



Lisbon School
of Economics
& Management
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

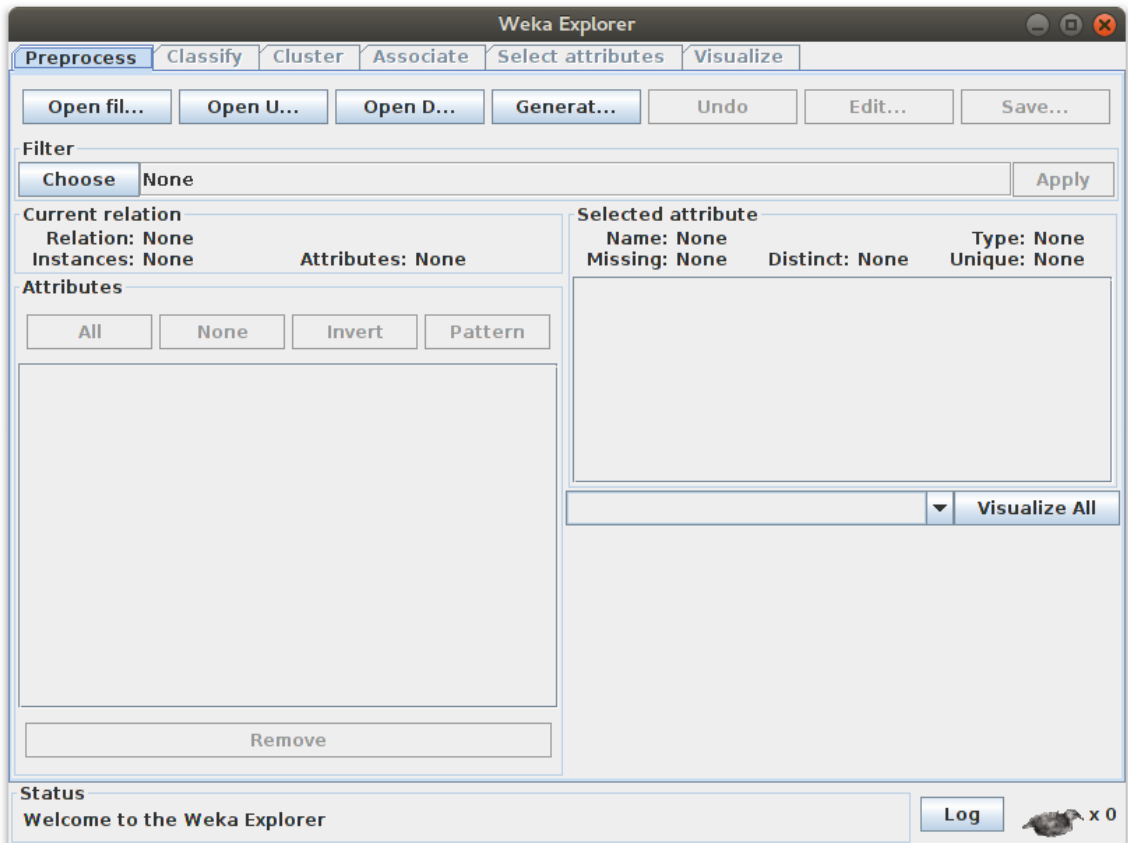


THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

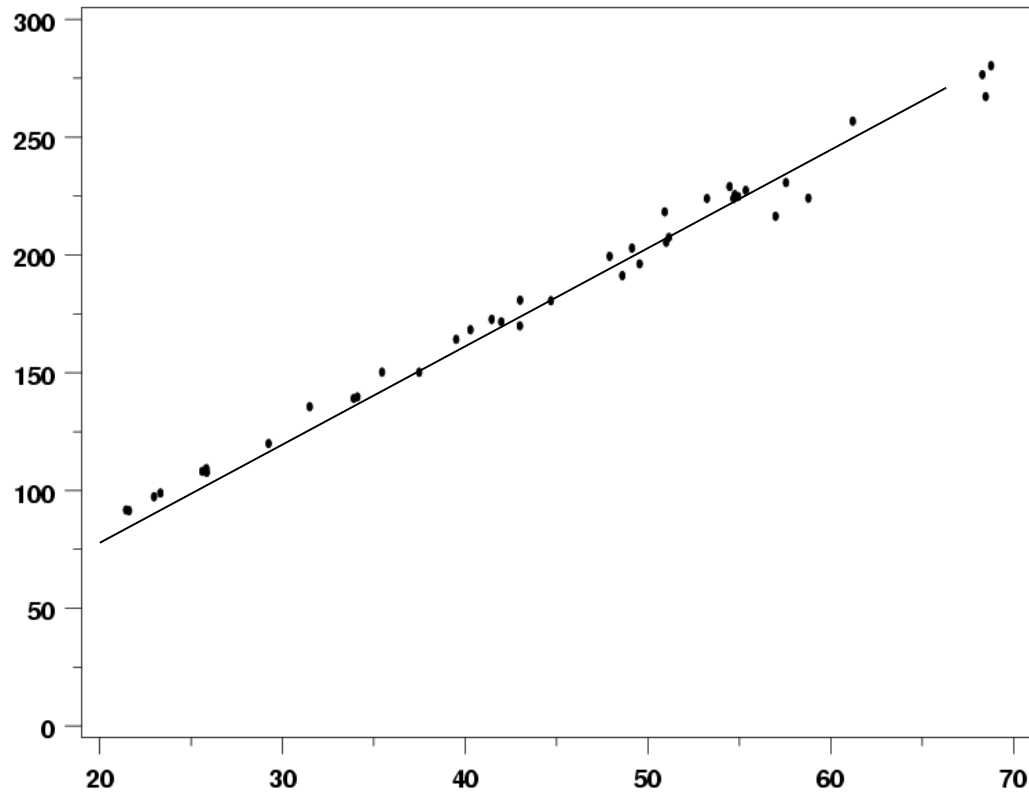
- 2006, Pentaho Acquired Weka Project
- 2015, Pentaho was acquired by Hitachi Data Systems
- 2017, Pentaho is incorporated in Hitachi Vantara



