DATA MINING CON Rweka

Grupo de Usuarios de R de Madrid

Mauricio Beltrán Pascual

Madrid, 28 de febrero de 2017

Índice

- ¿Qué es WEKA? (Waikato Environment for Knowledge Analysis)
  - Metodología de Minería de datos.
  - Técnicas y algoritmos de Minería de datos.
- RWeka
- Recursos de aprendizaje

Package 'RWeka'

January 23, 2017

Version 0.4-31

Title R/Weka Interface

Description An R interface to Weka (Version 3.9.1).
Weka is a collection of machine learning algorithms for data mining tasks written in Java, containing tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Package 'RWeka' contains the interface code, the Weka jar is in a separate package 'RWekajars'. For more information on Weka see <http://www.cs.waikato.ac.nz/ml/weka/>.

Depends R (>= 2.6.0)

Imports RWekajars (>= 3.9.1), rJava (>= 0.6-3), graphics, stats, utils, grid

Suggests partykit (>= 0.8.0), mlbench, e1071

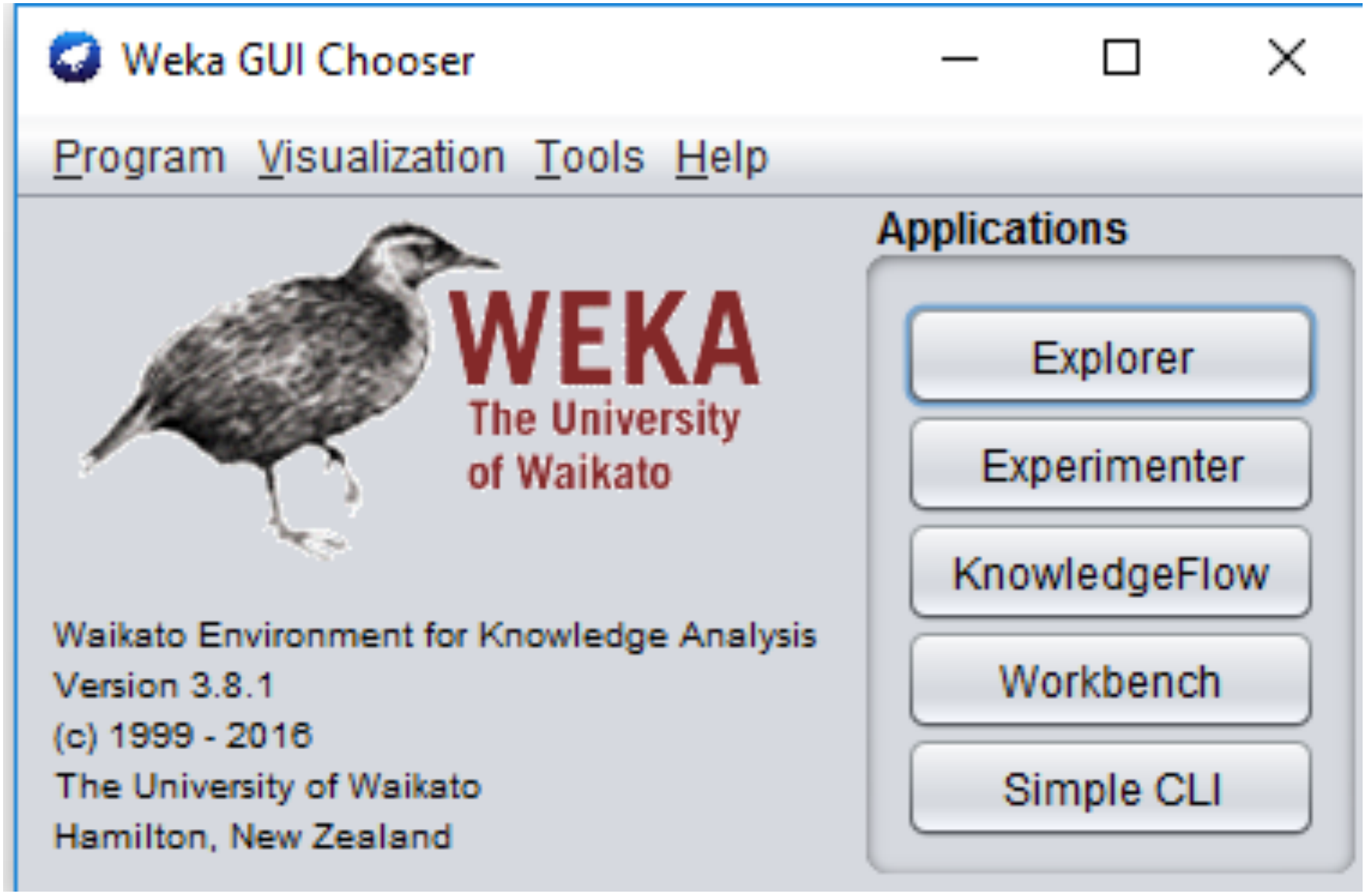
SystemRequirements Java (>= 7.0)

