



LISBON
**SCHOOL OF
ECONOMICS &
MANAGEMENT**
UNIVERSIDADE DE LISBOA

Carlos J. Costa

Weka

version 2020





- Waikato Environment for Knowledge Analysis
- data mining/machine learning tool
- developed by Department of Computer Science, University of Waikato, New Zealand
- Written in JAVA
- <https://www.cs.waikato.ac.nz/ml/weka/>



• Download:

Carlos J. Costa (ISEG)

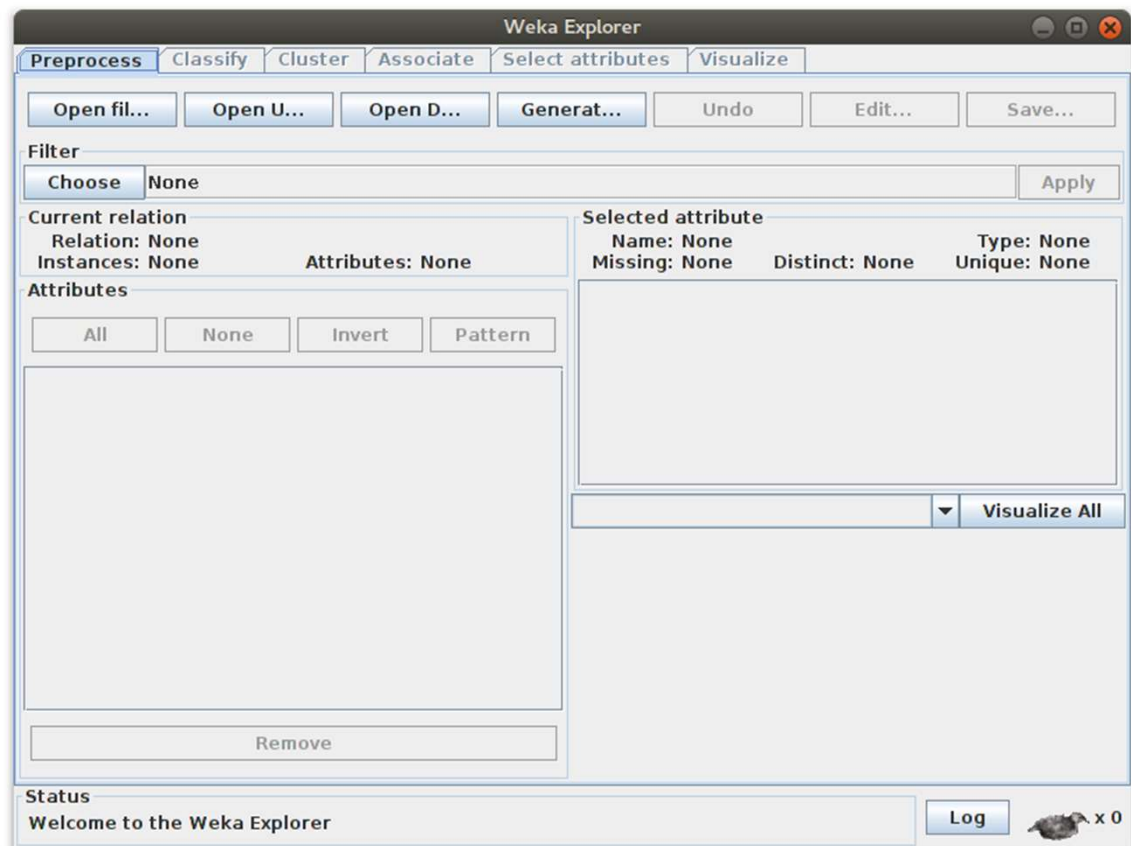
<https://sourceforge.net/projects/weka/>



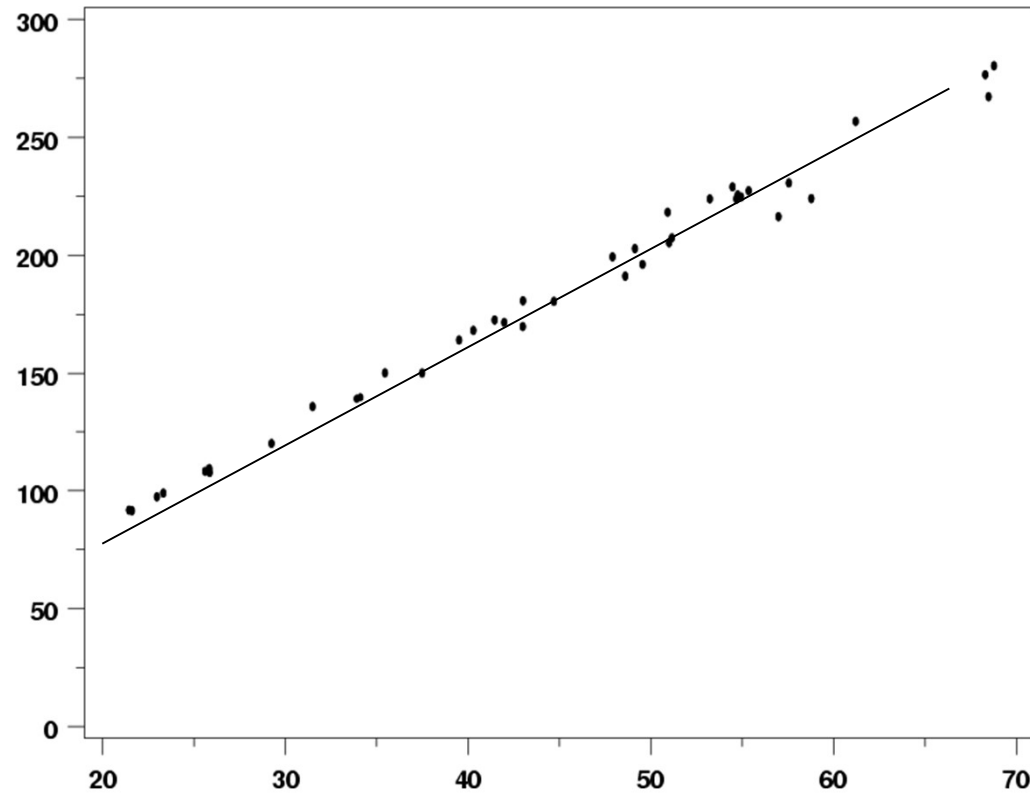
- 2006, Pentaho Acquired Weka Project
- 2015, Pentaho was acquired by Hitachi Data Systems
- 2017, Pentaho is incorporated in Hitachi Vantara



