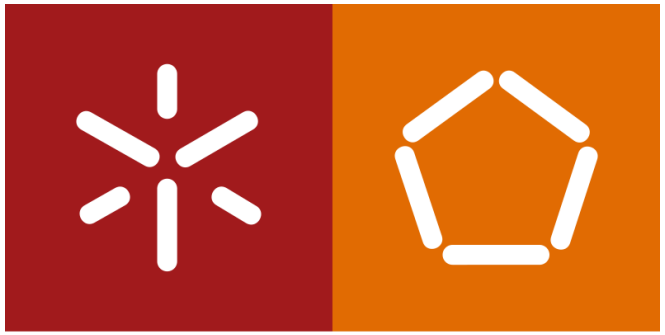


Universidade do Minho
Escola de Engenharia

S istemas de **A** prendizagem e **E** xtração de **C** onhecimento

José Machado

Diana Ferreira

PROGRAMA PRÁTICO

1 METODOLOGIA CRISP-DM

2 WEKA

3 PROCESSO DE DATA MINING

4 TRABALHOS PRÁTICOS

WEKA

Waikato Environment for Knowledge Analysis (WEKA):
É um *software* que permite pré-processar grandes volumes de dados, aplicar diferentes algoritmos de Machine Learning e comparar vários *outputs*.

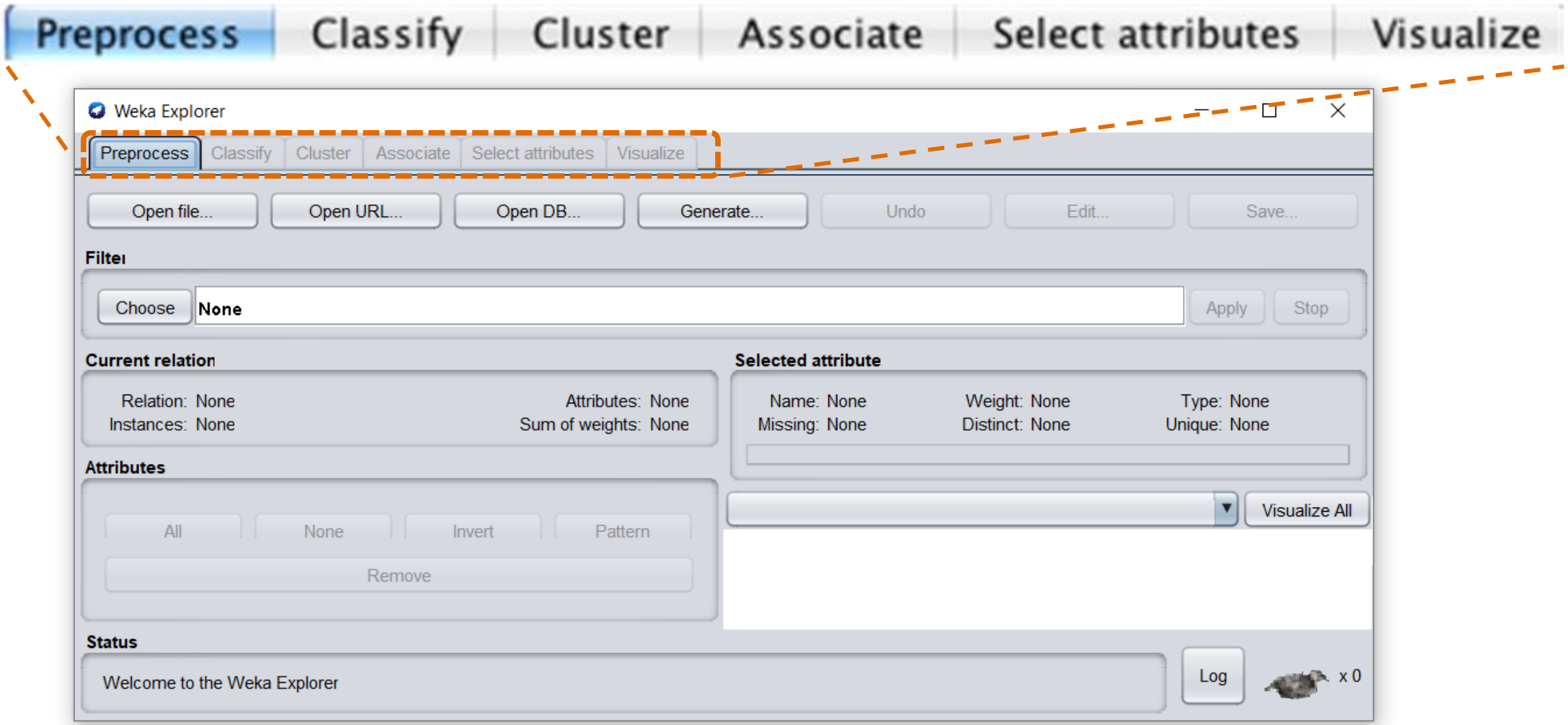
DOWNLOAD:

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

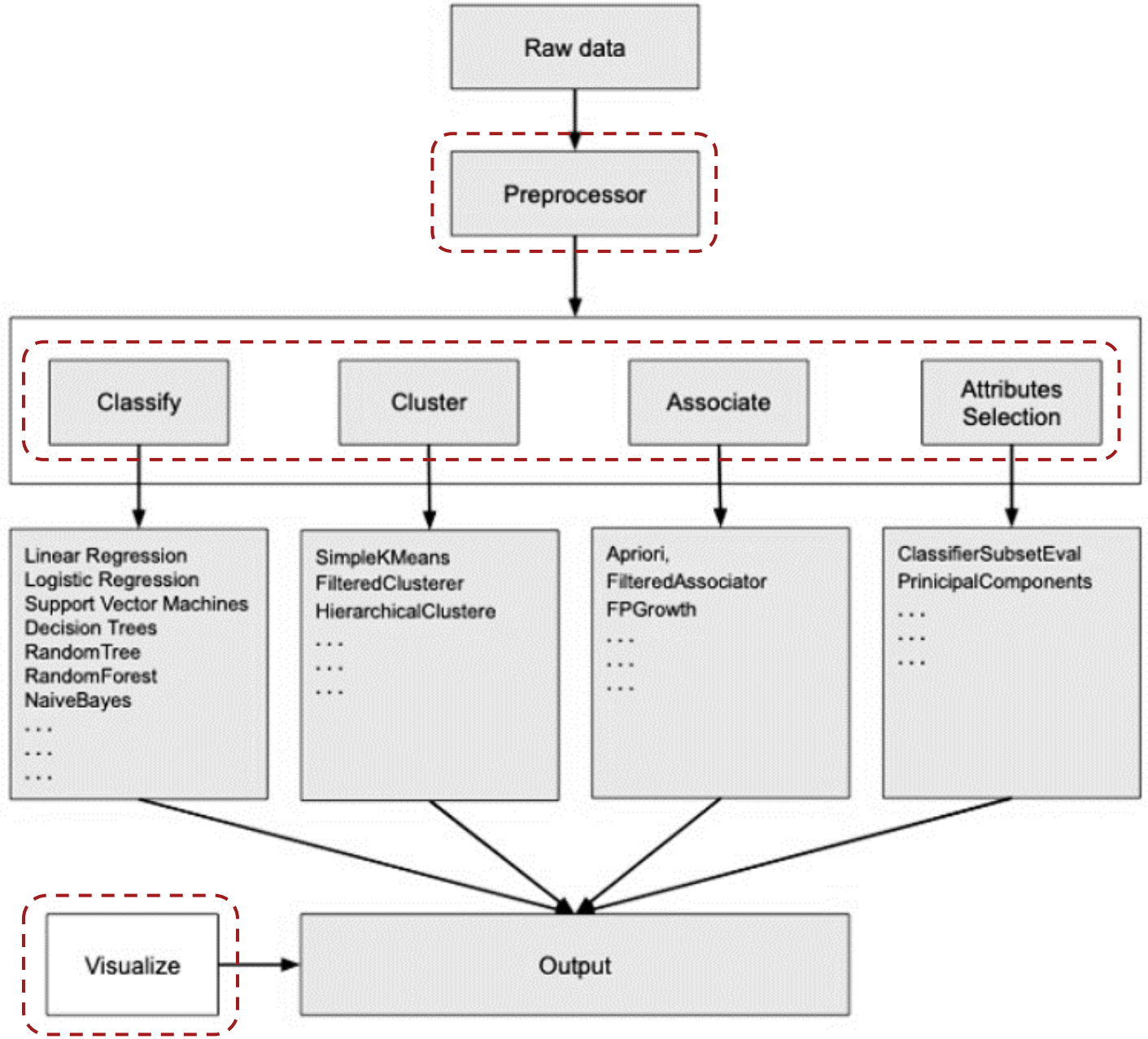
WEKA


Weka GUI Chooser
Program Visualization Tools Help
Applications
Explorer
Experimenter
KnowledgeFlow
Workbench
Simple CLI
WEKA
The University of Waikato
Waikato Environment for Knowledge Analysis
Version 3.8.4
(c) 1999 - 2019
The University of Waikato
Hamilton, New Zealand

