Lisbon School of Economics & Management
Universidade de Lisboa

LISBOA
UNIVERSIDADE DE LISBOA

Carlos J. Costa

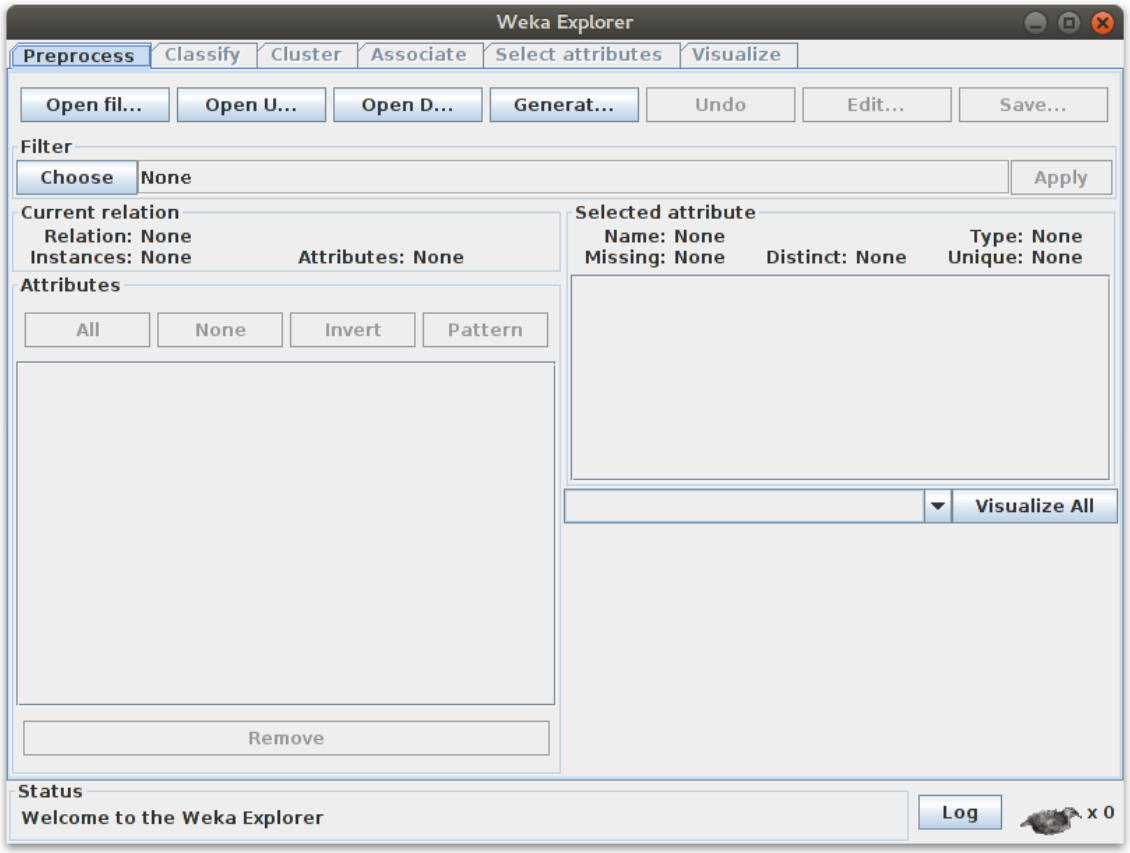
Weka

- 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: https://sourceforge.net/projects/weka/
- Licence GNU GPL
- bird found only in New Zealand

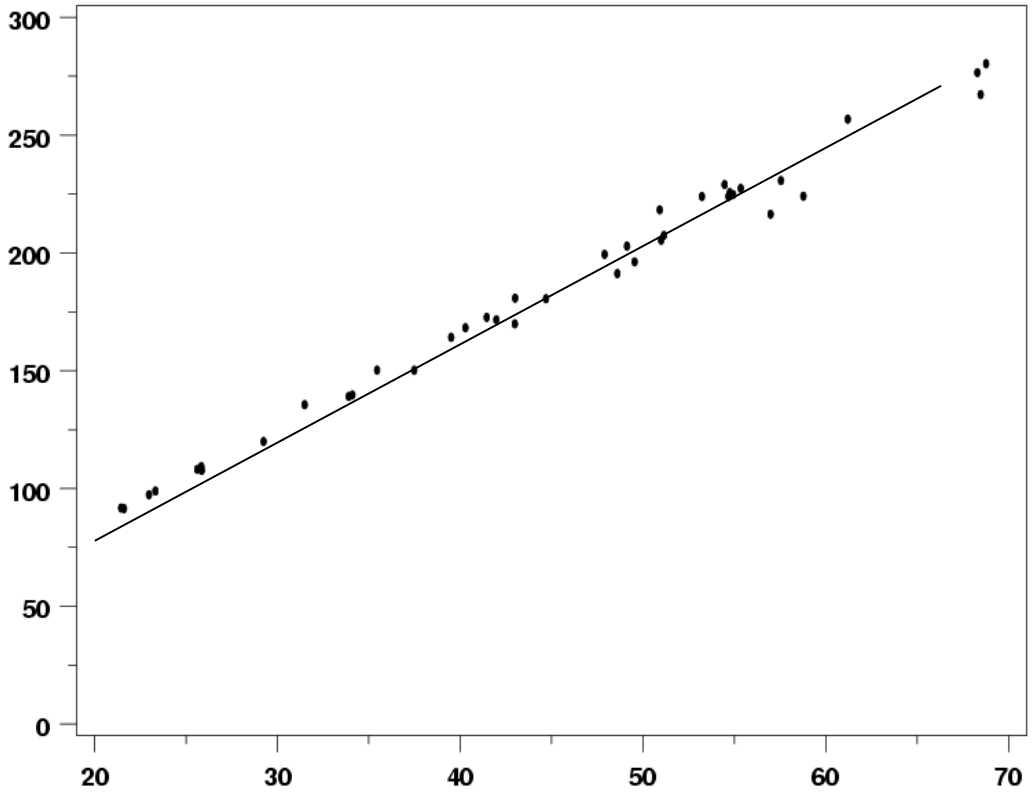
- 2006, Pentaho Acquired Weka Project
- 2015, Pentaho was acquired by Hitachi Data Systems
- 2017, Pentaho is incorporated in Hitachi Vantara


Weka GUI Chooser
Program
Visualization
Tools
Help
WEKA
The University of Waikato
Waikato Environment for Knowledge Analysis
Version 3.6.13
(c) 1999 - 2015
The University of Waikato
Hamilton, New Zealand
Applications
Explorer
Experimenter
KnowledgeFlow
Simple CLI