- Preprocess
- Classify
- Cluster
- Associate
- Select Attributes
- Visualize



Regression



Regression

$$\text{HDI} = a_1 * \text{GNI} + a_2 * \text{LEB} + a_3 * \text{MYS}$$

HDI – Human Development Index

GNI – GNI per Capita

LEB – Life Expectancy at Birth

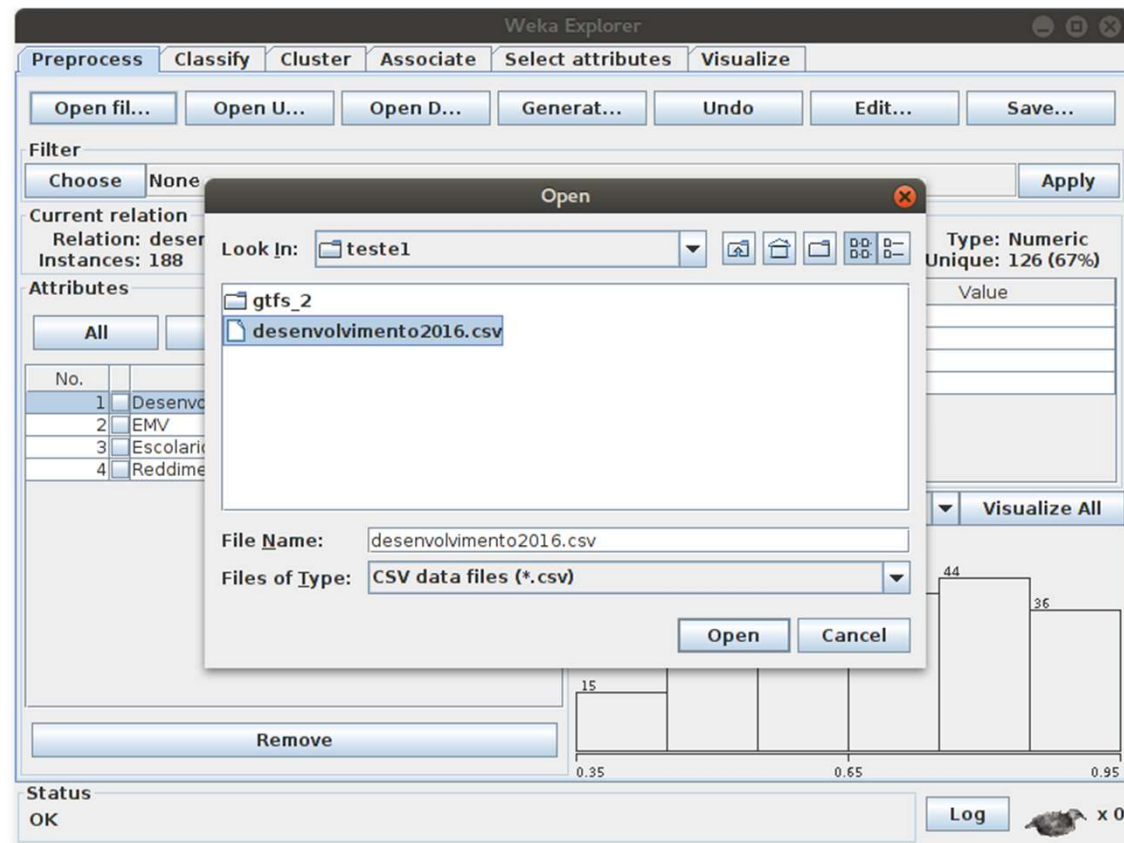
MYS – Mean years of Schooling



Desenvolvimento	EMV	Escolaridade	Rendimento Nacional Bruto
0.949	81.7	12.7	67614
0.939	82.5	13.2	42822
0.939	83.1	13.4	56364
0.926	81.1	13.2	45000
0.925	80.4	12.7	44519
0.925	83.2	11.6	78162
0.924	81.7	11.9	46326
0.923	81.1	12.3	43798
0.921	82.7	12.2	37065
0.920	82.2	13.1	42582
0.920	79.2	13.2	53245
0.917	84.2	11.6	54265
0.915	82.0	12.5	32870
0.913	82.3	12.3	46251
0.912	80.2	12.4	75065
0.909	80.8	13.3	37931
0.903	83.7	12.5	37268
0.901	82.1	12.2	34541
0.899	82.6	12.8	31215