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



Regression



Regression

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

HDI – Human Development Index

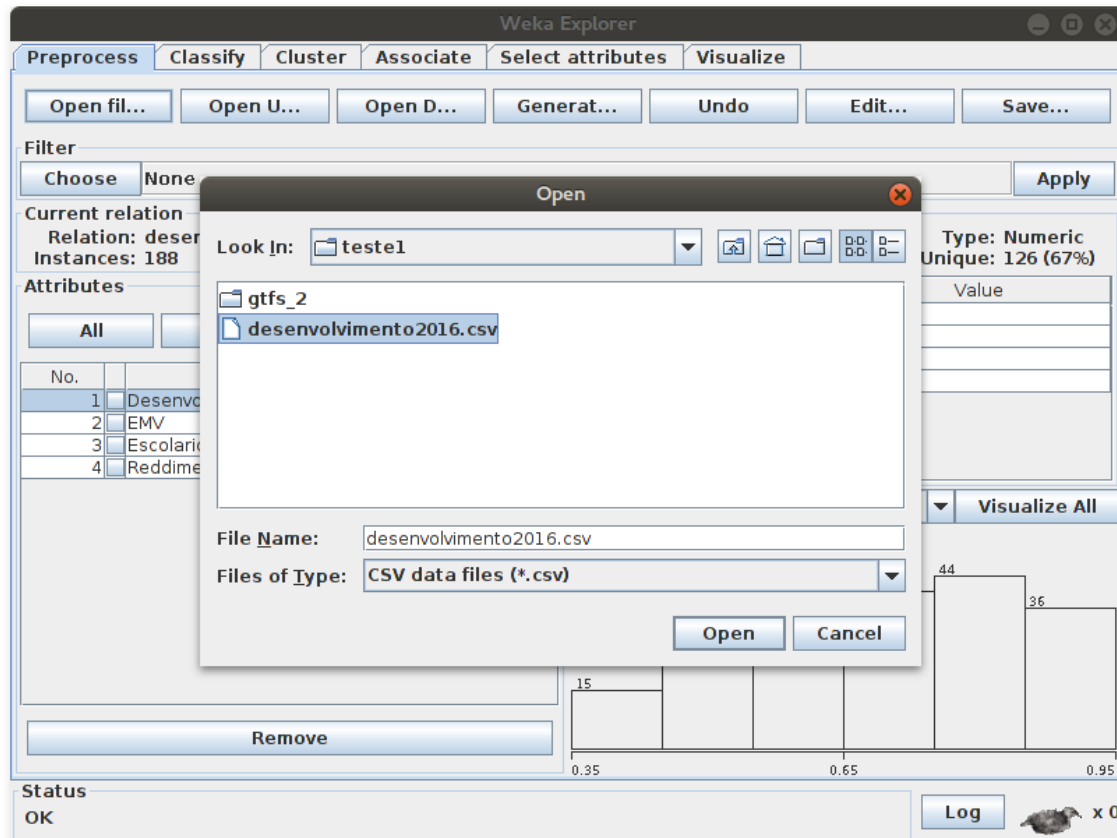
GNI – GNI per Capita

LEB – Life Expectancy at Birth

MYS – Mean years of Schooling

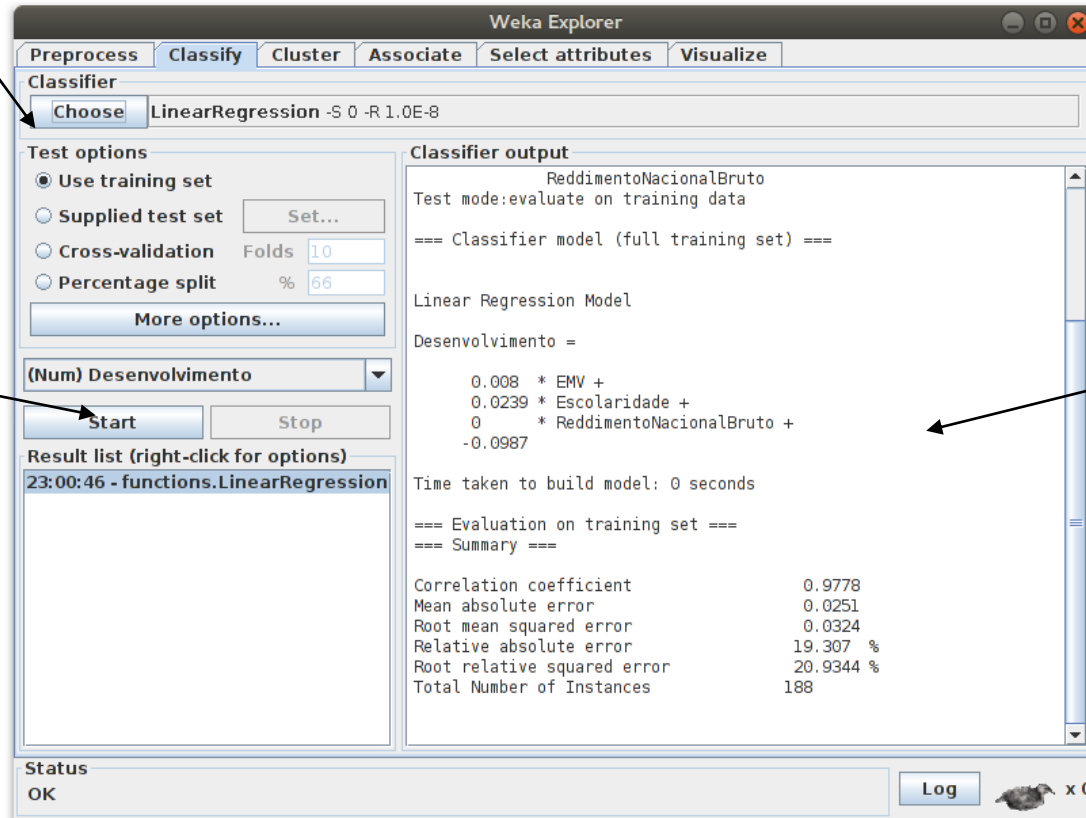
Desenvolvimento	EMV	Escolaridade	Rendimento Nacional Bruto
0.949	81.7	12.7	67614
0.939	82.5	13.2	42822
0.939	83.1	13.4	56364
0.926	81.1	13.2	45000
0.925	80.4	12.7	44519
0.925	83.2	11.6	78162
0.924	81.7	11.9	46326
0.923	81.1	12.3	43798
0.921	82.7	12.2	37065
0.920	82.2	13.1	42582
0.920	79.2	13.2	53245
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