- Preprocess
- Classify
- Cluster
- Associate
- Select Attribute
- Visualize



Regression



Regression

HDI=a1*GNI+a2*LEB+a3*MYS

HDI – Human Development Index

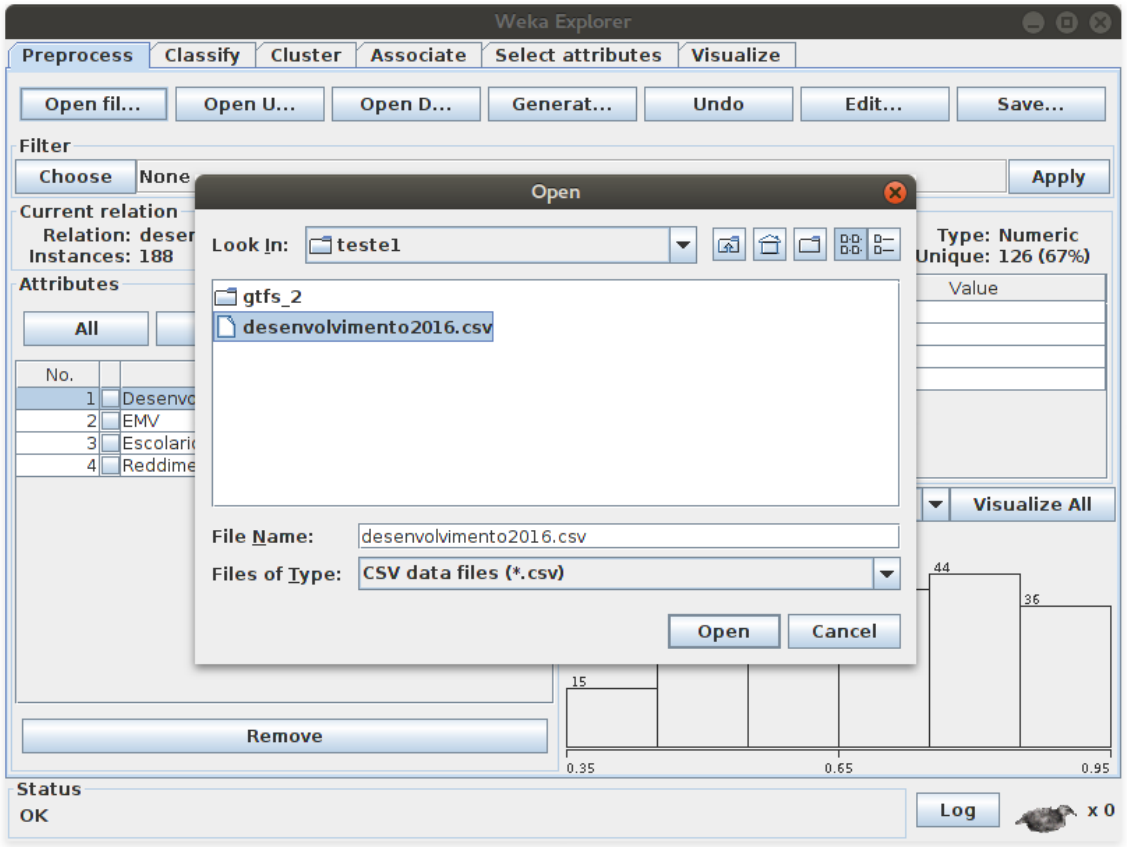
GNI – GNI per Capita

LEB – Life Expectancy at Birth

MYS – Mena years of Schooling

Desenvolvimento	EMV	Escolaridade	Rendimento NacionalBruto
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

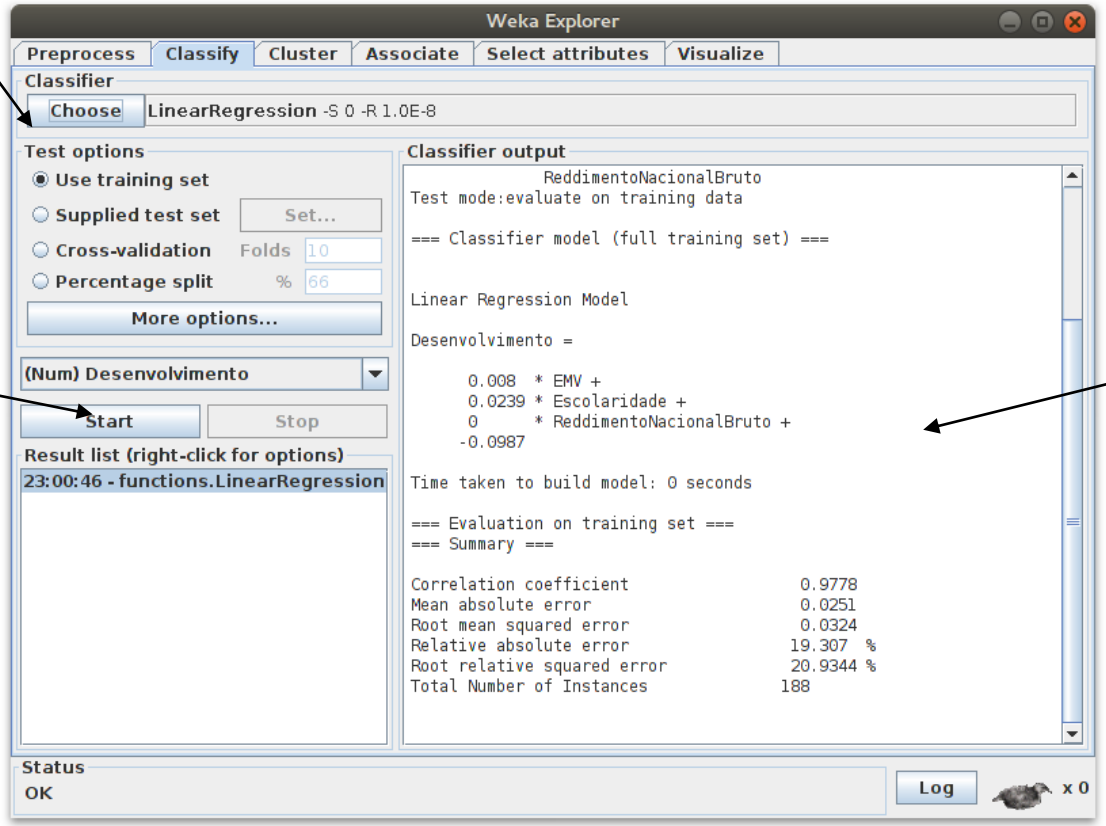

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open fil...
Open U...
Open D...
Generat...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: deser
Instances: 188
Attributes
All
No.
1 Desenvo
2 EMV
3 Escolari
4 Reddime
Remove
Open
Look In:
teste1
gtfs_2
desenvolvimento2016.csv
File Name:
desenvolvimento2016.csv
Files of Type:
CSV data files (*.csv)
Open
Cancel
Type: Numeric
Unique: 126 (67%)
Value
Visualize All
44
36
15
0.35
0.65
0.95
Status
OK
Log
x 0

