ".

[53] Artículo completo en el *ANEXO IV*.- Publicaciones. "*Combining Benford's Law and Machine Learning to detect Money Laundering. An actual Spanish Court case".*

(fraude demostrado) y sobre el resto de empresas proveedoras de las que no se dispone información (fraude sospechoso), para la detección y reconocimiento de patrones en un proceso supervisado de dos fases: la fase de entrenamiento (aprendizaje) y la fase de clasificación (comprobación).

El diseño de un sistema automático de detección de patrones esencialmente trata los siguientes tres aspectos, aspectos abordados en sucesivos capítulos de este trabajo: el acceso a la información y el pre-procesamiento de datos (Capítulo II), la aplicación de la metodología y la representación de los resultados (Capítulo III) y la toma de decisiones (Capítulo IV) (Li, 2005).

En el contexto del trabajo pericial realizado, el pre-procesamiento de datos cobra especial importancia en tanto que este análisis previo del contenido supuso alrededor del 65% del consumo de tiempo del trabajo forense. Este proceso mejora la exactitud de la clasificación de los modelos y reduce la cantidad de datos necesarios para obtener el nivel de rendimiento deseado, reduciendo los recursos computacionales necesarios e incorporando al estudio la mayor información disponible (García *et al*., 2015).

Aunque la bibliografía relacionada en esta área todavía profundiza poco en la heterogeneidad de las distintas áreas de conocimiento en las que se aplica (García *et al.*, 2015), en este trabajo se recopila la suficiente información para estructurar el proceso de gestión de las bases de datos en diferentes pasos que recogen y describen los diferentes procesos que requiere la preparación de los datos.

De este modo, el proceso de gestión de datos o de pre-procesamiento ha sido dividido en seis etapas: el acceso a la información (*Data Engineering*), la recopilación de los datos (*Data Scraping*), la limpieza de datos (*Data Cleansing*), el tratamiento de datos faltantes (*Missing Values*), la detección de valores atípicos (*Outliers*) y la transformación de los mismos (*Feature Engineering*).

Una vez realizado el proceso de gestión y depuración de las bases de datos son aplicadas las metodologías de aprendizaje propuestas a la muestra disponible. Aunque la oferta de modelos es muy amplia, en el Capítulo III se describen las cuatro metodologías de aprendizaje automático y ensamble de modelos (*ensemble models*) finalmente aplicadas

para la detección de patrones de forma sucinta y la justificación de su elección en este trabajo, (1) la Regresión Logística (LG), (2) el Árbol de Decisión (DT), (3) la Red Neuronal Artificial (NN) y (4) el Bosques Aleatorio (RF).

Desde un prisma global, las ventajas que ofrecen las nuevas metodologías de ensamble de modelos, especialmente la metodología de Bosque Aleatorio, sobre los modelos de clasificación independientes, han sido especialmente exitosas. En el resultado de esta investigación, el Bosque Aleatorio C4.5 aplicado sobre los estimadores de la Ley de Benford obtiene los mejores resultados, alcanzando un 96,15% de acierto de empresas fraudulentas y un 94,98% de acierto de empresas sospechosas (fraude sospechoso) [54].

La Ley de Benford es utilizada en su aplicación propuesta por Nigrini (1996), como medida para la detección de datos anómalos. Estos valores son utilizados en las 4 metodologías anteriormente descritas (segundo enfoque).

Esta ley afirma que algunos conjuntos de datos numéricos tienen una distribución no uniforme de los diferentes dígitos. En concreto, la Ley de Benford postula que existe un patrón de comportamiento para cada uno de los dígitos, siguiendo estos una ley logarítmica concreta (Benford, 1938). Como los datos económicos mayoritariamente siguen la ley de Benford se dispone de una herramienta para realizar comparaciones y poder detectar potenciales anomalías en los datos (Nigrini, 2011; Torres *et al.*, 2007; Bologna y Lindquist, 1995; Nigrini, 1996; Thomas, 1989; York, 2000). Sin embargo, el incumplimiento de la ley de Benford no es un delito en sí mismo, únicamente es evidencia de que los valores podrían presentar irregularidades.

En el caso que nos ocupa, y para clasificar a las empresas en legales o ilegales en función de las anomalías detectadas en su contabilidad, se establecen contrastes para determinar el grado de cumplimiento de la ley en cada empresa. En este sentido se propone un nuevo estadístico, el OverBenford Test, que se empleará de forma paralela con el Estadístico Z (Nigrini, 2012) para conocer los p-valores de ajuste a esta ley de los primeros y segundos dígitos de la variable "Importe de las Operaciones".

