MACHINE LEARNING: WEKA

Paz Rico Belda
Departamento de Análisis Económico
Universidad de Valencia

Making Learning: Weka

1. Motivación
2. Minería de datos
3. Weka
4. Aplicación
5. Conclusiones

Motivación

* Noticia: La consultora Multinacional Everis precisa contratar a jóvenes recién **titulados especialistas** en áreas tecnológicas como el **Big Data y Business Inteligence**

* **Motivar** a nuestros alumnos en las técnicas estadísticas y/o econométricas

* Utilización de herramientas/**programas** de **libre acceso**

Minería de datos

* **Ingentes** cantidades de **datos**
* Necesidad de extraer **información útil** para la toma de decisiones. Estrategia competitiva
* Minería de datos es una **tarea** dentro del **proceso KDD** (Knowledge Discovery in Database)
* La **estadística** es una herramienta **fundamental** en la Minería de datos

Minería de datos vs Estadística

* No hay línea divisora entre estadística y minería. Es una línea continua.
* Muchas técnicas son estadísticas y la mayor parte de los algoritmos utilizan test estadísticos para construir normas o validar/evaluar los resultados.
* Diferencia fundamental con la estadística es el tamaño del conjunto de datos.
  * Estadística: Cientos o miles de datos
  * Minería de datos: millones o miles de millones de datos

Minería de datos

Etapas extracción información útil:

1. Identificar objetivo/selección de datos
2. Pre-procesamiento de datos: 60%-70% del tiempo total
3. Minería de datos
4. Análisis de resultados
5. Asimilación de conocimiento

Minería de datos

Cosas que permite la minería de datos:

1. Predicción de comportamientos. Ejemplo: Marketing dirigido
2. Predicción de tendencias. Ejemplo: Predicción ventas
3. Descubrimiento de comportamientos desconocidos. Ejemplo: relaciones entre variables

Minería de datos

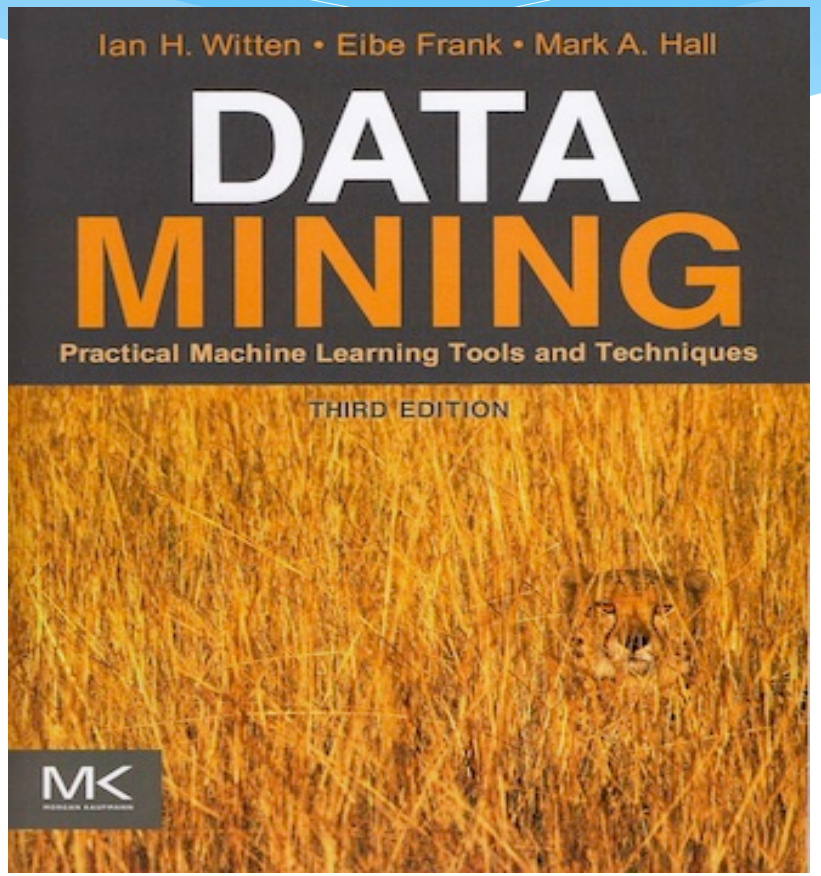
Técnicas de la minería de datos:

1. Redes neuronales
2. Árboles de decisión
3. Algoritmos
4. Regresión
5. Clustering

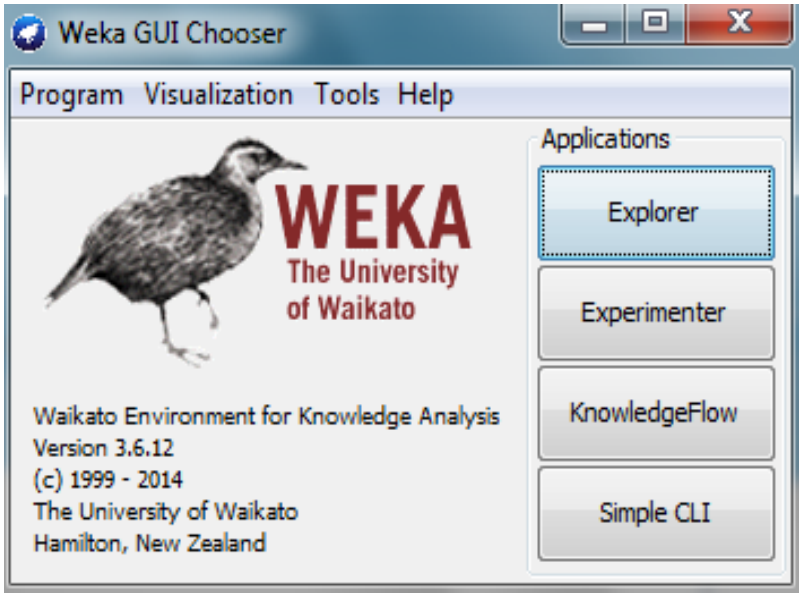
WEKA

* Waikato Environment for Knowledge Analysis
* Colección de algoritmos con licencia GNU-GPL
* 1993, primera versión en C
* 1997, versión en Java
* http://www.cs.waikato.ac.nz/ml/weka
* Windows, Mac y Linux
* Ave endémica de Nueva Zelanda

WEKA


Ian H. Witten • Eibe Frank • Mark A. Hall
DATA
MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK

WEKA



Cuatro entornos de trabajo:

1. **Explorer**: Uso de paquetes
2. Experimenter: Comparación de predicción
3. KnowledgeFlow: Aprendizaje incremental
4. Simple CLI: Comandos

Aplicación. Clasificación

PARADO	PRESTAM	RENTA	VIVPROP
0	0	6.943	1
0	0	17.889	1
1	0	1.138	1
1	0	8.093	0
1	0	4.116	1
1	0	6.785	1
1	0	7.278	1
1	0	2.749	1
1	0	6.433	1
1	0	8.359	1
1	0	7.174	0
1	0	5.004	0
0	0	5.558	1
1	0	4.511	1
1	0	6.286	1
1	0	5.965	1
1	0	7.729	1
1	0	4.882	0
1	0	9.261	0
0	1	14.219	1
0	0	14.996	1

* Datos utilizados antigua licenciatura ADE
* 1000 observaciones
* ¿Reciben o no préstamo bancario?
* Situación laboral
* Renta
* Propietario o no de vivienda

Aplicación. Clasificación

Dependent Variable: PRESTAM
Method: ML - **Binary Logit** (Quadratic hill climbing)
Sample: 1 1000
Included observations: 1000
Convergence achieved after 6 iterations
Covariance matrix computed using second derivatives

	Coefficient	Std. Error	z-Statistic	Prob.
C	-4.214824	0.516033	-8.167736	0.0000
RENTA	0.142950	0.018250	7.832665	0.0000
PARADO	-0.969263	0.270492	-3.583334	0.0003
VIVPROP	0.719559	0.488548	1.472853	0.1408

McFadden R-squared	0.215358	Mean dependent var	0.086000
S.D. dependent var	0.280504	S.E. of regression	0.253246
Akaike info criterion	0.468089	Sum squared resid	63.87690
Schwarz criterion	0.487720	Log likelihood	-230.0446
Hannan-Quinn criter.	0.475550	Restr. log likelihood	-293.1843
LR statistic	126.2794	Avg. log likelihood	-0.230045
Prob(LR statistic)	0.000000		

Obs with Dep=0	914	Total obs	1000
Obs with Dep=1	86		

Resultados del modelo:

* Renta incrementa probabilidad de recibir préstamo
* Estar parado disminuye la probabilidad de recibir préstamo
* Poseer vivienda propia no afecta