Regression

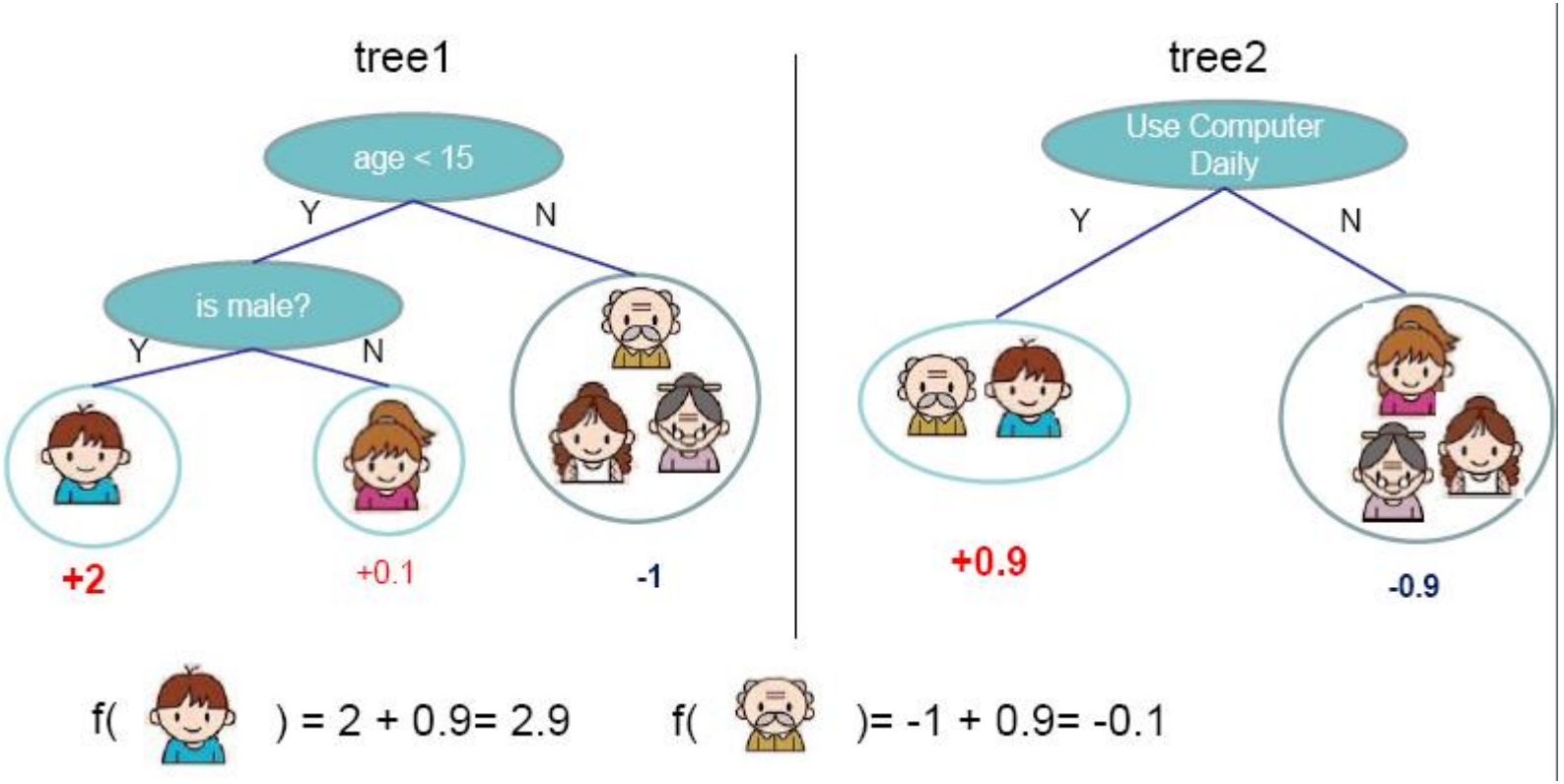
Choose/Weka/classifiers/Functions/LinearRegression