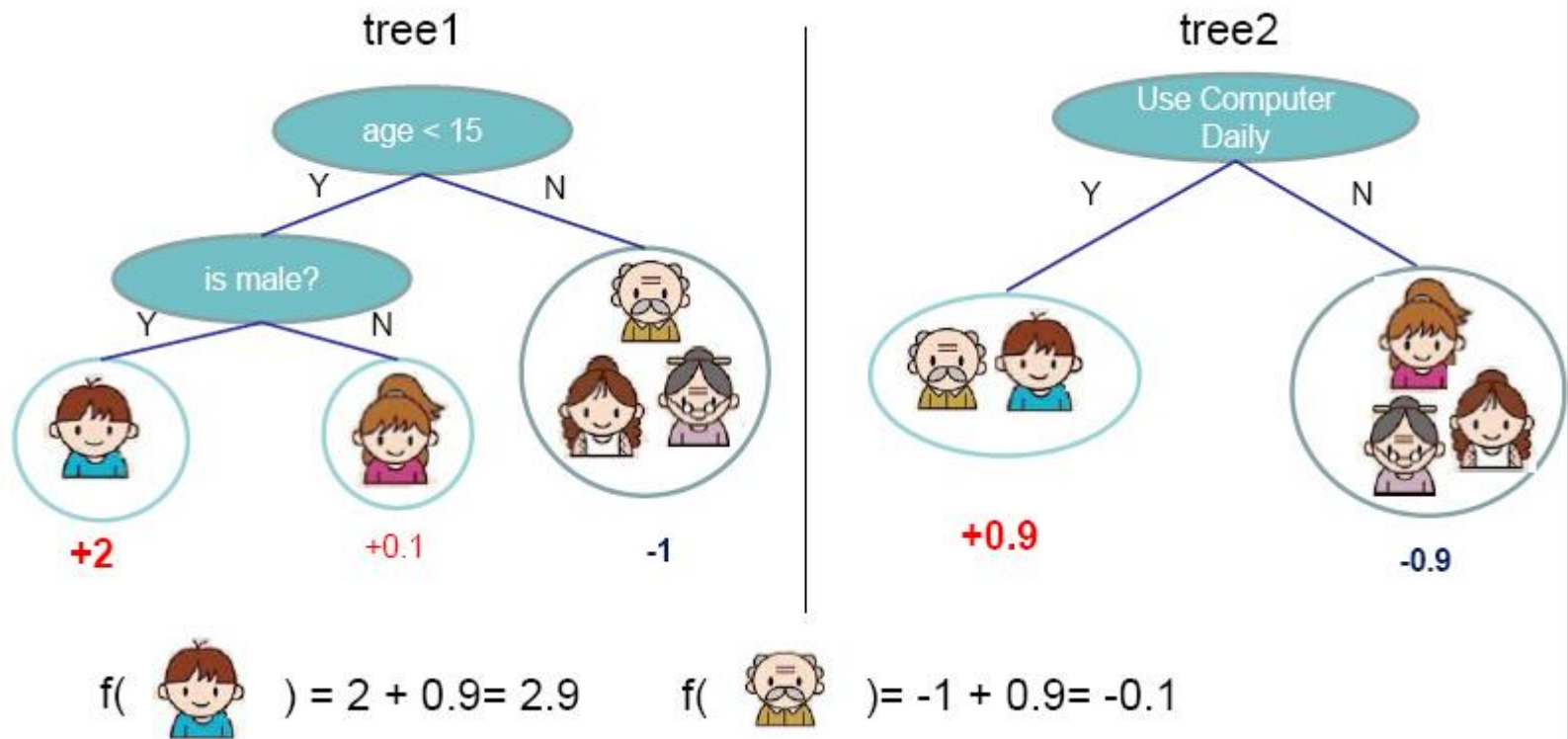
Choose/Weka/classifiers/Functions/LinearRegression



Start to Run

Results

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

The screenshot shows the Weka Explorer application window. The 'Open' dialog box is open, displaying the contents of the 'teste1' directory. The files listed are 'bmw.csv', 'filmes.csv', 'filmes1.arff', 'films.arff', and 'films.zip'. The 'File Name' field is set to 'filmes.csv' and the 'Files of Type' dropdown is set to 'All Files'. The 'Open' button is highlighted. In the background, the Weka Explorer interface shows the 'Preprocess' tab selected, and the 'Attributes' table is visible.

No.	Name
1	CountryOfOrigin
2	BigStar
3	Genre
4	Result

Classification



Weka Explorer

Preprocess | **Classify** | Cluster | Associate | Select attributes | Visualize

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

Filter: **Discretize** -B 1.0 -M -1.0 -R first-last Apply

Current relation
Relation: film_success
Instances: 10 Attributes: 4

Attributes: All | None | Invert | Pattern

No.	Name
1	<input checked="" type="checkbox"/> CountryOfOrigin
2	<input type="checkbox"/> BigStar
3	<input type="checkbox"/> Genre
4	<input type="checkbox"/> Result

Selected attribute
Name: CountryOfOrigin Type: Nominal
Missing: 0 (0%) Distinct: 4 Unique: 2 (20%)