WEKA

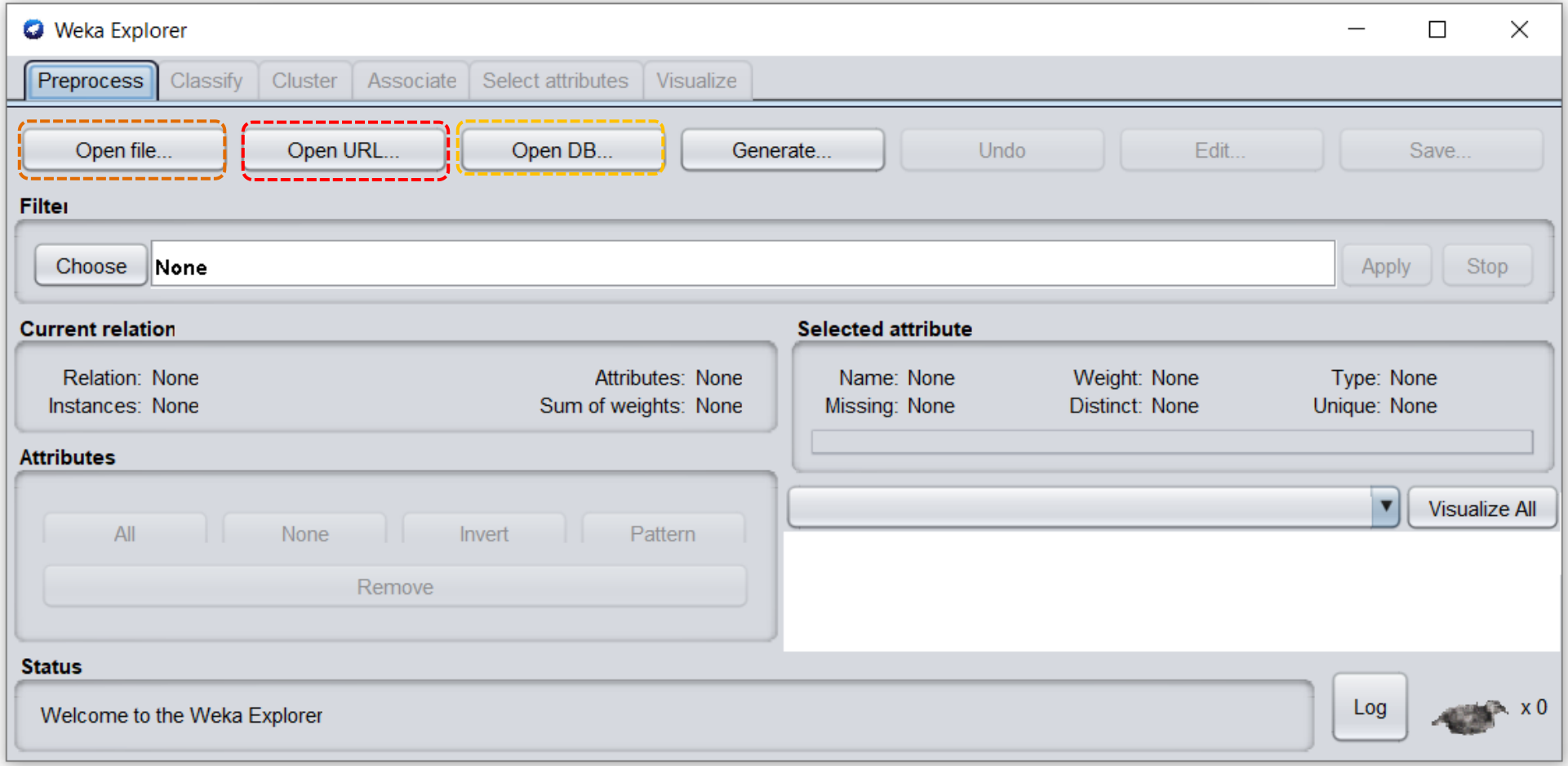

Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Stop
Current relation
Relation: None
Instances: None
Attributes: None
Sum of weights: None
Selected attribute
Name: None
Missing: None
Weight: None
Distinct: None
Type: None
Unique: None
Attributes
All
None
Invert
Pattern
Remove
Visualize All
Status
Welcome to the Weka Explorer
Log
x 0