Regression

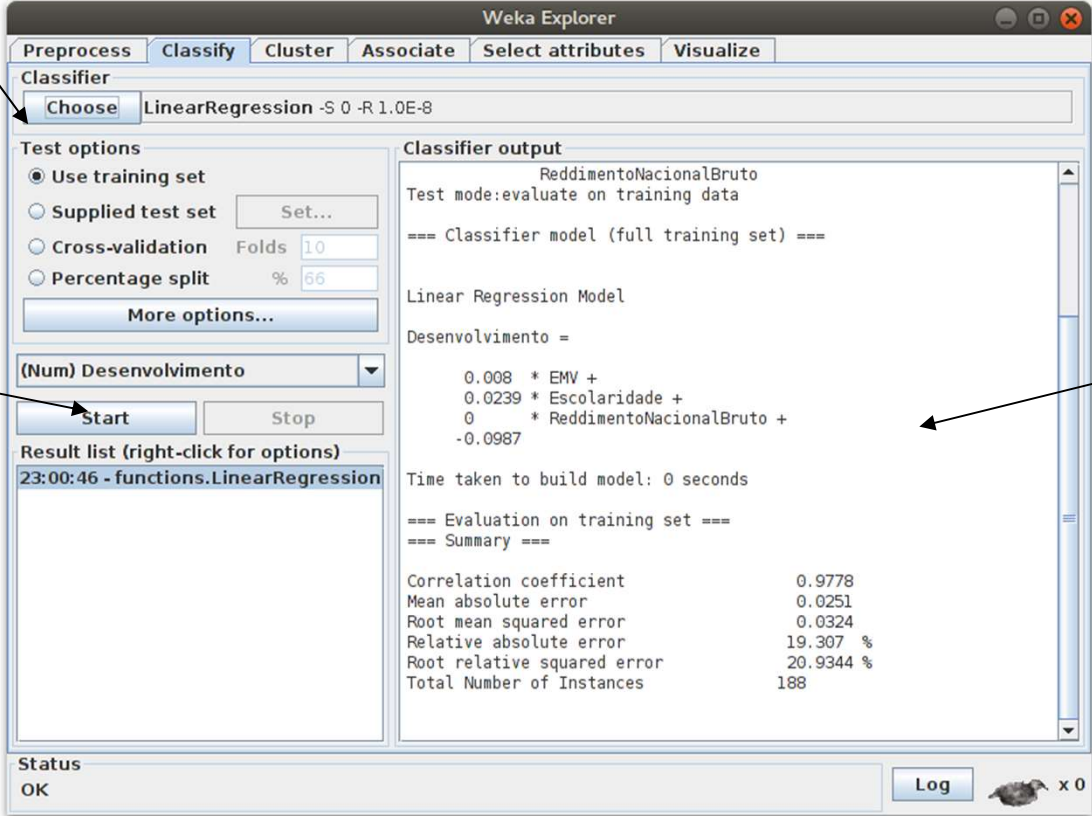


Regression

Choose/Weka/classifiers/Functions/LinearRegression

Start to Run

Results



The screenshot shows the Weka Explorer window with the 'Classify' tab selected. The 'Classifier' dropdown is set to 'LinearRegression -S 0 -R 1.0E-8'. The 'Test options' section has 'Use training set' selected. The 'Classifier output' pane displays the following results:

```
RedditoNacionalBruto
Test mode: evaluate on training data

=== Classifier model (full training set) ===

Linear Regression Model

Desenvolvimento =

0.008 * EMV +
0.0239 * Escolaridade +
0 * RedditoNacionalBruto +
-0.0987

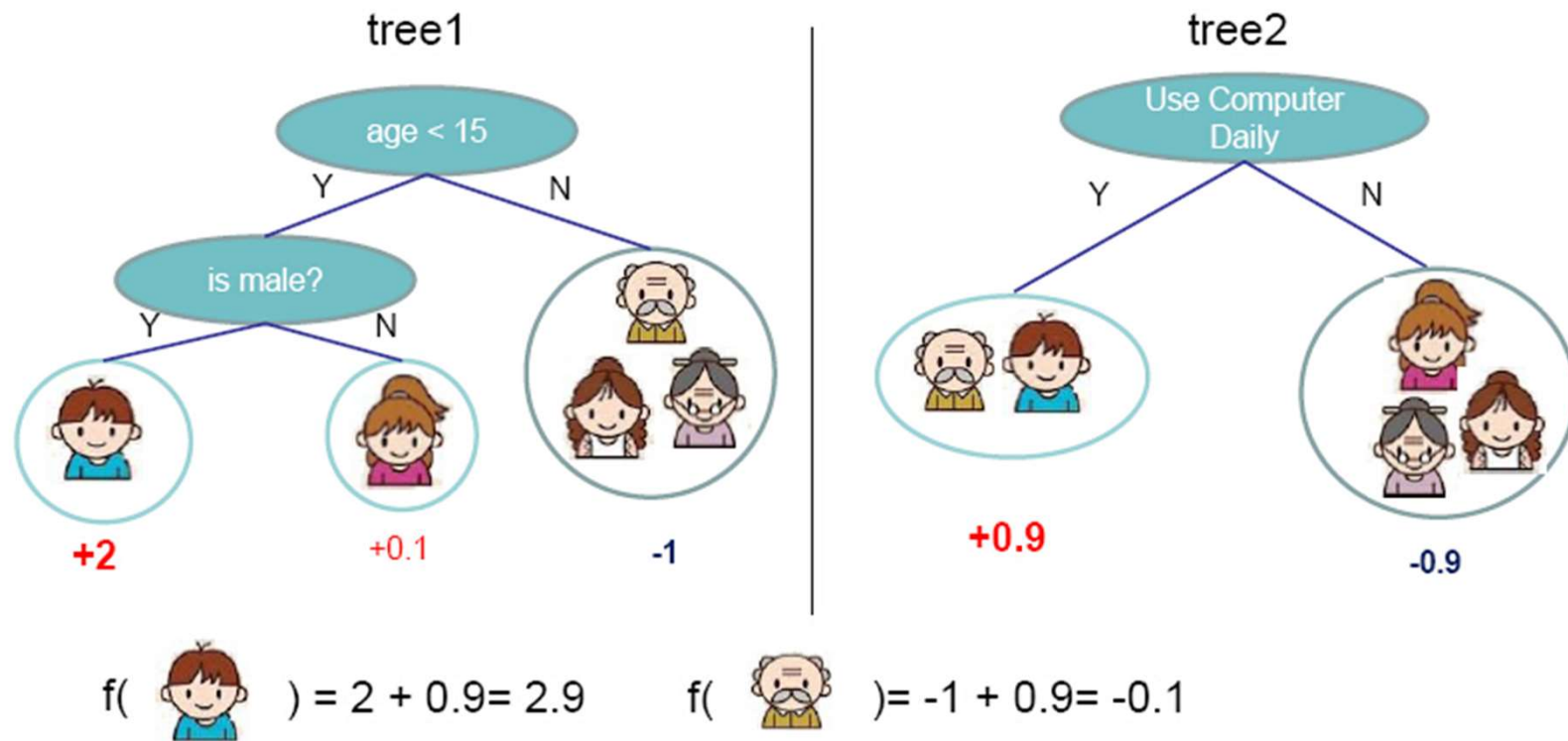
Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===

Correlation coefficient      0.9778
Mean absolute error         0.0251
Root mean squared error     0.0324
Relative absolute error     19.307 %
Root relative squared error  20.9344 %
Total Number of Instances   188
```

