Universidade Federal do Paraná (UFPR)
Especialização em Engenharia Industrial 4.0

Introdução ao Weka

Data Mining with Open Source Machine Learning Software in Java

David Menotti
www.inf.ufpr.br/menotti/am-17

Hoje

- Weka
  - Introdução
  - Como instalar
  - Datasets
  - Usando algoritmos de:
    - Classificação
    - Clustering
    - Regressão

Introdução

- Weka é uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados. Ele contém ferramentas para preparação de dados, classificação, regressão, agrupamento, mineração de regras de associação e visualização.

- Encontrada apenas nas ilhas da Nova Zelândia, a **Weka** é uma ave que não voa e tem uma natureza inquisitiva. O nome é pronunciado assim, e o pássaro soa assim.

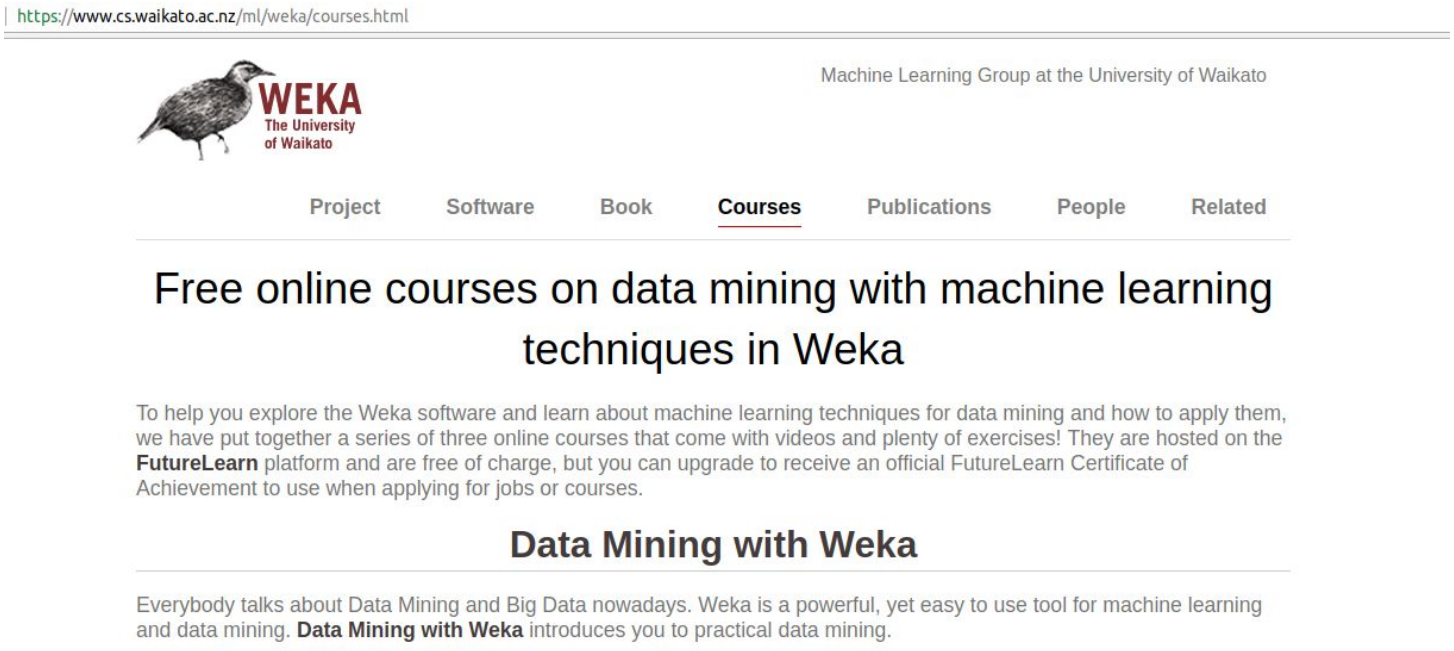
Introdução

- Weka é um software de código aberto emitido sob a ***GNU General Public License***.

- Sim, é possível aplicar a Weka para processar **big data** e realizar aprendizado profundo (**deep learning**)!

Introdução

- No site, existem vários cursos on-line gratuitos que ensinam aprendizado de máquina e mineração de dados usando o Weka. Confira no site os cursos para detalhes sobre quando e como se inscrever. Os vídeos dos cursos estão disponíveis no Youtube.

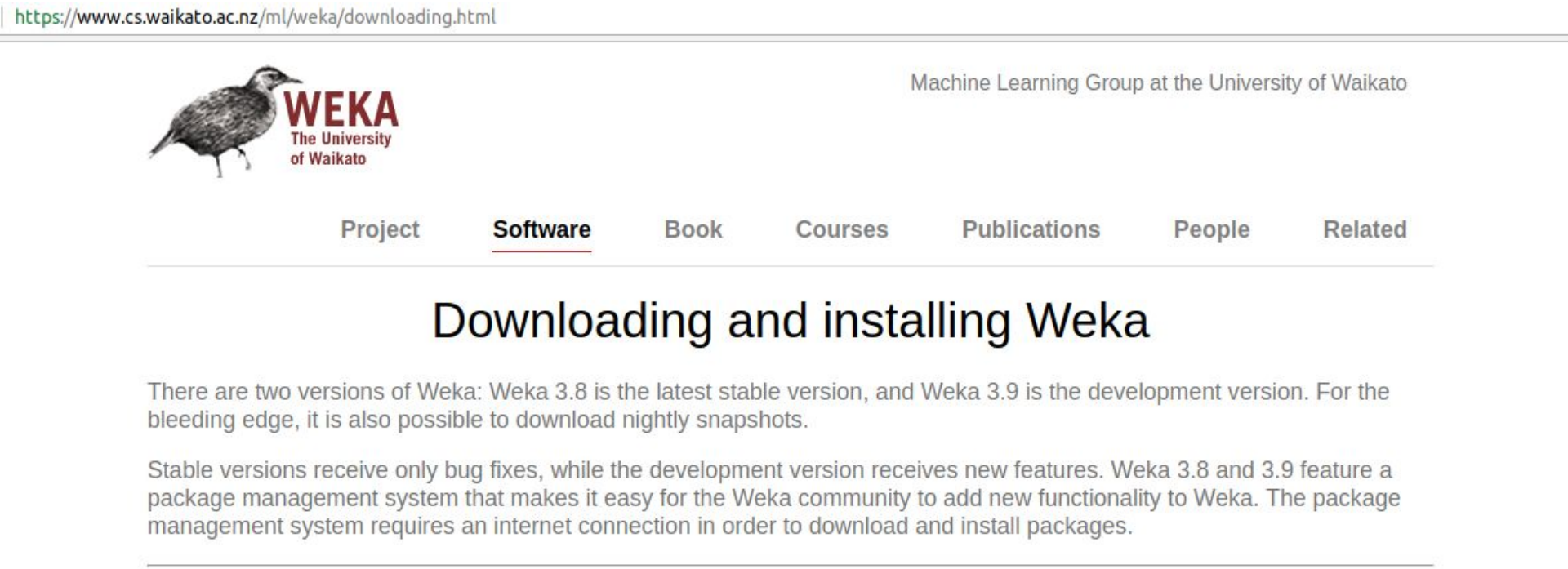


Manual

- Weka Manual
  - (v3-**6**-8) 03/05/2012
    - http://www.nilc.icmc.usp.br/elc-ebralc2012/minicursos/WekaManual-3-6-8.pdf
  - (v3-**7**-8) 21/01/2013
    - http://statweb.stanford.edu/~lpekelis/13_datafest_cart/WekaManual-3-7-8.pdf

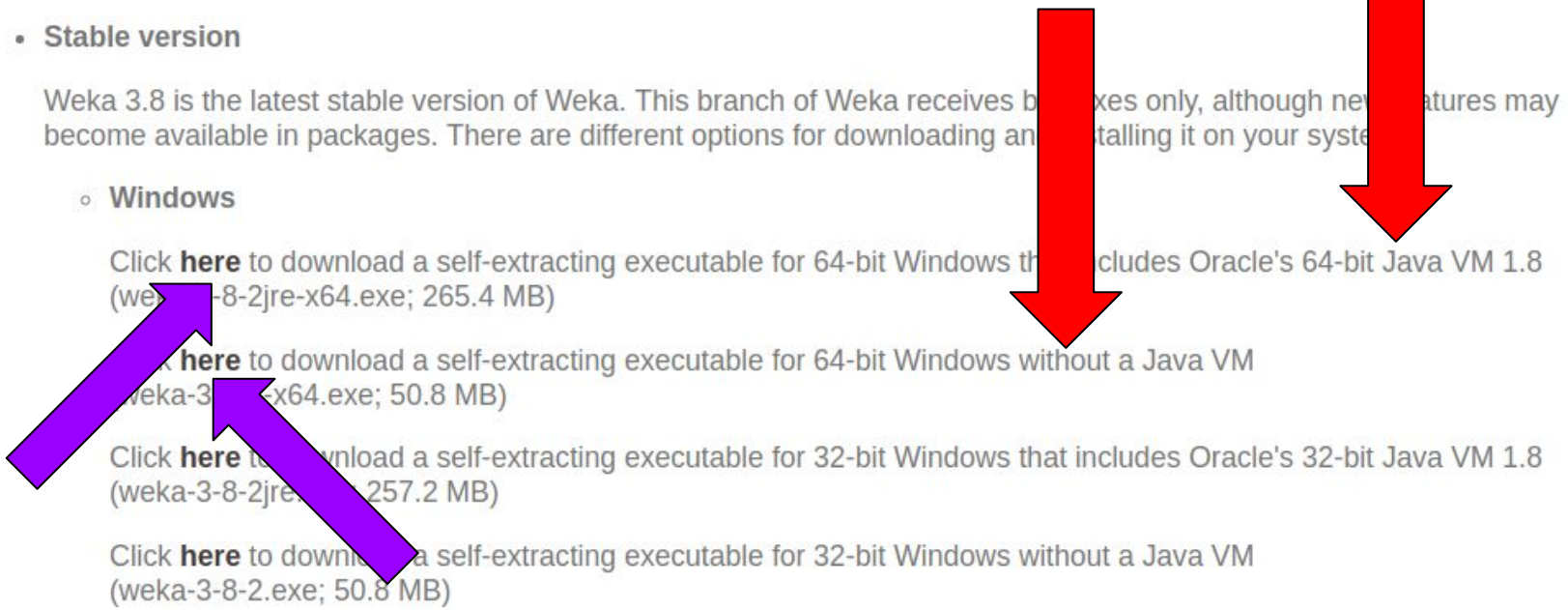
Como Instalar

- Weka website (latest version 3.8/3.9)
  - https://www.cs.waikato.ac.nz/ml/weka/



Como Instalar

https://www.cs.waikato.ac.nz/ml/weka/downloadi .html



- **Stable version**

  Weka 3.8 is the latest stable version of Weka. This branch of Weka receives b xes only, although ne atures may become available in packages. There are different options for downloading an talling it on your syste

  - **Windows**

    Click **here** to download a self-extracting executable for 64-bit Windows th cludes Oracle's 64-bit Java VM 1.8 (we -8-2jre-x64.exe; 265.4 MB)

    here to download a self-extracting executable for 64-bit Windows without a Java VM eka-3 -x64.exe; 50.8 MB)

    Click **here** t nload a self-extracting executable for 32-bit Windows that includes Oracle's 32-bit Java VM 1.8 (weka-3-8-2jre 257.2 MB)

    Click **here** to downl a self-extracting executable for 32-bit Windows without a Java VM (weka-3-8-2.exe; 50.8 MB)

Basta baixar / executar Weka-3-8-2jre-x64.exe

ou

Basta baixar / executar Weka-3-8-2-x64.exe

Como Instalar

https://www.cs.waikato.ac.nz/ml/weka/downloading.html

- **Mac OS X**

  Click **here** to download a disk image for OS X that contains a Mac application including Oracle's Java 1.8 JVM (weka-3-8-2-oracle-jvm.dmg; 124.2 MB)

- **Other platforms (Linux, etc.)**

  Click **here** to download a zip archive containing Weka (wek 8-2.zip; 51.2 MB)

  unzip the zip file. This will create a new directory called weka-3-8-2. To run Weka, change into that directory nd type