WEKA

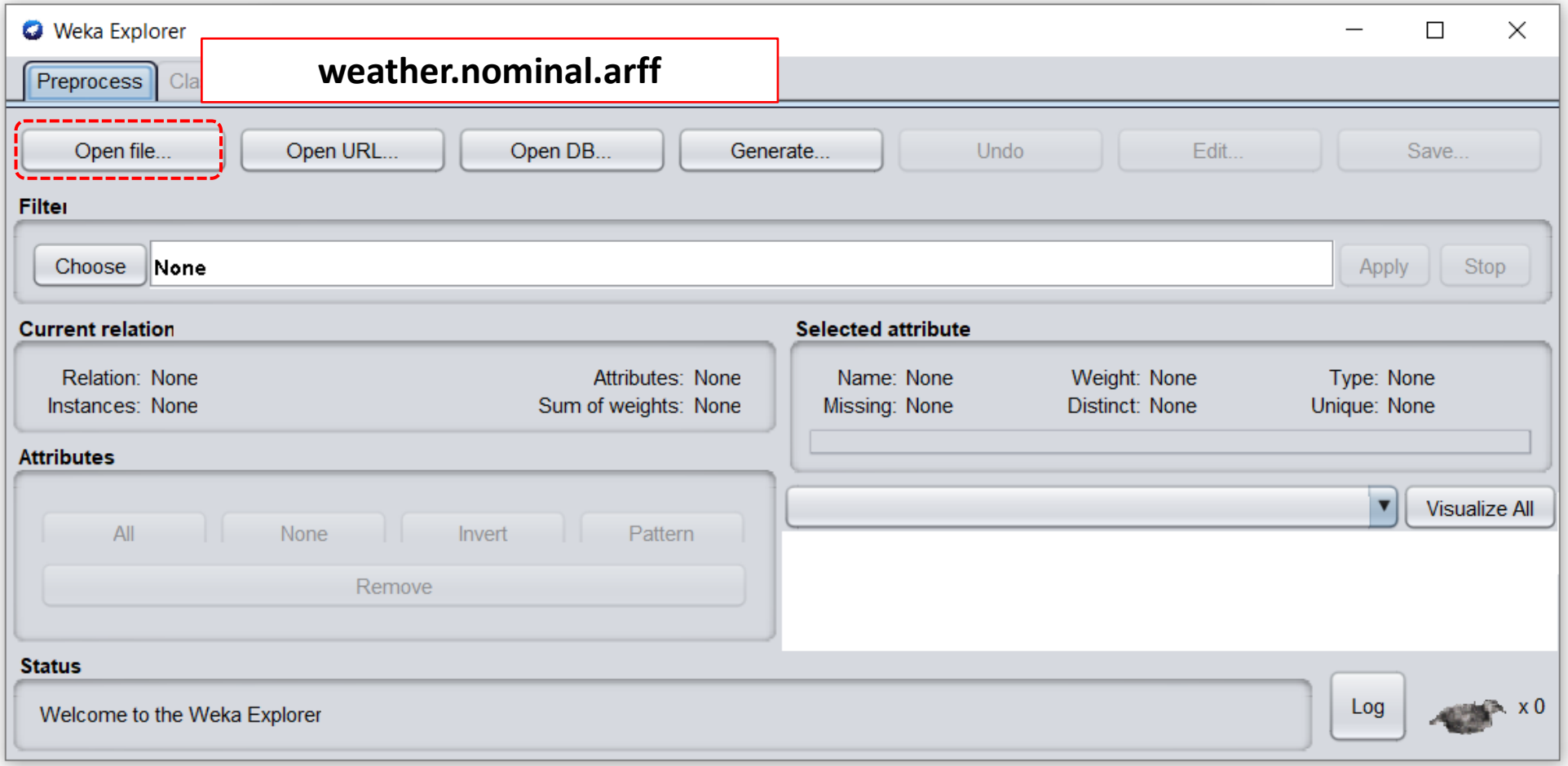

Raw data
Preprocessor
Classify
Cluster
Associate
Attributes Selection
Linear Regression
Logistic Regression
Support Vector Machines
Decision Trees
RandomTree
RandomForest
NaiveBayes
...
SimpleKMeans
FilteredClusterer
HierarchicalClustere
...
Apriori,
FilteredAssociator
FPGrowth
...
ClassifierSubsetEval
PrinicipalComponents
...
Visualize
Output