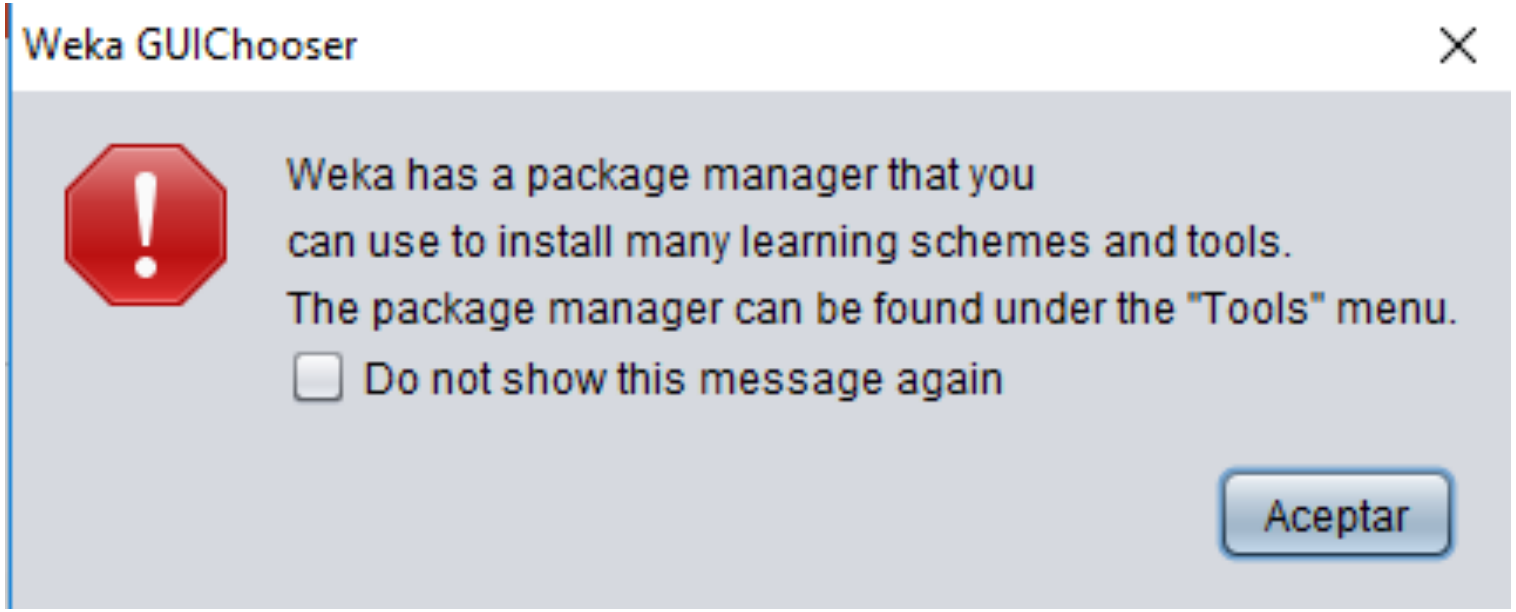
License GPL-2

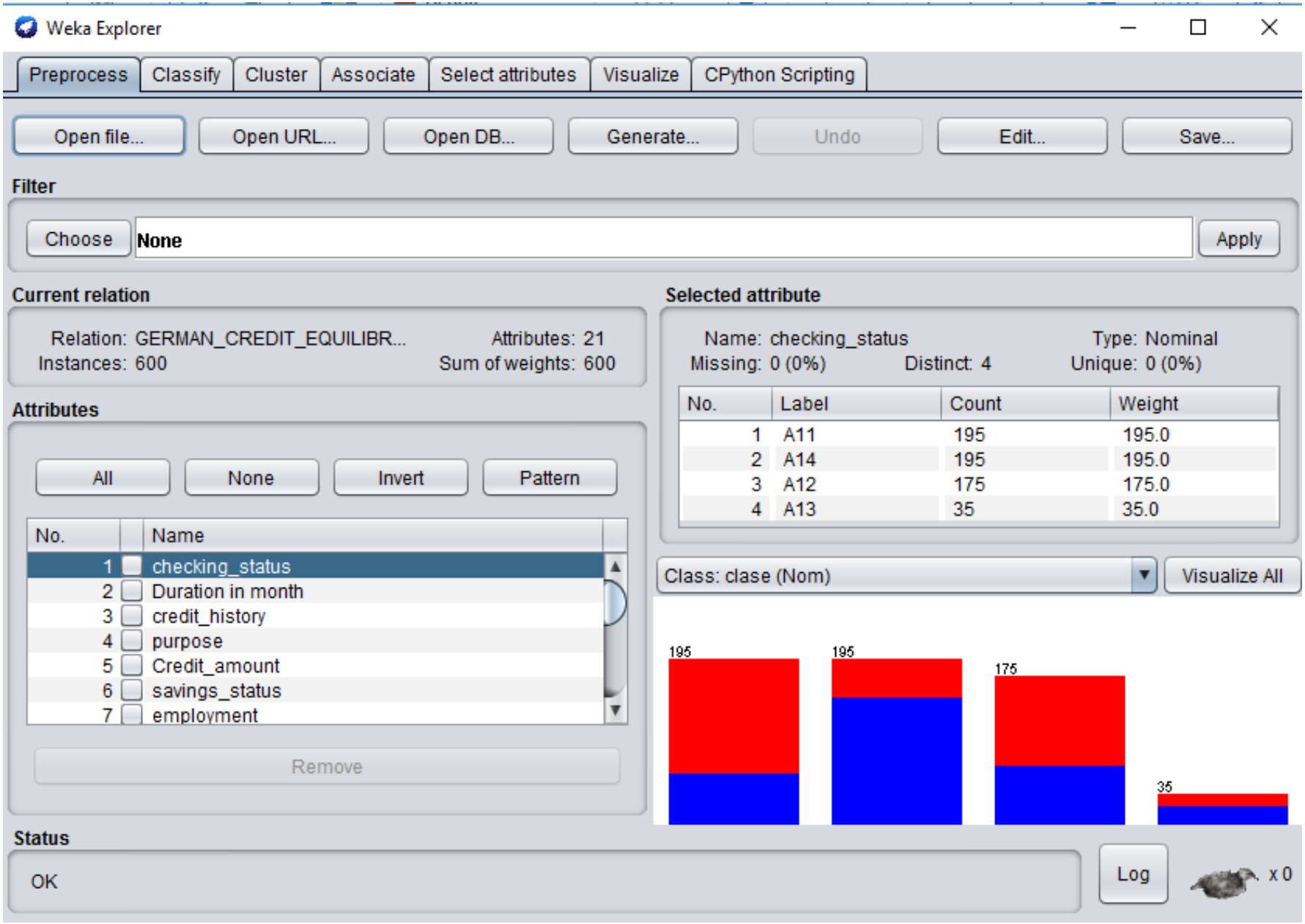
NeedsCompilation no

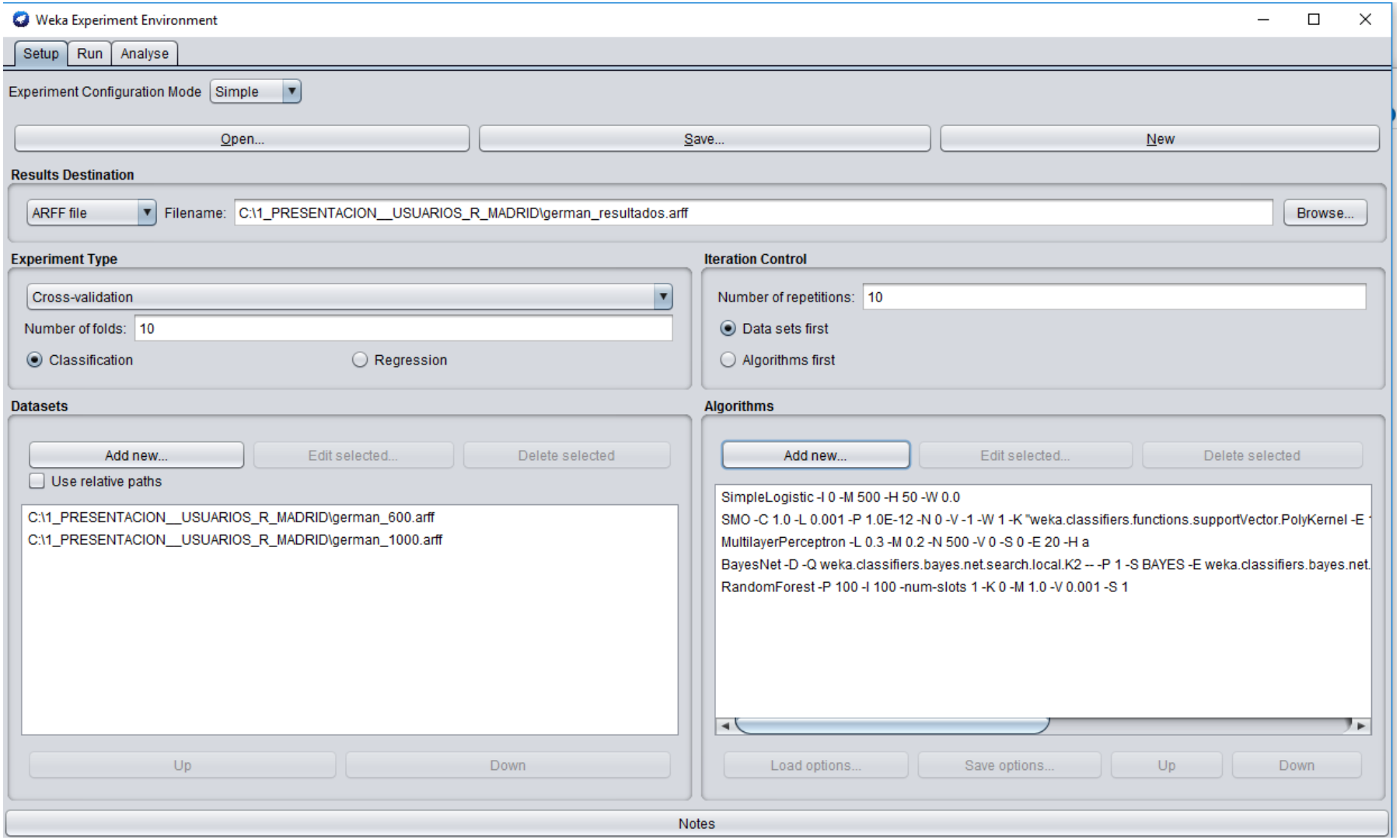
Author Kurt Hornik [aut, cre],
Christian Buchta [ctb],
Torsten Hothorn [ctb],
Alexandros Karatzoglou [ctb],
David Meyer [ctb],
Achim Zeileis [ctb]