Start to Run

Results



Classification

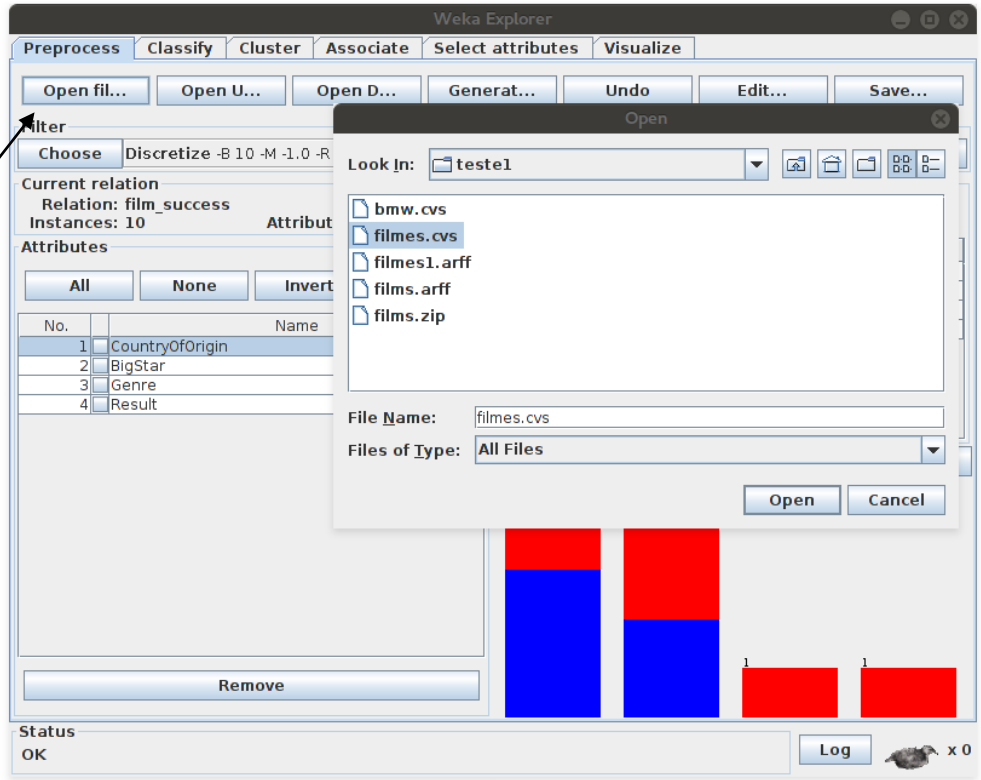


Classification

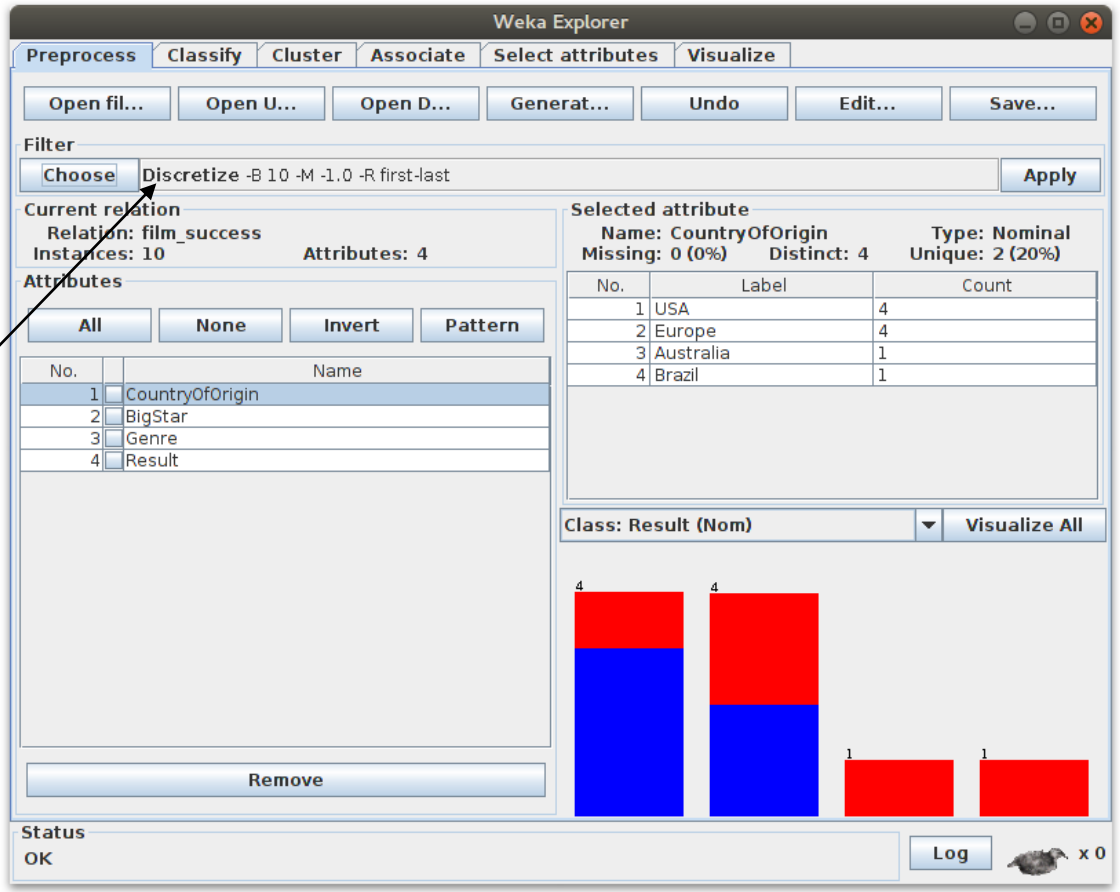

```
CountryOfOrigin,BigStar,Genre,Result
USA,yes,scifi,Success
USA,no,comedy,Failure
USA,yes,comedy,Success
Europe,no,comedy,Success
Europe,yes,scifi,Failure
Europe,yes,romance,Failure
Australia,yes,comedy,Failure
Brazil,no,scifi,Failure
Europe,yes,comedy,Success
USA,yes,comedy,Success
```