WEKA – carregar dados

- Ficheiro Local
- Web
- Base de Dados



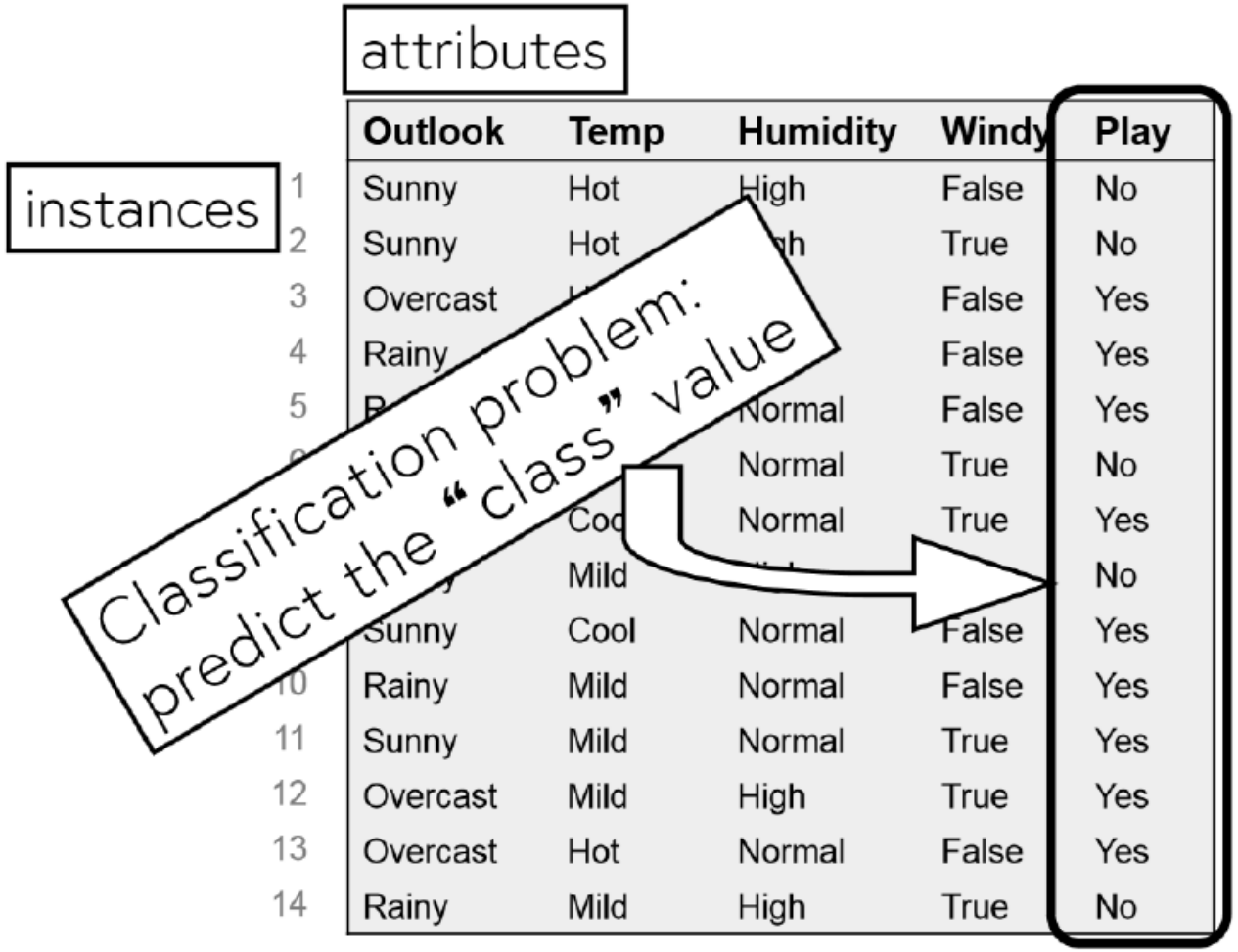
WEKA – carregar dados



*Os *datasets* estão guardados na pasta Data que está dentro da pasta de instalação do *software*
C:\Program Files\Weka-3-8-4\data