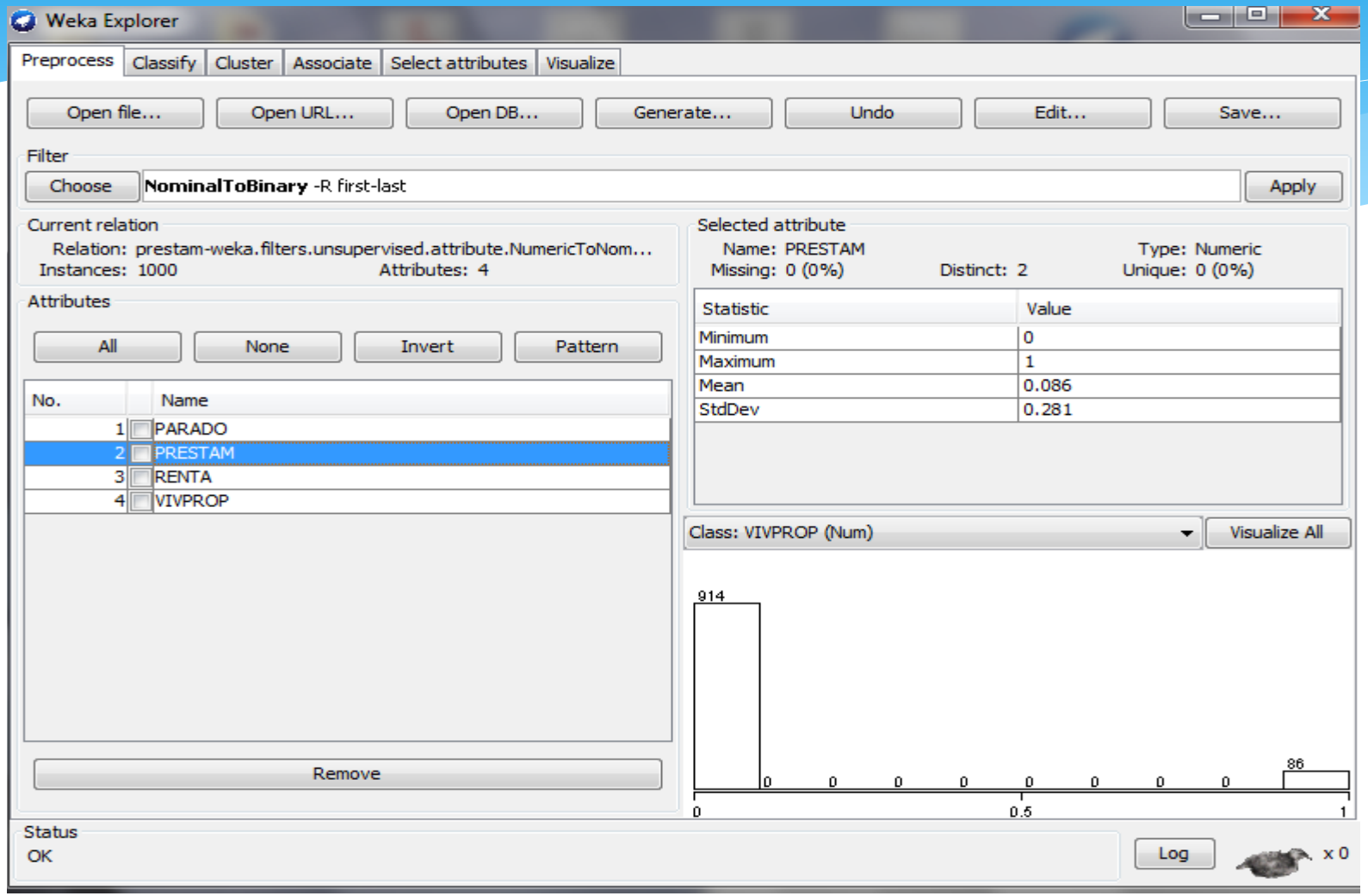
Aplicación. Clasificación

Dependent Variable: PRESTAM
Method: ML - Binary Logit (Quadratic hill climbing)

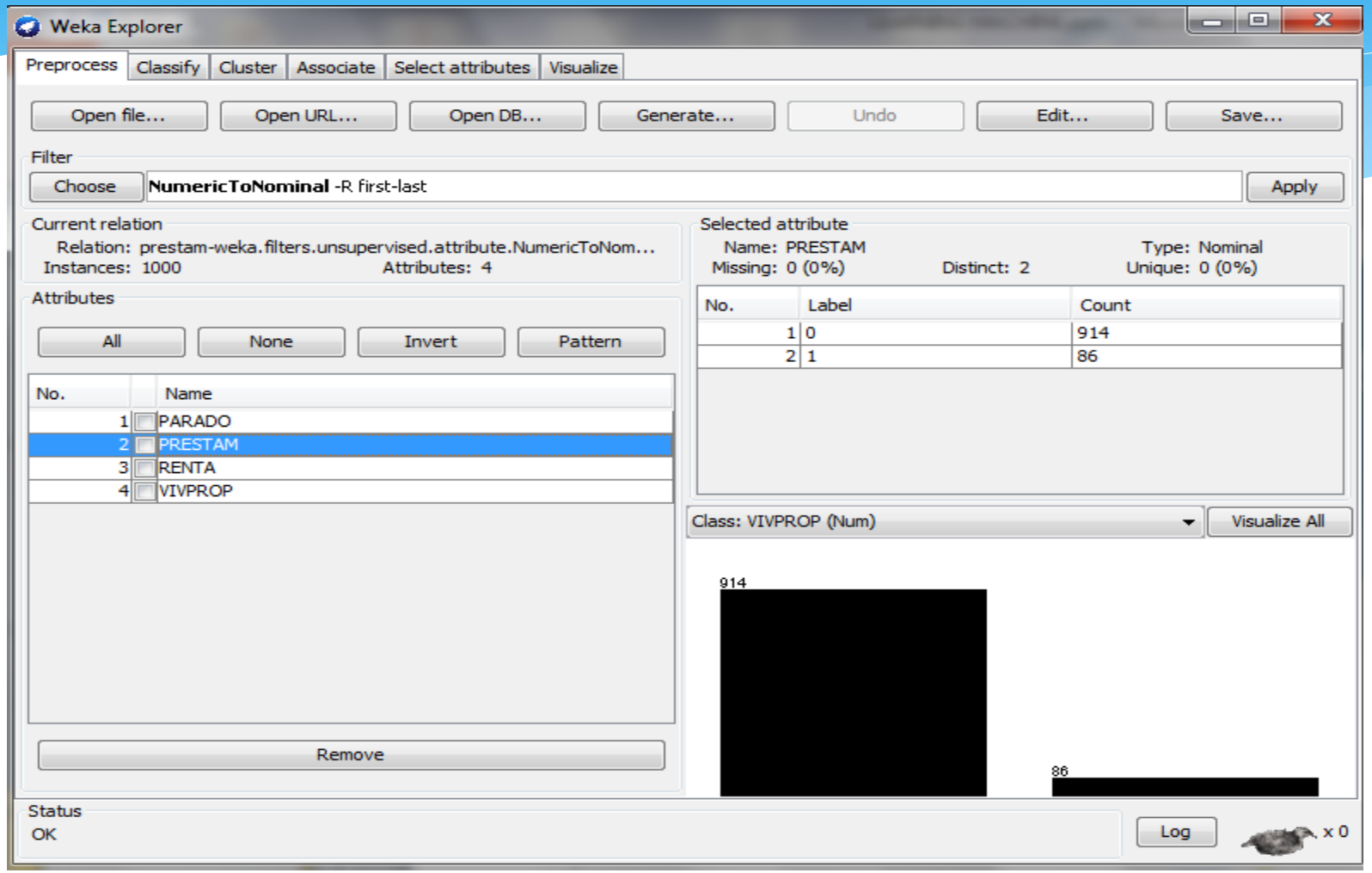
Sample: 1 1000.Included observations: 1000
Prediction Evaluation (success cutoff C = 0.5)

	Estimated Equation			**Constant Probability**		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
P(Dep=1)<=C	901	72	973	914	86	1000
P(Dep=1)>C	13	14	27	0	0	0
Total	914	86	1000	914	86	1000
Correct	901	14	915	914	0	914
% Correct	98.58	16.28	**91.50**	100.00	0.00	**91.40**
% Incorrect	1.42	83.72	8.50	0.00	100.00	8.60
Total Gain*	-1.42	16.28	0.10			
Percent Gain**	NA	16.28	1.16			