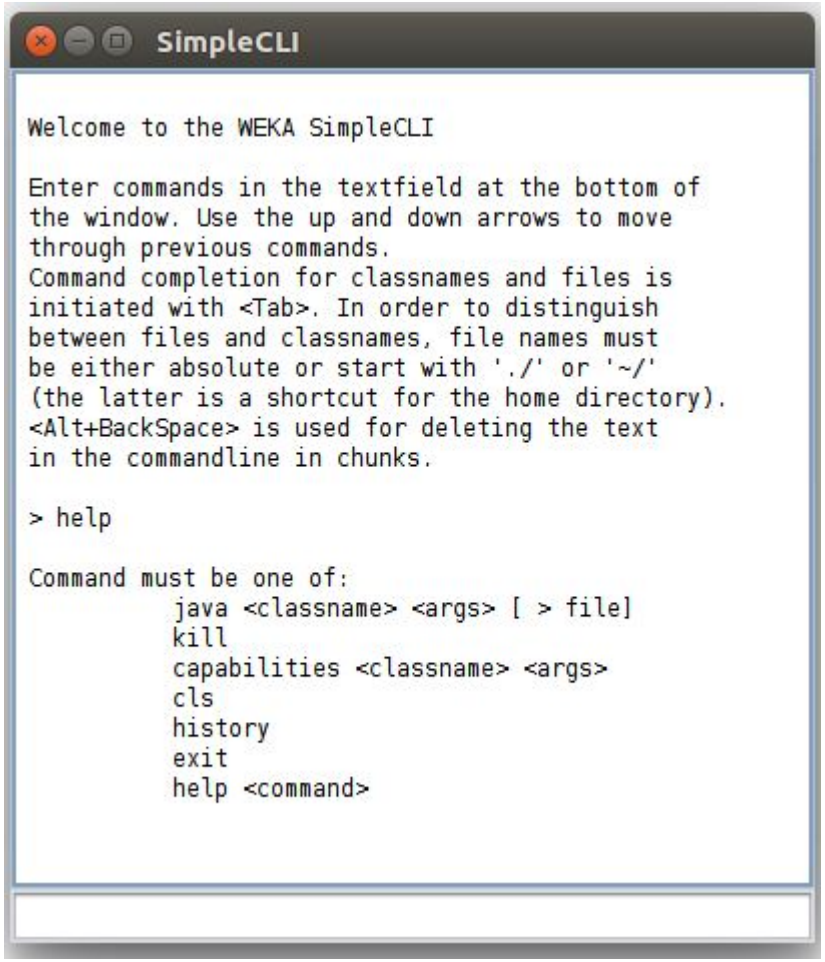
  ```
  java -jar weka.jar
  ```

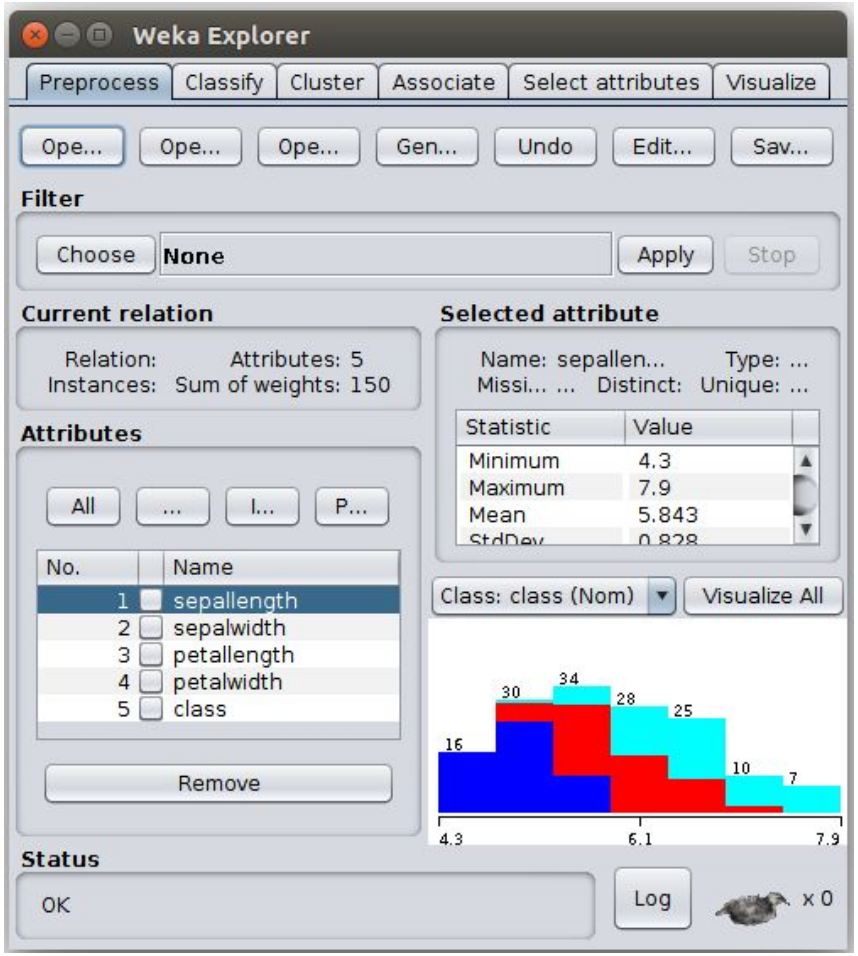
  Note that Java needs to be installed on your system for this to work. Also note, that using `-jar` will override your current CLASSPATH variable and only use the `weka.jar`.

CLI vs GUI

- Início

CLI vs GUI


SimpleCLI
Welcome to the WEKA SimpleCLI
Enter commands in the textfield at the bottom of
the window. Use the up and down arrows to move
through previous commands.
Command completion for classnames and files is
initiated with <Tab>. In order to distinguish
between files and classnames, file names must
be either absolute or start with './' or '~/'
(the latter is a shortcut for the home directory).
<Alt+BackSpace> is used for deleting the text
in the commandline in chunks.
> help
Command must be one of:
java <classname> <args> [> file]
kill
capabilities <classname> <args>
cls
history
exit
help <command>


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Ope...
Ope...
Ope...
Gen...
Undo
Edit...
Sav...
Filter
Choose
None
Apply
Stop
Current relation
Relation:
Attributes: 5
Instances:
Sum of weights: 150
Selected attribute
Name: sepallen...
Type: ...
Missi... ...
Distinct:
Unique: ...
Statistic
Value
Minimum
4.3
Maximum
7.9
Mean
5.843
Attributes
All
...
I...
P...
No.
Name
1 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Remove
Class: class (Nom)
Visualize All
16
30
34
28
25
10
7
4.3
6.1
7.9
Status
OK
Log
x 0

Atributos

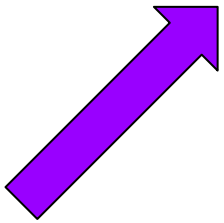
- *Nominal*: um de uma lista predefinida de valores
  – e.g. vermelho, azul, amarelo
- *Numérico*: Um número real ou inteiro
- *String*: delimitada por “aspas duplas”
- *Data*
- *Relational*

Arquivos ARFF

- A representação das instâncias

- Consiste em:
  - Um cabeçalho (header): Descreve os tipos de atributos e seus valores
  - Seção de dados: lista de dados separada por vírgula

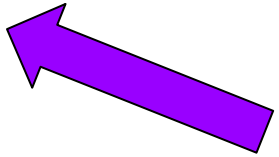
Exemplo de Arquivo ARFF



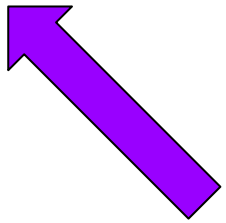
Comment

```
% This is a toy example, the UCI weather dataset
% Any relation to real weather is purely coincidental

@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no
```

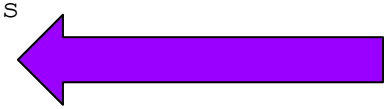
Nome do Dataset



Atributos

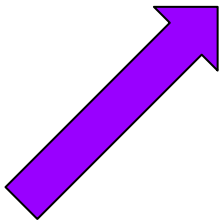


Classe / Meta



Dados/Valores

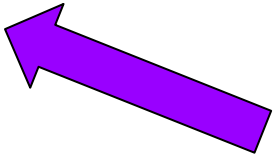
ARFF



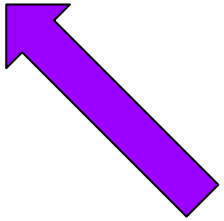
Comment

```
% This is a toy example, the UCI weather dataset
% Any relation to real weather is purely coincidental

@relation weather.symbolic

@attribute outlook {sunny, overcast, rainy}
@attribute temperature {hot, mild, cool}
@attribute humidity {high, normal}
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}

@data
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yesrainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no
```