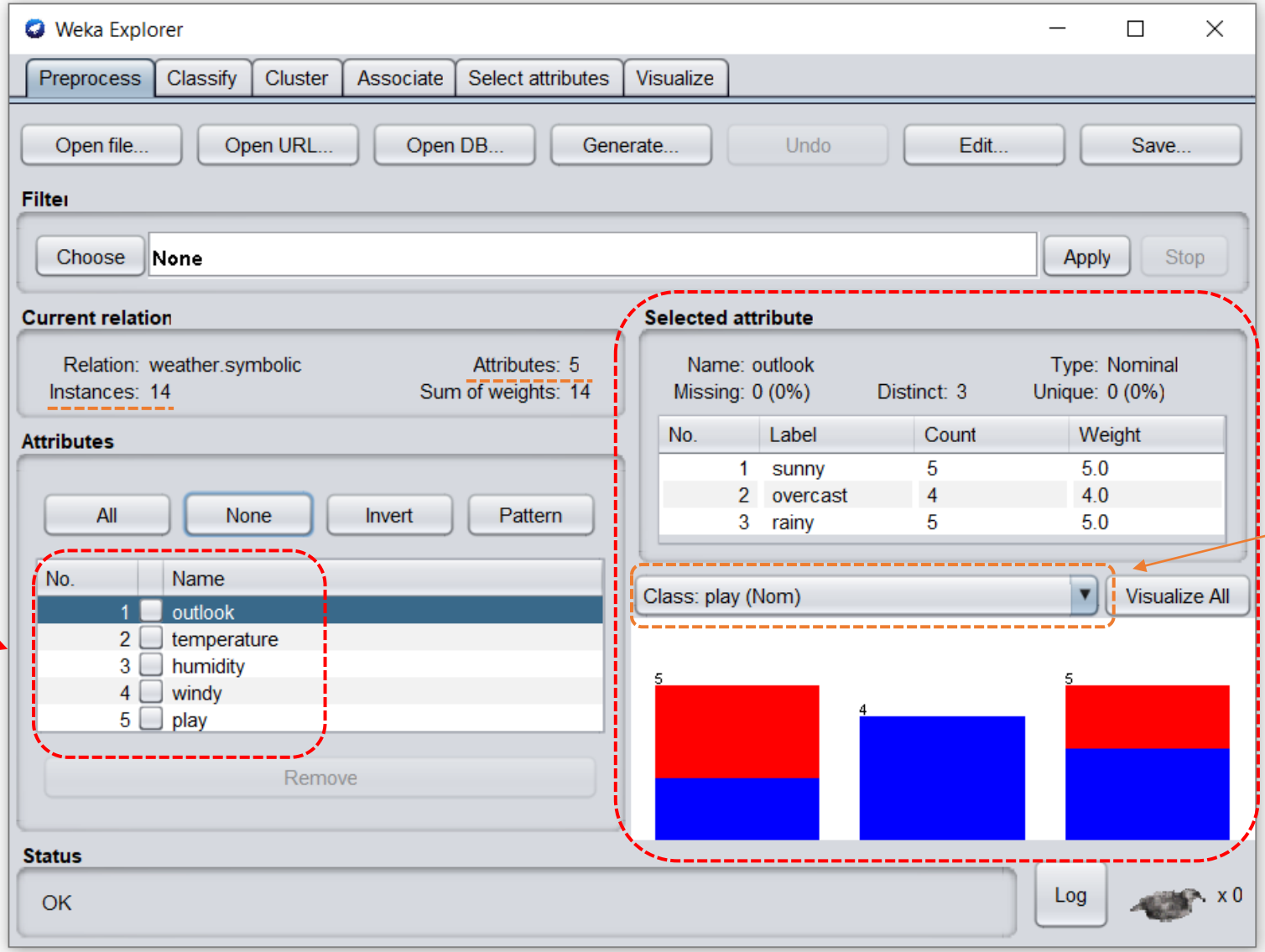
WEKA – carregar dados

Weather.arff



	Outlook	Temp	Humidity	Windy	Play
1	Sunny	Hot	High	False	No
2	Sunny	Hot	h	True	No
3	Overcast			False	Yes
4	Rainy			False	Yes
5			Normal	False	Yes
			Normal	True	No
		Coo	Normal	True	Yes
		Mild			No
	Sunny	Cool	Normal	False	Yes
10	Rainy	Mild	Normal	False	Yes
11	Sunny	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Rainy	Mild	High	True	No