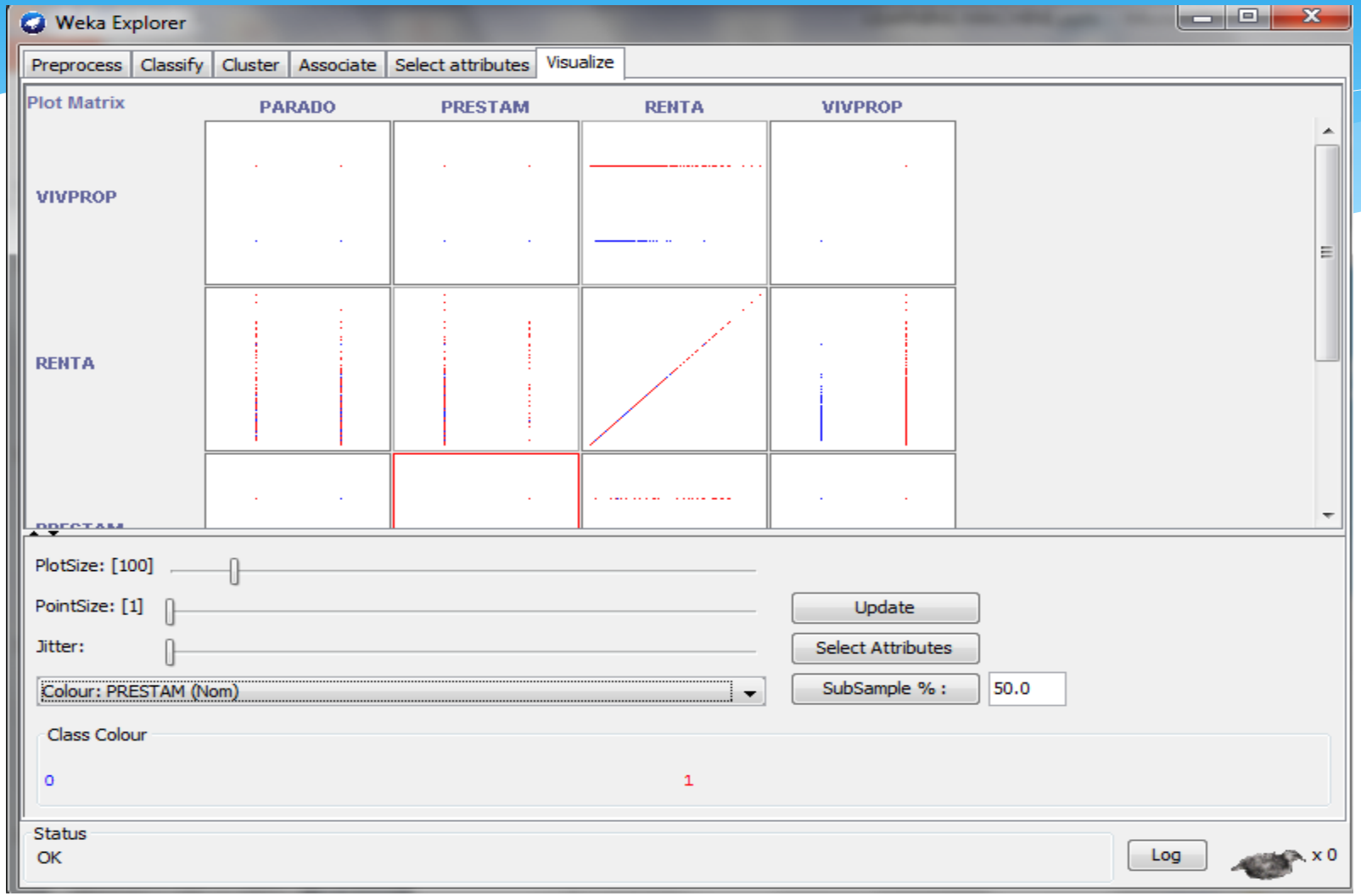
WEKA. Clasificación



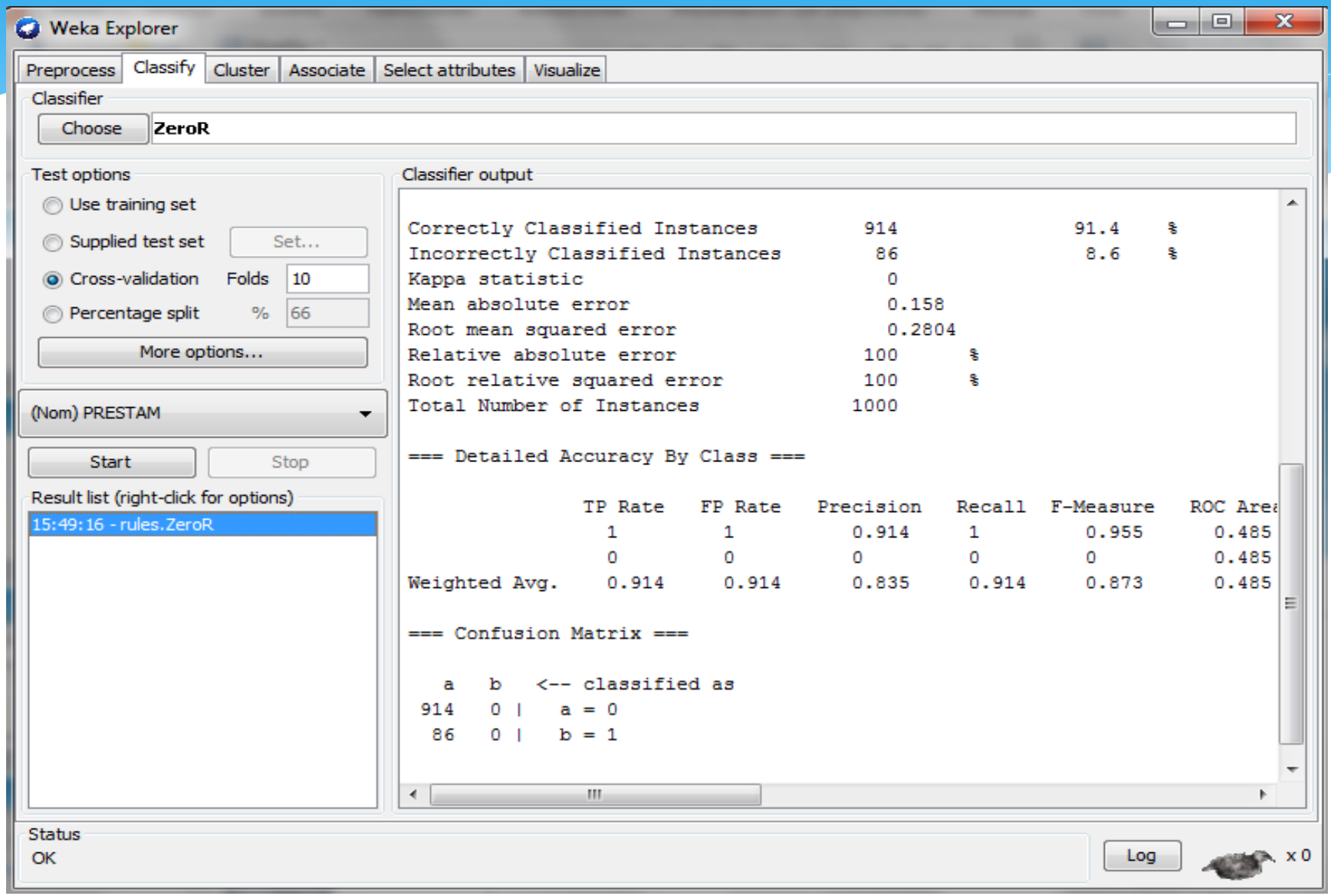
WEKA. Clasificación



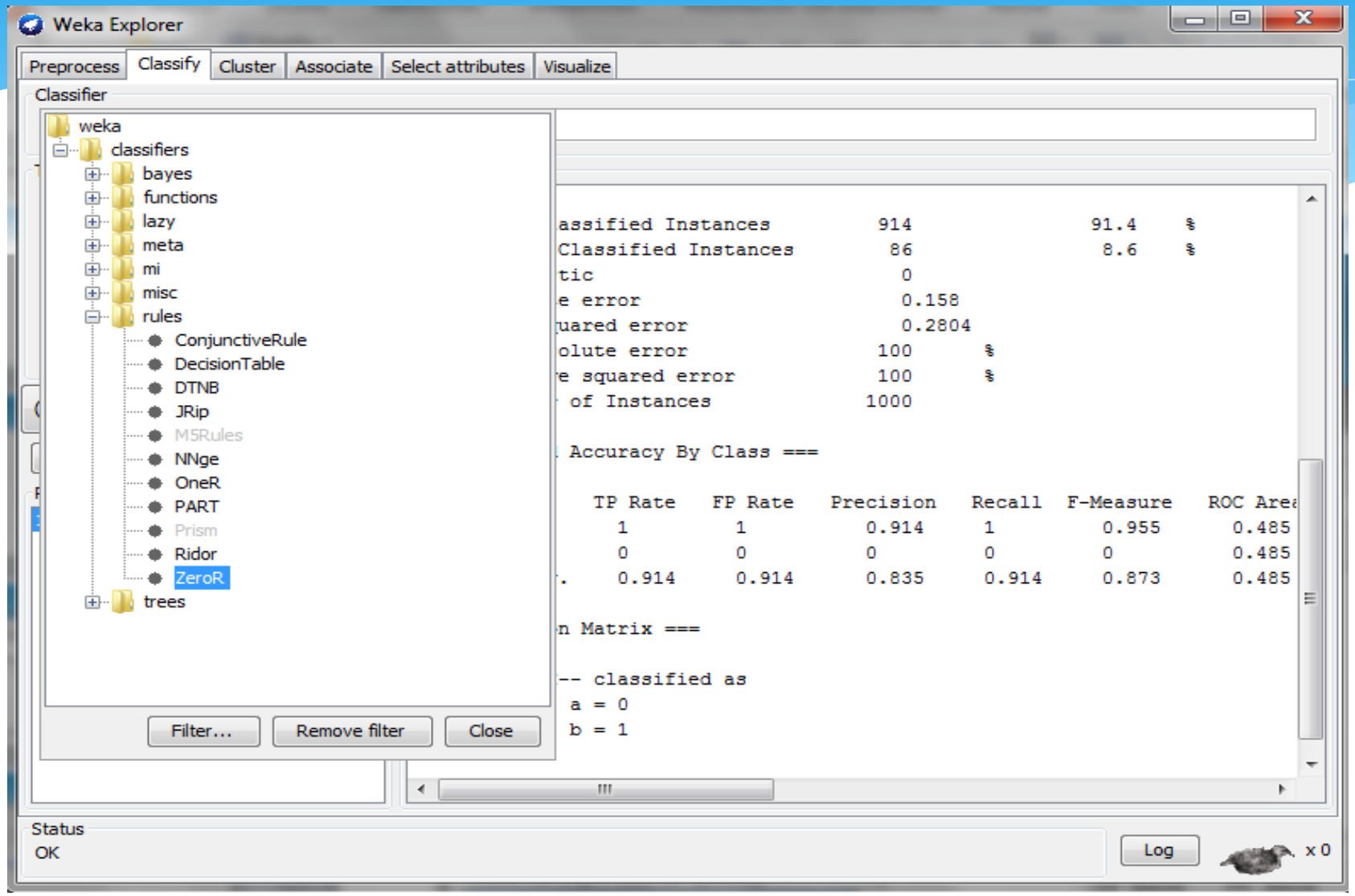
WEKA. Clasificación