[54] Bosque Aleatorio C4.5 con balanceado de datos mediante la técnica SMOTE.

Especialmente cuando se aplican técnicas de aprendizaje automático en bases de datos reales, un problema importante es la existencia de conjuntos de datos desequilibrados (He y García, 2009; Chawla, 2005). Los conjuntos de datos desequilibrados son bastante habituales en los casos reales presentados en la literatura científica y, como en el caso expuesto, normalmente la categoría que es más relevante para el análisis es la que tiene una proporción menor de instancias (Japkowicz, 2000b; Chawla *et al.* 2003b; Chawla *et al.*, 2004; Dietterich *et al.*, 2003).

En el proceso judicial que nos ocupa la base de datos está claramente desequilibrada, del total de las 643 empresas que forman parte del entramado de blanqueo sólo se tiene certeza de que 26 de ellas son fraudulentas, es decir, las operaciones que realizan no se ajustan a la legalidad. Por tanto, de la información disponible *a priori*[55] sólo se tiene certeza de que el 4% de las empresas son fraudulentas. En este sentido, es empleado un método de generación sintética de datos (*Synthetic Minority Oversampling Technique*-SMOTE) y otro de aprendizaje sensible a costes (Matriz de Costes) para estudiar la mejora que se produce en relación con un análisis sin ninguna transformación de los datos originales (*Raw Data*).

Por último, y para comprobar la sensibilidad y rendimiento de los clasificadores, son propuestas diferentes estrategias de evaluación. En el primer enfoque se utiliza la matriz de confusión para evaluar el rendimiento y la sensibilidad de la red sobre 100 conjuntos de entrenamiento extraídos aleatoriamente. En el segundo enfoque se emplea validación cruzada para determinar la precisión de los modelos sobre 10 conjuntos de entrenamiento seleccionados al azar, en esta aproximación las medidas seleccionadas para evaluar la precisión son: el área de la curva ROC, el estadístico Kappa y el estadístico RMSE (*Root Mean Squared Error*).

Las técnicas de evaluación se emplean para comparar las diferentes metodologías de aprendizaje automático y analizar las diferencias de resultados entre los datos de entrenamiento (explicación) y los datos de comprobación (predicción). Con ello se

[55] Información facilitada por los investigadores policiales en el marco del proceso judicial.

determina la capacidad de los modelos, lo que ofrece suficiente información para analizar los patrones de comportamiento seguidos por cada empresa del entramado.

Con ello, las técnicas de aprendizaje automático propuestas en este trabajo representan una nueva herramienta eficiente y objetiva de detección de patrones de comportamiento fraudulentos para la investigación de delitos de blanqueo de capitales, permitiendo a los investigadores policiales priorizar los limitados recursos económicos y humanos disponibles en los procesos judiciales hacia aquellas empresas sospechosas con un patrón de comportamiento similar al de empresas fraudulentas previamente reconocidas.

La Tesis Doctoral se estructura en dos partes. En la primera parte, integrada por tres Capítulos, se establece el marco teórico sobre el que se sustenta la investigación. El primer Capítulo enmarca el concepto de blanqueo de capitales y estudia la tendencia de este delito en España. En el Capítulo II se describe el proceso de gestión y acceso a la información previa a la aplicación de las técnicas propuestas (pre-procesamiento de datos). A continuación, en el Capítulo III, se especifica la metodología basada en técnicas de aprendizaje automático aplicada para la detección de patrones de blanqueo de capitales.

La segunda parte se dedica a la exposición del proceso judicial y al análisis de los resultados. Tras una breve descripción del proceso judicial objeto de estudio y de la muestra disponible, en el Capítulo IV se presentan los resultados obtenidos en la aplicación de las técnicas de aprendizaje automático propuestas en los dos enfoques. La tesis doctoral finaliza con las conclusiones.

CONCLUSIONES

El análisis sobre la evolución del blanqueo de capitales (Apartado I.6.2) pone de manifiesto una tendencia creciente de este grave delito en España durante las últimas décadas. Las nuevas tecnologías, las transferencias electrónicas y los servicios de pago mediante plataforma *on-line* son excelentes alternativas de "*cyber laundering methodologies*" que dificultan la identificación de la trazabilidad de los capitales ilícitos (Souto, 2013). En este contexto, los investigadores deberían de disponer de herramientas automáticas y efectivas para su detección que garantizase un sistema judicial y de prevención eficiente (FATF, 2012).