No.	Label	Count
1	USA	4
2	Europe	4
3	Australia	1
4	Brazil	1

Class: Result (Nom) Visualize All

Remove

Status: OK Log x 0

Discretize

Classification



Choose/Weka/classifiers/Tree/J48

Start to Run

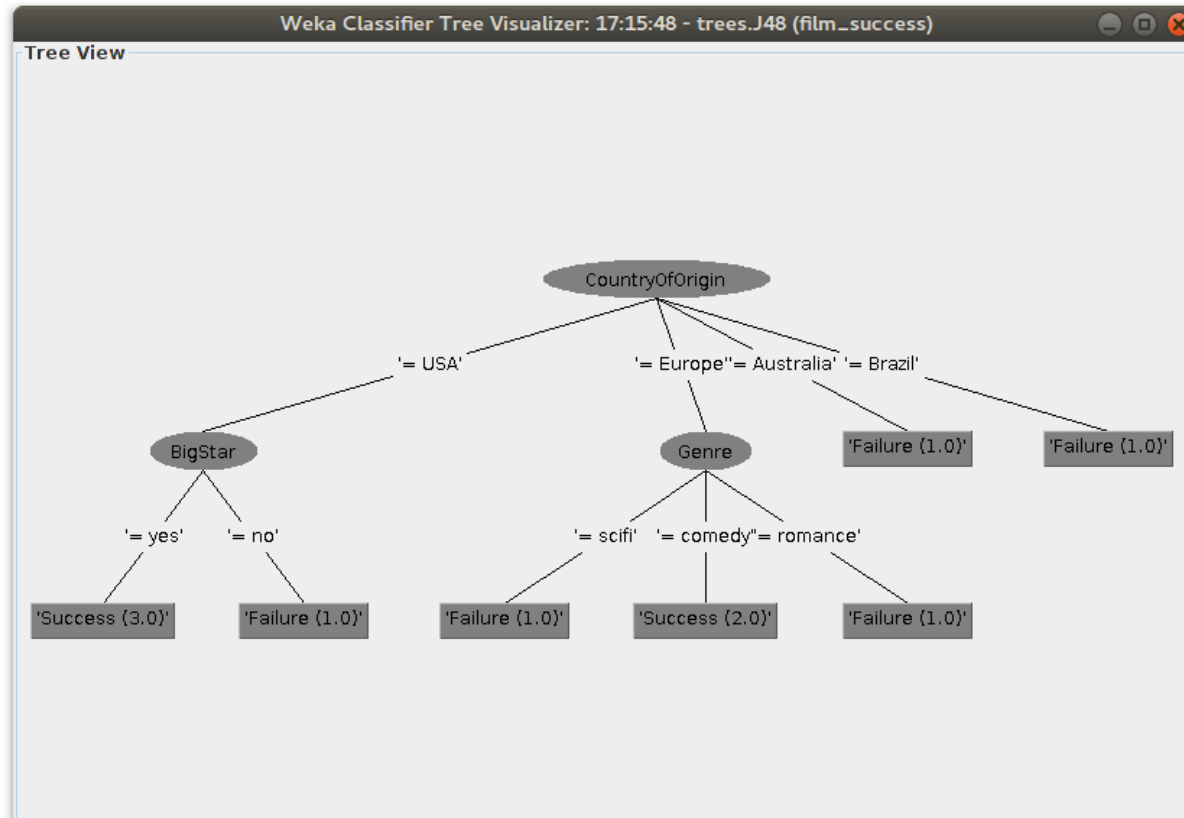
Right button of Mouse
to visualization tree

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The 'Classifier' dropdown is set to 'J48 - C 0.25 - M 1'. The 'Test options' section has 'Use training set' selected. The 'Start' button is highlighted. A 'weka.gui.GenericObjectEditor' dialog box is open, showing the configuration for 'weka.classifiers.trees.J48'. The dialog includes fields for 'binarySplits' (False), 'confidenceFactor' (0.25), 'debug' (False), 'minNumObj' (1), 'numFolds' (3), 'reducedErrorPruning' (False), 'saveInstanceData' (False), 'seed' (1), 'subtreeRaising' (True), 'unpruned' (False), and 'useLaplace' (False). The 'Classifier output' pane shows a summary of the model's performance, including 'Time taken to build model', 'Evaluation on training set', and a 'Confusion Matrix'.