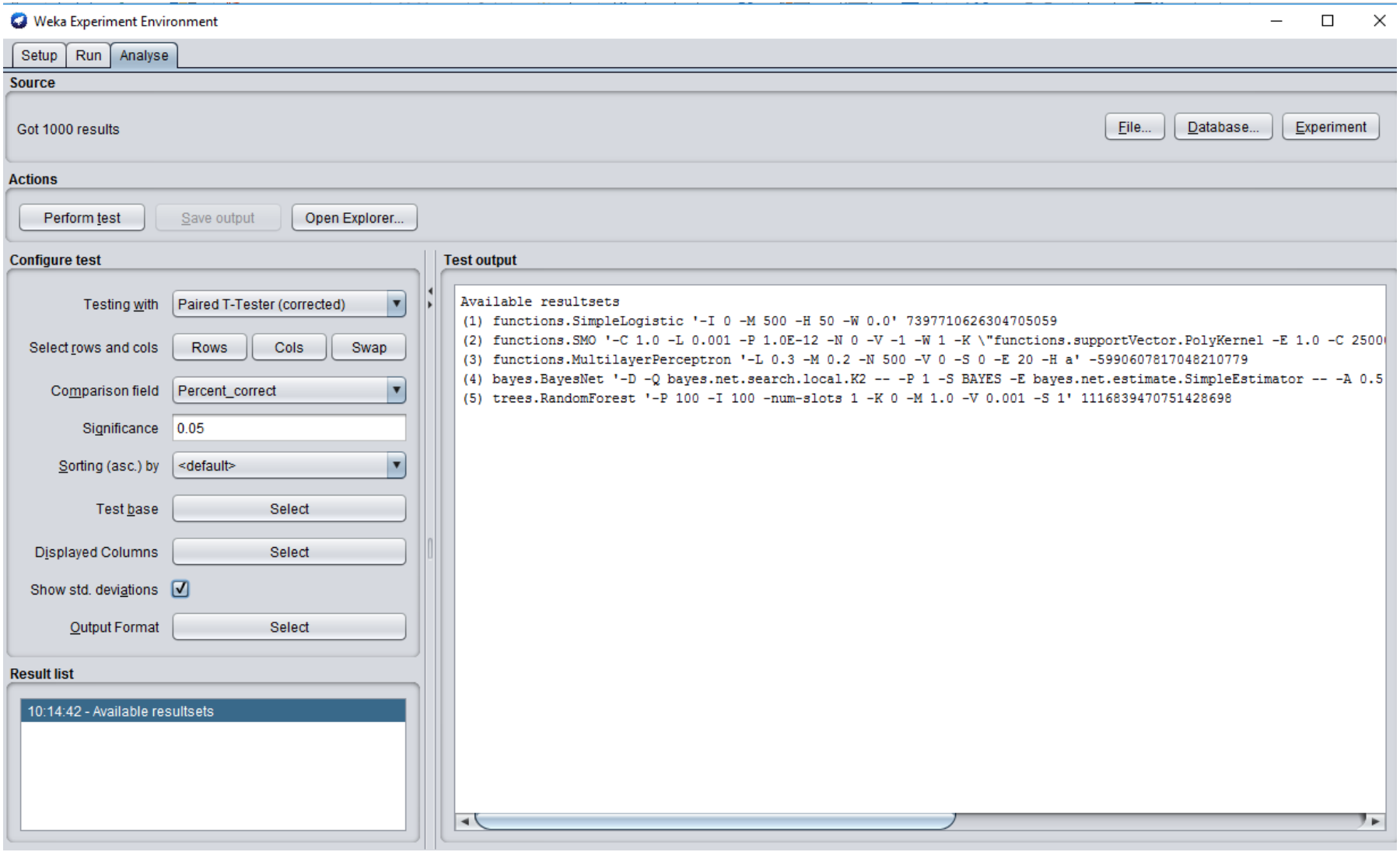
Maintainer Kurt Hornik <Kurt.Hornik@R-project.org>

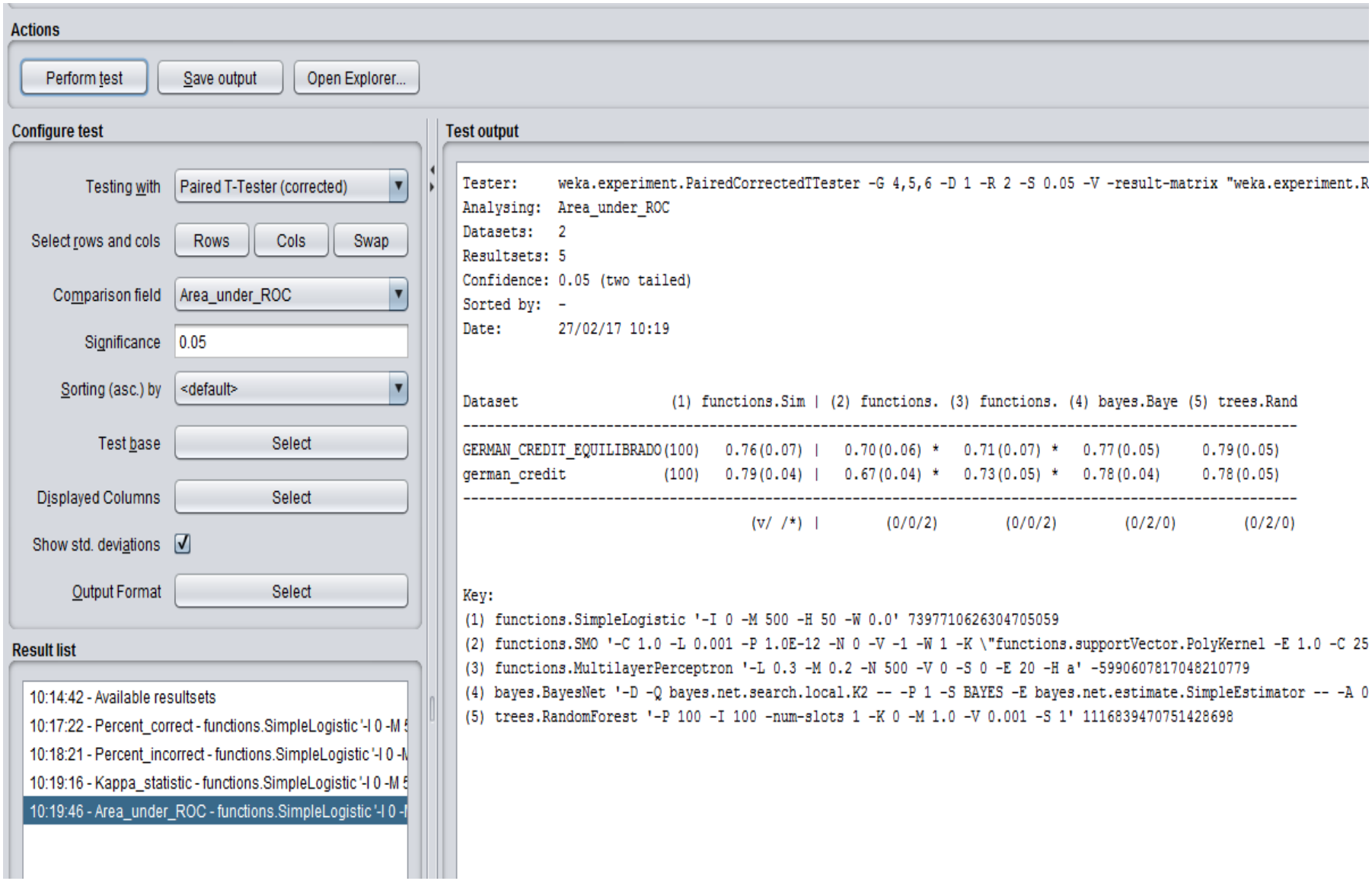

Weka GUI Chooser
Program Visualization Tools Help
WEKA
The University of Waikato
Waikato Environment for Knowledge Analysis
Version 3.8.1
(c) 1999 - 2016
The University of Waikato
Hamilton, New Zealand
Applications
Explorer
Experimenter
KnowledgeFlow
Workbench
Simple CLI