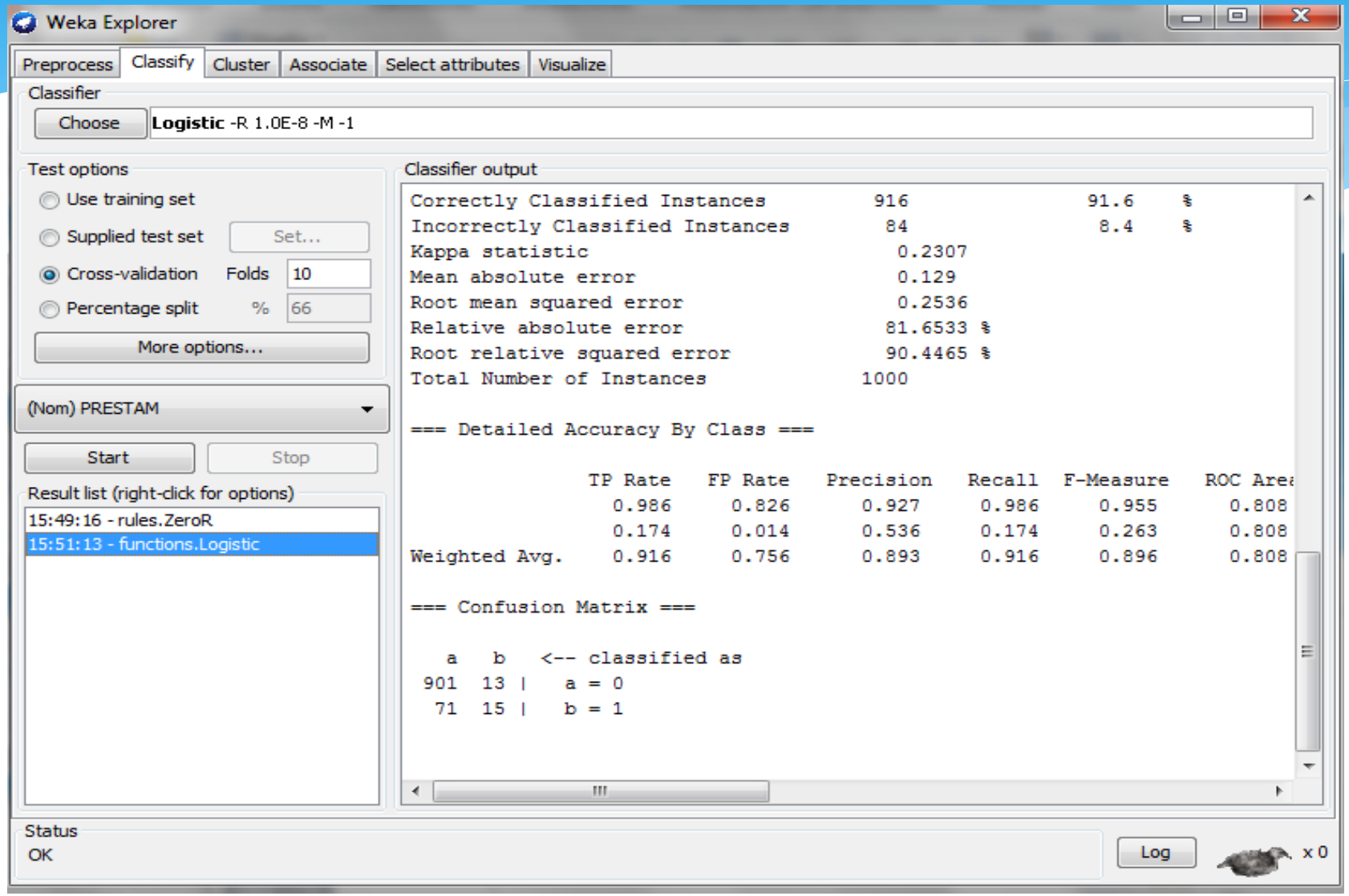
WEKA. Clasificación



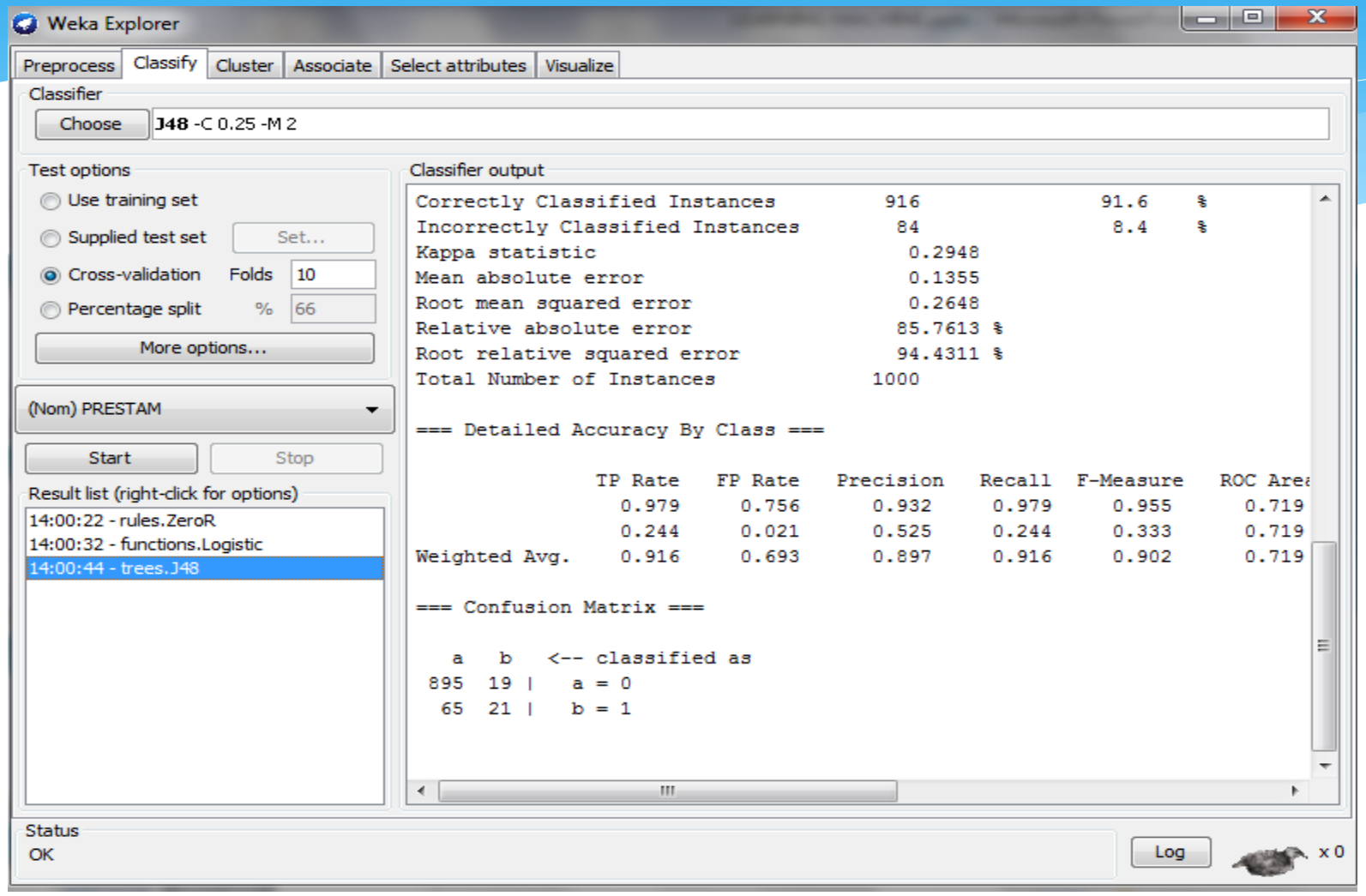
WEKA. Clasificación



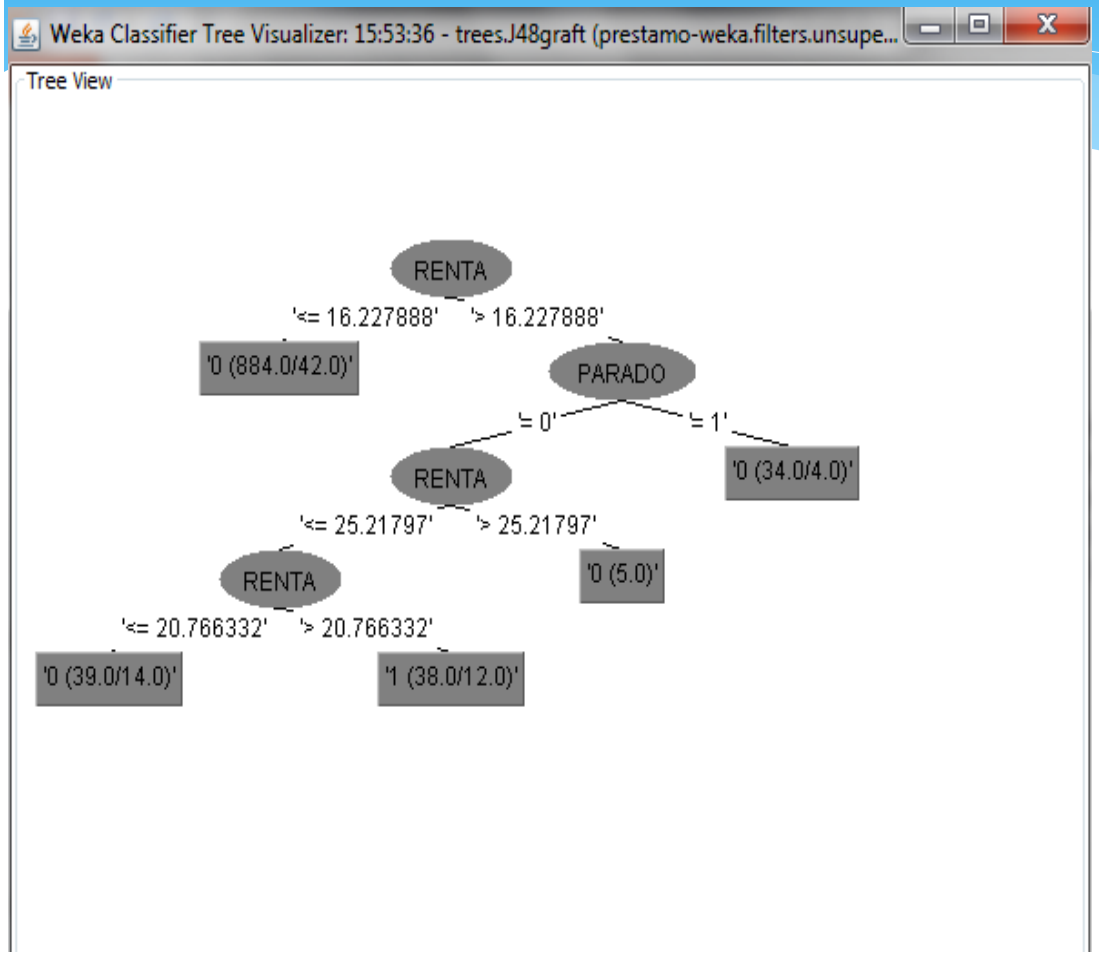
WEKA. Clasificación