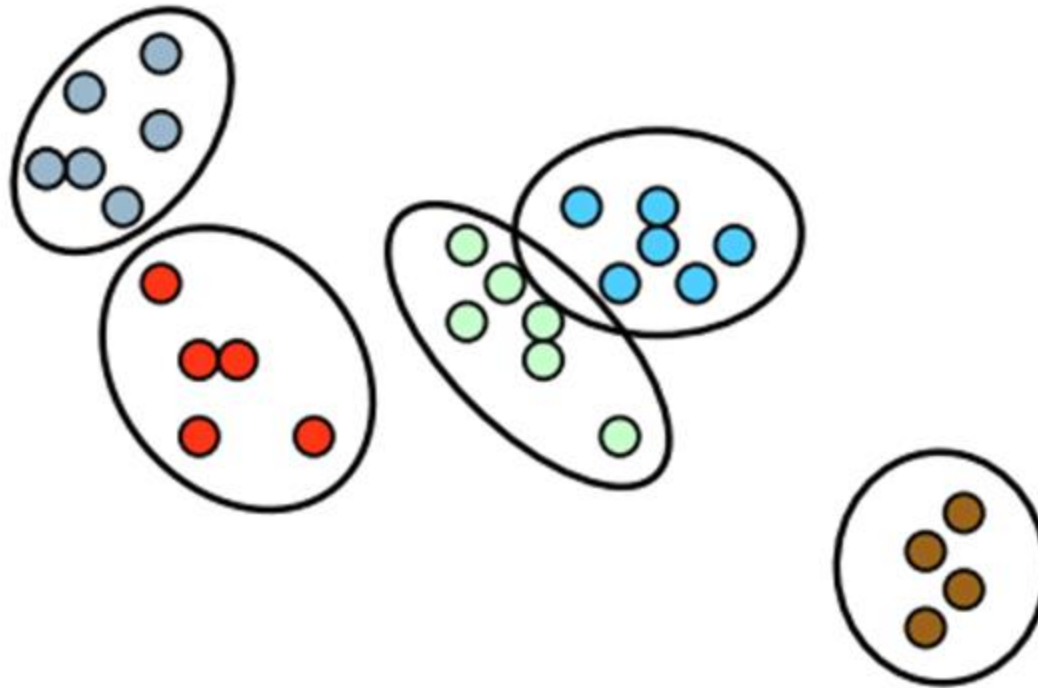
```
==== Detailed Accuracy Evaluation Summary ====
Time taken to build model: 0.000000000
==== Evaluation on training set ====
Summary
Correctly Classified Instances: 1
Incorrectly Classified Instances: 0
Kappa statistic: 1
Mean absolute error: 0
Root mean squared error: 0
Relative absolute error: 0
Root relative squared error: 0
Total Number of Instances: 1

==== Confusion Matrix ====
a b <- classified as
5 0 | a = Success
0 5 | b = Failure
```