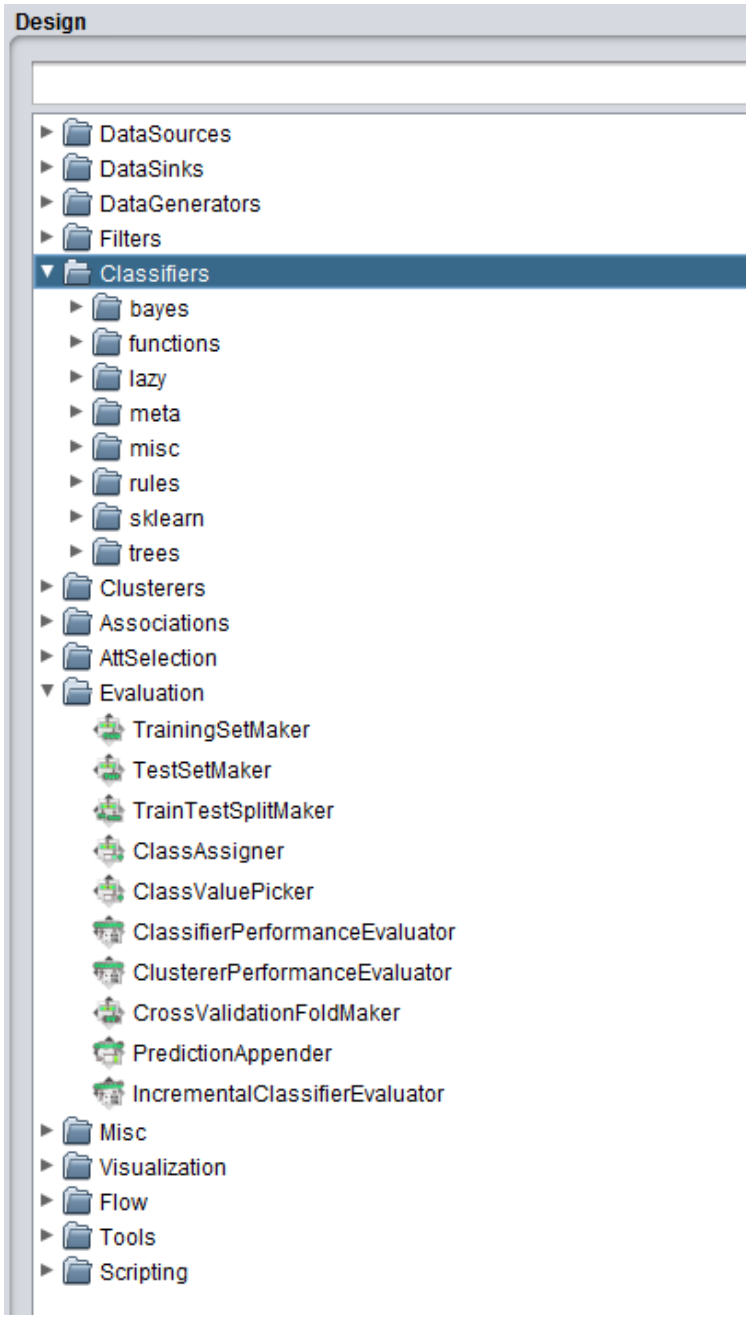

Weka GUIChooser
Weka has a package manager that you
can use to install many learning schemes and tools.
The package manager can be found under the "Tools" menu.
Do not show this message again
Aceptar


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
CPython Scripting
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: GERMAN_CREDIT_EQUILIBR...
Attributes: 21
Instances: 600
Sum of weights: 600
Selected attribute
Name: checking_status
Type: Nominal
Missing: 0 (0%)
Distinct: 4
Unique: 0 (0%)
No. Label Count Weight
1 A11 195 195.0
2 A14 195 195.0
3 A12 175 175.0
4 A13 35 35.0
Attributes
All
None
Invert
Pattern
No. Name
1 checking_status
2 Duration in month
3 credit_history
4 purpose
5 Credit_amount
6 savings_status
7 employment
Remove
Class: clase (Nom)
Visualize All
195
195
175
35
Status
OK
Log
x 0

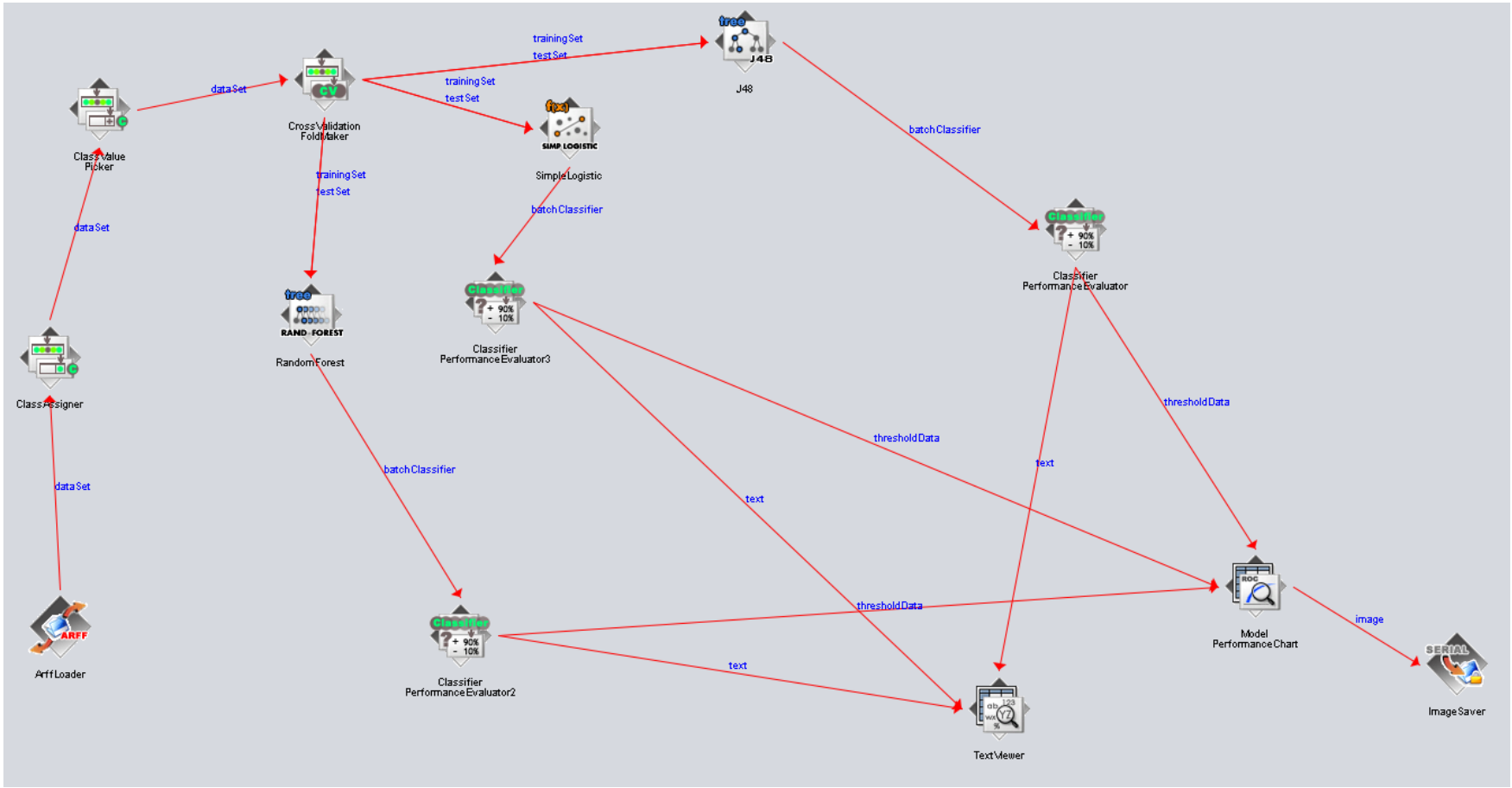

Weka Experiment Environment
Setup
Run
Analyse
Experiment Configuration Mode
Simple
Open...
Save...
New
Results Destination
ARFF file
Filename:
C:\1_PRESENTACION__USUARIOS_R_MADRID\german_resultados.arff
Browse...
Experiment Type
Cross-validation
Number of folds:
10
Classification
Regression
Iteration Control
Number of repetitions:
10
Data sets first
Algorithms first
Datasets
Add new...
Edit selected...
Delete selected
Use relative paths
C:\1_PRESENTACION__USUARIOS_R_MADRID\german_600.arff
C:\1_PRESENTACION__USUARIOS_R_MADRID\german_1000.arff
Up
Down
Algorithms
Add new...
Edit selected...
Delete selected
SimpleLogistic -I 0 -M 500 -H 50 -W 0.0
SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E
MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a
BayesNet -D -Q weka.classifiers.bayes.net.search.local.K2 -- -P 1 -S BAYES -E weka.classifiers.bayes.net.
RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
Load options...
Save options...
Up
Down
Notes