Este estudio ofrece a los investigadores policiales una forma objetiva de identificar a empresas que presenten un patrón fraudulento dentro de una organización empresarial sospechosa de blanqueo de capitales. De este modo, la metodología aplicada ayuda a priorizar los recursos de investigación disponibles hacia aquellas empresas que presenten un mayor número de operaciones sospechosas.

Es de esperar que los modelos de aprendizaje automático propuestos sean tan buenos como el "ojo policial" y que aporten una nueva forma de detectar un mayor número de operaciones de fraude. Por tanto, los modelos implementados tratan de obtener correctamente el mayor número de operaciones fraudulentas posible, siendo importante minimizar la proporción de operaciones fraudulentas mal clasificadas (falsos positivos).

En el proceso judicial objeto de estudio se tiene la certeza de que las empresas proveedoras fraudulentas realmente lo son (fraude demostrado), sin embargo, del resto de empresas proveedoras se desconoce si participaron en esta compleja trama de blanqueo de capitales o no (fraude sospechoso). Detectar potenciales empresas defraudadoras por su similitud con las ya investigadas es uno de los objetivos de este trabajo.

Para ello, se desarrolla un proceso de análisis de información evaluable que favorece la interpretación de los resultados y que lo avala en el curso del proceso judicial. La investigación pericial se inicia con un completo pre-procesamiento de datos que dota de riqueza y robustez a los modelos. Una vez depurada la base de datos y verificadas las variables disponibles, se introduce la información en los modelos de clasificación.

Finalmente, los clasificadores son sometidos a análisis de sensibilidad que permiten la evaluación de los resultados.

Aunque en problemas "reales" donde se depende del criterio de un experto para determinar la etiqueta de la variable objetivo nunca se puede estar seguro del algoritmo de clasificación a utilizar (Bishop, 2006). En el primer enfoque se propone la exploración de los modelos de red neuronal como herramienta de trabajo en esta actividad pericial, con la que se han obtenido unos resultados notables. Por un lado, ha sido posible incorporar toda la información (variables) disponible en la estimación del modelo. Por otro lado, la rapidez con la que se ha obtenido el ajuste ha sido destacable.

En el modelo ajustado con datos de entrenamiento sin balancear se observa que la tasa de ajuste es mejorable, con un 72,16% de operaciones bien clasificadas. Esta tasa desciende al 17,63% cuando se trata de operaciones de fraude bien clasificadas. Por su parte, la tasa de falsos negativos (superior al 80%) hay que interpretarla con cuidado, ya que incluye en su recuento operaciones no irregulares de empresas identificadas por la policía como fraudulentas.

Además, la estrategia de muestreo seguida para la selección del conjunto de datos de entrenamiento no es neutral, tal como se constata con el experimento realizado con 100 conjuntos de entrenamiento distintos para detectar la sensibilidad del ajuste a cambios en el conjunto de entrenamiento. Se observa que la distribución de verdaderos positivos es asimétrica. La asimetría indicaría que la exclusión/inclusión de casos en el conjunto de entrenamiento no es neutral, por lo que la estrategia de selección aleatoria deja margen para la mejora.

El experimento anterior y la propia metodología de ajuste de la estructura de red utilizada llevan a aplicar una estrategia de mejora con datos de entrenamiento balanceados utilizando la técnica SMOTE. Se observa cómo, a pesar de descender la tasa de aciertos globales, la capacidad de detección de operaciones fraudulentas mejora hasta alcanzar casi un 70%, que prácticamente alcanza la tasa de aciertos globales del modelo sin balancear.

Habiéndose conseguido un ajuste notable se detectaron importantes oportunidades de mejora a través de la revisión de la estrategia de selección de casos para el conjunto de entrenamiento. Como la implementación de otras estructuras de red u otras metodologías de clasificación, e incluso incorporar otros sistemas de balanceado.

En el segundo enfoque se experimentan diferentes estructuras de clasificación: Regresión Logística Ridge (LG), Red Neuronal Artificial (NN), Árbol de Decisión C4.5 (DT) y Bosque Aleatorio (RF), que combinadas con la Ley de Benford son empleadas para la detección de patrones de blanqueo. Se incorpora la Matriz de Costes y la técnica SMOTE como técnicas de balanceo de datos.