Classification

Import Data



Classification



Discretize

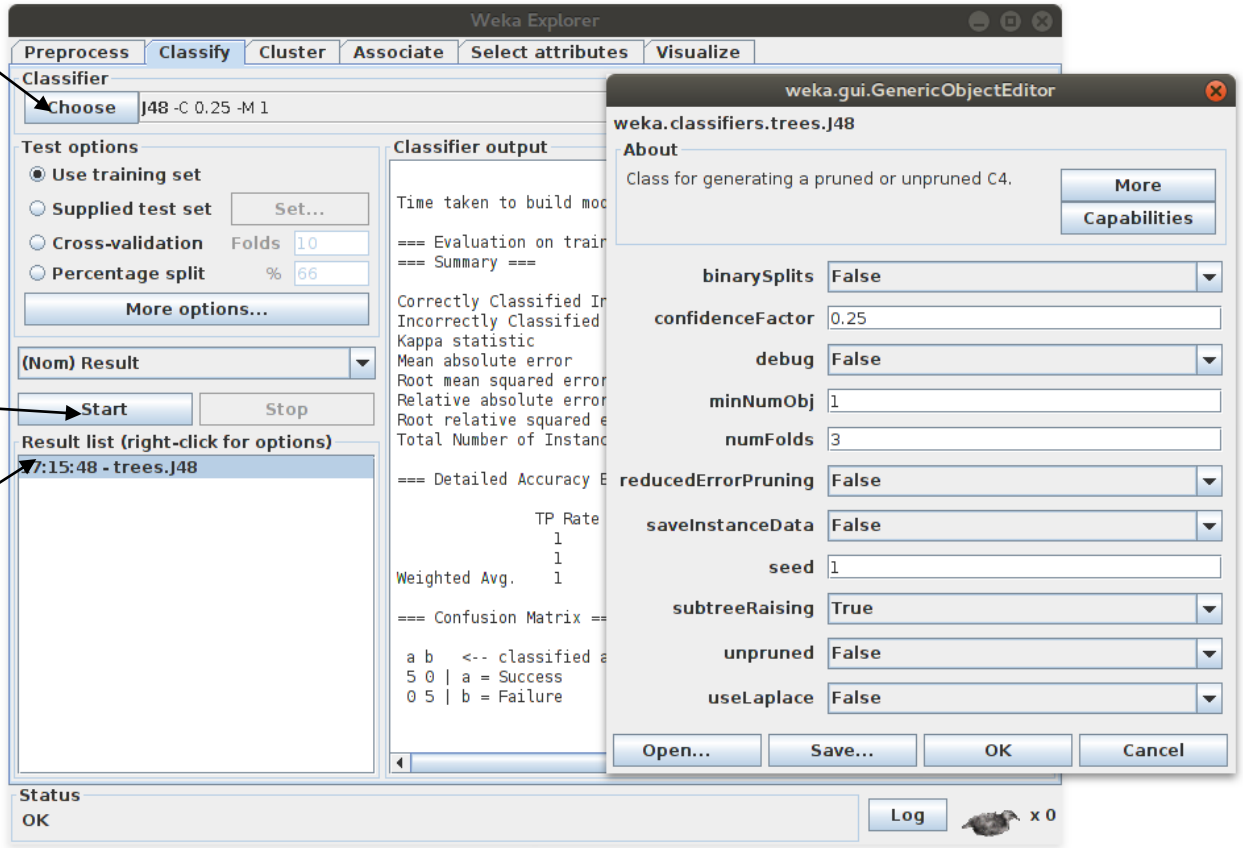
Classification

Choose/Weka/classifiers/Tree/J48

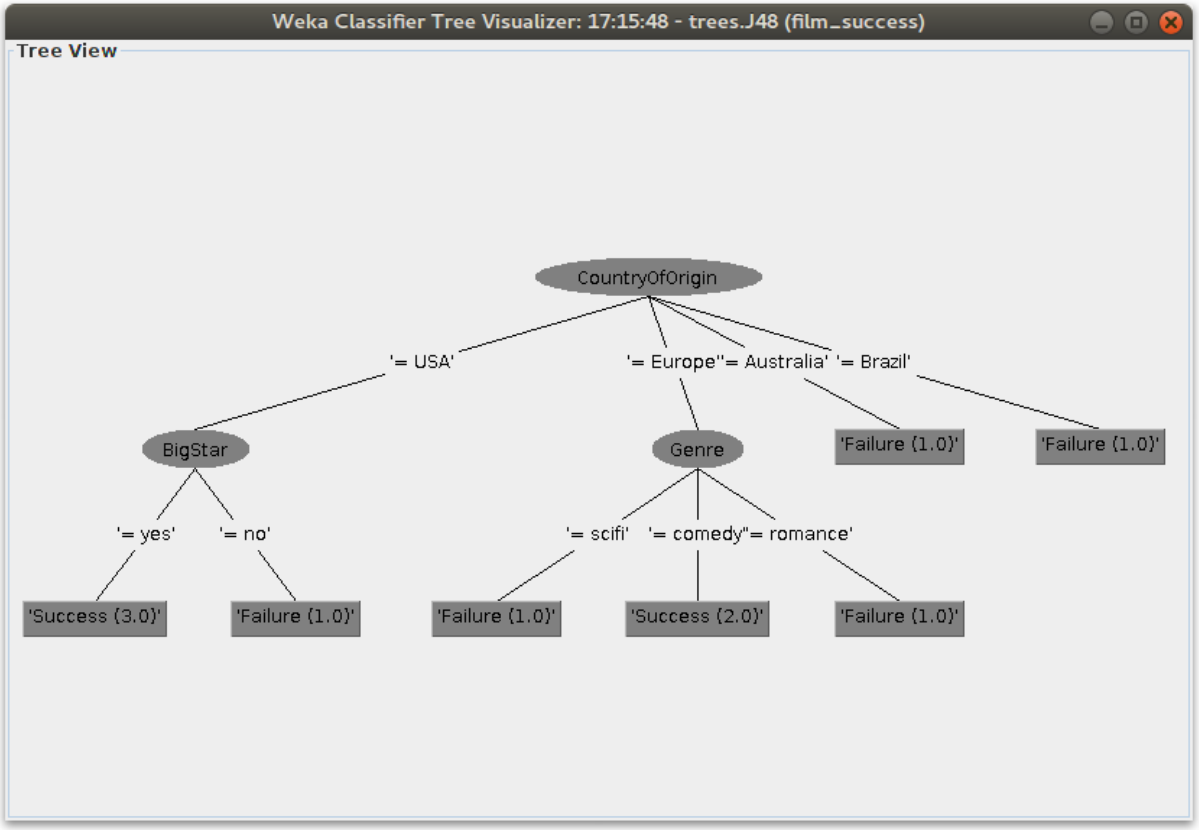
Start to Run

Right button of Mouse

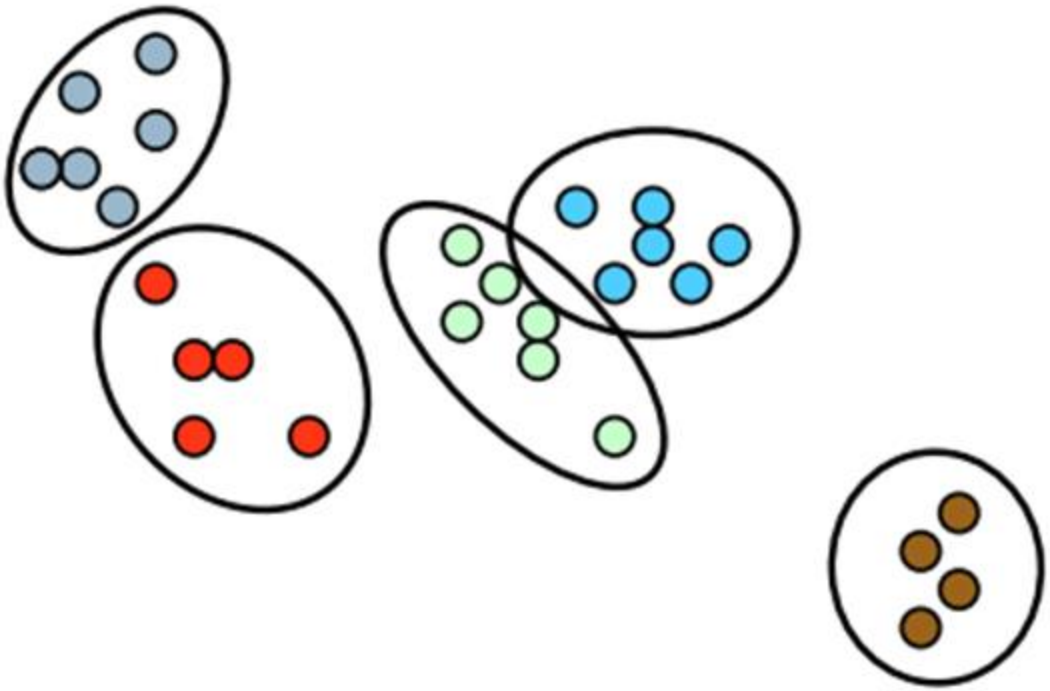
to visualization tree



Classification



Clusters

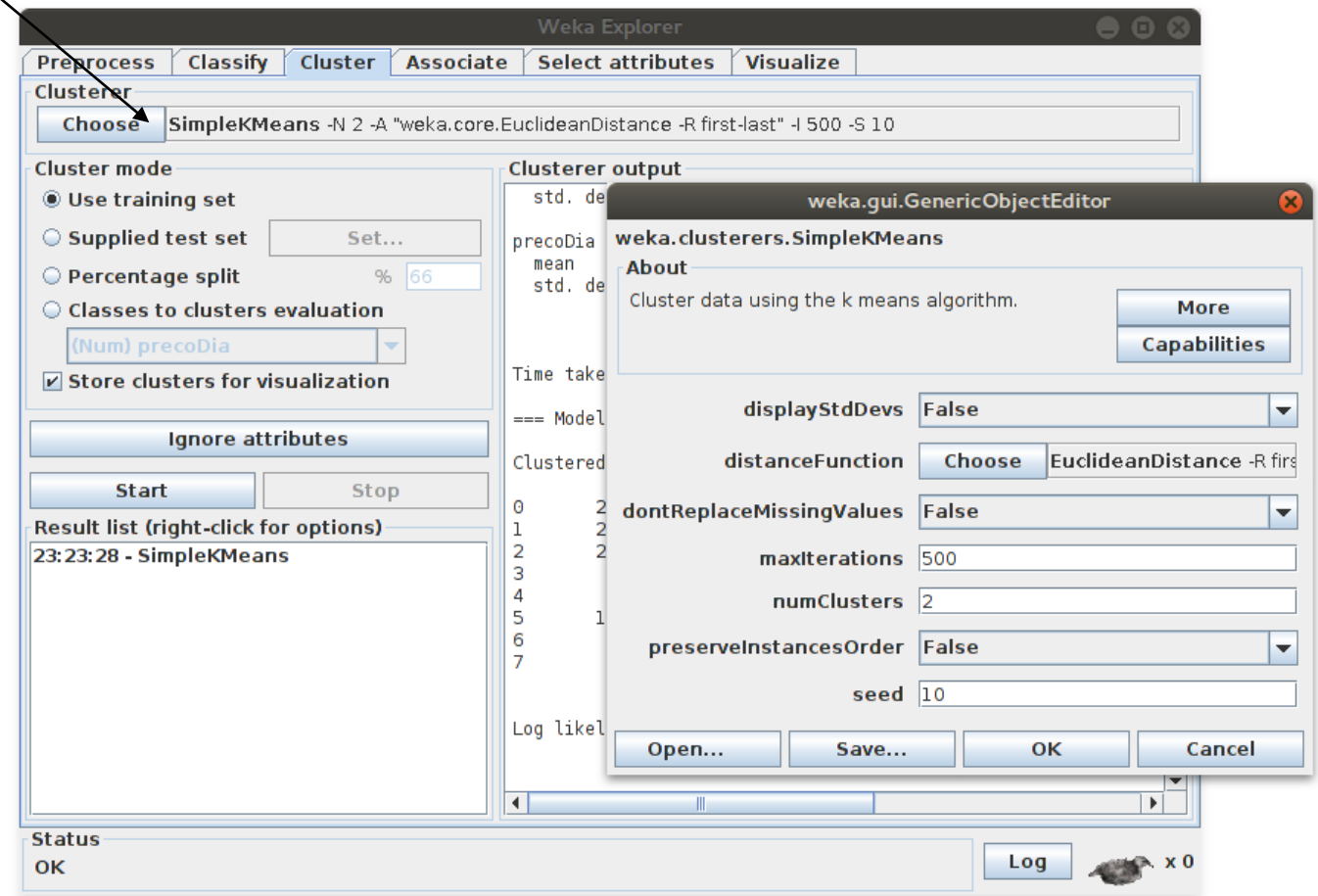