Weka Experiment Environment
Setup
Run
Analyse
Source
Got 1000 results
File...
Database...
Experiment
Actions
Perform test
Save output
Open Explorer...
Configure test
Testing with
Paired T-Tester (corrected)
Select rows and cols
Rows
Cols
Swap
Comparison field
Percent_correct
Significance
0.05
Sorting (asc.) by
<default>
Test base
Select
Displayed Columns
Select
Show std. deviations
Output Format
Select
Result list
10:14:42 - Available resultsets
Test output
Available resultsets
(1) functions.SimpleLogistic '-I 0 -M 500 -H 50 -W 0.0' 7397710626304705059
(2) functions.SMO '-C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"functions.supportVector.PolyKernel -E 1.0 -C 2500
(3) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5990607817048210779
(4) bayes.BayesNet '-D -Q bayes.net.search.local.K2 -- -P 1 -S BAYES -E bayes.net.estimate.SimpleEstimator -- -A 0.5
(5) trees.RandomForest '-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1' 1116839470751428698


Actions
Perform test
Save output
Open Explorer...
Configure test
Testing with
Paired T-Tester (corrected)
Select rows and cols
Rows
Cols
Swap
Comparison field
Area_under_ROC
Significance
0.05
Sorting (asc.) by
<default>
Test base
Select
Displayed Columns
Select
Show std. deviations
Output Format
Select
Result list
10:14:42 - Available resultsets
10:17:22 - Percent_correct - functions.SimpleLogistic '-I 0 -M 5
10:18:21 - Percent_incorrect - functions.SimpleLogistic '-I 0 -M
10:19:16 - Kappa_statistic - functions.SimpleLogistic '-I 0 -M 5
10:19:46 - Area_under_ROC - functions.SimpleLogistic '-I 0 -M
Test output
Tester: weka.experiment.PairedCorrectedTTester -G 4,5,6 -D 1 -R 2 -S 0.05 -V -result-matrix "weka.experiment.R
Analysing: Area_under_ROC
Datasets: 2
Resultsets: 5
Confidence: 0.05 (two tailed)
Sorted by: -
Date: 27/02/17 10:19
Dataset (1) functions.Sim | (2) functions. (3) functions. (4) bayes.Baye (5) trees.Rand
GERMAN_CREDIT_EQUILIBRADO(100) 0.76(0.07) | 0.70(0.06) * 0.71(0.07) * 0.77(0.05) 0.79(0.05)
german_credit (100) 0.79(0.04) | 0.67(0.04) * 0.73(0.05) * 0.78(0.04) 0.78(0.05)
(v/ /*) | (0/0/2) (0/0/2) (0/2/0) (0/2/0)
Key:
(1) functions.SimpleLogistic '-I 0 -M 500 -H 50 -W 0.0' 7397710626304705059
(2) functions.SMO '-C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K \"functions.supportVector.PolyKernel -E 1.0 -C 25
(3) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5990607817048210779
(4) bayes.BayesNet '-D -Q bayes.net.search.local.K2 -- -P 1 -S BAYES -E bayes.net.estimate.SimpleEstimator -- -A 0
(5) trees.RandomForest '-P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1' 1116839470751428698

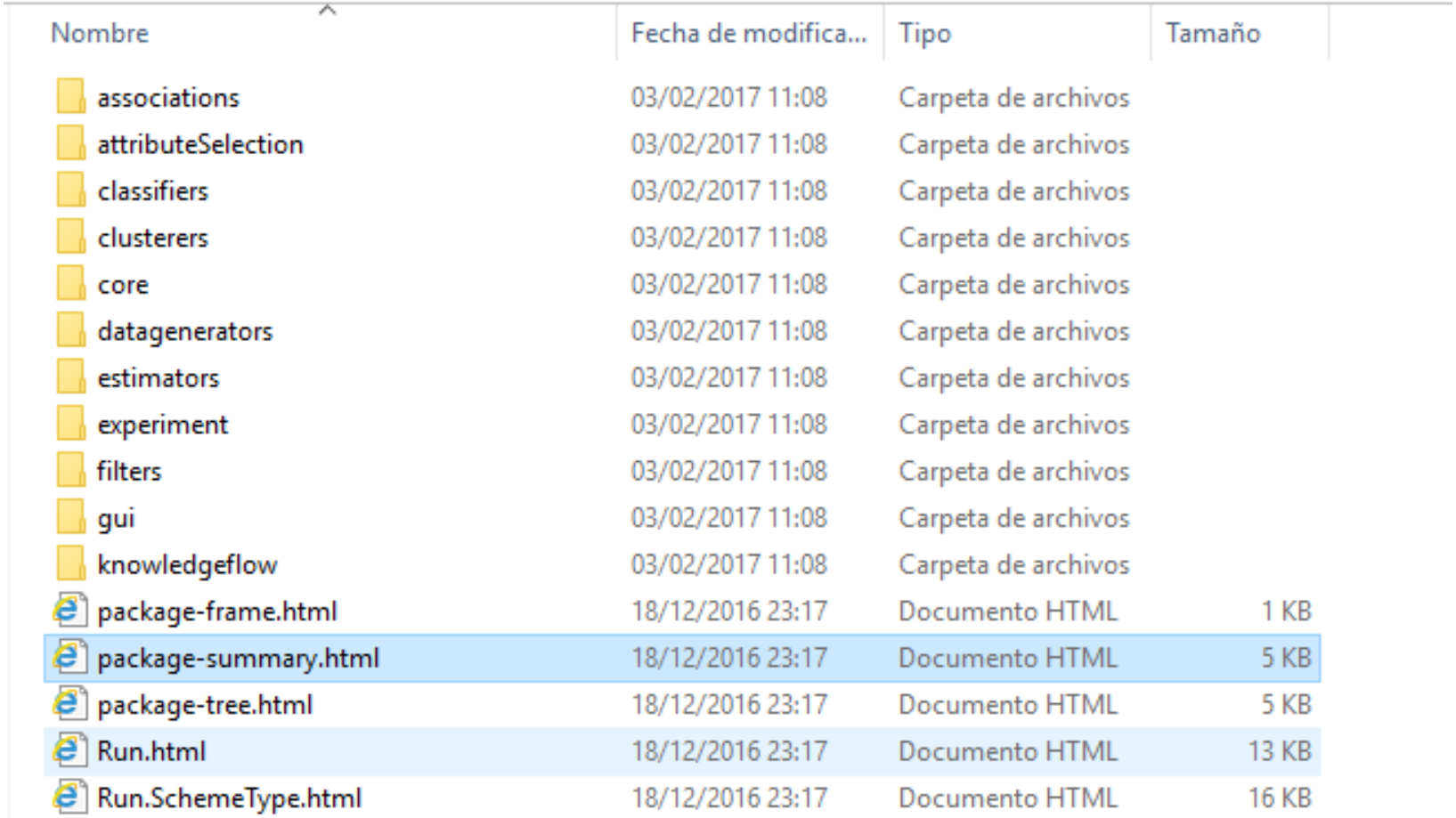

Design
DataSources
DataSinks
DataGenerators
Filters
Classifiers
bayes
functions
lazy
meta
misc
rules
sklearn
trees
Clusterers
Associations
AttSelection
Evaluation
TrainingSetMaker
TestSetMaker
TrainTestSplitMaker
ClassAssigner
ClassValuePicker
ClassifierPerformanceEvaluator
ClustererPerformanceEvaluator
CrossValidationFoldMaker
PredictionAppender
IncrementalClassifierEvaluator
Misc
Visualization
Flow
Tools
Scripting