Nome do Dataset



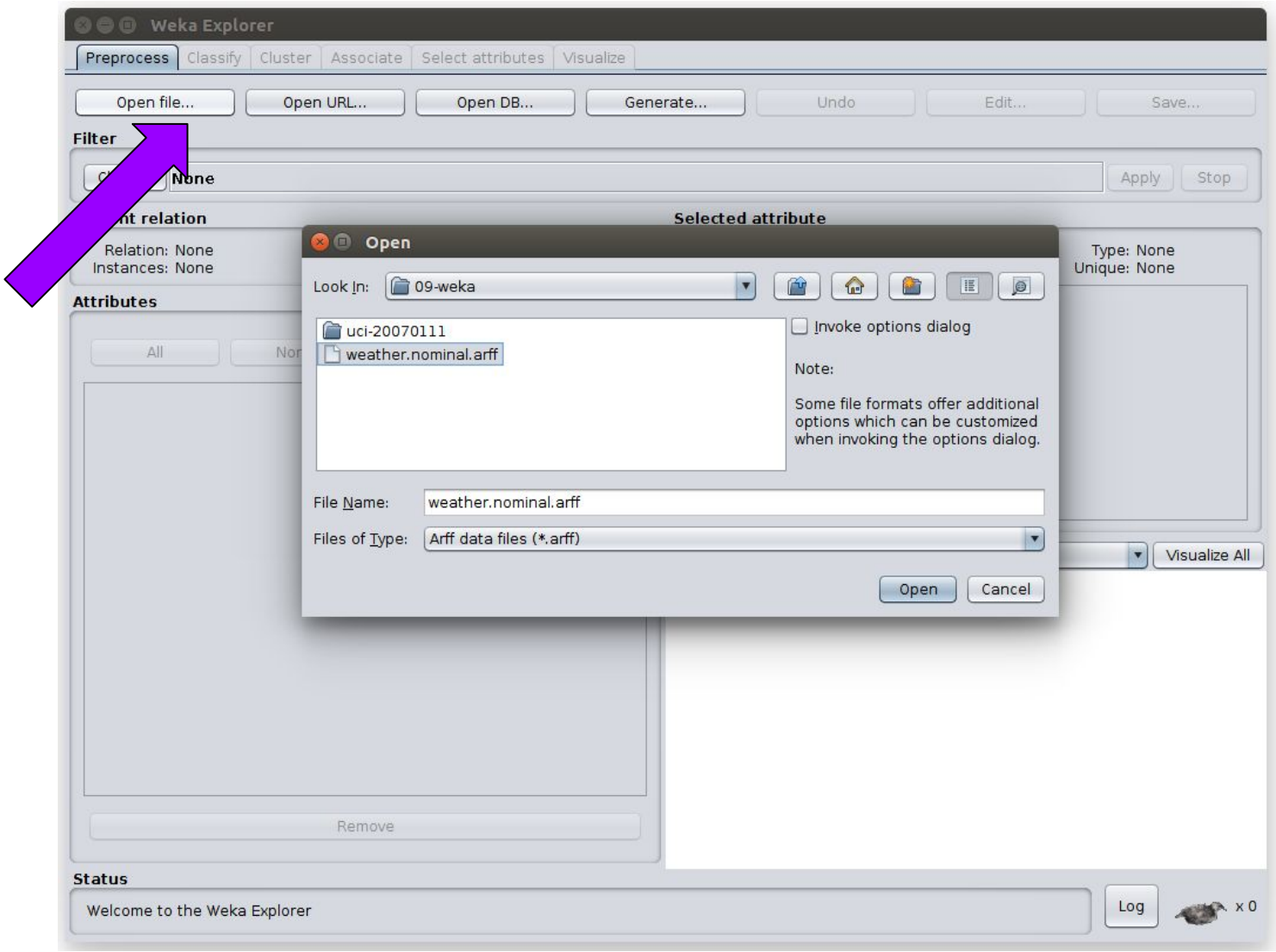
Atributos



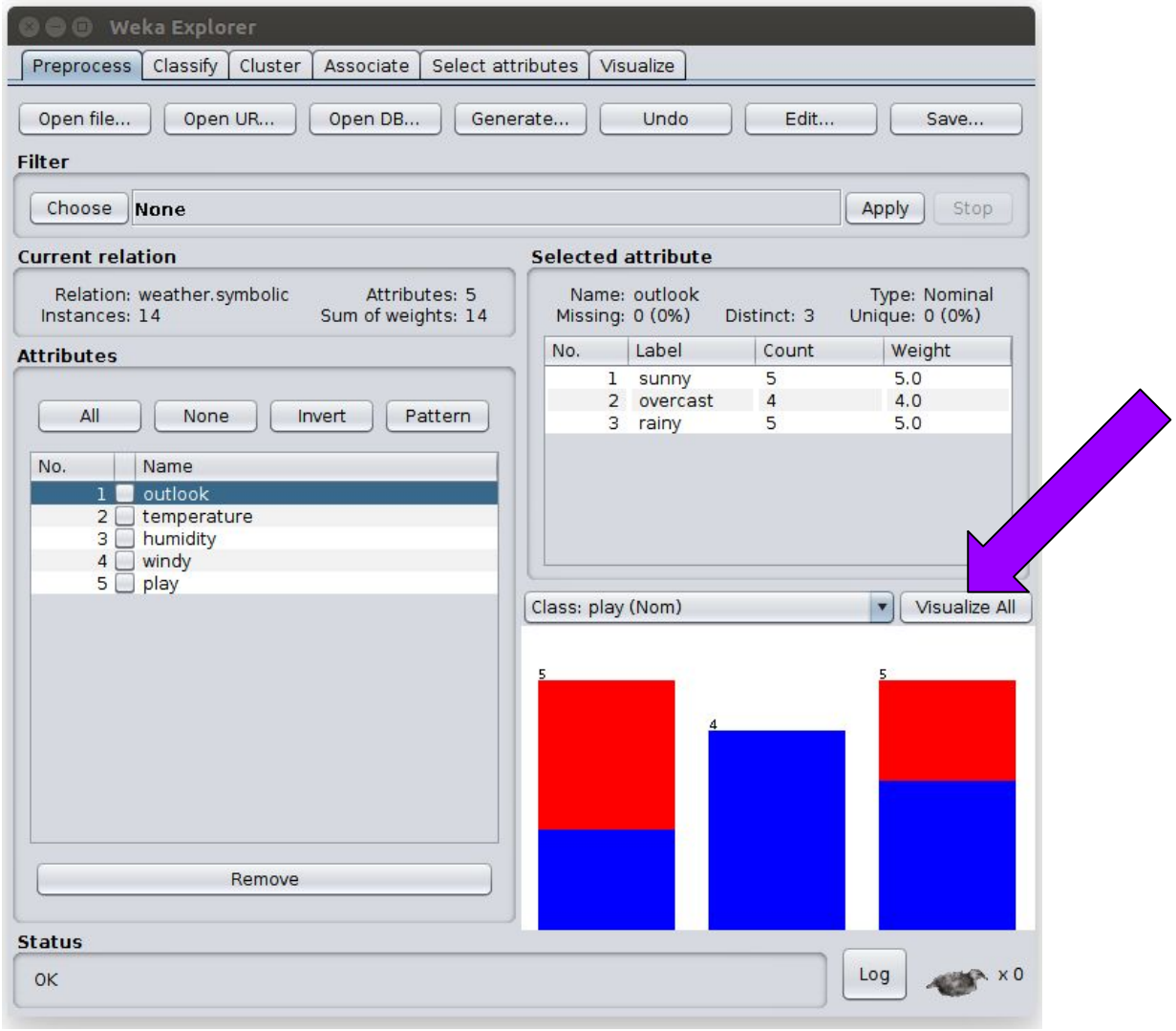
Classe / Meta

Dados/Valores

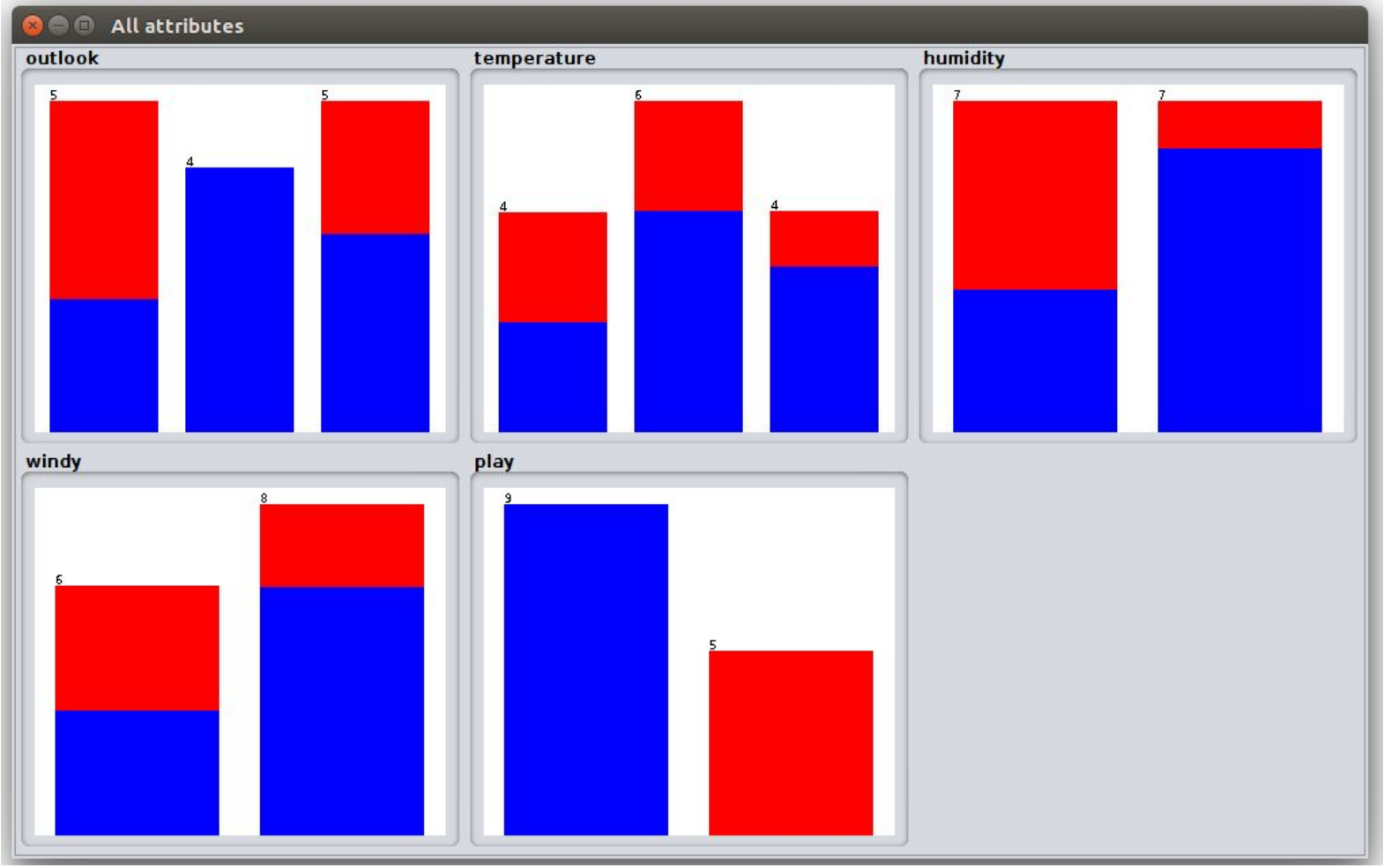
Abrindo um Dataset



Visualizando



Visualizando



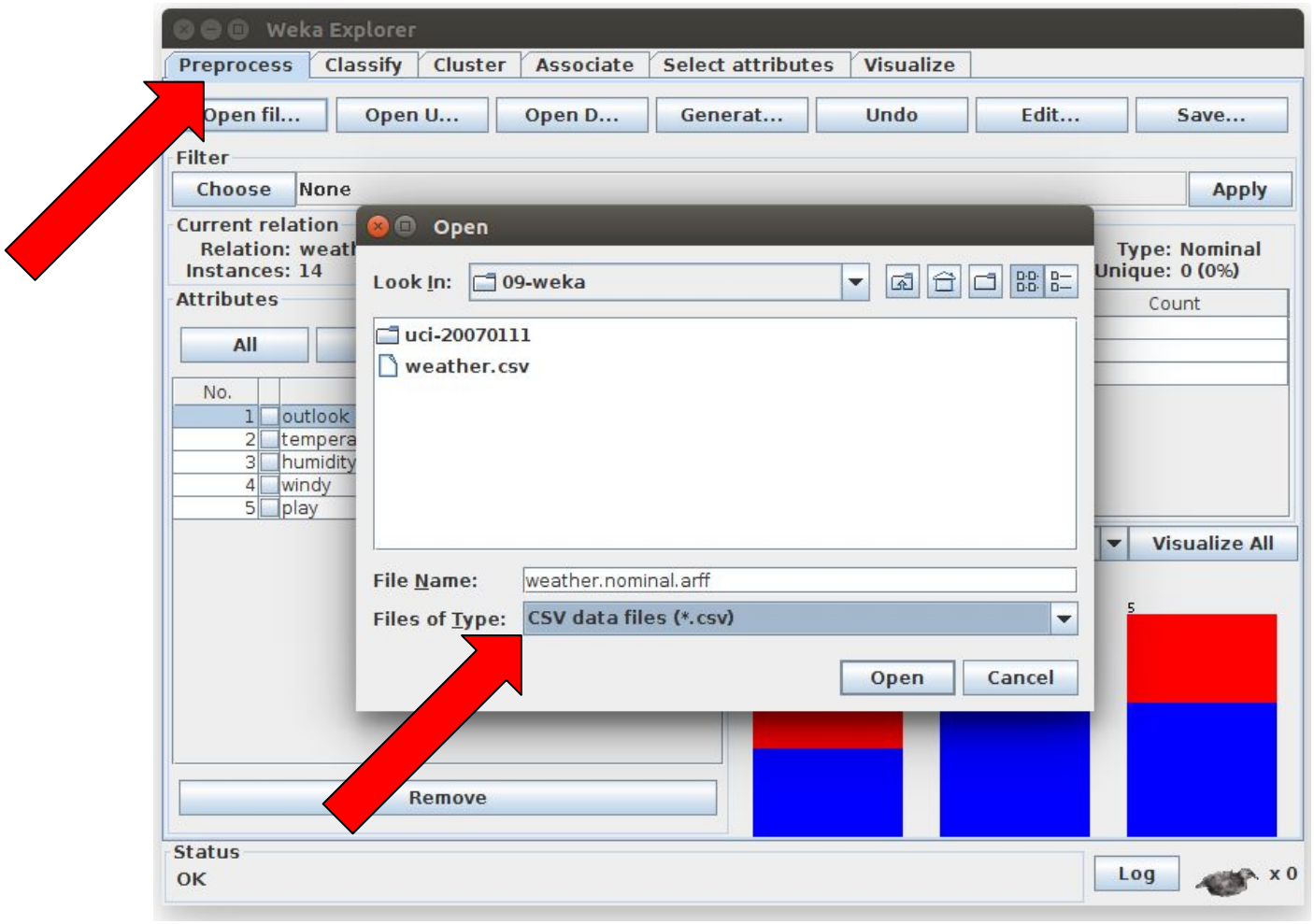
Excel => CSV

	A	B	C	D	E
1	outlook	temperature	humidity	windy	play
2	sunny	hot	high	FALSE	no
3	sunny	hot	high	TRUE	no
4	overcast	hot	high	FALSE	yes
5	rainy	mild	high	FALSE	yes
6	rainy	cool	normal	FALSE	yes
7	rainy	cool	normal	TRUE	no
8	overcast	cool	normal	TRUE	yes
9	sunny	mild	high	FALSE	no
10	sunny	cool	normal	FALSE	yes
11	rainy	mild	normal	FALSE	yes
12	sunny	mild	normal	TRUE	yes
13	overcast	mild	high	TRUE	yes
14	overcast	hot	normal	FALSE	yes
15	rainy	mild	high	TRUE	no