WEKA – preprocess



Atributos

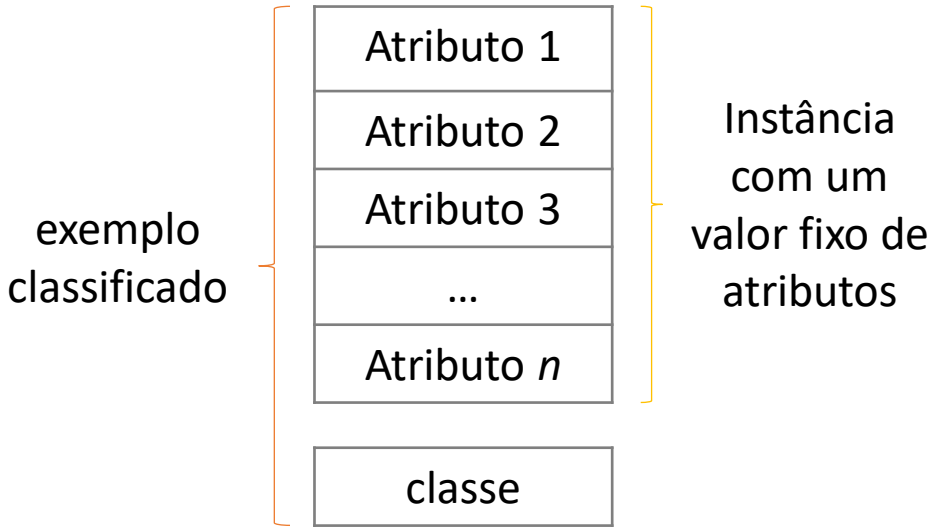
Classe

WEKA – preprocess

PROBLEMA DE CLASSIFICAÇÃO (supervised learning)