Classification



Clusters

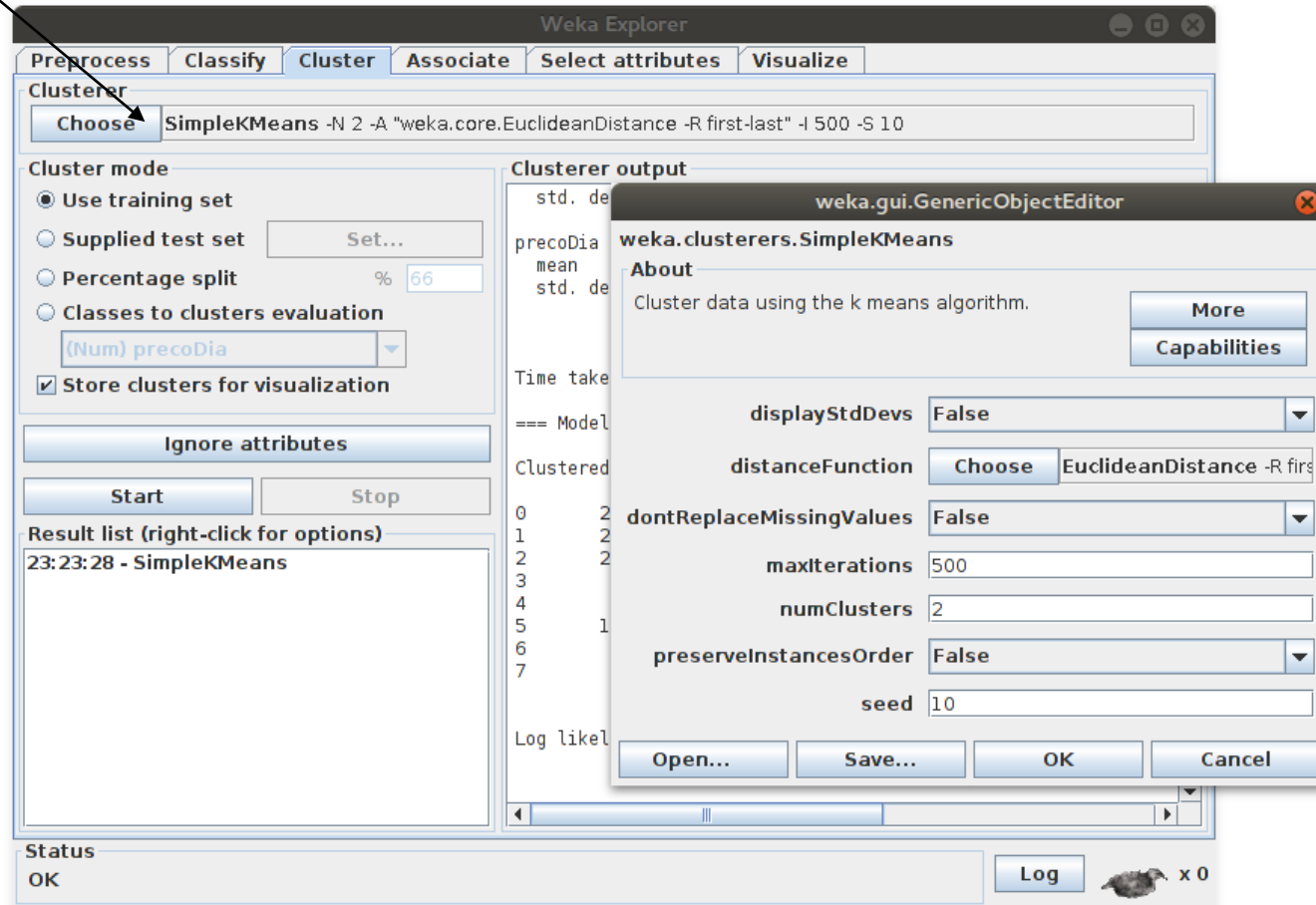


Clusters

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

The screenshot shows the Weka Explorer interface with the Clusterer tab selected. The Clusterer window displays the EM algorithm settings and its output. The settings include using the training set, a percentage split of 66%, and storing clusters for visualization. The output shows 6 iterations, a sum of squared errors of 1069.526871172735, and two clusters with centroids.

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

0	161 (15%)
1	945 (85%)