Los resultados obtenidos en este enfoque muestran de nuevo que el algoritmo SMOTE obtiene mejores resultados sobre los verdaderos positivos que la muestra sin balancear, y por encima de la aplicación de la Matriz de Costes. Sin embargo, la precisión general del modelo es muy similar, por lo que la proporción de falsos positivos crecerá con la metodología SMOTE.

De forma análoga, aplicando la Matriz de Costes para el balanceo de los datos se identifican menos positivos de forma global. Consecuentemente, serían menos las empresas indicadas como potenciales defraudadores, y disminuirían las empresas investigadas a posteriori por la autoridad policial.

Seleccionar un número elevado de empresas a investigar tendría dos puntos negativos. El primero, sería el incremento de los costes de la investigación: más empresas investigadas supondrá más tiempo, más personal y más medios. El segundo, considerando las molestias causadas a las empresas legales que tendrán que soportar una investigación y sus consecuencias.

De forma global, el Bosque Aleatorio obtiene los mejores resultados con la transformación SMOTE en los conjuntos de entrenamiento, obteniendo un 96,15% de verdaderos negativos y un 94,98% de verdaderos positivos. Sin duda la capacidad de clasificación de esta metodología es muy elevada.

Las ventajas que ofrecen las nuevas metodologías de ensamble de modelos, particularmente la metodología de Bosque Aleatorio, sobre los modelos de clasificación

independientes, son especialmente exitosas, ya que (1) principalmente mejoran la precisión obtenida y reducen el sesgo del clasificador y (2) dotan de robustez a los modelos.

Sin embargo, la capacidad predictiva de los modelos con la transformación SMOTE se ha visto reducida cuando se utiliza un conjunto de datos para el entrenamiento y otro para la comprobación. Mientras que los resultados obtenidos por validación cruzada para el total de datos no difiere sustancialmente de los resultados obtenidos sólo con el 70% de los datos, los resultados de la predicción con el conjunto de comprobación son inferiores.

No obstante, los resultados que se obtienen en las 10 repeticiones son bastante estables, por lo que los modelos planteados en el segundo enfoque no son muy sensibles a la división que se haya efectuado de los datos. Tanto la capacidad descriptiva del modelo como la capacidad predictiva presentan rangos estrechos, lo que confirmaría que los modelos son estables.

Debe destacarse que la rapidez con la que evolucionan las técnicas de aprendizaje automático y la gran variedad de técnicas de aprendizaje automático disponibles ofrecen a los investigadores y forenses contables diferentes algoritmos que podrían incrementar el rendimiento y la eficiencia de los modelos de clasificación aquí propuestos. Por ejemplo, la experimentación con otras estructuras de Red Neuronal más complejas u otros métodos de aprendizaje (Ripley, 2007; Bishop, 2006), e incluso su integración en una estrategia tipo *Boosting* (Freund y Schapire, 1996) junto con la técnica SMOTE (SMOTE-Boost de Chawla *et al*., 2003b), u otras técnicas de ensamblaje de modelos (Wang y Shao, 2009).

Una solución óptima sería utilizar diferentes métodos y algoritmos para evaluar enfoques adicionales. De este modo se ofrecería a las autoridades policiales la mayor información posible para analizar correctamente los resultados que se derivan de aplicar las técnicas de aprendizaje automático. Por ejemplo, los datos sin transformación clasifican muy bien los verdaderos negativos, y esto también podría ser aprovechado si se aplican métodos de ensamble de modelos para la clasificación final.

Por tanto, los resultados aquí obtenidos abren un amplio abanico de posibilidades a la mejora del trabajo pericial en la detección del fraude financiero y el blanqueo de capitales. El uso de este tipo de herramientas predictivas junto con los análisis de sensibilidad y rendimiento de las técnicas de aprendizaje automático facilitan la detección/identificación de patrones de fraude que puedan describirse *a priori*. Esto refuerza el papel del contable forense en este tipo de investigaciones, ya que uno de los principales problemas que los investigadores de procesos judiciales de blanqueo de capitales afrontan es la incapacidad para traducir la gran cantidad de información, derivada de las complejas estructuras empresariales, en patrones de comportamiento de los individuos claramente fraudulentos.

Como resultado de este trabajo, del listado de empresas no categorizadas por la Policía Judicial como fraudulentas en la investigación previa (fraude sospechoso), los sistemas expertos aplicados han sido exitosamente capaces de ofrecer información sobre ciertas empresas proveedoras que presentaban patrones de comportamiento similares a las empresas categorizadas como fraudulentas en la información disponible *a priori*. Desafortunadamente, la realidad del proceso judicial (todavía *sub judice*) no permite realizar una rigurosa valoración *a posteriori* de las indicaciones que se ofrecieron a las autoridades policiales.