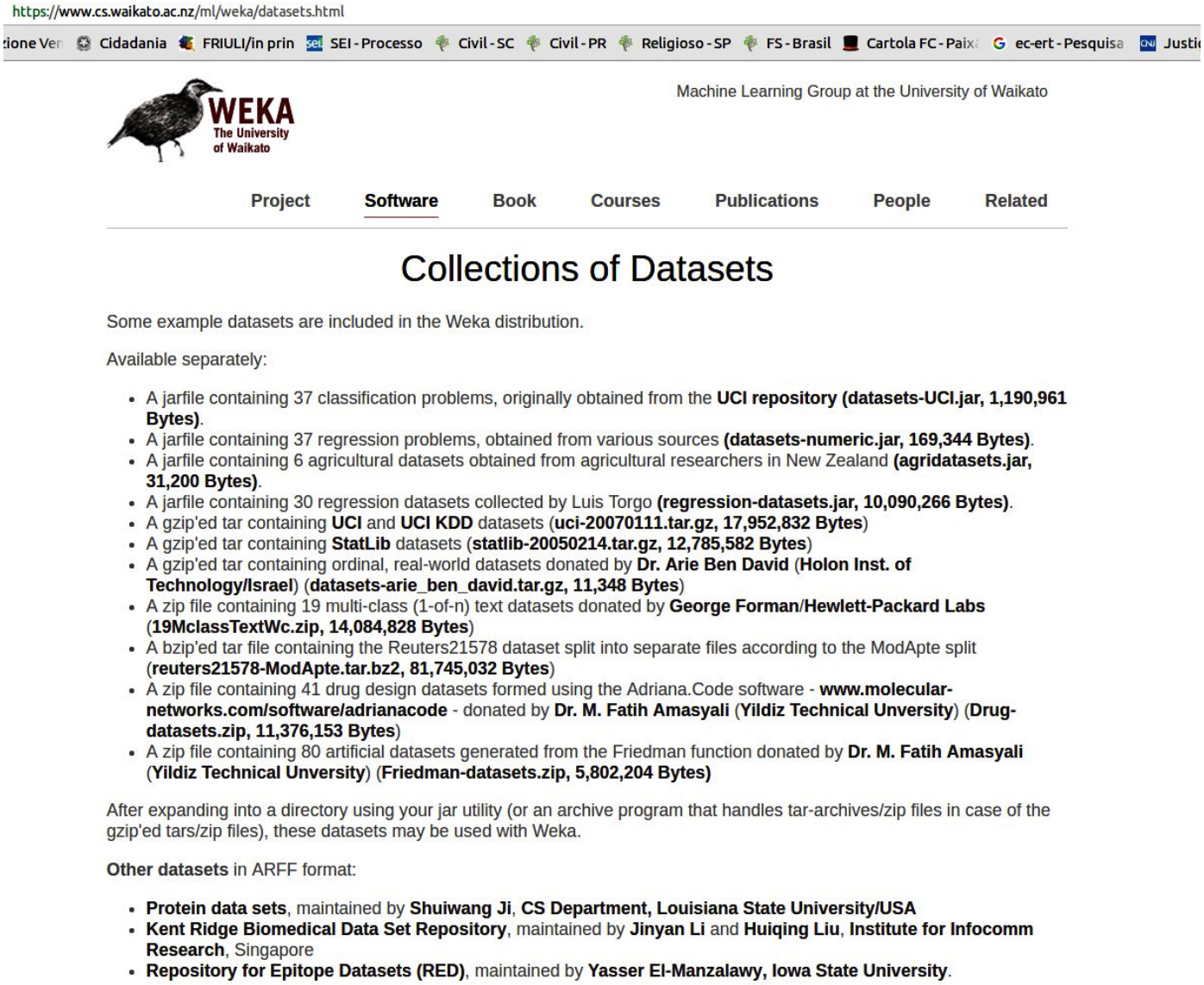
weather.csv

```
outlook,temperature,humidity,windy,play
sunny,hot,high,FALSE,no
sunny,hot,high,TRUE,no
overcast,hot,high,FALSE,yes
rainy,mild,high,FALSE,yes
rainy,cool,normal,FALSE,yes
rainy,cool,normal,TRUE,no
overcast,cool,normal,TRUE,yes
sunny,mild,high,FALSE,no
sunny,cool,normal,FALSE,yes
rainy,mild,normal,FALSE,yes
sunny,mild,normal,TRUE,yes
overcast,mild,high,TRUE,yes
overcast,hot,normal,FALSE,yes
rainy,mild,high,TRUE,no
```

Excel => CSV


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: weath
Instances: 14
Attributes
All
No.
1 outlook
2 tempera
3 humidity
4 windy
5 play
Type: Nominal
Unique: 0 (0%)
Count
Visualize All
Remove
Status
OK
Log
x 0
Open
Look In:
09-weka
uci-20070111
weather.csv
File Name:
weather.nominal.arff
Files of Type:
CSV data files (*.csv)
Open
Cancel

Software > Datasets



https://www.cs.waikato.ac.nz/ml/weka/datasets.html

Cidadania FRIULI/in prin SEI - Processo Civil - SC Civil - PR Religioso - SP FS - Brasil Cartola FC - Paix ec-ert - Pesquisa

WEKA The University of Waikato

Machine Learning Group at the University of Waikato

Project Software Book Courses Publications People Related

Collections of Datasets

Some example datasets are included in the Weka distribution.

Available separately:

- A jarfile containing 37 classification problems, originally obtained from the **UCI repository (datasets-UCI.jar, 1,190,961 Bytes)**.
- A jarfile containing 37 regression problems, obtained from various sources **(datasets-numeric.jar, 169,344 Bytes)**.
- A jarfile containing 6 agricultural datasets obtained from agricultural researchers in New Zealand **(agridatasets.jar, 31,200 Bytes)**.
- A jarfile containing 30 regression datasets collected by Luis Torgo **(regression-datasets.jar, 10,090,266 Bytes)**.
- A gzip'ed tar containing **UCI** and **UCI KDD** datasets (**uci-20070111.tar.gz, 17,952,832 Bytes**)
- A gzip'ed tar containing **StatLib** datasets (**statlib-20050214.tar.gz, 12,785,582 Bytes**)
- A gzip'ed tar containing ordinal, real-world datasets donated by **Dr. Arie Ben David** (**Holon Inst. of Technology/Israel**) (**datasets-arie_ben_david.tar.gz, 11,348 Bytes**)
- A zip file containing 19 multi-class (1-of-n) text datasets donated by **George Forman/Hewlett-Packard Labs** (**19MclassTextWc.zip, 14,084,828 Bytes**)
- A bzip'ed tar file containing the Reuters21578 dataset split into separate files according to the ModApte split (**reuters21578-ModApte.tar.bz2, 81,745,032 Bytes**)
- A zip file containing 41 drug design datasets formed using the Adriana.Code software - **www.molecular-networks.com/software/adrianacode** - donated by **Dr. M. Fatih Amasyali** (**Yildiz Technical Unversity**) (**Drug-datasets.zip, 11,376,153 Bytes**)
- A zip file containing 80 artificial datasets generated from the Friedman function donated by **Dr. M. Fatih Amasyali** (**Yildiz Technical Unversity**) (**Friedman-datasets.zip, 5,802,204 Bytes)**

After expanding into a directory using your jar utility (or an archive program that handles tar-archives/zip files in case of the gzip'ed tars/zip files), these datasets may be used with Weka.

Other datasets in ARFF format:

- **Protein data sets**, maintained by **Shuiwang Ji**, **CS Department, Louisiana State University/USA**
- **Kent Ridge Biomedical Data Set Repository**, maintained by **Jinyan Li** and **Huiqing Liu**, **Institute for Infocomm Research**, Singapore
- **Repository for Epitope Datasets (RED)**, maintained by **Yasser El-Manzalawy, Iowa State University**.

WEKA Datasets

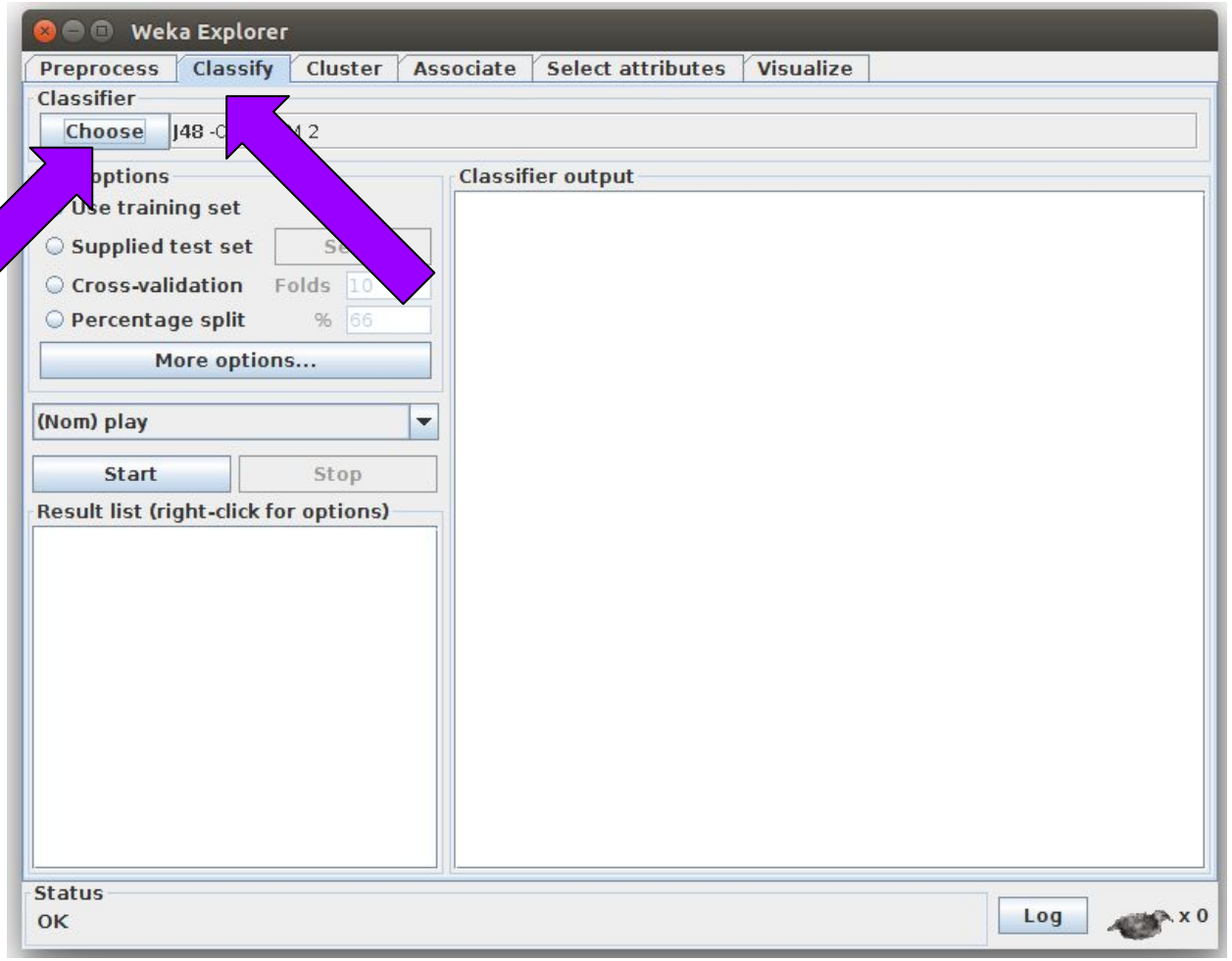
- Alguns datasets em formato **ARFF**
  http://storm.cis.fordham.edu/~gweiss/data-mining/datasets.html

- contact-lens.arff
- cpu.arff
- cpu.with-vendor.arff
- diabetes.arff
- glass.arff
- ionospehre.arff
- iris.arff
- labor.arff
- ReutersCorn-train.arff
- ReutersCorn-test.arff
- ReutersGrain-train.arff
- ReutersGrain-test.arff
- segment-challenge.arff
- segment-test.arff
- soybean.arff
- supermarket.arff
- vote.arff
- weather.arff
- weather.nominal.arff

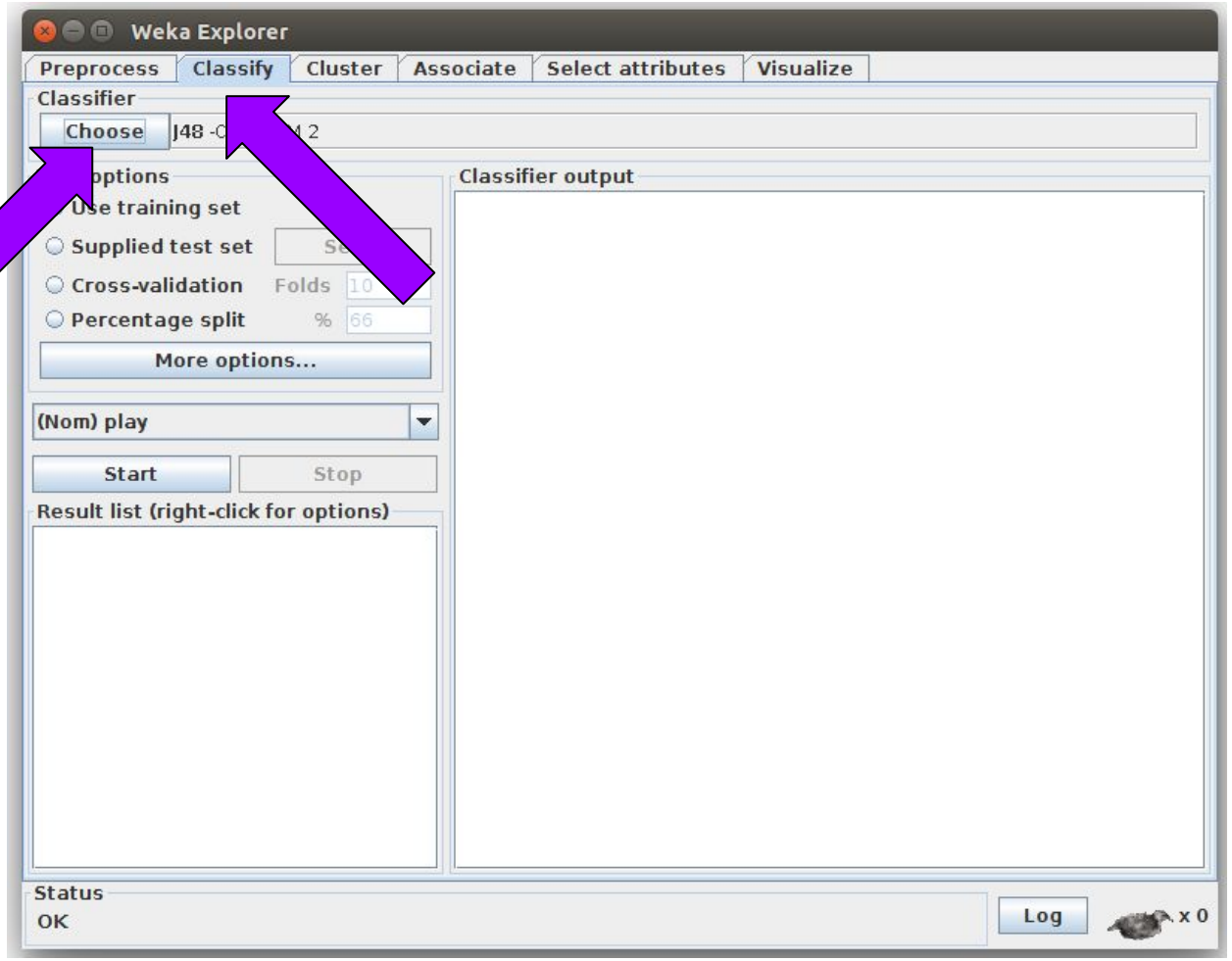
Classificação

- Como gerar:
  - uma árvore de decisão J48
  - um k-NN
  - Naive Bayes classifier
  - MLP
  - SVM
  - PCA