Status: OK

Classification



Classification



CountryOfOrigin, BigStar, Genre, Result

USA, yes, scifi, Success

USA, no, comedy, Failure

USA, yes, comedy, Success

Europe, no, comedy, Success

Europe, yes, scifi, Failure

Europe, yes, romance, Failure

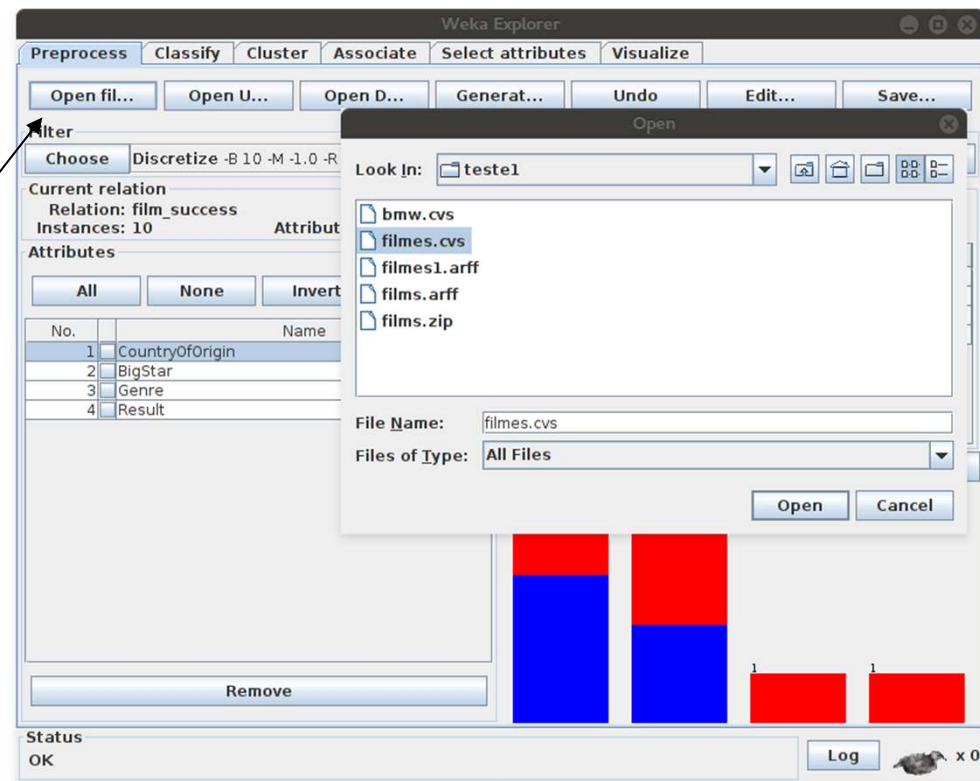


Australia, yes, comedy, Failure

Classification