KnowledgeFlow



Control de paquetes de WEKA

C:\Program Files\Weka-3-8\doc\weka\package-summary.html

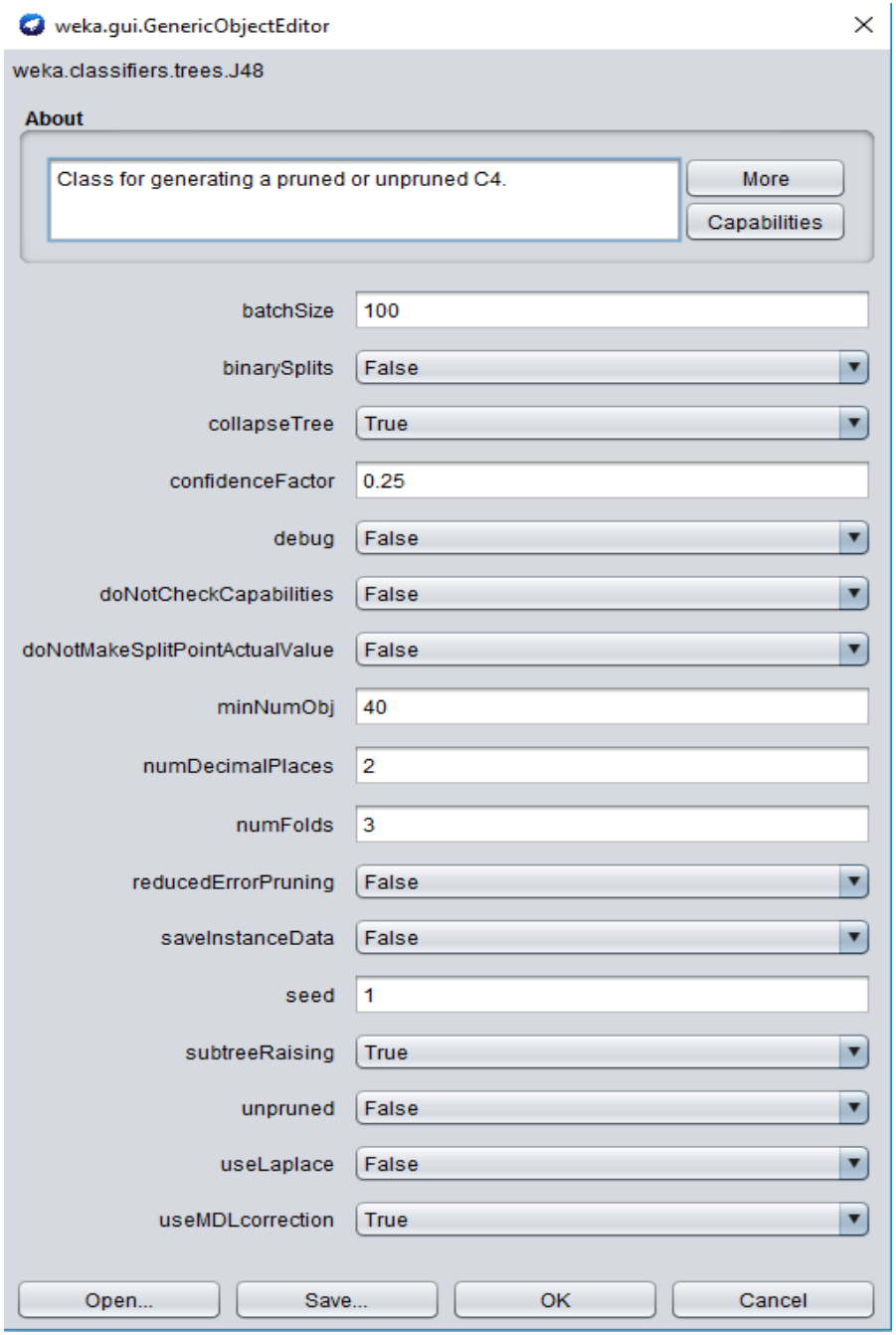


Nombre	Fecha de modifica...	Tipo	Tamaño
associations	03/02/2017 11:08	Carpeta de archivos	
attributeSelection	03/02/2017 11:08	Carpeta de archivos	
classifiers	03/02/2017 11:08	Carpeta de archivos	
clusterers	03/02/2017 11:08	Carpeta de archivos	
core	03/02/2017 11:08	Carpeta de archivos	
datagenerators	03/02/2017 11:08	Carpeta de archivos	
estimators	03/02/2017 11:08	Carpeta de archivos	
experiment	03/02/2017 11:08	Carpeta de archivos	
filters	03/02/2017 11:08	Carpeta de archivos	
gui	03/02/2017 11:08	Carpeta de archivos	
knowledgeflow	03/02/2017 11:08	Carpeta de archivos	
package-frame.html	18/12/2016 23:17	Documento HTML	1 KB
package-summary.html	18/12/2016 23:17	Documento HTML	5 KB
package-tree.html	18/12/2016 23:17	Documento HTML	5 KB
Run.html	18/12/2016 23:17	Documento HTML	13 KB
Run.SchemeType.html	18/12/2016 23:17	Documento HTML	16 KB


OVERVIEW PACKAGE CLASS TREE DEPRECATED INDE
PREV NEXT FRAMES NO FRAMES ALL CLASSES
Packages
Package
weka
weka.associations
weka.attributeSelection
weka.classifiers
weka.classifiers.bayes
weka.classifiers.bayes.net
weka.classifiers.bayes.net.estimate
weka.classifiers.bayes.net.search
weka.classifiers.bayes.net.search.ci
weka.classifiers.bayes.net.search.fixed
weka.classifiers.bayes.net.search.global
weka.classifiers.bayes.net.search.local
weka.classifiers.evaluation
weka.classifiers.evaluation.output.prediction
weka.classifiers.functions
weka.classifiers.functions.neural
weka.classifiers.functions.supportVector
weka.classifiers.lazy
weka.classifiers.lazy.kstar