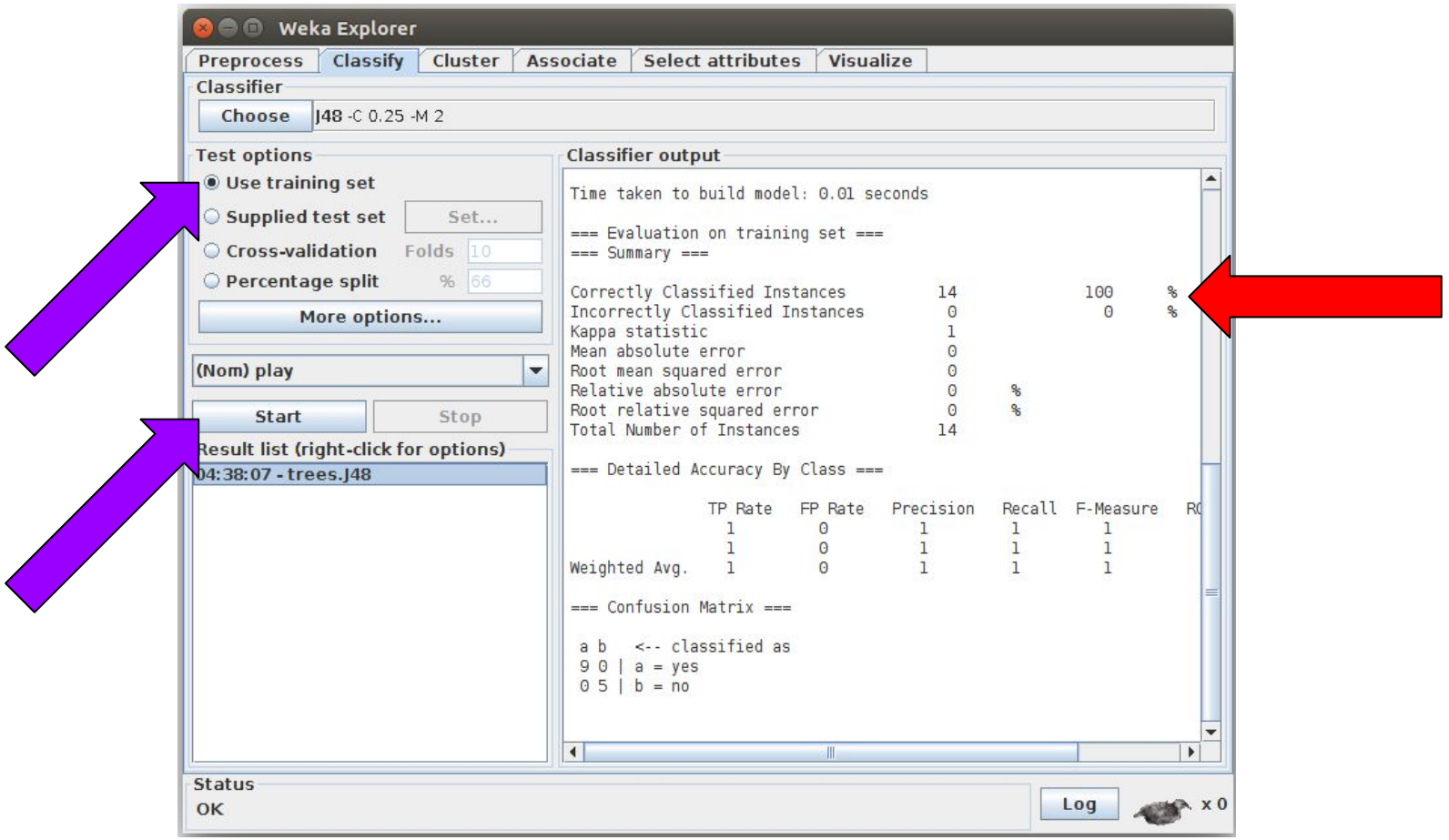
Classify > Choose



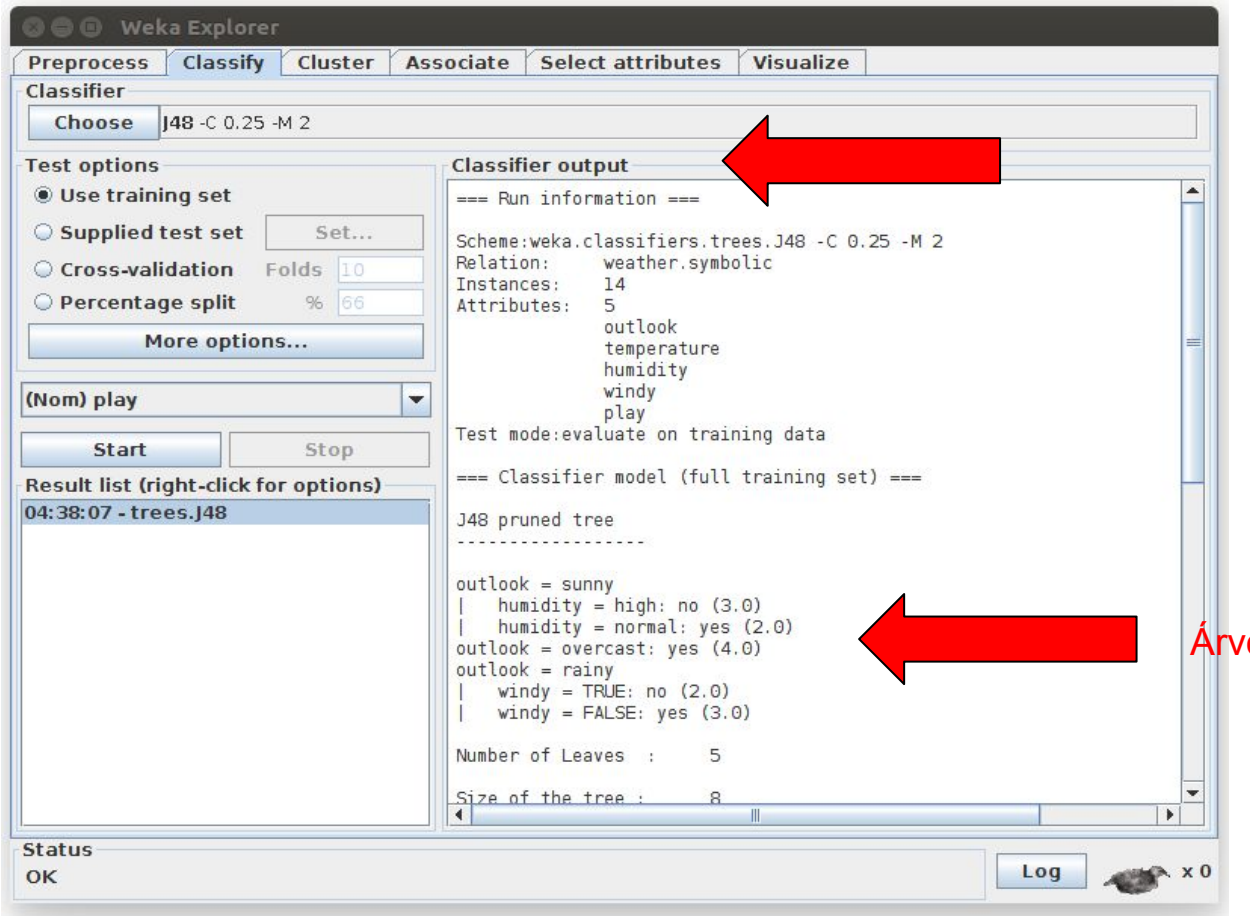
Classify > tree > J48



Classify > tree > J48

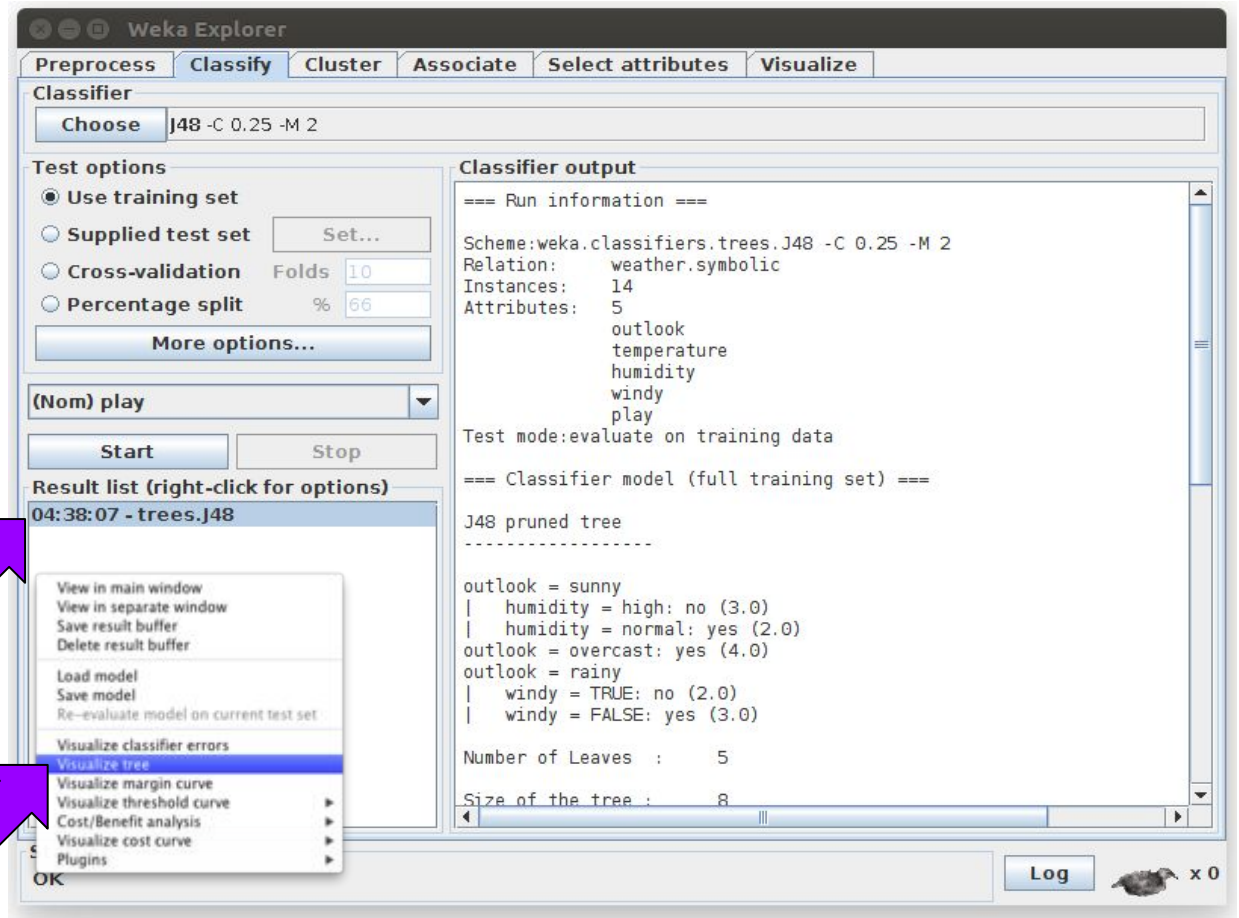


Classifier output

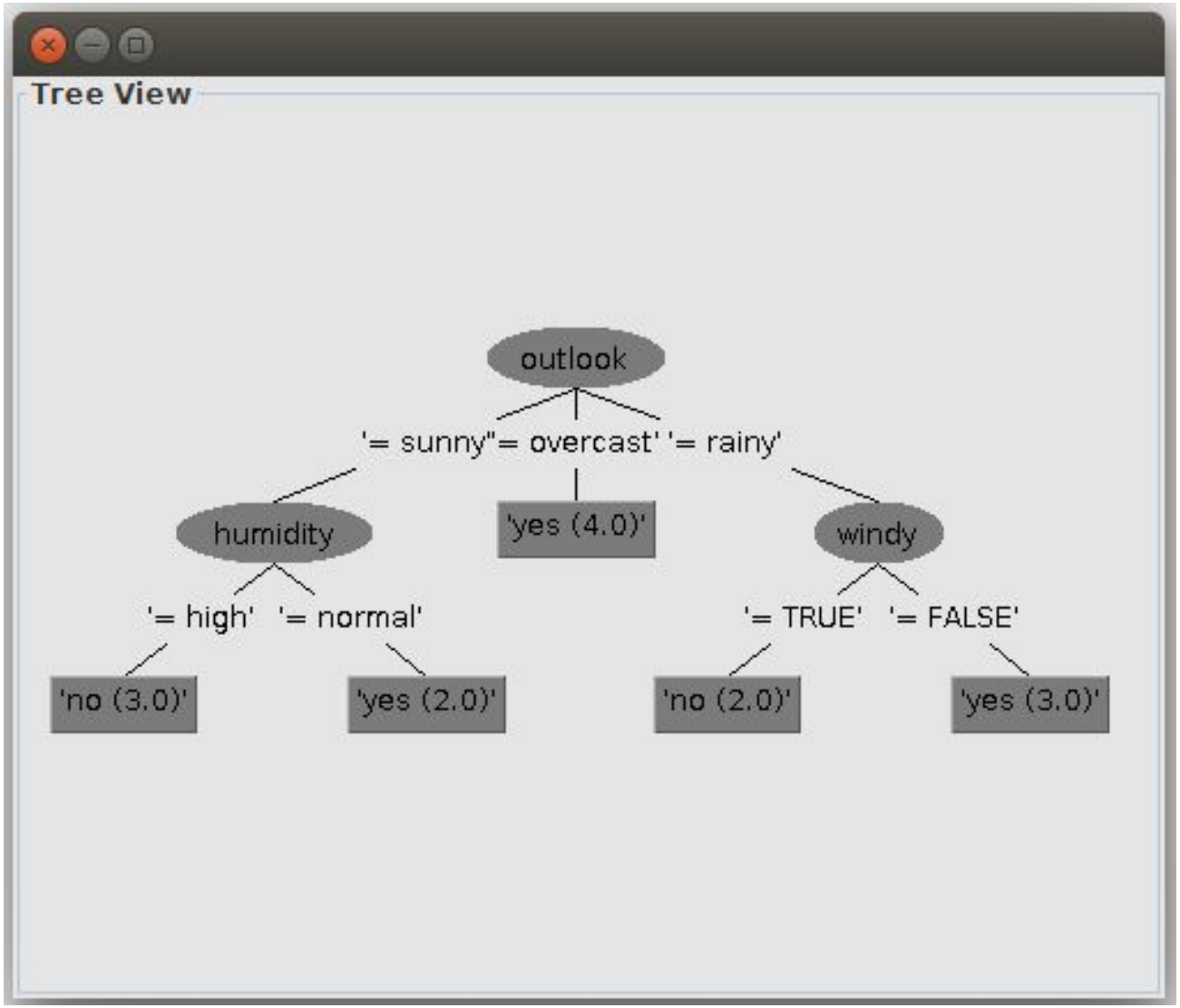


Árvore / Regras Geradas

Visualize


Clique com botão direito
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) play
Start
Stop
Result list (right-click for options)
04:38:07 - trees.J48
View in main window
View in separate window
Save result buffer
Delete result buffer
Load model
Save model
Re-evaluate model on current test set
Visualize classifier errors
Visualize tree
Visualize margin curve
Visualize threshold curve
Cost/Benefit analysis
Visualize cost curve
Plugins
Classifier output
=== Run information ===
Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: weather.symbolic
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
play
Test mode:evaluate on training data
=== Classifier model (full training set) ===
J48 pruned tree
outlook = sunny
| humidity = high: no (3.0)
| humidity = normal: yes (2.0)
outlook = overcast: yes (4.0)
outlook = rainy
| windy = TRUE: no (2.0)
| windy = FALSE: yes (3.0)
Number of Leaves : 5
Size of the tree : 8
OK
Log
x 0

Visualize


Tree View
outlook
'= sunny"= overcast' '= rainy'
humidity
'yes (4.0)'
windy
'= high' '= normal'
'= TRUE' '= FALSE'