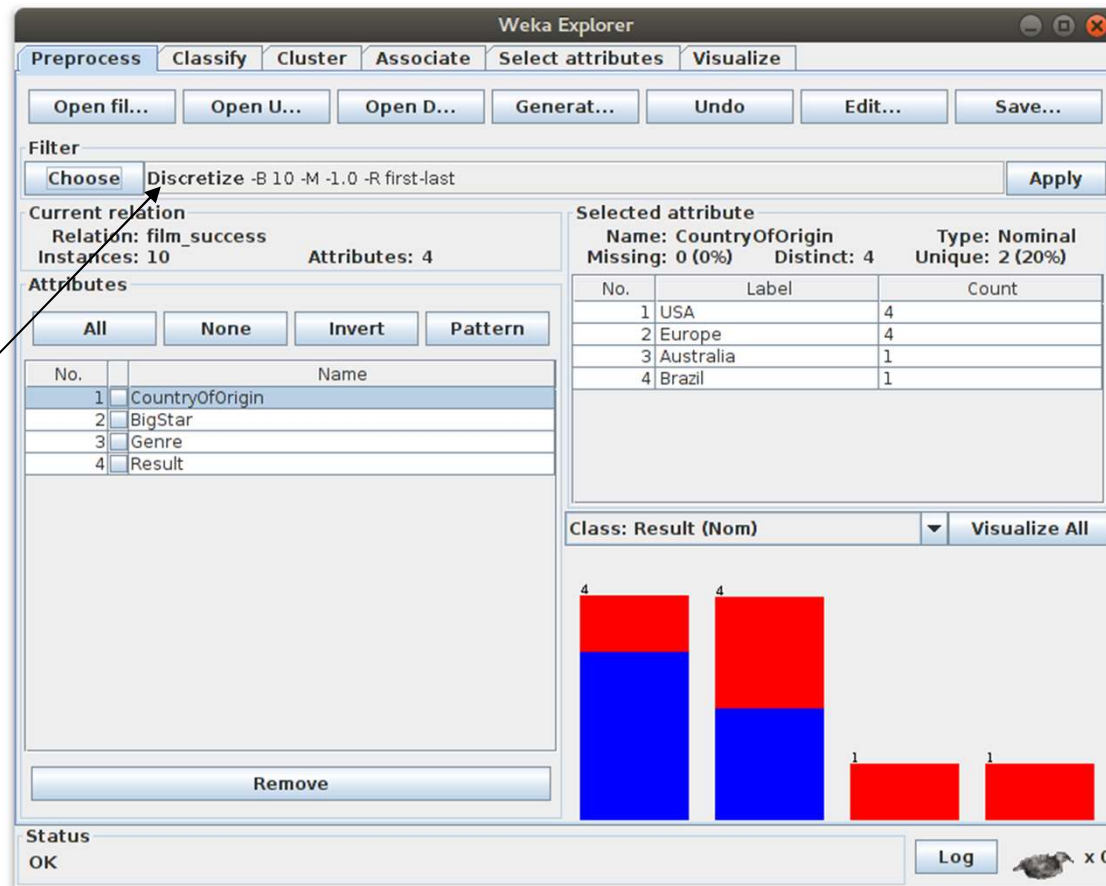
Import Data



Classification



Discretize



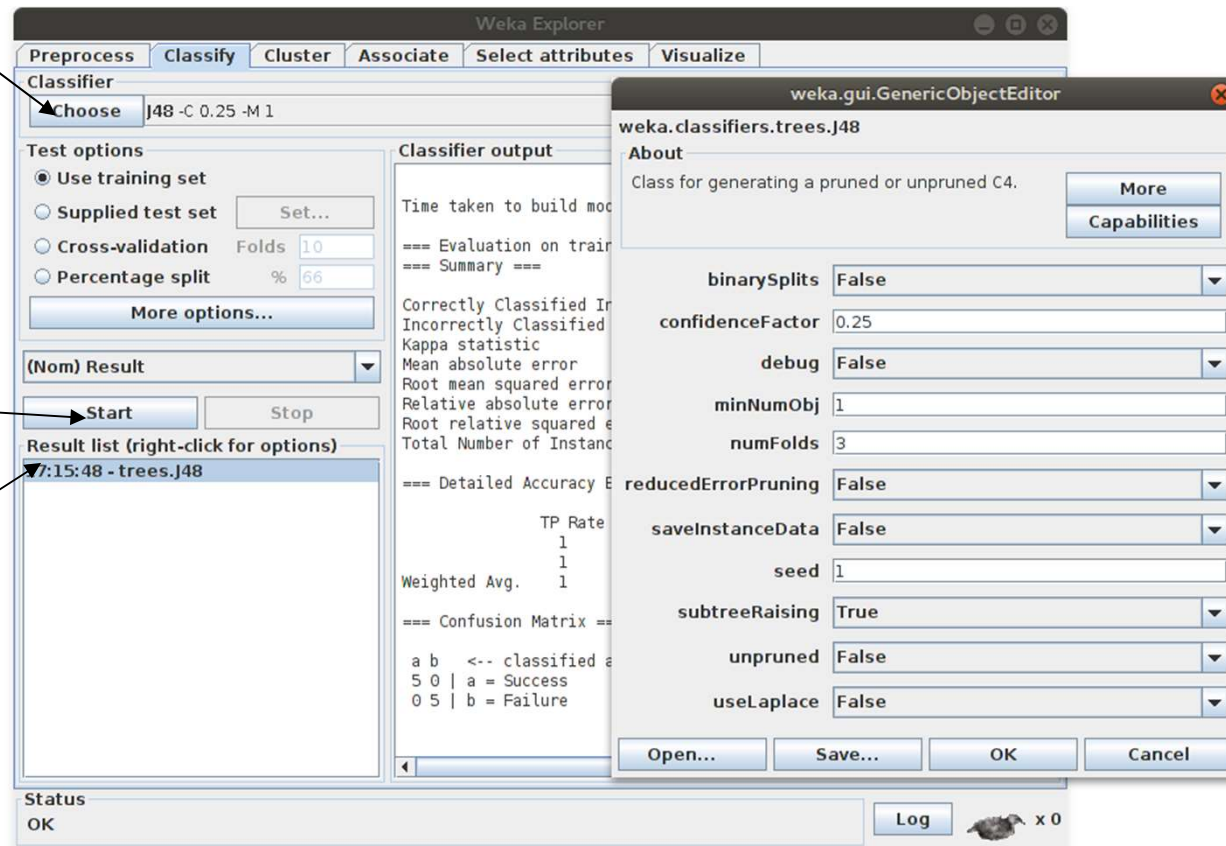
Classification



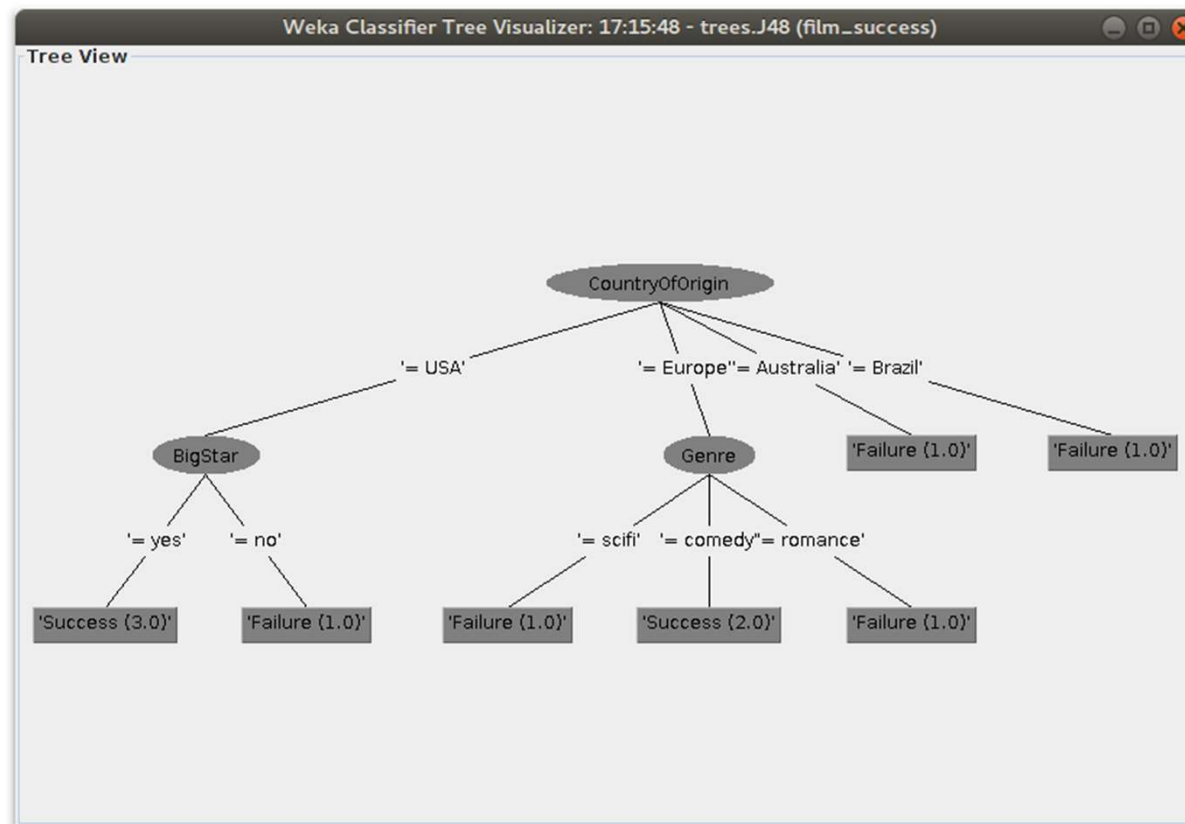
Choose/Weka/classifiers/Tree/J48

Start to Run

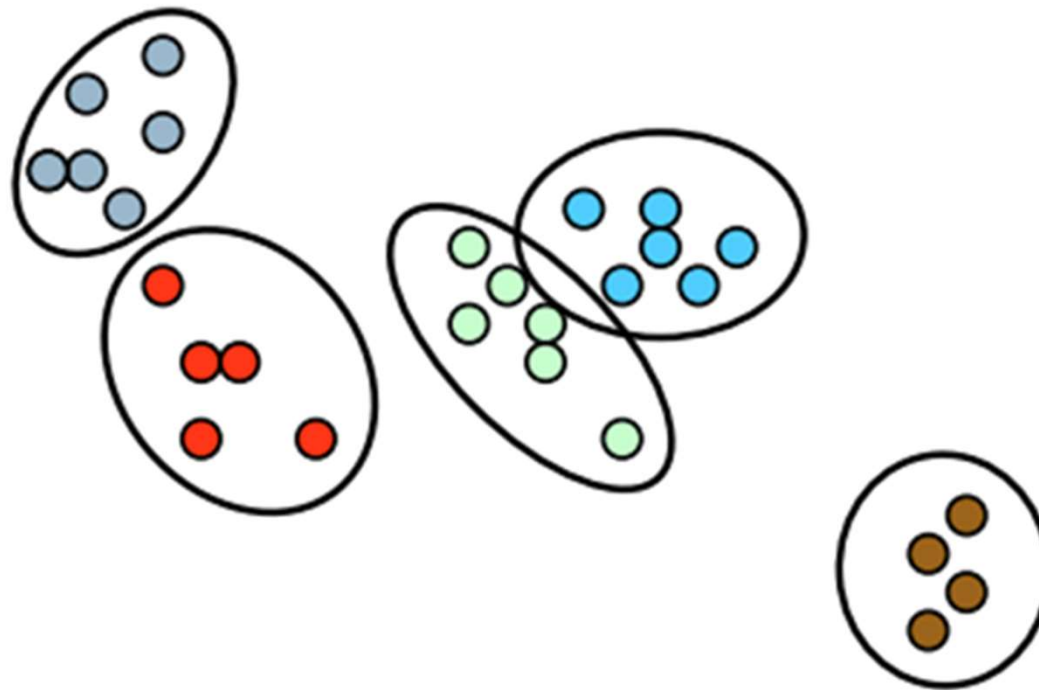
Right button of Mouse
to visualization tree



Classification