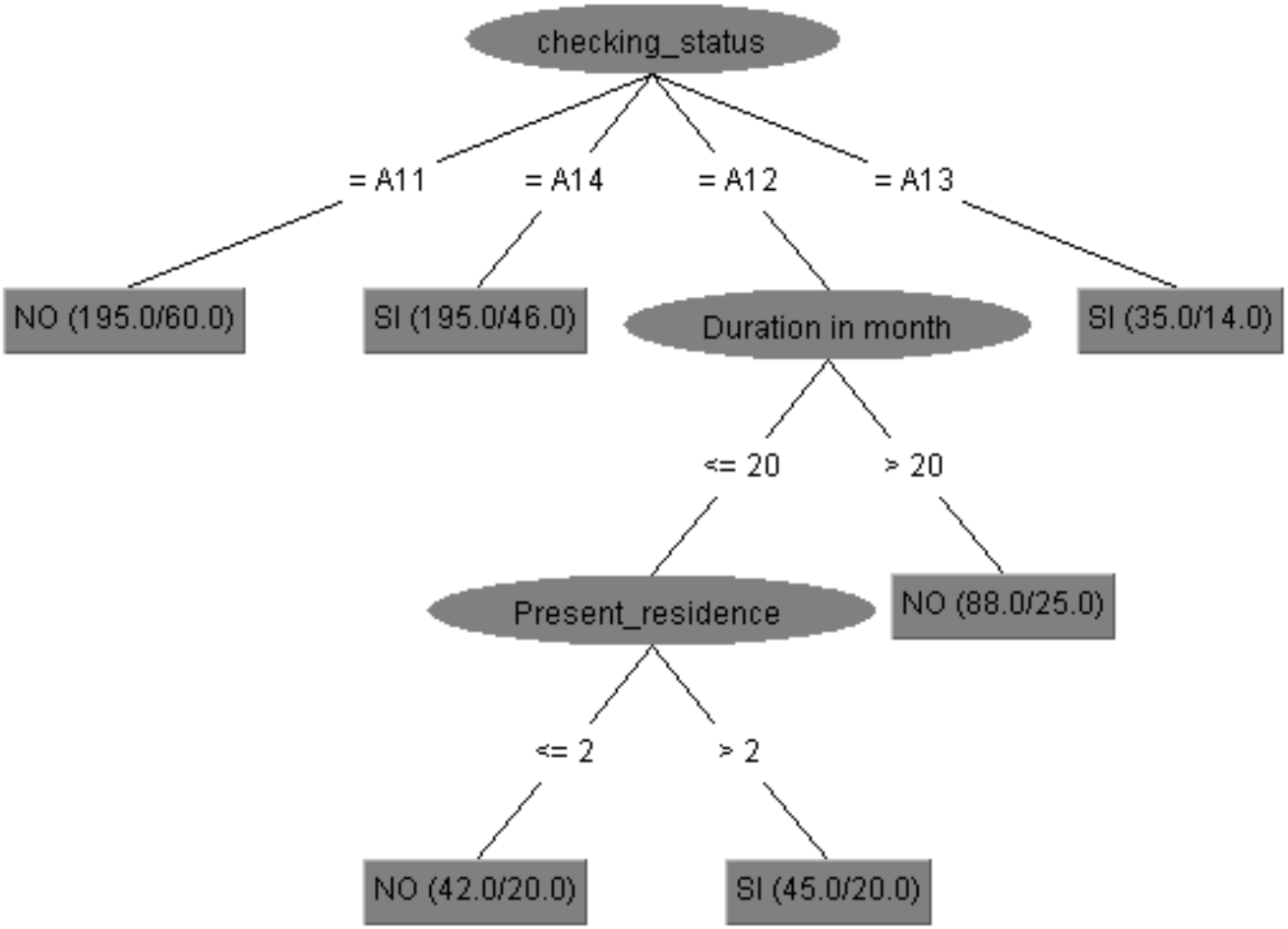
Class Summary

Class	Description
ADTree	Class for generating an alternating decision tree.
BFTree	Class for building a best-first decision tree classifier.
DecisionStump	Class for building and using a decision stump.
FT	Classifier for building 'Functional trees', which are classification trees that could have logistic regressior
Id3	Class for constructing an unpruned decision tree based on the ID3 algorithm.
J48	Class for generating a pruned or unpruned C4.5 decision tree.
J48graft	Class for generating a grafted (pruned or unpruned) C4.5 decision tree.
LADTree	Class for generating a multi-class alternating decision tree using the LogitBoost strategy.
LMT	Classifier for building 'logistic model trees', which are classification trees with logistic regression functic
M5P	M5Base.
NBTree	Class for generating a decision tree with naive Bayes classifiers at the leaves. For more information, see Ron Kohavi: Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid.
RandomForest	Class for constructing a forest of random trees. For more information see: Leo Breiman (2001).
RandomTree	Class for constructing a tree that considers K randomly chosen attributes at each node.
REPTree	Fast decision tree learner.
SimpleCart	Class implementing minimal cost-complexity pruning. Note when dealing with missing values, use "fractional instances" method instead of surrogate split met For more information, see: Leo Breiman, Jerome H.
UserClassifier	Interactively classify through visual means.


weka.gui.GenericObjectEditor
weka.classifiers.trees.J48
About
Class for generating a pruned or unpruned C4.
More
Capabilities
batchSize
100
binarySplits
False
collapseTree
True
confidenceFactor
0.25
debug
False
doNotCheckCapabilities
False
doNotMakeSplitPointActualValue
False
minNumObj
40
numDecimalPlaces
2
numFolds
3
reducedErrorPruning
False
saveInstanceData
False
seed
1
subtreeRaising
True
unpruned
False
useLaplace
False
useMDLcorrection
True
Open...
Save...
OK
Cancel

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

J48 pruned tree
------------------

checking_status = A11: NO (195.0/60.0)
checking_status = A14: SI (195.0/46.0)
checking_status = A12
|   Duration in month <= 20
|   |   Present_residence  <= 2: NO (42.0/20.0)
|   |   Present_residence  > 2: SI (45.0/20.0)
|   Duration in month > 20: NO (88.0/25.0)
checking_status = A13: SI (35.0/14.0)

Number of Leaves  :     6

Size of the tree :      9


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         398               66.3333 %
Incorrectly Classified Instances       202               33.6667 %
Kappa statistic                          0.3267
Mean absolute error                      0.4234
Root mean squared error                  0.4639
Relative absolute error                 84.6831 %
Root relative squared error             92.7731 %
Total Number of Instances              600
```

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,650    0,323    0,668      0,650    0,659      0,327    0,693     0,666     SI
                 0,677    0,350    0,659      0,677    0,668      0,327    0,693     0,635     NO
Weighted Avg.    0,663    0,337    0,663      0,663    0,663      0,327    0,693     0,651

=== Confusion Matrix ===

   a   b   <-- classified as
 195 105 |   a = SI
  97 203 |   b = NO
```


checking_status
= A11
= A14
= A12
= A13
NO (195.0/60.0)
SI (195.0/46.0)
Duration in month
SI (35.0/14.0)
<= 20
> 20
Present_residence
NO (88.0/25.0)
<= 2
> 2
NO (42.0/20.0)
SI (45.0/20.0)

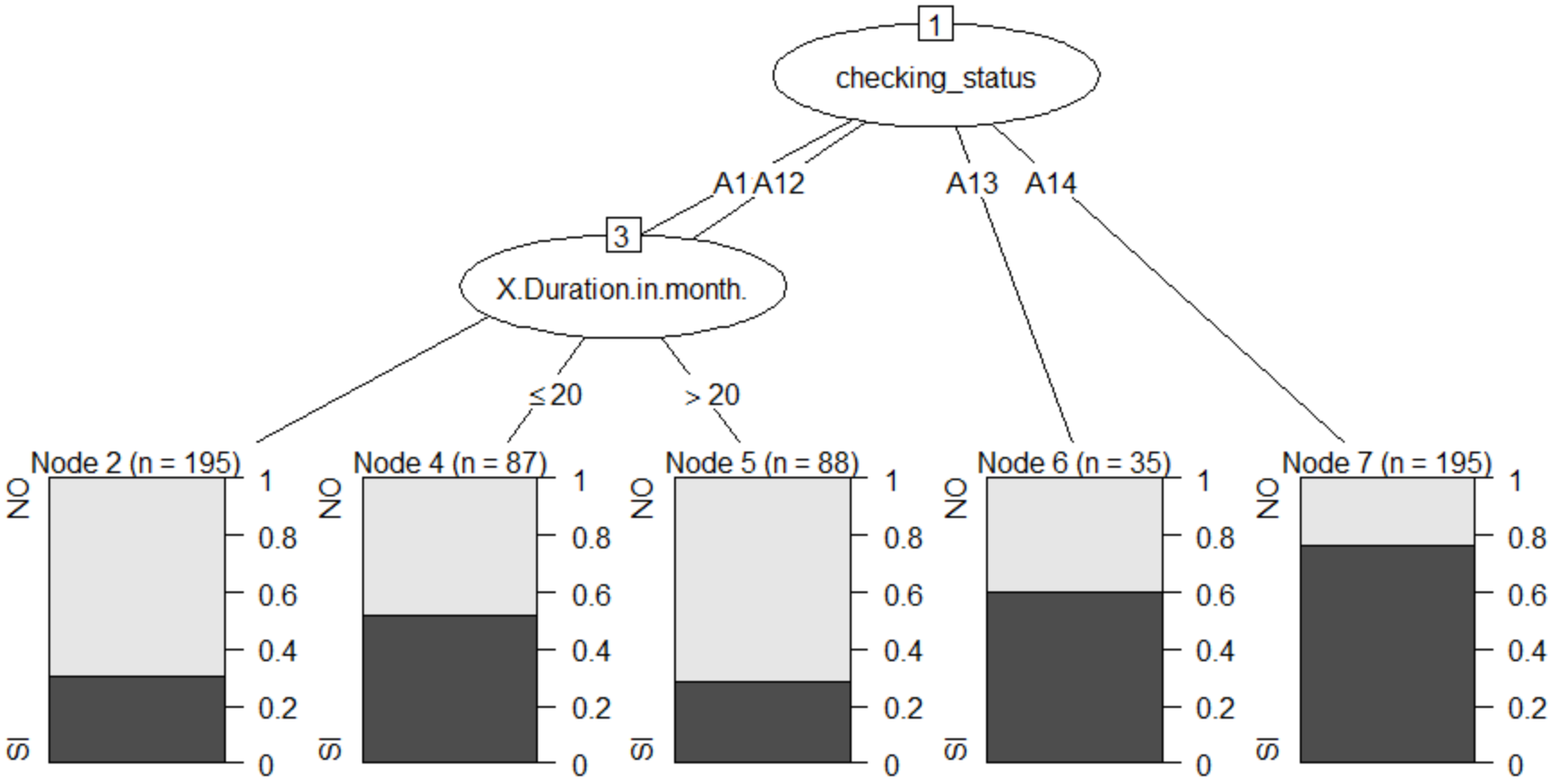
Árboles de decisión con RWeka

```
Library(Rweka)

WOW(J48)

m <- J48(clase ~ ., data = german_600, control = Weka_control(R = TRUE, M = 40))

m <- J48(clase ~ ., data = german_600, control = Weka_control(M = 40, C=0.2))

summary(m)

plot(m)

objects(modelo)

table(m$predictions,german_600$clase)
```

```
> WOW(J48)
-U      Use unpruned tree.
-O      Do not collapse tree.
-C <pruning confidence>
        Set confidence threshold for pruning.  (default 0.25)
        Number of arguments: 1.
-M <minimum number of instances>
        Set minimum number of instances per leaf.  (default 2)
        Number of arguments: 1.
-R      Use reduced error pruning.
-N <number of folds>
        Set number of folds for reduced error pruning. One fold is used as pruning set.
        (default 3)
        Number of arguments: 1.
-B      Use binary splits only.
-S      Do not perform subtree raising.
-L      Do not clean up after the tree has been built.
-A      Laplace smoothing for predicted probabilities.
-J      Do not use MDL correction for info gain on numeric attributes.
-Q <seed>
        Seed for random data shuffling (default 1).
        Number of arguments: 1.
-doNotMakeSplitPointActualValue
        Do not make split point actual value.
-output-debug-info
        If set, classifier is run in debug mode and may output additional info to the
        console
-do-not-check-capabilities
        If set, classifier capabilities are not checked before classifier is built (use
        with caution).
-num-decimal-places
        The number of decimal places for the output of numbers in the model (default
        2).
        Number of arguments: 1.
-batch-size
        The desired batch size for batch prediction (default 100).
        Number of arguments: 1.
```