'no (3.0)'
'yes (2.0)'
'no (2.0)'
'yes (3.0)'

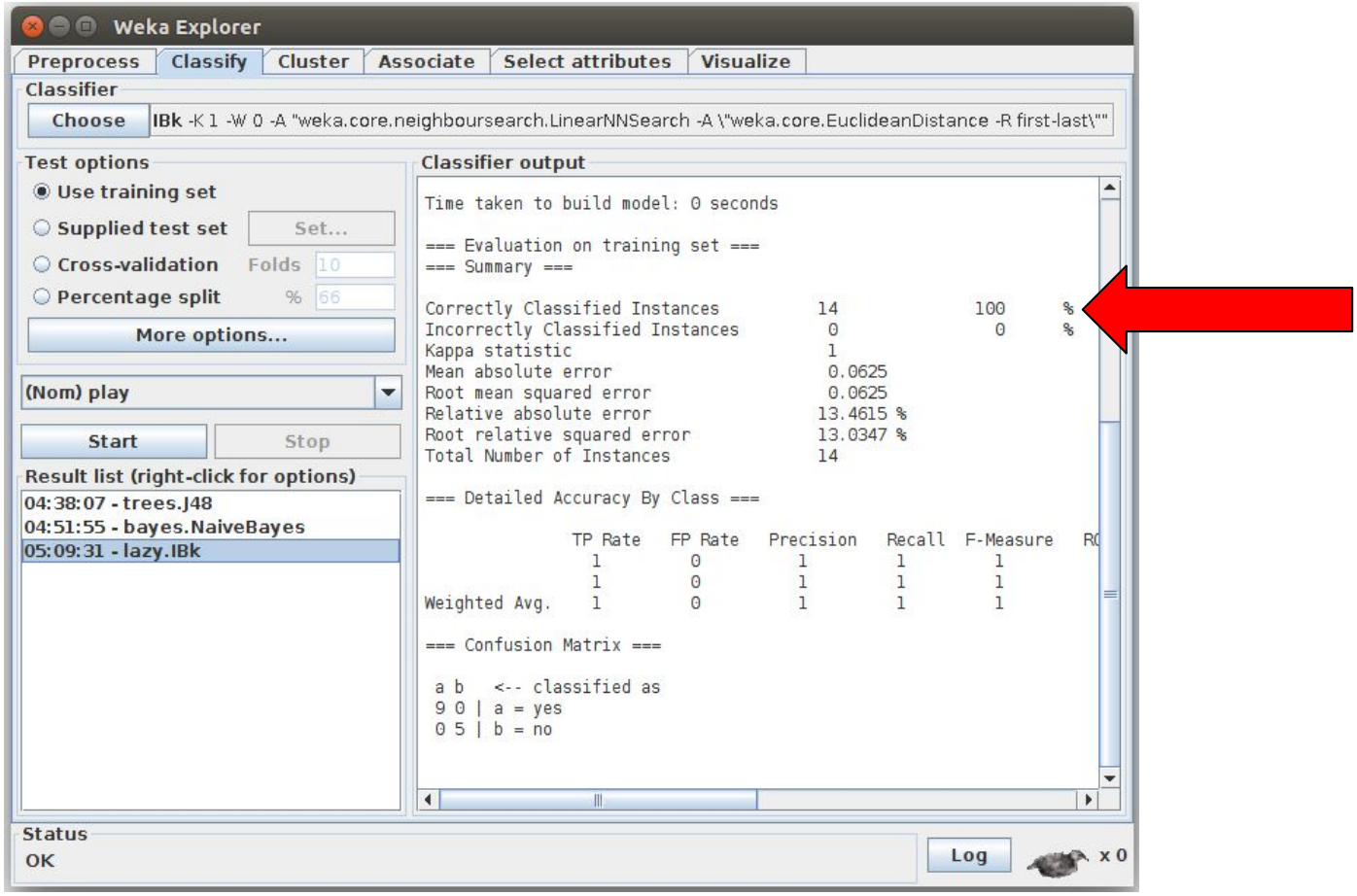
Código em Java

```
import java.awt.BorderLayout;
import java.io.BufferedReader;
import java.io.FileReader;

import weka.classifiers.*;
import weka.classifiers.trees.J48;
import weka.core.Instances;
import weka.gui.treevisualizer.PlaceNode2;
import weka.gui.treevisualizer.TreeVisualizer;

public class WekaJ48 {
public static void main(String args[]) throws Exception {
    // train classifier
    J48 cls = new J48();
    Instances data = new Instances(new BufferedReader(new FileReader("D:\\sample.arff")));
    data.setClassIndex(data.numAttributes() - 1);
    cls.buildClassifier(data);
```

Código em Java

```
    // display classifier
    final javax.swing.JFrame jf =
      new javax.swing.JFrame("Weka Classifier Tree Visualizer: J48");
    jf.setSize(500,400);
    jf.getContentPane().setLayout(new BorderLayout());
    TreeVisualizer tv = new TreeVisualizer(null,
        cls.graph(),
        new PlaceNode2());
    jf.getContentPane().add(tv, BorderLayout.CENTER);
    jf.addWindowListener(new java.awt.event.WindowAdapter() {
      public void windowClosing(java.awt.event.WindowEvent e) {
        jf.dispose();
      }
    });

    jf.setVisible(true);
    tv.fitToScreen();
  }
}
```

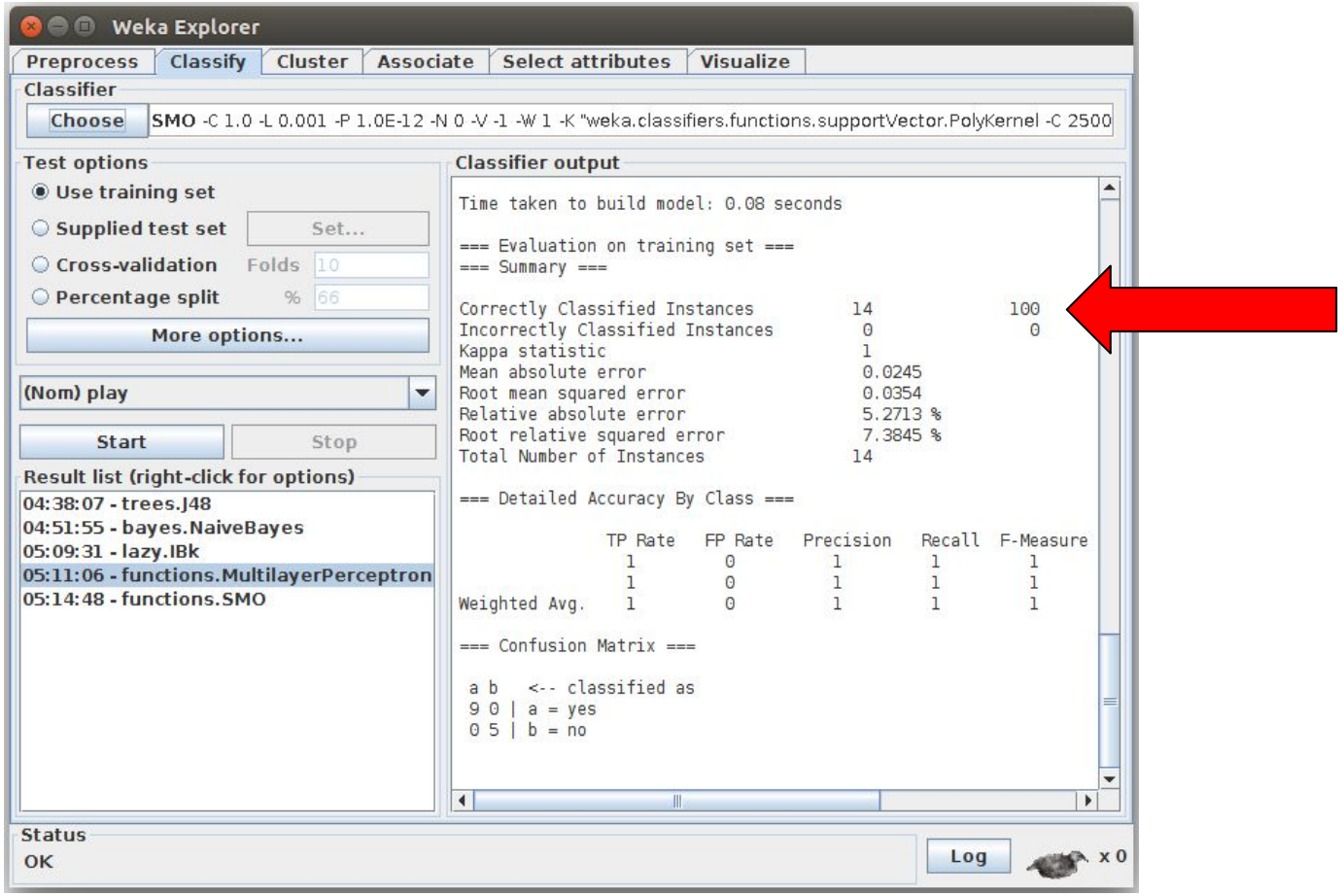
Classify > Lazy > k-NN (IBk)