Clusters

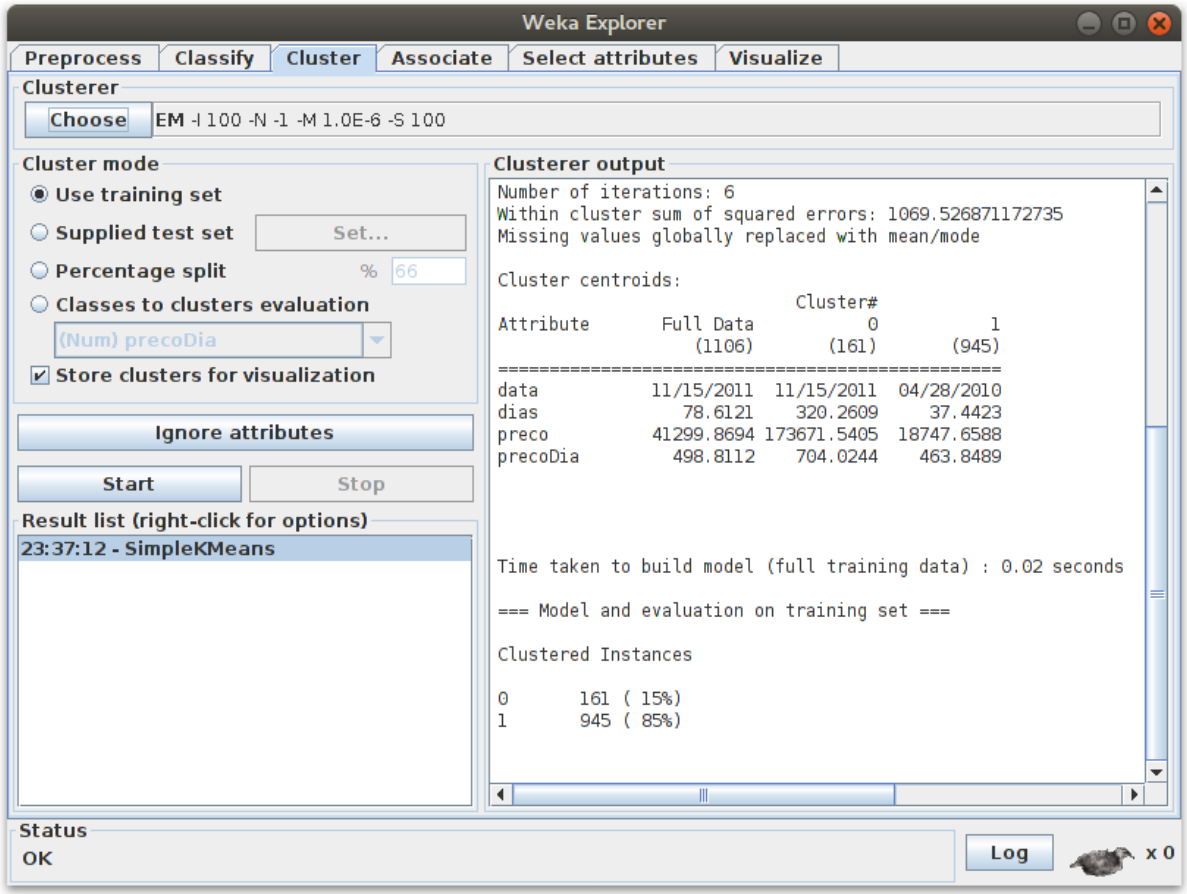
data	dias	preco	precoDia
10/04/2016	21	5172.35	246.302
09/30/2016	90	13345	148.278
09/30/2016	60	9470	157.833
09/30/2016	30	11455	381.833
09/30/2016	45	9636.25	214.139
09/29/2016	45	21247.25	472.161
09/29/2016	365	5540	15.178
09/29/2016	55	13205	240.091
09/29/2016	60	8240	137.333