WEKA. Clasificación

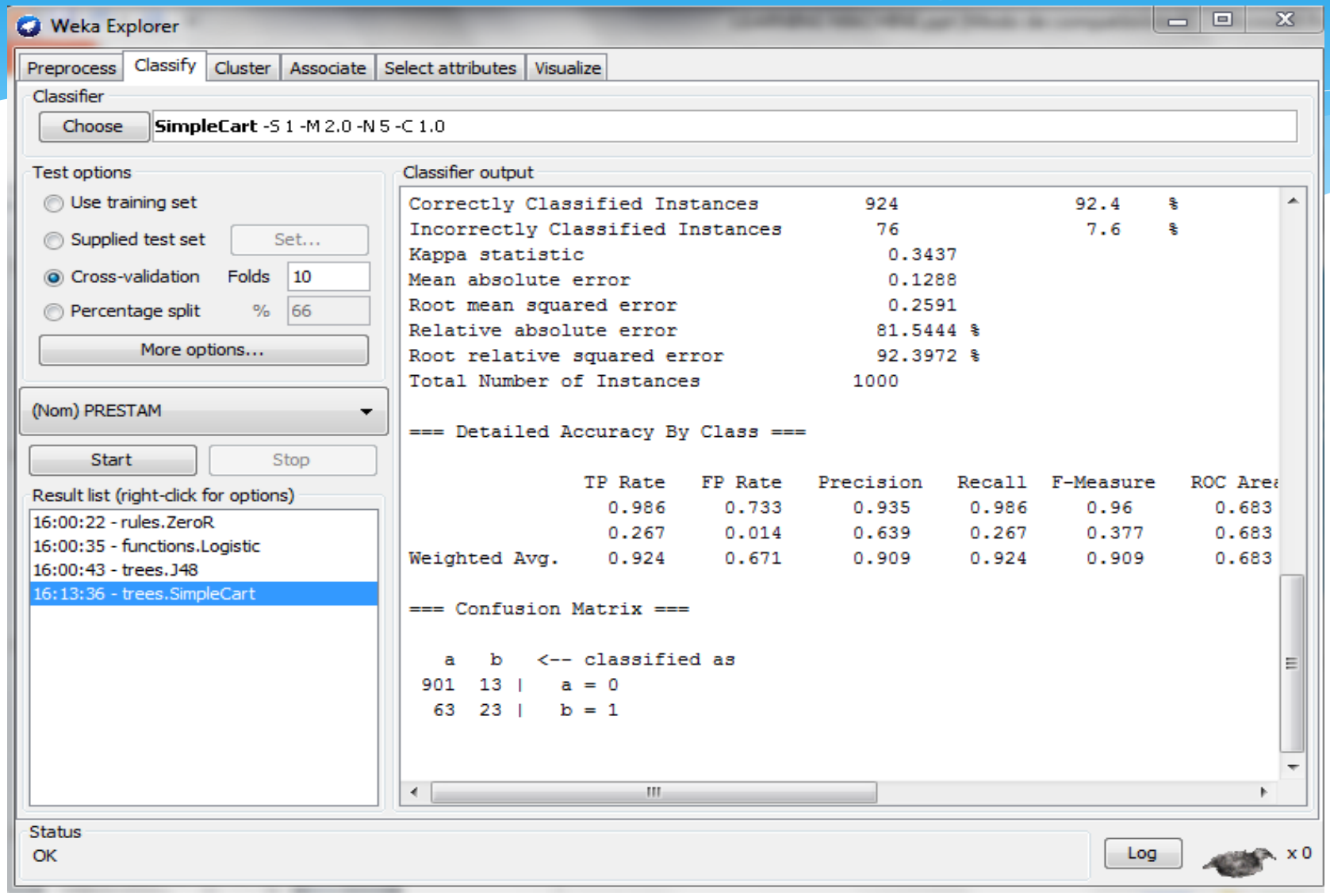


WEKA. Clasificación



Reciben préstamo bancario quienes tienen rentas superiores a 20,7 y no están en desempleo