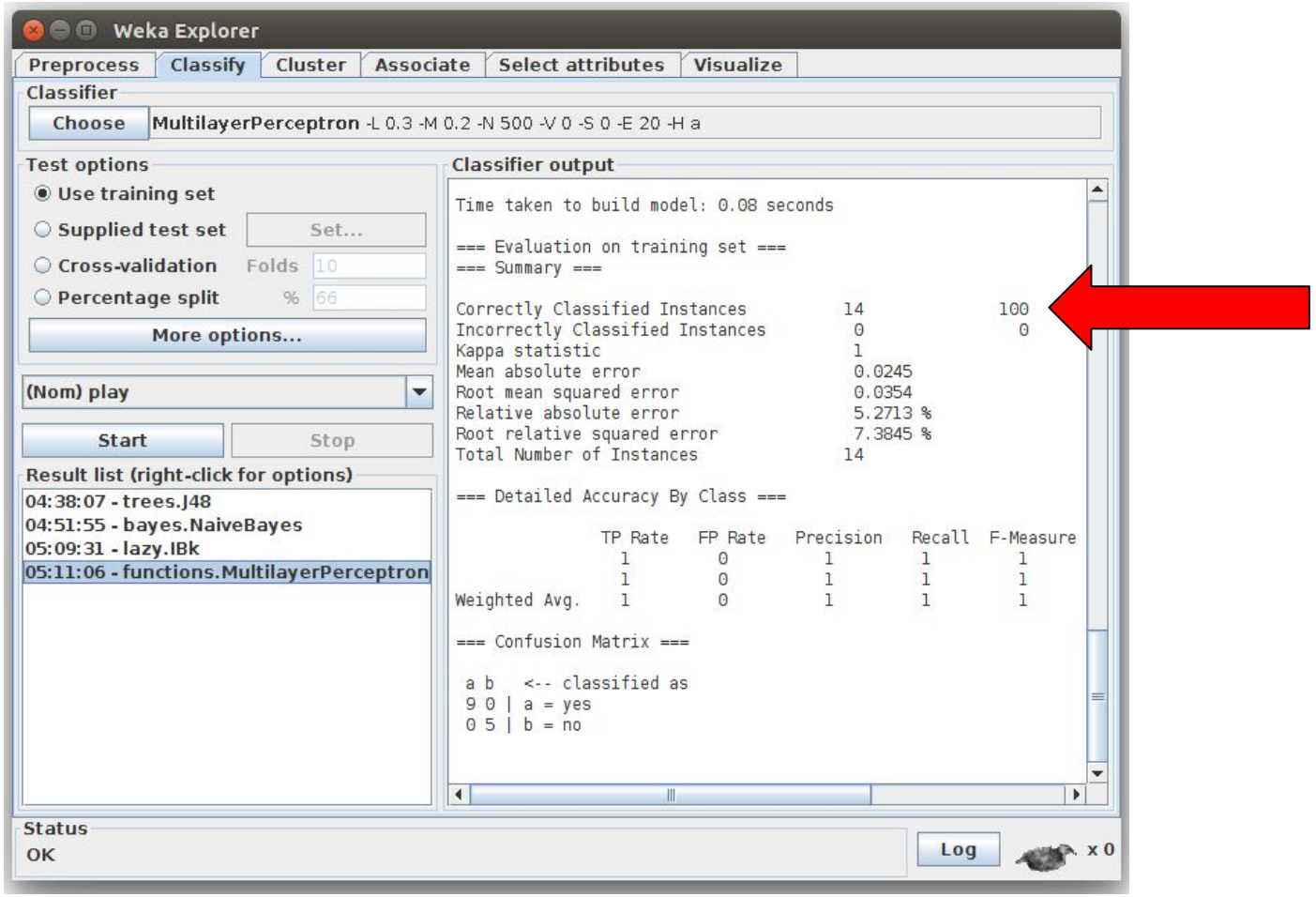
Classify > function > SVM

- Deve-se instalar o LibSVM
  - LIBSVM - A Library for Support Vector Machines
  - https://www.csie.ntu.edu.tw/~cjlin/libsvm/