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:
Cluster#
Attribute Full Data 0 1
(1106) (161) (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
0 161 (15%)
1 945 (85%)
Status
OK
Log
x 0

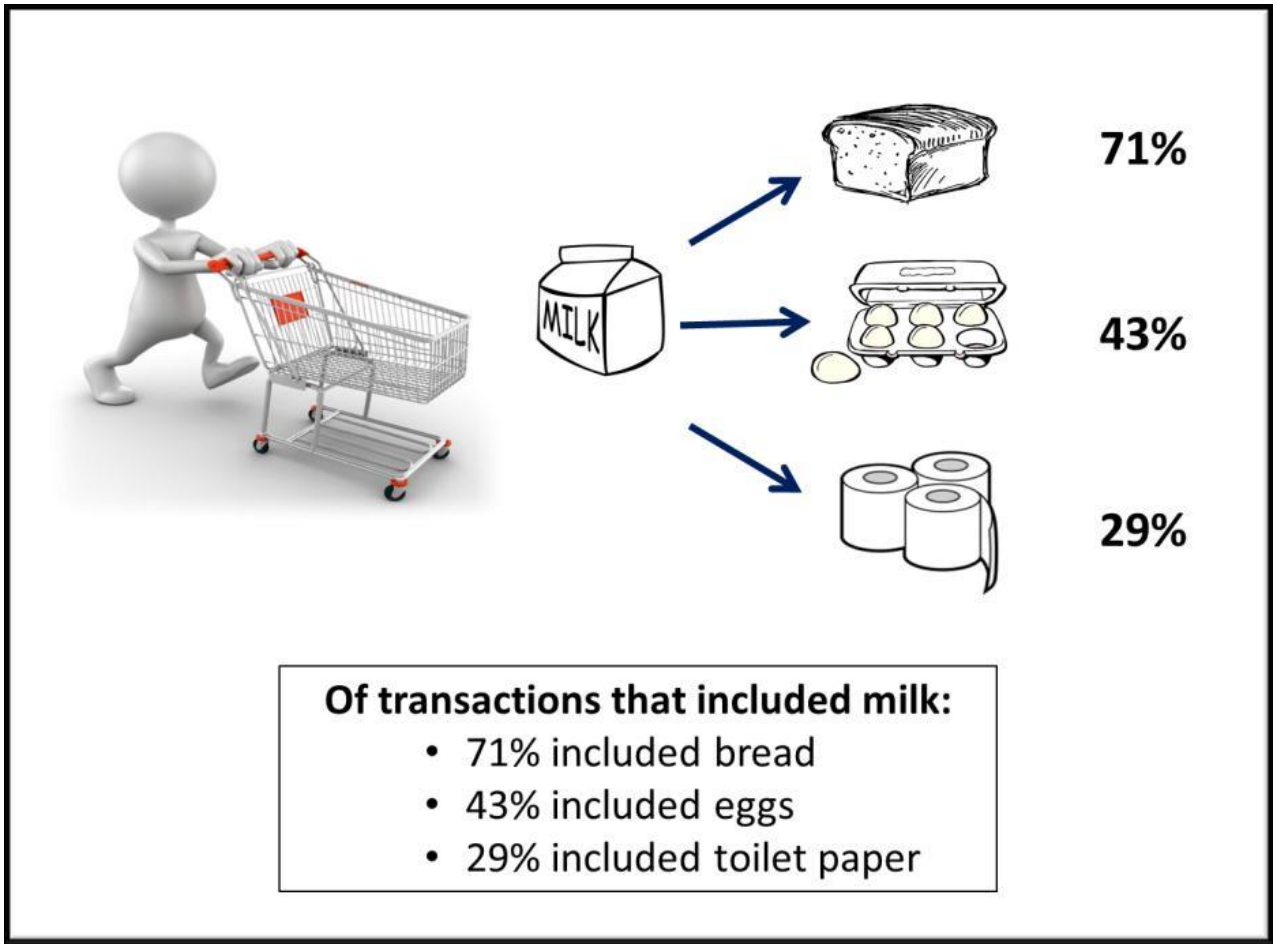
Clusters

Number of Clusters?