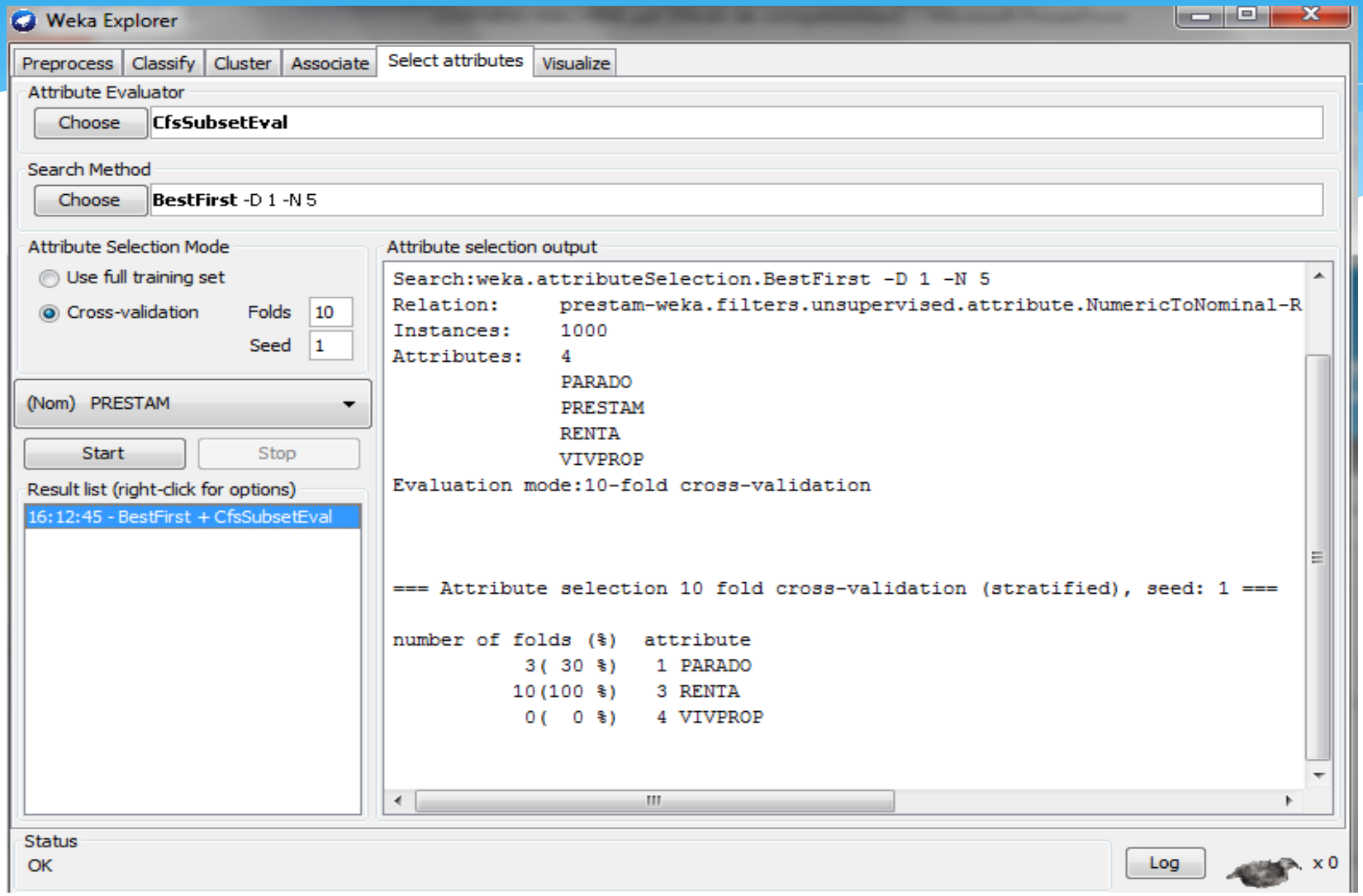
WEKA. Clasificación



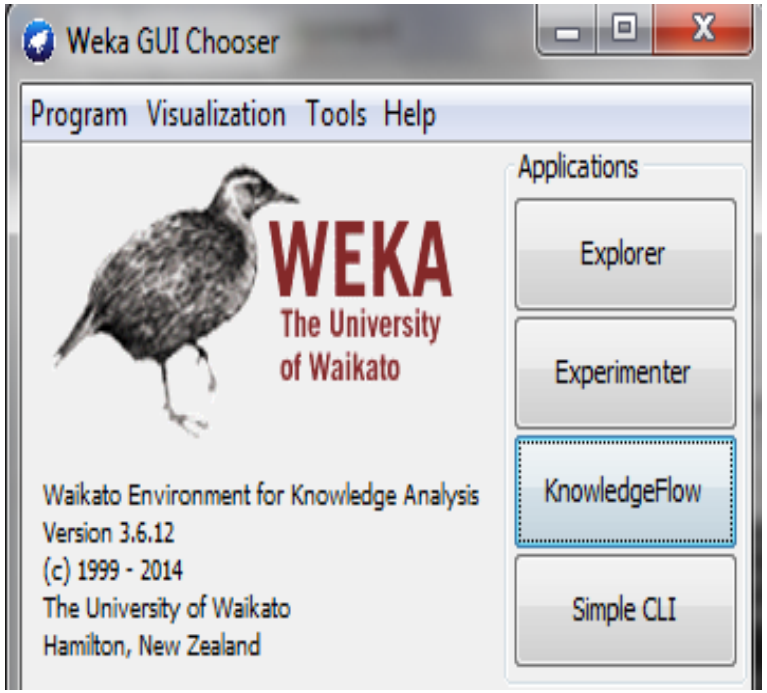
WEKA. Clasificación

```
CART Decision Tree

RENTA < 16.421594499999998: 0(842.0/42.0)
RENTA >= 16.421594499999998
|   PARADO=(0)
|   |   RENTA < 25.2920725
|   |   |   RENTA < 21.075501
|   |   |   |   RENTA < 17.166837: 1(2.0/0.0)
|   |   |   |   RENTA >= 17.166837
|   |   |   |   |   RENTA < 20.3251475
|   |   |   |   |   |   RENTA < 19.2279765
|   |   |   |   |   |   |   RENTA < 17.666596: 0(2.0/0.0)
|   |   |   |   |   |   |   RENTA >= 17.666596
|   |   |   |   |   |   |   |   RENTA < 18.845913
|   |   |   |   |   |   |   |   |   RENTA < 18.242503499999998: 0(8.0/4.0)
|   |   |   |   |   |   |   |   |   RENTA >= 18.242503499999998: 1(2.0/0.0)
|   |   |   |   |   |   |   |   RENTA >= 18.845913: 0(2.0/0.0)
|   |   |   |   |   |   RENTA >= 19.2279765: 1(2.0/0.0)
|   |   |   |   |   RENTA >= 20.3251475: 0(13.0/4.0)
|   |   |   RENTA >= 21.075501: 1(26.0/12.0)
|   |   RENTA >= 25.2920725: 0(5.0/0.0)
|   PARADO!=(0): 0(30.0/4.0)

Number of Leaf Nodes: 11

Size of the Tree: 21
```

WEKA. Clasificación



WEKA. Clasificación



EXPERIMENTER:

Comparar varios algoritmos de clasificación:

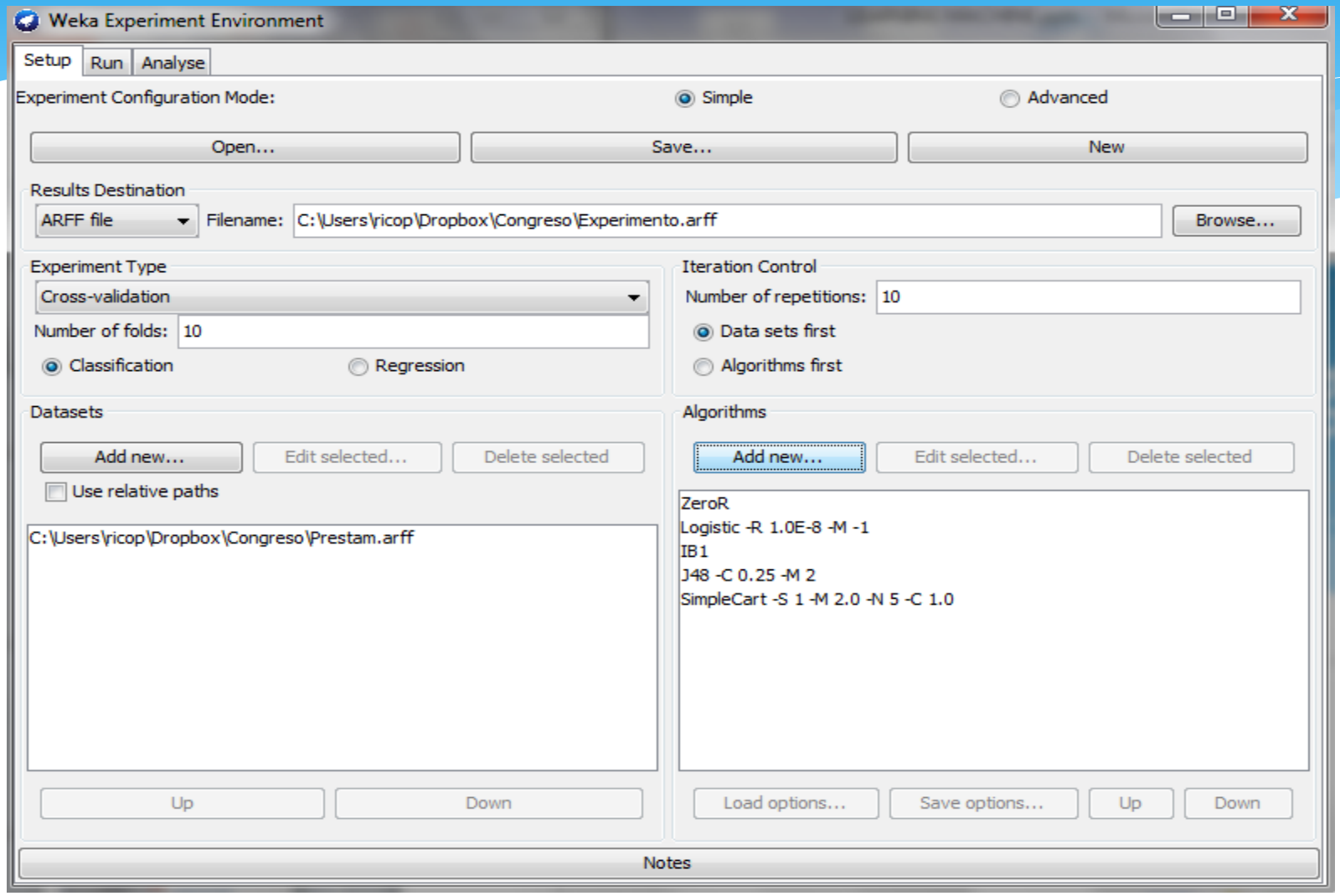
ZeroR

Logit

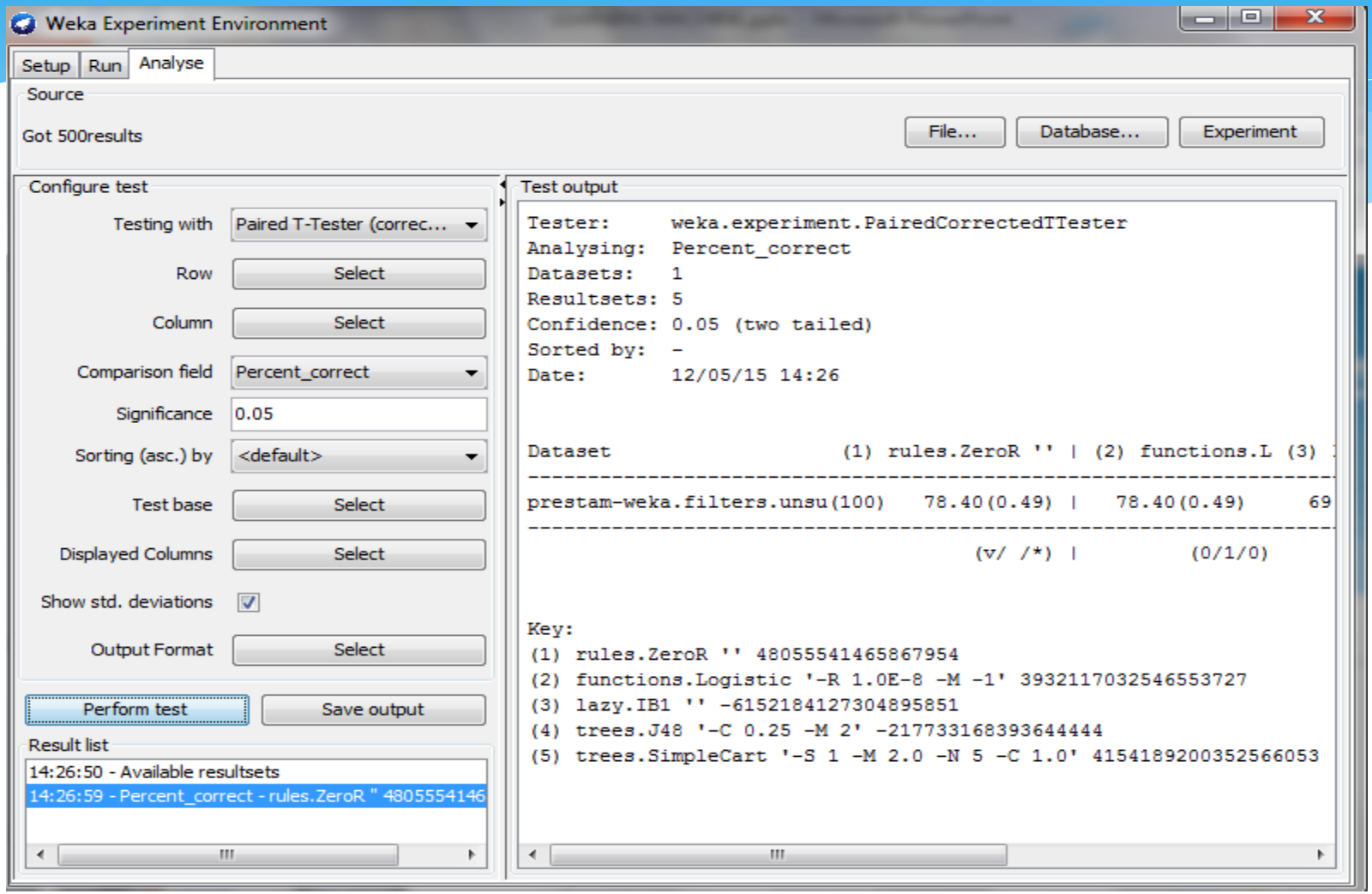
Tree

Lazy

WEKA. Clasificación



WEKA. Clasificación



WEKA. Clasificación

Dataset	(1) rules.ZeroR ''	(2) functions.L	(3) lazy.IB1 ''	(4) trees.J48 '	(5) trees.Simpl
prestam-weka.filters.unsu(100)	78.40(0.49)	78.40(0.49)	69.99(4.27) *	78.40(0.49)	78.37(0.56)
(v/ /*)		(0/1/0)	(0/0/1)	(0/1/0)	(0/1/0)

Key:
(1) rules.ZeroR '' 48055541465867954
(2) functions.Logistic '-R 1.0E-8 -M -1' 3932117032546553727
(3) lazy.IB1 '' -6152184127304895851
(4) trees.J48 '-C 0.25 -M 2' -217733168393644444
(5) trees.SimpleCart '-S 1 -M 2.0 -N 5 -C 1.0' 4154189200352566053

Aplicación. Cluster

V1	V2	V3	V4	V5	V6
6	4	7	3	2	3
2	3	1	4	5	4
7	2	6	4	1	3
4	6	4	5	3	6
1	3	2	2	6	4
6	4	6	3	3	4
5	3	6	3	3	4
7	3	7	4	1	4
2	4	3	3	6	3
3	5	3	6	4	6
1	3	2	3	5	3
5	4	5	4	2	4
2	2	1	5	4	4
4	6	4	6	4	7
6	5	4	2	1	4
3	5	4	6	4	7
4	4	7	2	2	5
3	7	2	6	4	3
4	6	3	7	2	7
2	3	2	4	7	2

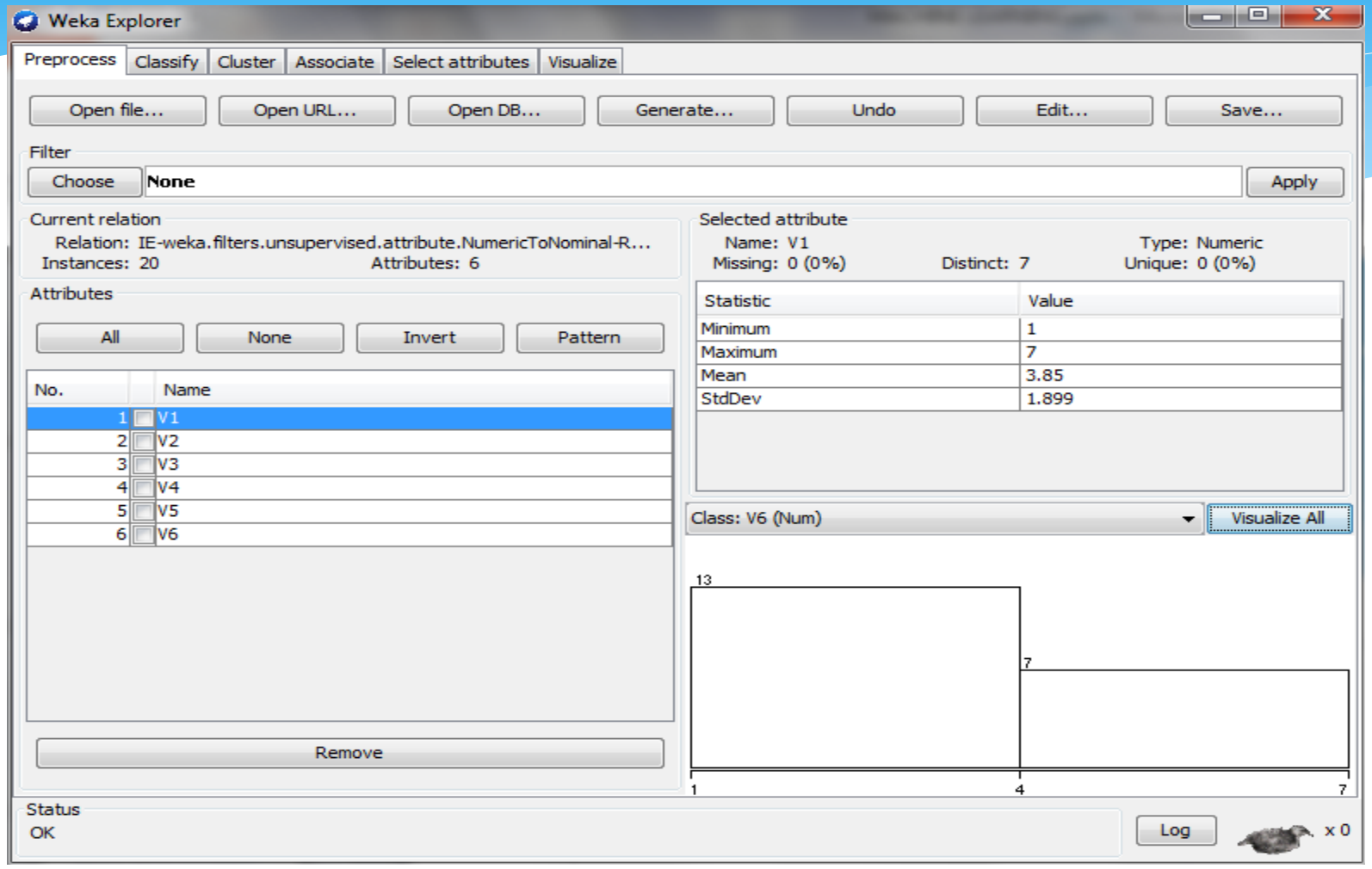
* V1: Salir de compras es divertido
* V2: Salir de compras afecta al presupuesto
* V3: Combinar salir de compras con comida fuera
* V4: Salir de compras para hacer las mejores compras
* V5: No me importa salir de compras
* V6: Se puede ahorrar dinero comparando precios