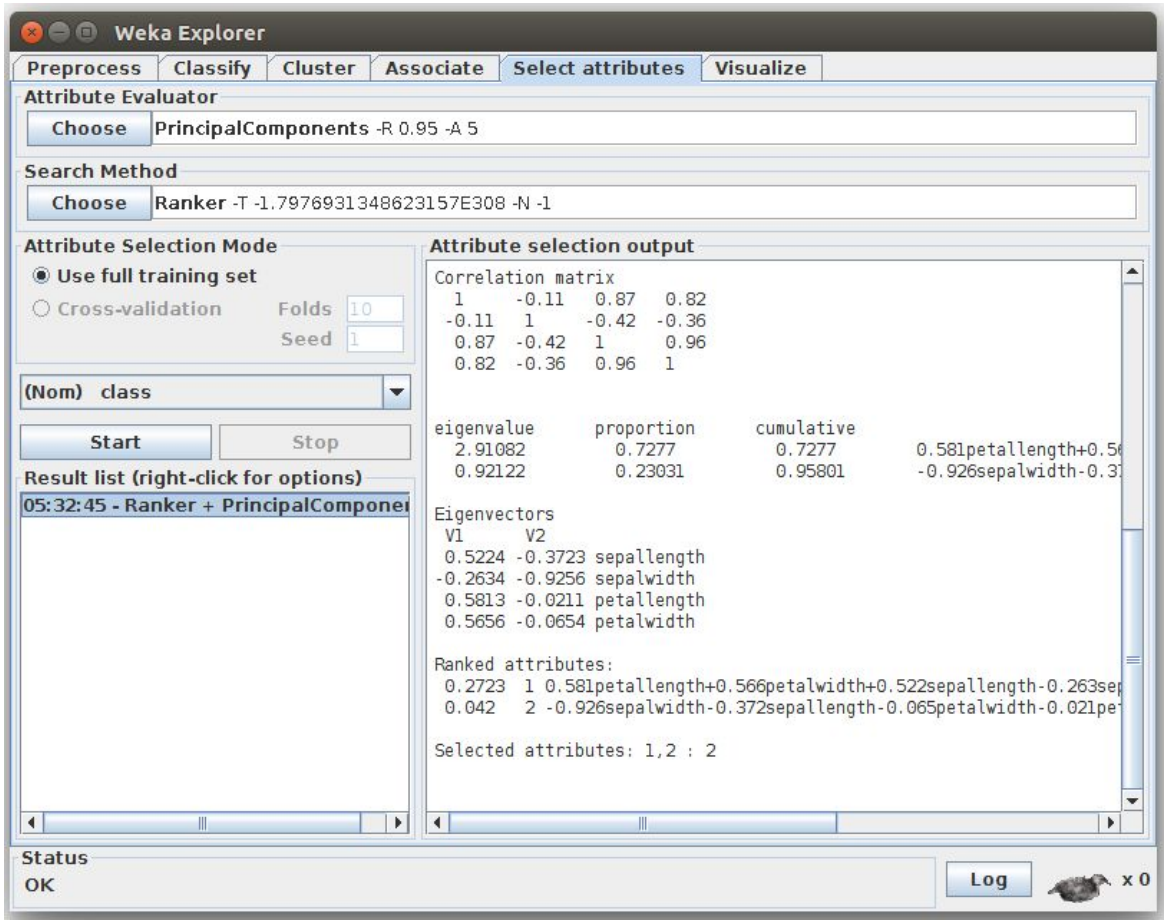
Classify > function > SMO



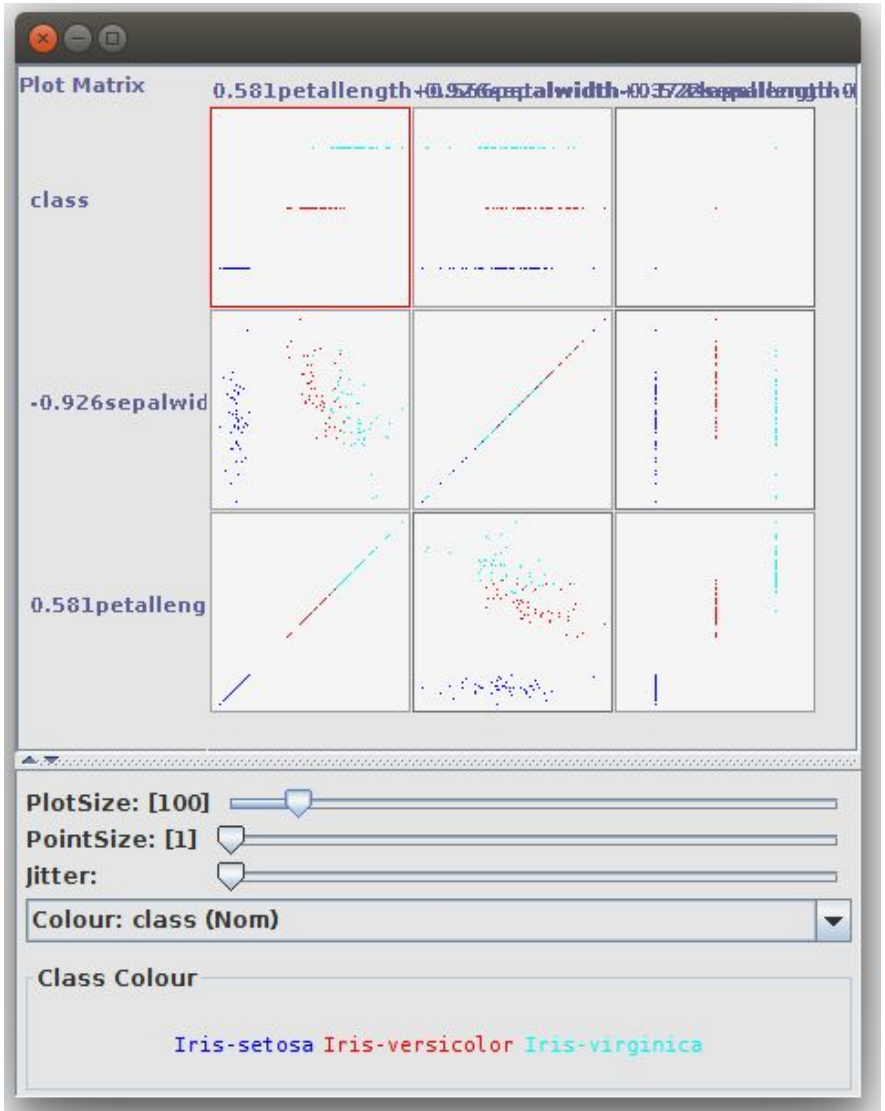
Classify > function > MLP



Select Attributes > PCA



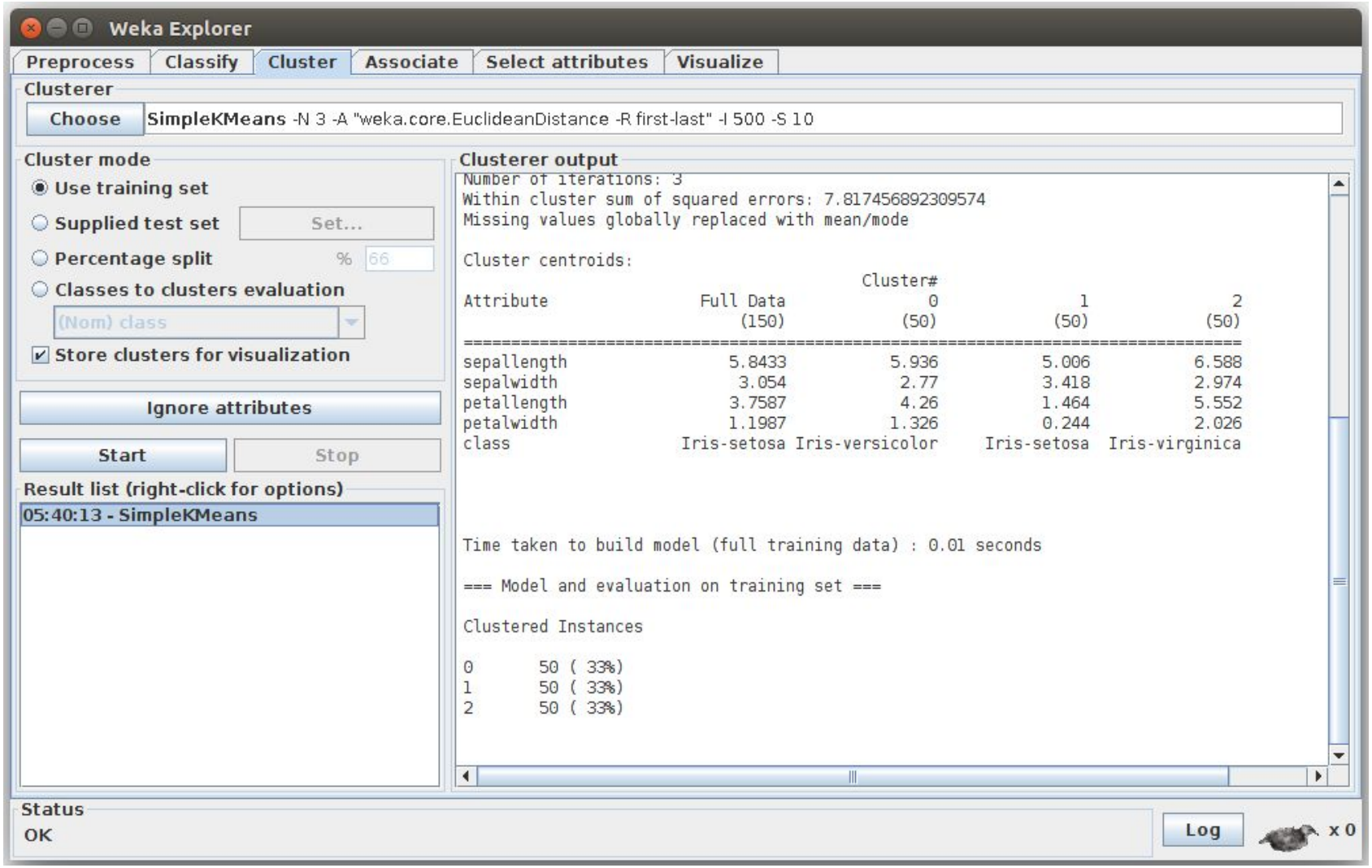
Visualizar


Plot Matrix
0.581petallength
class
-0.926sepalwid
0.581petalleng
PlotSize: [100]
PointSize: [1]
Jitter:
Colour: class (Nom)
Class Colour
Iris-setosa Iris-versicolor Iris-virginica

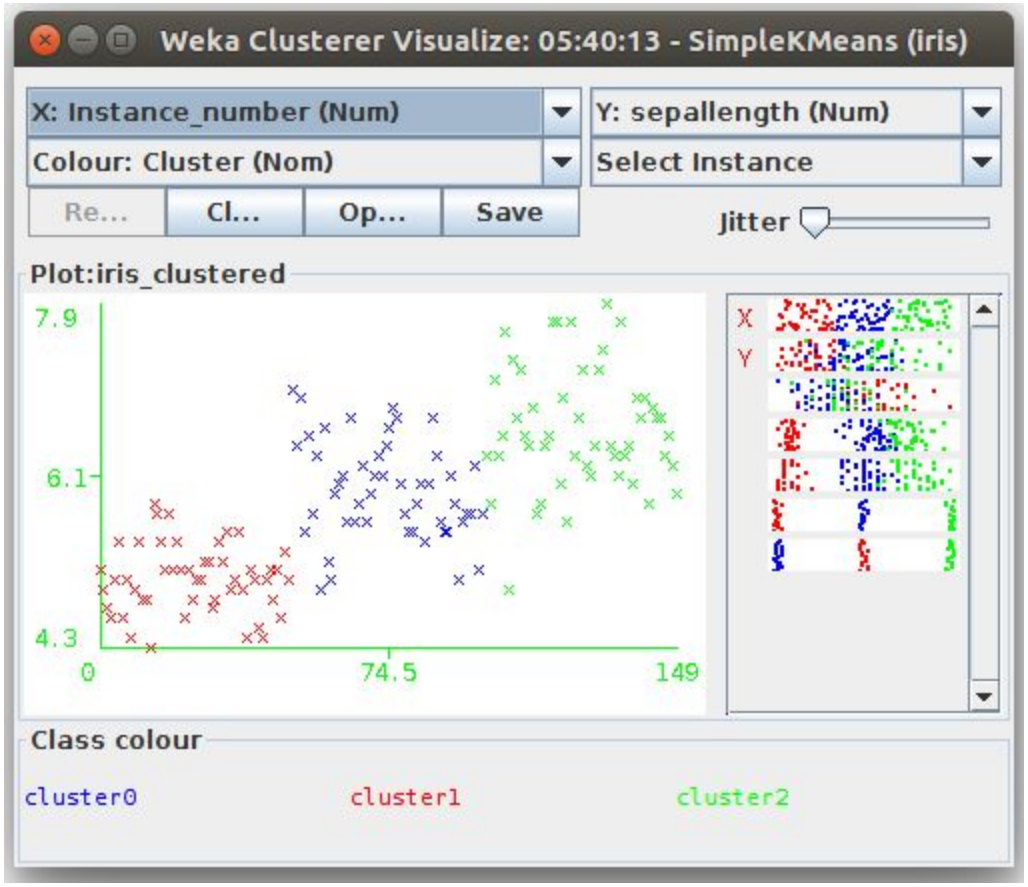
Clustering & Regressão

- Como gerar:
  - Um kMeans
  - Uma regressão linear

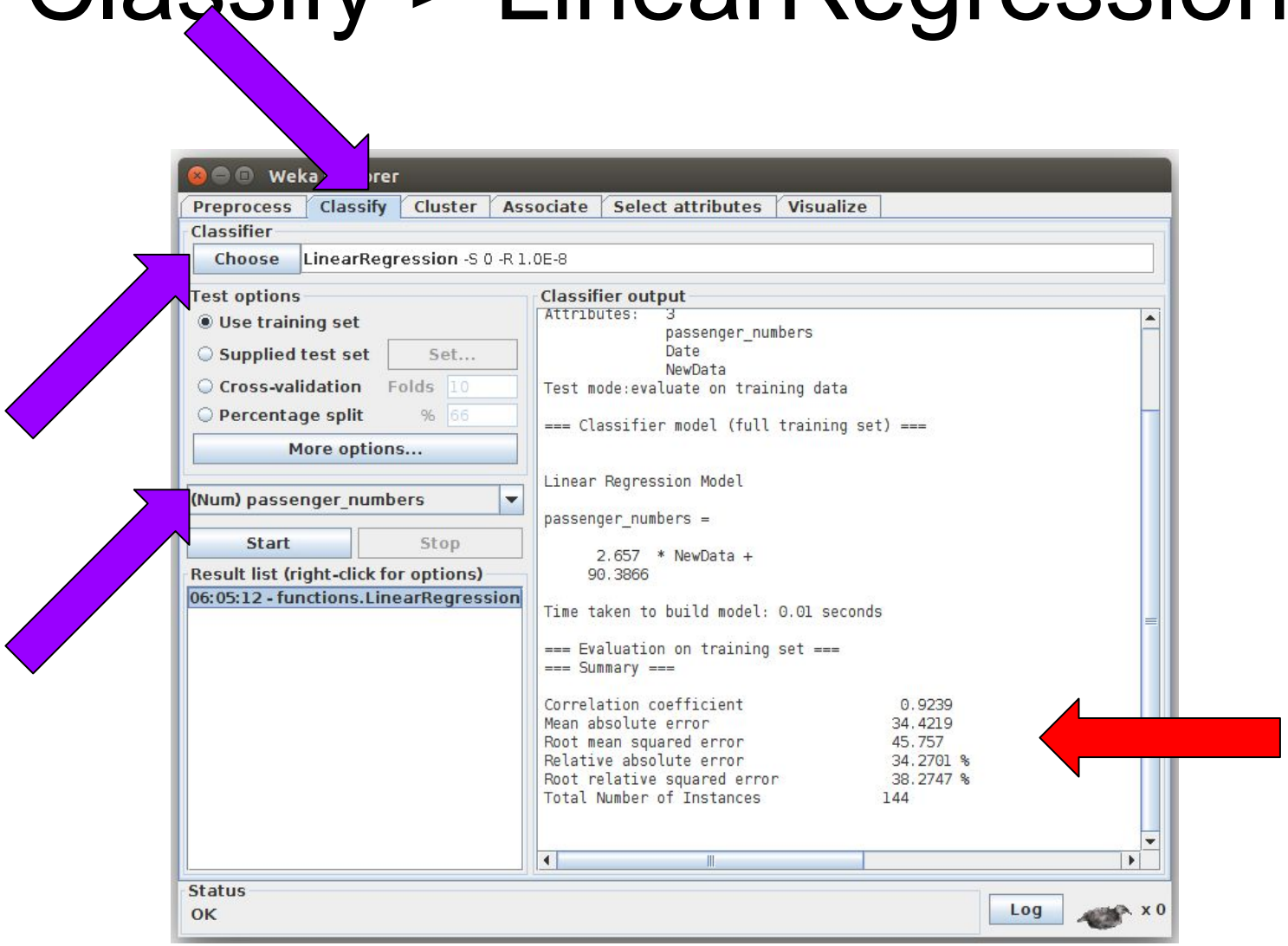
Cluster > SimpleKmeans



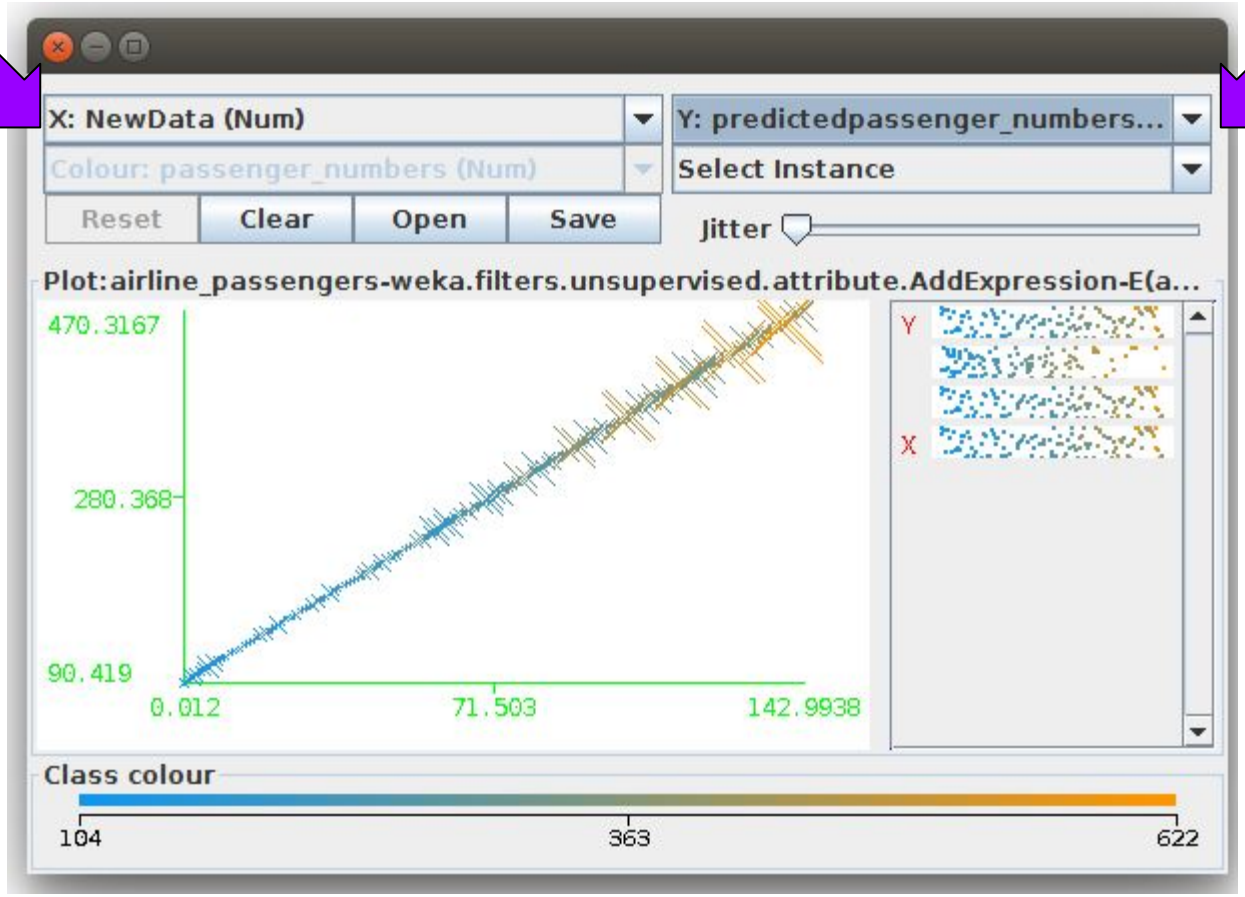
Cluster > SimpleKmeans



Classify > LinearRegression



Classify > LinearRegression



References

- Weka 3: Data Mining Software in Java
  - https://www.cs.waikato.ac.nz/ml/weka/

- Weka Datasets
  - http://storm.cis.fordham.edu/~gweiss/data-mining/datasets.html