Clusters

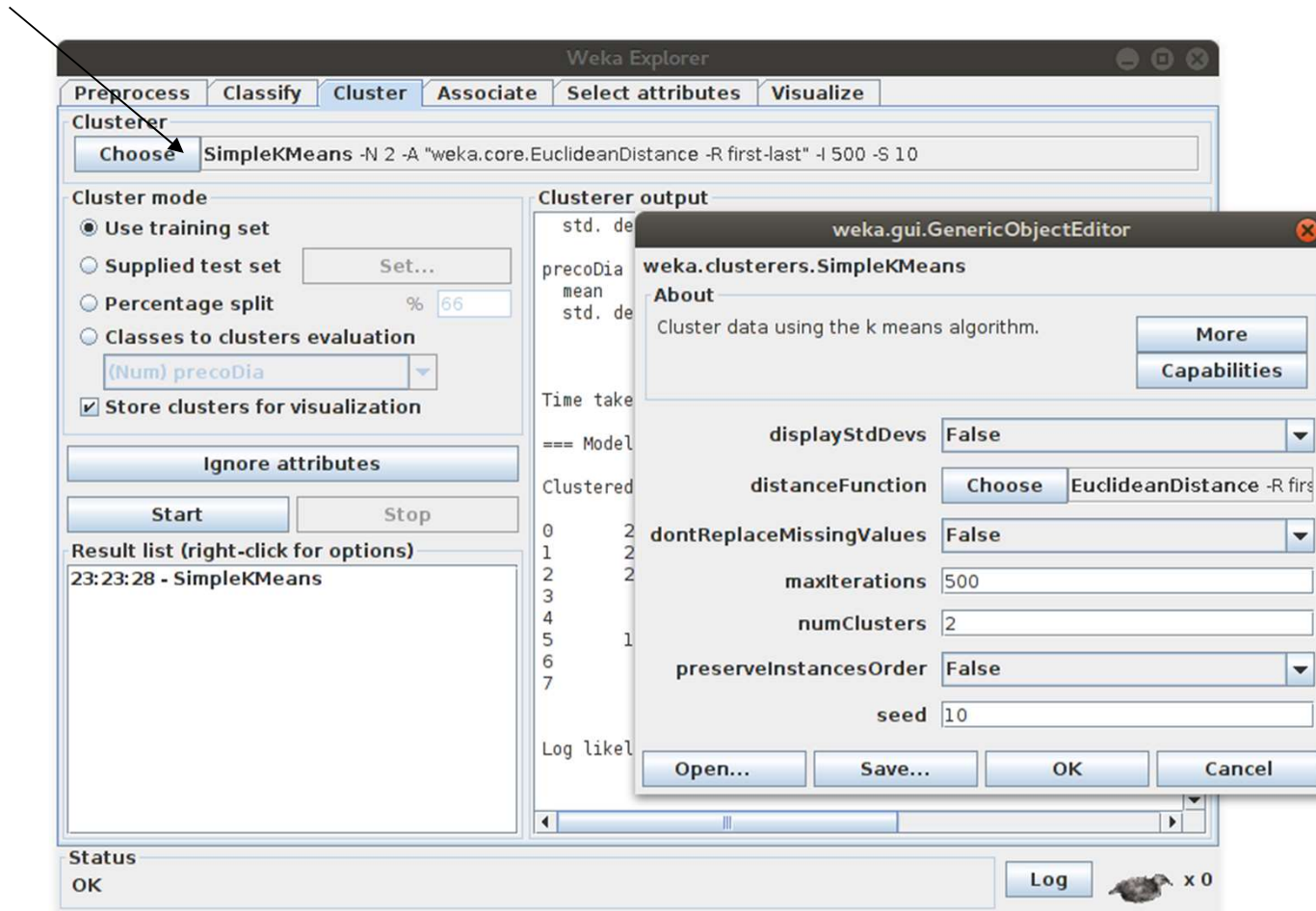


Clusters

data	dias	preco	precoDia
10/04/2016	21	5172.35	246.302
09/30/2016	90	13345	148.278
09/30/2016	60	9470	157.833
09/30/2016	30	11455	381.833
09/30/2016	45	9636.25	214.139
09/29/2016	45	21247.25	472.161
09/29/2016	365	5540	15.178
09/29/2016	55	13205	240.091
09/29/2016	60	8240	137.333