Actualmente, la detección de patrones de comportamientos fraudulentos por este tipo de sistemas expertos no supone un delito en sí mismo, sin embargo, es un indicio más que complementa la información de la que disponen los investigadores policiales. La dificultad para describir los patrones de fraude detectados (modelos de cajas negras) por las técnicas de aprendizaje automático imposibilita una explicación detallada y coherente de los mismos, por lo que este hecho no representa un indicio lo suficientemente consolidado para ser aportado en el proceso judicial como evidencia.

Fundamentado en la experiencia de esta investigación, la cual está enmarcada en un proceso judicial actual y circunscrita en la pionera colaboración con la Policía Judicial Española, Brigada de Blanqueo de Capitales, se obtiene un resultado que indicaría la viabilidad de este tipo de modelos como herramienta en la planificación y priorización de las tareas de investigación policial.

Con todo ello, esta interpretación ha evidenciado que las técnicas de aprendizaje automático representan una nueva herramienta que es capaz de orientar eficazmente y de forma objetiva a los investigadores hacia aquellas empresas que presentan mayor probabilidad de estar blanqueando capitales.

Sin embargo, al ser la primera vez que se utilizan técnicas de aprendizaje automático en un caso real para la detección del fraude y el blanqueo de capitales, y bajo las limitaciones de tiempo para su implementación, son muchas las cuestiones que se podrían estudiar y analizar en investigaciones futuras como se expone seguidamente.

Se podrían usar nuevas metodologías de aprendizaje automático como el *Deep Learning* o el E*xtreme Gradient Boosting* (algoritmo *XGBoost*). En ambos casos el tiempo de cómputo se incrementa considerablemente en comparación con las metodologías utilizadas en esta Tesis Doctoral. No obstante, los resultados obtenidos en otros campos alientan su utilización para estas labores, por lo que se podría profundizar en su utilización y aplicación en el sector financiero.

También se debería estudiar la posibilidad de incluir información adicional sobre las empresas del entramado. La información contable de las empresas es muy importante para detectar ciertas irregularidades, pero es bien cierto, que en la mayoría de los casos es una contabilidad manipulada lo que dificulta que los forenses contables obtengan información. Incluir información de redes sociales, registros públicos, precios de los mercados financieros, etc. aportaría datos que mejoraría la fiabilidad de los modelos utilizados.

Además, a través de las transacciones comerciales entre empresas y mediante la Teoría de Grafos se podría conocer las relaciones dentro de un entramado empresarial fraudulento. Este análisis permitiría identificar comunidades de empresas que trabajan conjuntamente en el proceso de blanqueo. La Teoría de Grafos haría evidente la relación entre empresas que *a priori* no tuvieran conexión comercial directa, lo que facilitaría la trazabilidad de los capitales y la identificación de nuevas empresas potencialmente defraudadoras.

Asimismo, el conocimiento adquirido en este trabajo reforzaría el empleo de técnicas de aprendizaje automático a fin de analizar información empresarial (transacciones, organización de empresas, facturación,..) proveniente de instituciones como la Agencia Tributaria, entidades financieras o Unidades de Inteligencia Financiera. En todo caso, futuras líneas de investigación se deberían desarrollar en su aplicación como herramienta para el análisis de texto de informes de operaciones sospechosas (SARs), el objetivo sería codificarlas según su nivel de riesgo. Con ello, se agilizaría un proceso que actualmente consume excesivos recursos humanos y de tiempo, y facilitaría la clasificación de empresas o grupos empresariales sospechosos, ayudando a priorizar recursos disponibles.

Además, otra futura línea de investigación debería avanzar en su aplicación para la detección de patrones de blanqueo de capitales en mercados financieros con plataforma *on-line*. La existencia de monedas virtuales pone de manifiesto nuevos mecanismos de blanqueo que requieren el uso de técnicas de gestión de datos automáticas para su prevención, detección e investigación.

El empleo de herramientas automáticas como las propuestas en esta Tesis Doctoral, que supongan una mejora de la eficiencia de los mecanismos de lucha que se están llevando a cabo a nivel internacional contra este delito, incrementaría las oportunidades de combatir el blanqueo de capitales y reduciría los sofisticados sistemas por los que se financia el terrorismo (FATF, 2012; UNODC, 2008).