Status OK Log x 0

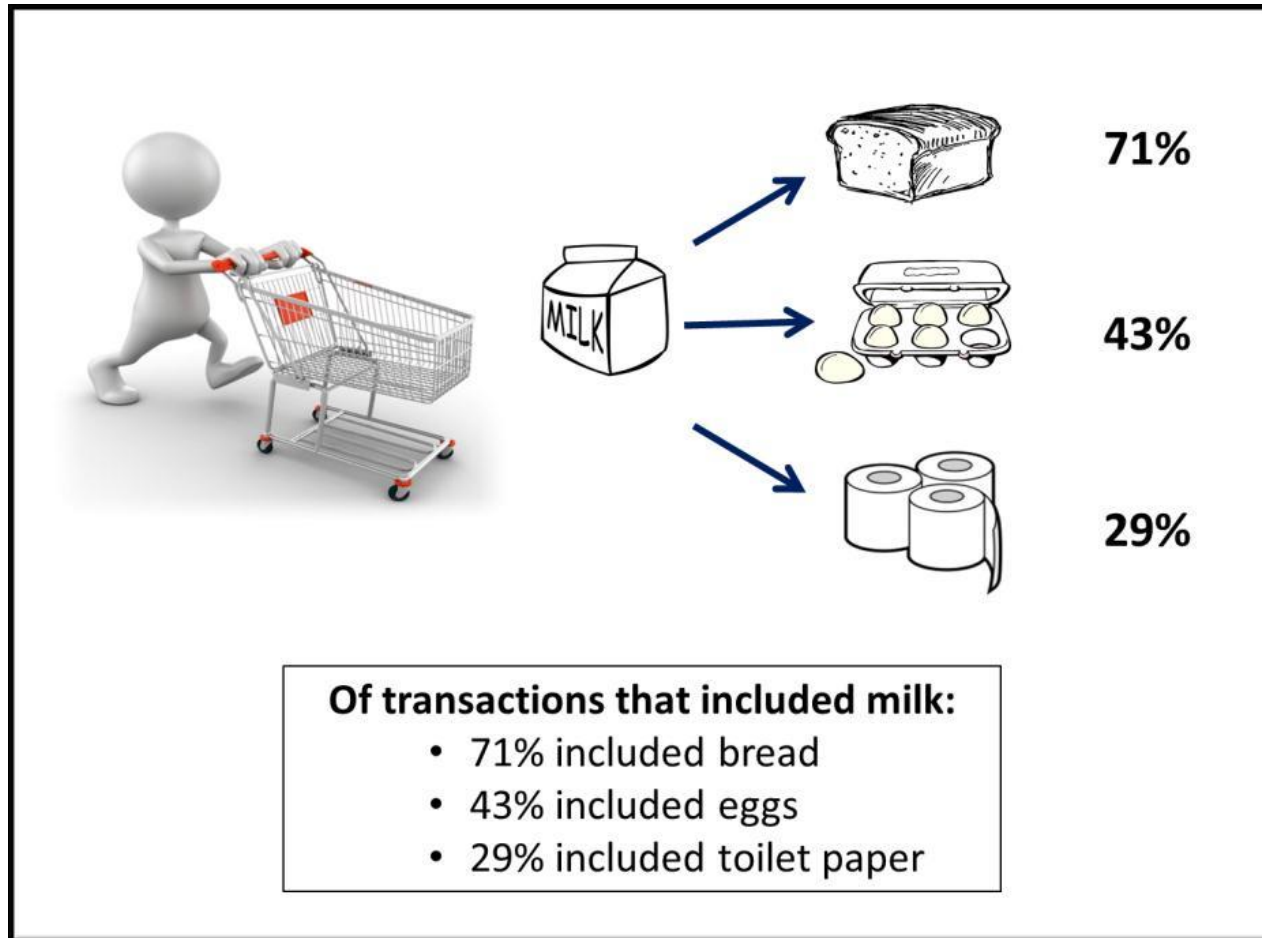
Clusters

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

The screenshot shows the Weka Explorer interface with the 'Associate' tab selected. The 'Filter' menu is open, and 'Discretize' is chosen. The 'Current relation' is 'bank-data-weka.filters.unsupervised...'. The 'Attributes' list shows 12 attributes, with 'current_act' selected. The 'Selected attribute' section shows 'Name: current_act', 'Type: Nominal', 'Missing: 0 (0%)', 'Distinct: 2', and 'Unique: 0 (0%)'. A table below shows the distribution of 'current_act':

No.	Label	Count
1	NO	145
2	YES	455

The 'Class: pep (Nom)' section shows a bar chart with two bars. The first bar (NO) has a height of 145, and the second bar (YES) has a height of 455. The bars are stacked with red on top and blue on the bottom.

Discretize

Association Rules

Apriori

The screenshot shows the Weka Explorer interface with the 'Associate' tab selected. The 'Associator' dropdown is set to 'Apriori' with parameters: `-N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1`. The 'Start' button has been clicked, and the 'Associator output' pane displays the following results:

Size of set of large itemsets L(1): 33
Size of set of large itemsets L(2): 161
Size of set of large itemsets L(3): 286
Size of set of large itemsets L(4): 171
Size of set of large itemsets L(5): 26

Best rules found:

1. children='(-inf-0.3]' save_act=YES mortgage=NO pep=NO 74 ==> married=YES 73 conf:(0.99)
2. sex=FEMALE children='(-inf-0.3]' mortgage=NO pep=NO 64 ==> married=YES 63 conf:(0.98)
3. children='(-inf-0.3]' current_act=YES mortgage=NO pep=NO 82 ==> married=YES 80 conf:(0.98)
4. children='(-inf-0.3]' mortgage=NO pep=NO 107 ==> married=YES 104 conf:(0.97)
5. children='(-inf-0.3]' car=NO mortgage=NO pep=NO 62 ==> married=YES 60 conf:(0.97)
6. married=YES children='(-inf-0.3]' save_act=YES current_act=YES 87 ==> pep=NO 80 conf:(0.9)
7. married=YES children='(-inf-0.3]' save_act=YES mortgage=NO 80 ==> pep=NO 73 conf:(0.91)
8. married=YES children='(-inf-0.3]' current_act=YES mortgage=NO 88 ==> pep=NO 80 conf:(0.91)
9. sex=FEMALE married=YES children='(-inf-0.3]' mortgage=NO 70 ==> pep=NO 63 conf:(0.9)

The 'Result list (right-click to open)' pane shows a single entry: `00:41:48 - Apriori`. The 'Status' bar at the bottom indicates 'OK'.

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/>