Árboles de decisión

```
=== Summary ===

Correctly Classified Instances         413               68.8333 %
Incorrectly Classified Instances       187               31.1667 %
Kappa statistic                          0.3767
Mean absolute error                      0.4157
Root mean squared error                  0.4559
Relative absolute error                 83.1394 %
Root relative squared error             91.1808 %
Total Number of Instances              600

=== Confusion Matrix ===

   a   b   <-- classified as
 198 102 |   a = NO
  85 215 |   b = SI
```


1
checking_status
A1 A12
A13
A14
3
X.Duration.in.month.
≤ 20
> 20
Node 2 (n = 195)
Node 4 (n = 87)
Node 5 (n = 88)
Node 6 (n = 35)
Node 7 (n = 195)
NO
SI
1
0.8
0.6
0.4
0.2
0

Package ‘Rgraphviz’ was removed from the CRAN repository.
Formerly available versions can be obtained from the archive.

```
## visualization
## use partykit package
if(require("partykit", quietly = TRUE))
plot(m)
## or Graphviz
write_to_dot(m)
## or Rgraphviz
## Not run:
library("Rgraphviz")
ff <- tempfile()
write_to_dot(m, ff)
plot(agread(ff))
```

Objetos del modelo

```
objects(m)

"classifier“
"handlers“
"levels"
"predictions"
"terms"
```

```
> table(m$predictions,german_600$clase)

     NO  SI
NO  198  85
SI  102 215
```

Características de los principales algoritmos

CARACTERÍSTICAS DE LOS PRINCIPALES ALGORITMOS DE ÁRBOLES DE DECISIÓN

Algoritmo	Variables predictoras	Tipo de división	Criterio de División	Casos *missing*	Método de Poda	Implementación
CART (1984)	Continuas/ Discretas	Binaria	Impureza (*Gini index*)	SI	Post-	Libre Comercial
ID3 (1979)	Discretas	*n*-aria	Ganancia de información (Entropía)	NO	NO	Comercial
C4.5 (1993)	Continuas/ Discretas	Binaria/ *n*-aria	*Gain ratio* (Entropía)	SI	Pre-/Post-	Libre Comercial
J4.8	Continuas/ Discretas	Binaria/ n-aria	*Gain ratio* (Entropía)	SI	Pre-/Post-	Libre (Weka)
C5.0	Continuas/ Discretas	Binaria/ n-aria	*Gain ratio* (Entropía)	SI	Pre-/Post-	Comercial
CHAID (1975)	Discretas	*n*-aria	X^2	SI	Pre- (nivel de significancia)	Comercial

Fuente: Pérez, J.M.(2006). Tesis doctoral. Universidad del País Vasco

R/Weka interfaces

Description

Create an R interface to an existing Weka learner, attribute evaluator or filter, or show the available interfaces.

Usage

```
make_Weka_associator(name, class = NULL, init = NULL)
make_Weka_attribute_evaluator(name, class = NULL, init = NULL)
make_Weka_classifier(name, class = NULL, handlers = list(),
                     init = NULL)
make_Weka_clusterer(name, class = NULL, init = NULL)
make_Weka_filter(name, class = NULL, init = NULL)
list_Weka_interfaces()
make_Weka_package_loader(p)
```

Arguments

`name`	a character string giving the fully qualified name of a Weka learner/filter class in JNI notation.
`class`	`NULL` (default), or a character vector giving the names of R classes the objects returned by the interface function should inherit from in addition to the default ones (for representing associators, classifiers, and clusterers).
`handlers`	a named list of special handler functions, see **Details**.
`init`	`NULL`, or a function with no arguments to be called when the interface is used for building the learner/filter, or queried for available options via `WOW`. Typically, this is used for loading Weka packages when interfacing functionality in these.
`p`	a character string naming a Weka package to be loaded via `WPM`.

list_weka_interfaces()

- Associators: Apriori, Tertius..
- Evaluators: GainRatioAttributeEval, InfoGainAttributeEval..

- Classifiers: AdaBoostM1, Bagging, CostSensitiveClassifier, DecisionStump, IBk, J48, JRip, LBR, LinearRegression, LMT, Logistic, LogitBoost, M5P, M5Rules, MultiBoostAB, MultilayerPerceptron, OneR, PART, RBFNetwork, SMO, Stacking….

- Clusterers: Cobweb, DBScan, FarthestFirst, SimpleKMeans, Xmeans…

- Loaders: C45Loader, XRFFLoader..

- Savers: C45Saver, XRFFSaver..

- Filters: Discretize, Normalize..

- Stemmers: IteratedLovinsStemmer, LovinsStemmer..

- Tokenizers: AlphabeticTokenizer, NGramTokenizer, WordTokenizer…

WOW(J48)

```
arbol <- make_Weka_classifier("weka/classifiers/trees/J48")

m <- arbol(clase ~ ., data = german_600,control=Weka_control())

modelo <- evaluate_Weka_classifier(m,newdata=german_600,numFolds=10,class=T)

print(modelo)

objects(modelo)
```

```
> print(modelo)
=== 10 Fold Cross Validation ===

=== Summary ===

Correctly Classified Instances         398               66.3333 %
Incorrectly Classified Instances       202               33.6667 %
Kappa statistic                          0.3267
Mean absolute error                      0.3868
Root mean squared error                  0.5106
Relative absolute error                 77.3679 %
Root relative squared error            102.1113 %
Total Number of Instances              600

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,700    0,373    0,652      0,700    0,675      0,328    0,669     0,627     NO
                 0,627    0,300    0,676      0,627    0,651      0,328    0,669     0,619     SI
Weighted Avg.    0,663    0,337    0,664      0,663    0,663      0,328    0,669     0,623

=== Confusion Matrix ===

   a   b   <-- classified as
 210  90 |   a = NO
 112 188 |   b = SI
```

objects(modelo)

- "confusionMatrix“
- "details“
- "detailsClass“
- "string"

An object of class Weka_classifier_evaluation, a list of the following components:

string character, concatenation of the string representations of the performance statis- tics.

details vector, base statistics, e.g., the percentage of instances correctly classified, etc.

detailsComplexity vector, entropy-based statistics (if selected).

detailsClass matrix, class statistics, e.g., the true positive rate, etc., for each level of the response variable (if selected).

confusionMatrix table, cross-classification of true and predicted classes.

```
> modelo$details
           pctCorrect          pctIncorrect       pctUnclassified                 kappa     meanAbsoluteError
            66.1666667            33.8333333             0.0000000             0.3233333             0.4281020
 rootMeanSquaredError  relativeAbsoluteError rootRelativeSquaredError
            0.4687977            85.6203996            93.7595437
```

```
> modelo$detailsClass
   falsePositiveRate falseNegativeRate precision    recall  fMeasure areaUnderROC
NO         0.3933333         0.2833333 0.6456456 0.7166667 0.6793049    0.6713889
SI         0.2833333         0.3933333 0.6816479 0.6066667 0.6419753    0.6713889
```