Clusters

Choose/Weka/Clusteres/SimpleKMeans



Clusters

Weka Explorer

Preprocess Classify **Cluster** Associate Select attributes Visualize

Clusterer

Choose EM -I 100 -N -1 -M 1.0E-6 -S 100

Cluster mode

- ☒ Use training set
- ☐ Supplied test set Set...
- ☐ Percentage split % 66
- ☐ Classes to clusters evaluation (Num) precoDia
- ☒ Store clusters for visualization

Ignore attributes

Start Stop

Result list (right-click for options)

23:37:12 - SimpleKMeans

Clusterer output

Number of iterations: 6
Within cluster sum of squared errors: 1069.526871172735
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (1106)	Cluster# 0 (161)	Cluster# 1 (945)
data	11/15/2011	11/15/2011	04/28/2010
dias	78.6121	320.2609	37.4423
preco	41299.8694	173671.5405	18747.6588
precoDia	498.8112	704.0244	463.8489

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

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

Clustered Instances

Cluster	Instances	Percentage
0	161	(15%)
1	945	(85%)

Status OK Log x 0

Clusters

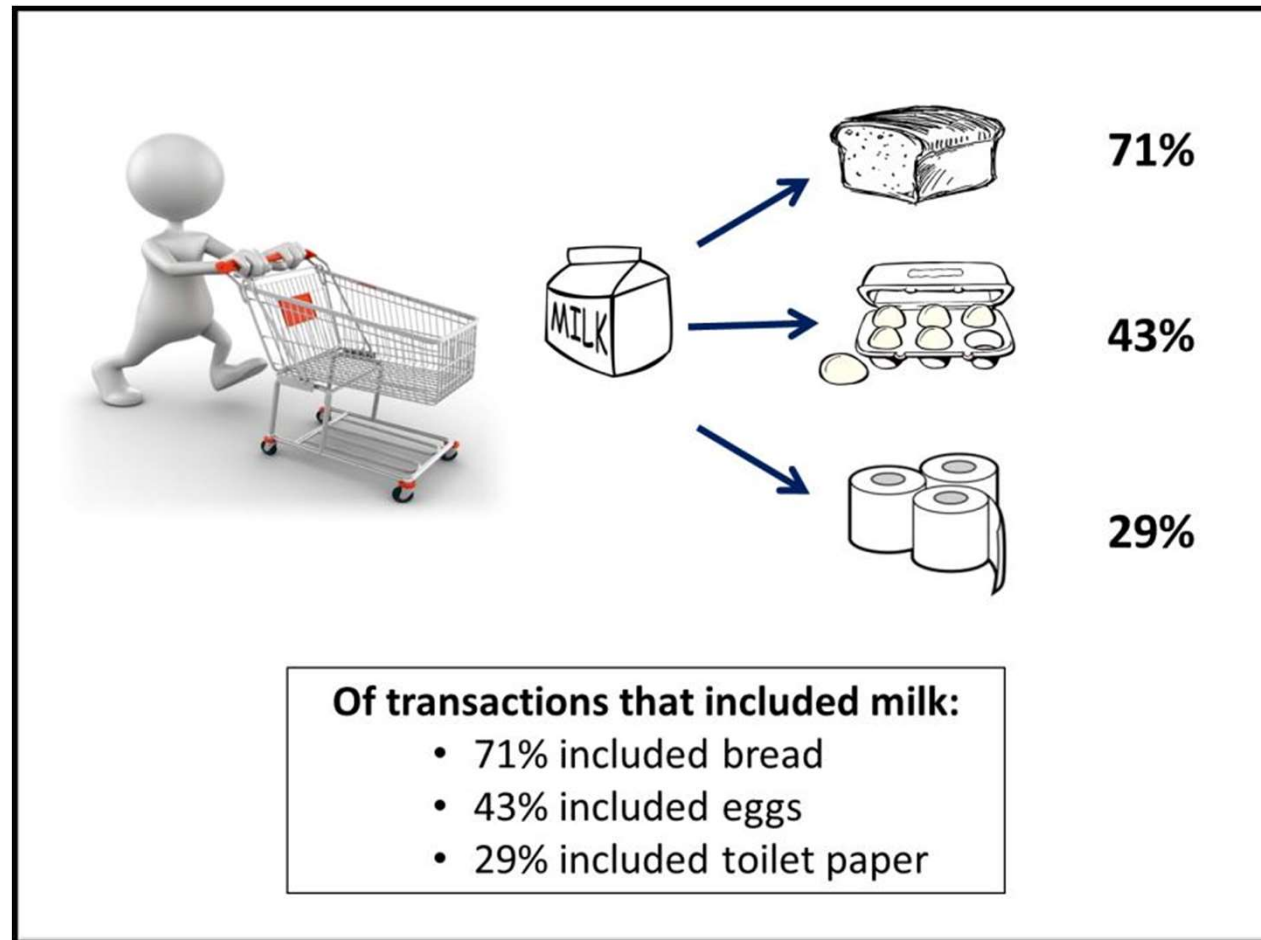
Number of Clusters?