Aplicación. Clúster

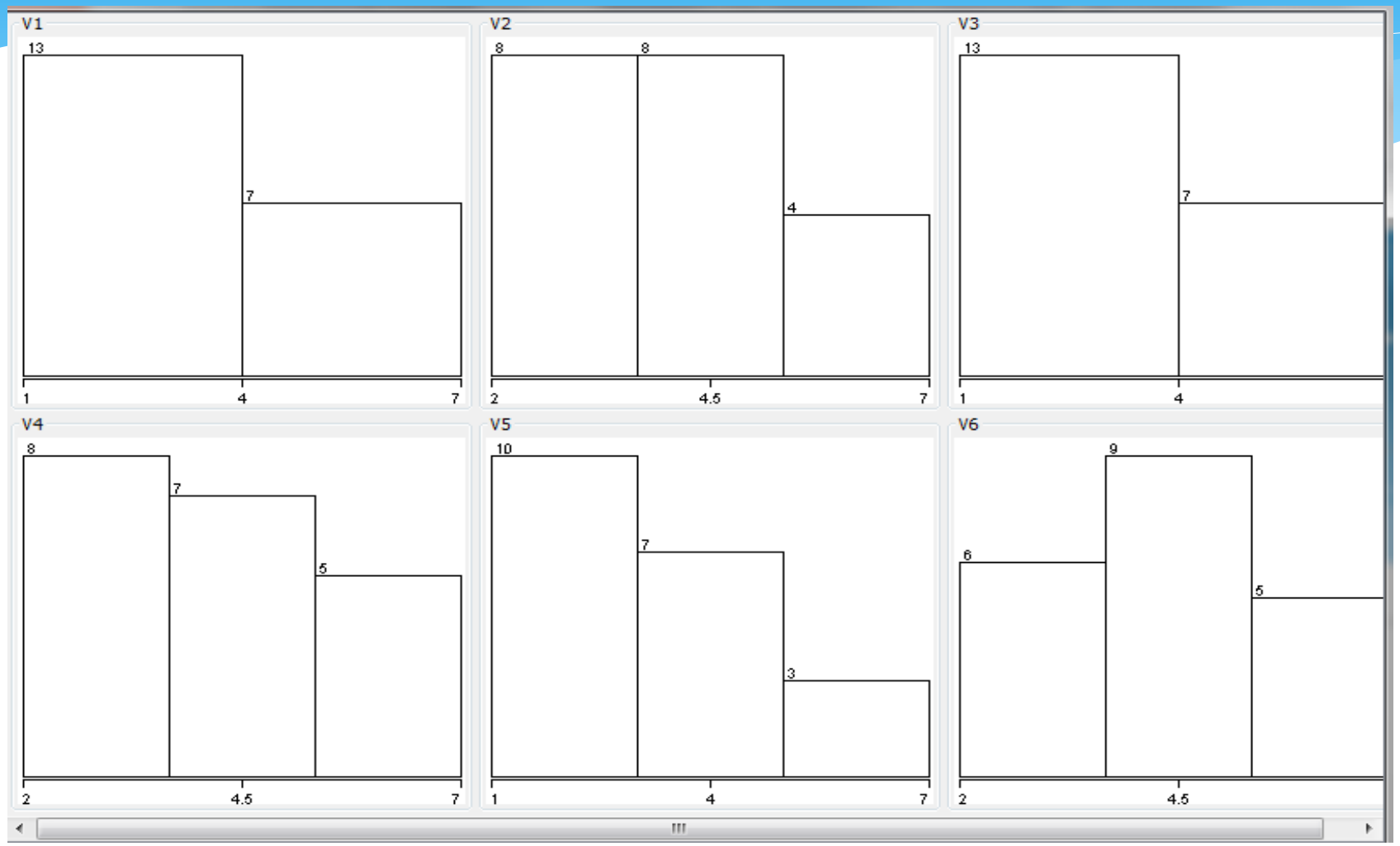
Objetivo: Agrupar consumidores homogéneos frente a su actitud hacia las compras

Fuente: “Análisis de conglomerados” Santiago de la Fuente Fernández. UAM, 2011

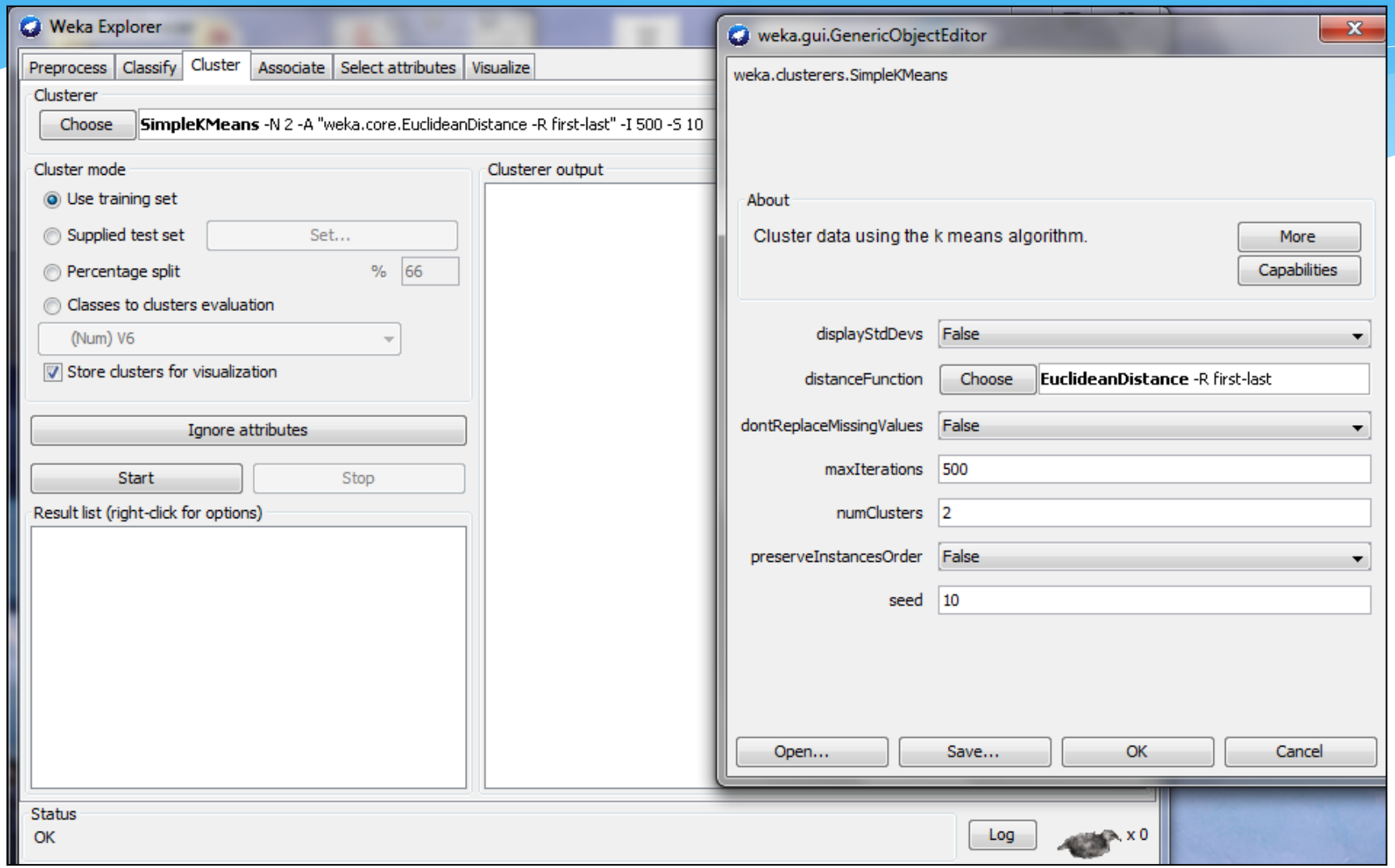
WEKA CLUSTERING



WEKA CLUSTERING



WEKA


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Clusterer
Choose
SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Cluster mode
Use training set
Supplied test set
Set...
Percentage split
%
66
Classes to clusters evaluation
(Num) V6
Store clusters for visualization
Ignore attributes
Start
Stop
Result list (right-click for options)
Clusterer output
Status
OK
Log
x 0
weka.gui.GenericObjectEditor
weka.clusterers.SimpleKMeans
About
Cluster data using the k means algorithm.
More
Capabilities
displayStdDevs
False
distanceFunction
Choose
EuclideanDistance -R first-last
dontReplaceMissingValues
False
maxIterations
500
numClusters
2
preserveInstancesOrder
False
seed
10
Open...
Save...
OK
Cancel

WEKA

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 6.09
Missing values globally replaced with mean/mode

Cluster centroids:
                            Cluster#
Attribute    Full Data          0           1
                  (20)        (12)         (8)
=============================================
V1                3.85        2.58        5.75
V2                4.10        4.42        3.62
V3                3.95        2.58        6
V4                4.10        4.75        3.12
V5                3.45        4.50        1.87
V6                4.35        4.67        3.87



Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       12 ( 60%)
1        8 ( 40%)
```

WEKA

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 2.72
Missing values globally replaced with mean/mode

Cluster centroids:
                            Cluster#
Attribute    Full Data          0          1          2
                  (20)        (6)        (8)        (6)
=============================================================
V1                3.85       3.50       5.75       1.67
V2                4.10       5.83       3.62       3
V3                3.95       3.33       6          1.83
V4                4.10       6          3.12       3.5
V5                3.45       3.5        1.87       5.5
V6                4.35       6          3.87       3.33



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0         6 ( 30%)
1         8 ( 40%)
2         6 ( 30%)
```

WEKA

- Primer clúster corresponde a consumidores ahorrativos: V2, V4 y V6 elevados
- Segundo clúster corresponde a consumidores que disfrutan con las compras: V1 y V3 elevados
- Tercer clúster corresponde a consumidores apáticos: V1 y V3 bajos y V5 elevado

- V1: Salir de compras es divertido
- V2: Salir de compras afecta al presupuesto
- V3: Combinar salir de compras con comida fuera
- V4: Salir de compras para hacer las mejores compras
- V5: No me importa salir de compras
- V6: Se puede ahorrar dinero comparando precios

CONCLUSIONES

* Alumnos: mejorar las competencias en estadística
* Reflexionar sobre planes de estudios en ciencias económicas y empresariales
* Fomentar el uso de software libre entre el alumnado

MUCHAS GRACIAS POR VUESTRA ATENCIÓN