Graphic Analysis

Choose EM Clusterer

Association Rules

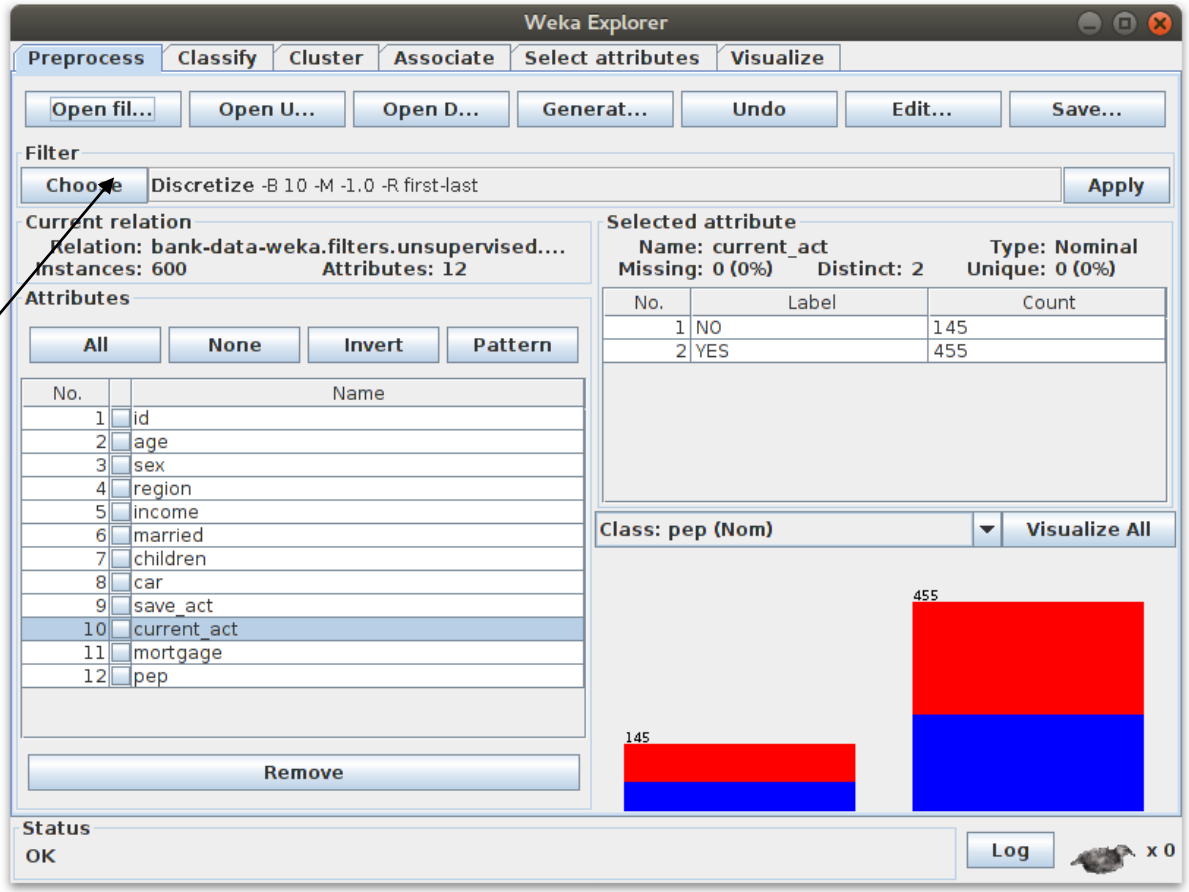


Association Rules

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

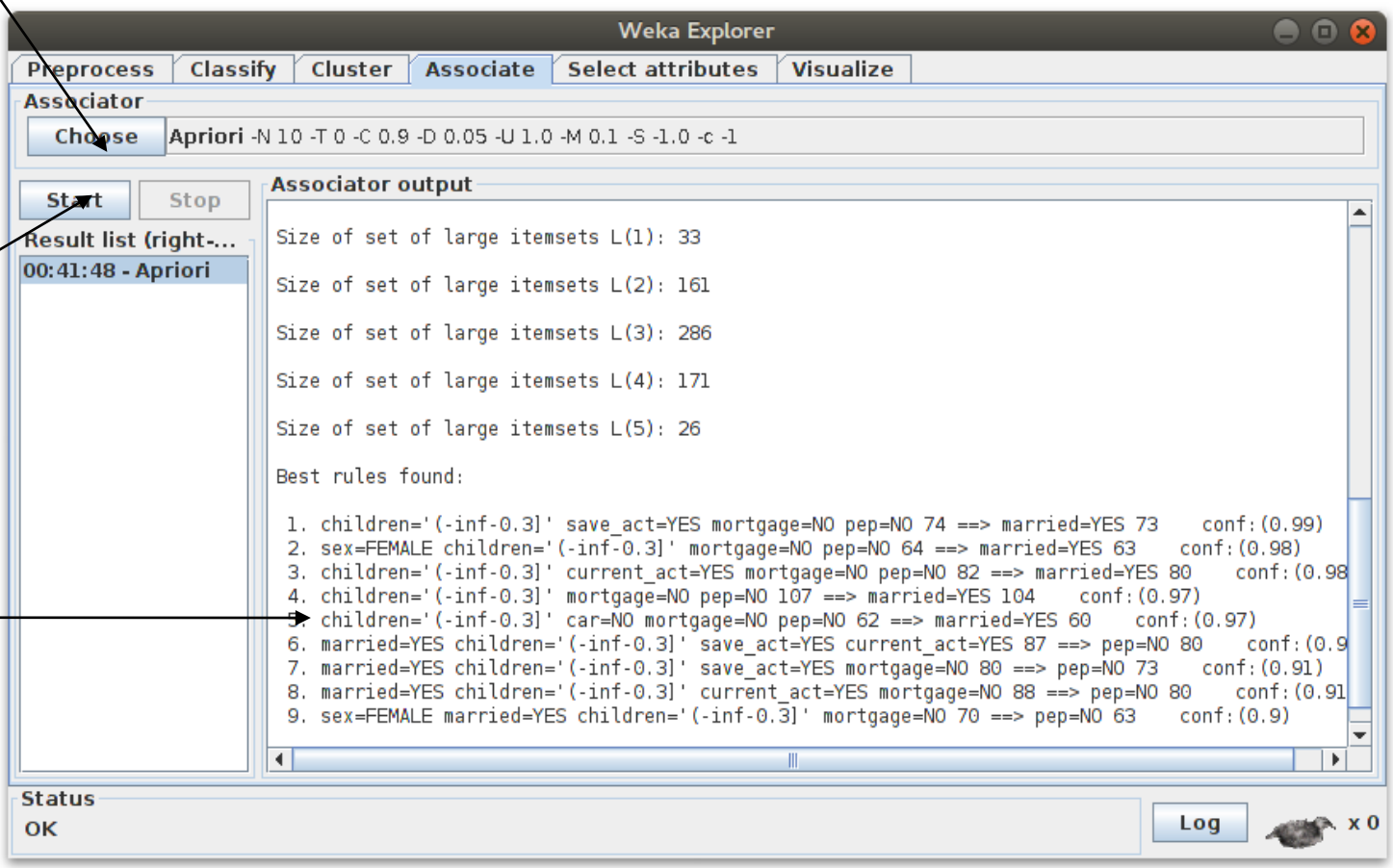


Association Rules

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/