Graphic Analysis

Choose EM Clusterer



Association Rules



Association Rules

id	age	sex	region	income	married	children	car	save_act	current_act	mortgage	pep
ID12101	48	FEMALE	INNER_CITY	17546	NO	1	NO	NO	NO	NO	YES
ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO	YES	YES	NO
ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES	YES	NO	NO
ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO	YES	NO	NO
ID12105	57	FEMALE	RURAL	50576.3	YES	0	NO	YES	NO	NO	NO
ID12106	57	FEMALE	TOWN	37869.6	YES	2	NO	YES	YES	NO	YES
ID12107	22	MALE	RURAL	8877.07	NO	0	NO	NO	YES	NO	YES
ID12108	58	MALE	TOWN	24946.6	YES	0	YES	YES	YES	NO	NO
ID12109	37	FEMALE	SUBURBAN	25304.3	YES	2	YES	NO	NO	NO	NO
ID12110	54	MALE	TOWN	24212.1	YES	2	YES	YES	YES	NO	NO
ID12111	66	FEMALE	TOWN	59803.9	YES	0	NO	YES	YES	NO	NO

Association Rules

Discretize

The screenshot shows the Weka Explorer interface with the 'Preprocess' tab selected. The 'Filter' section shows the 'Discretize' filter applied to the 'current_act' attribute. The 'Attributes' list on the left shows 'current_act' selected. The 'Selected attribute' section on the right shows the distribution of 'current_act' with two labels: 'NO' (145 instances) and 'YES' (455 instances). The 'Class' is set to 'pep (Nom)' and the 'Visualize All' button is visible.

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Open fil... Open U... Open D... Generat... Undo Edit... Save...

Filter

Choose Discretize -B 1.0 -M -1.0 -R first-last Apply

Current relation
Relation: bank-data-weka.filters.unsupervised....
Instances: 600 Attributes: 12

Attributes

All None Invert Pattern

No.	Name
1	id
2	age
3	sex
4	region
5	income
6	married
7	children
8	car
9	save_act
10	current_act
11	mortgage
12	pep

Remove

Selected attribute
Name: current_act
Missing: 0 (0%) Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	NO	145
2	YES	455

Class: pep (Nom) Visualize All

Status
OK

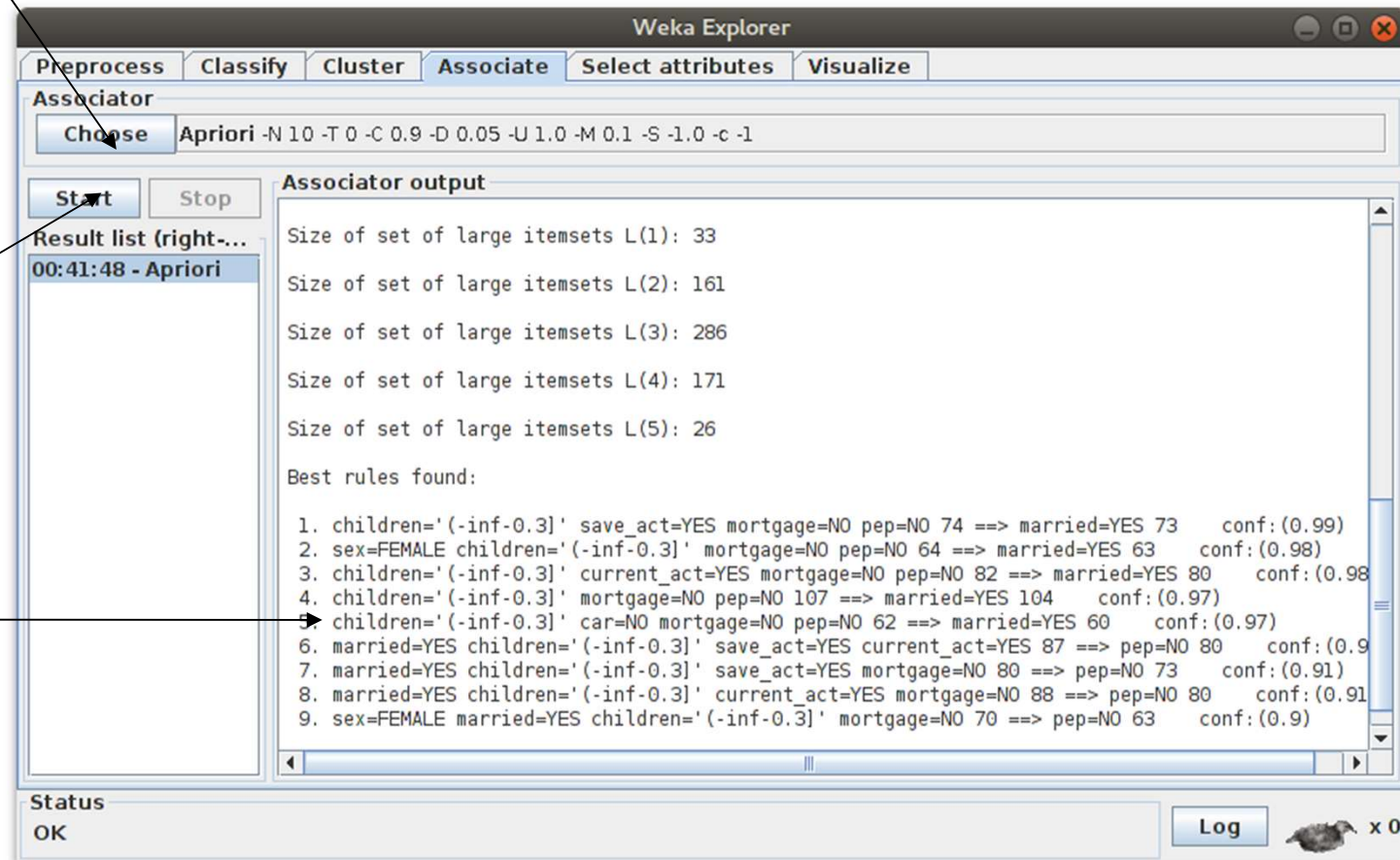
Log x 0

Association Rules

Apriori

Start

Output



Bibliography

Abernethy, M (2010) Classification and clustering. Data mining with WEKA, Part 2. Retrieved December 1, 2017, from <http://www.ibm.com/developerworks/library/os-weka2/index.html>

Abernethy, M (2010) Introduction and regression. Data mining with WEKA, Part 1. Retrieved December 1, 2017, from <https://www.ibm.com/developerworks/opensource/library/os-weka1/index.html>

Weka 3 - Data Mining with Open Source Machine Learning Software in Java. (n.d.). Retrieved December 4, 2017, from <https://www.cs.waikato.ac.nz/ml/weka/>