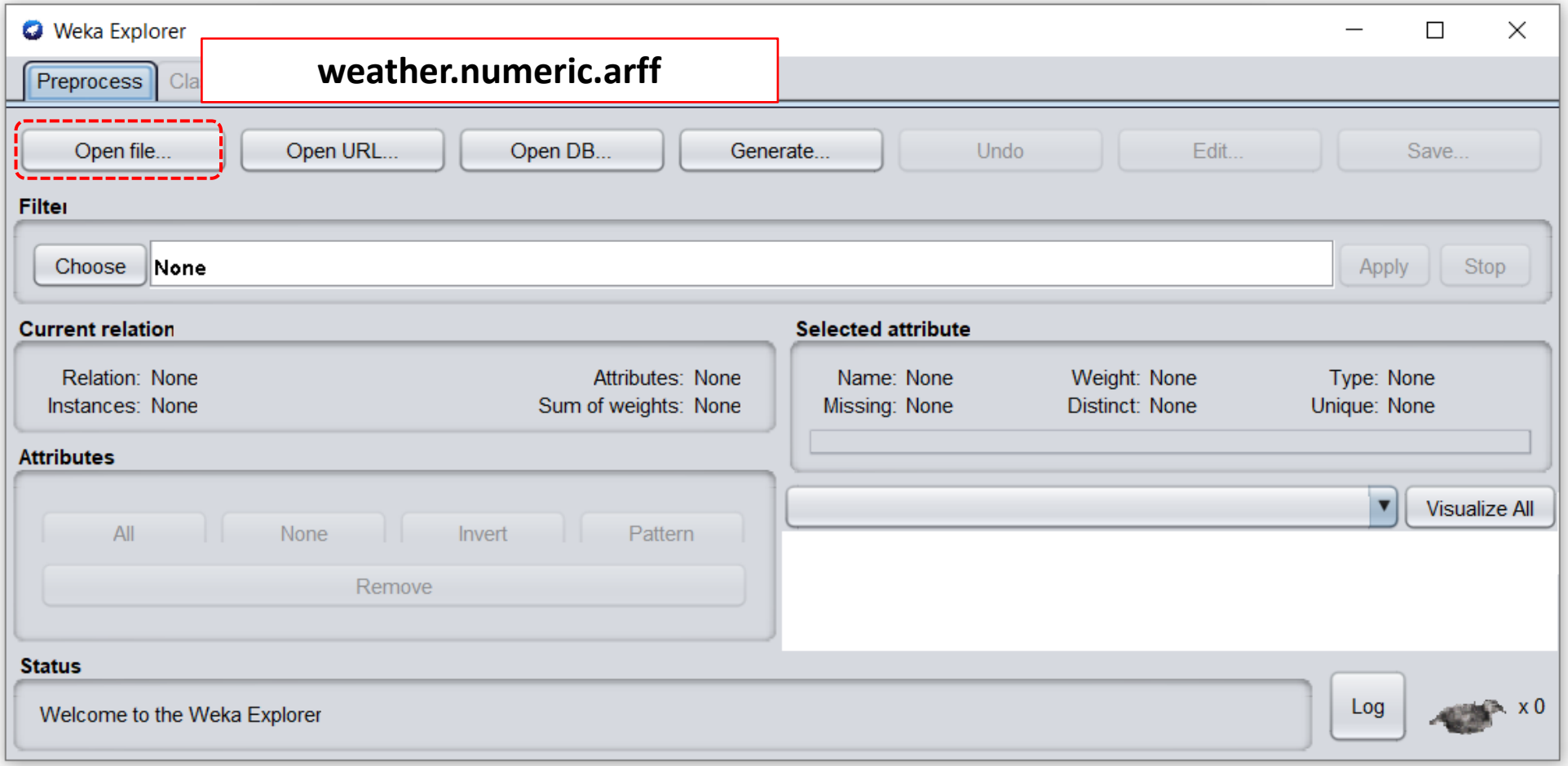
Dataset -> exemplos classificados

➡ Criar modelos que classifiquem novos exemplos



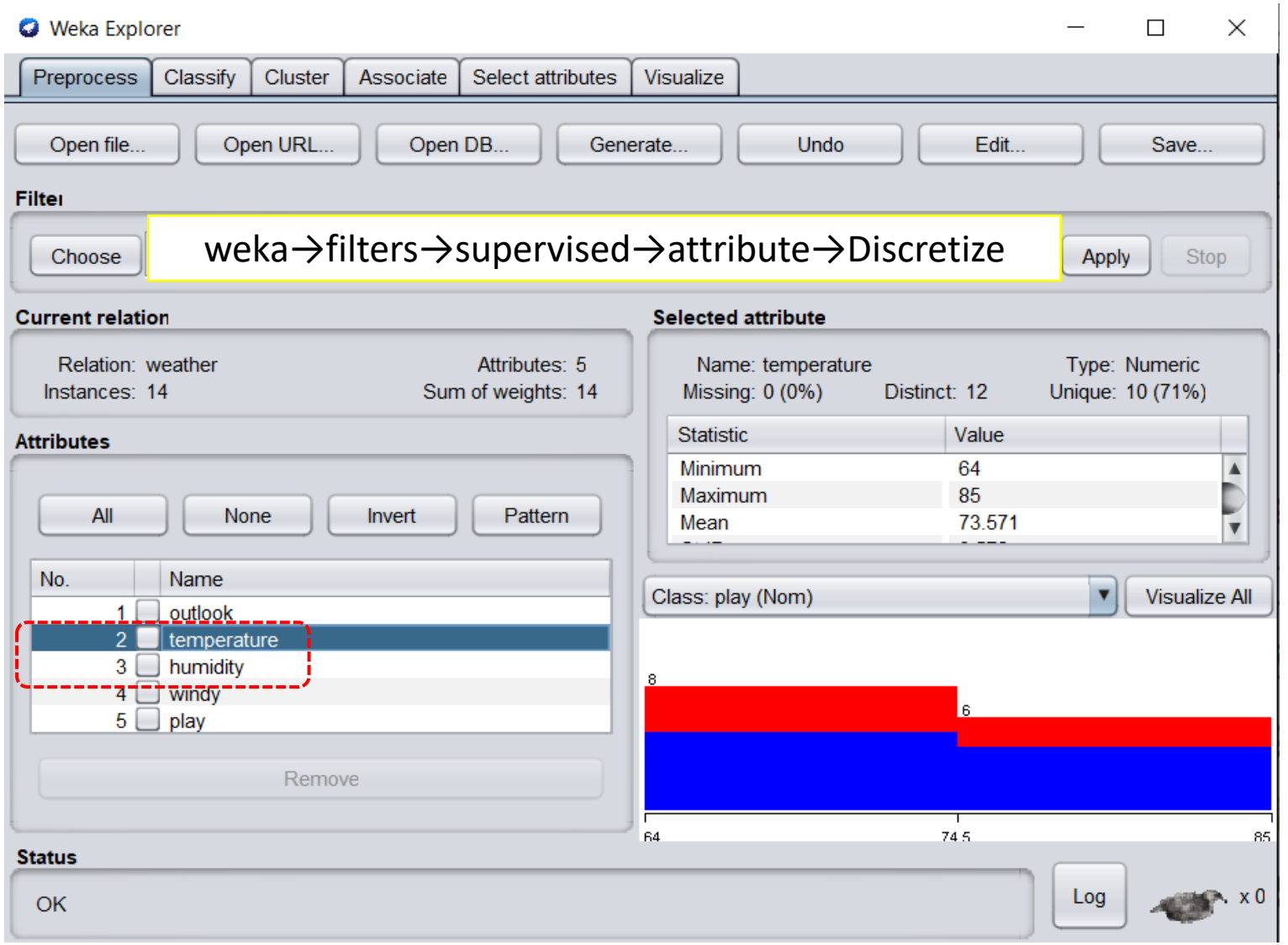
Discreto -> nominal -> problema de classificação
Contínuo -> numérico -> problema de regressão

WEKA – carregar dados



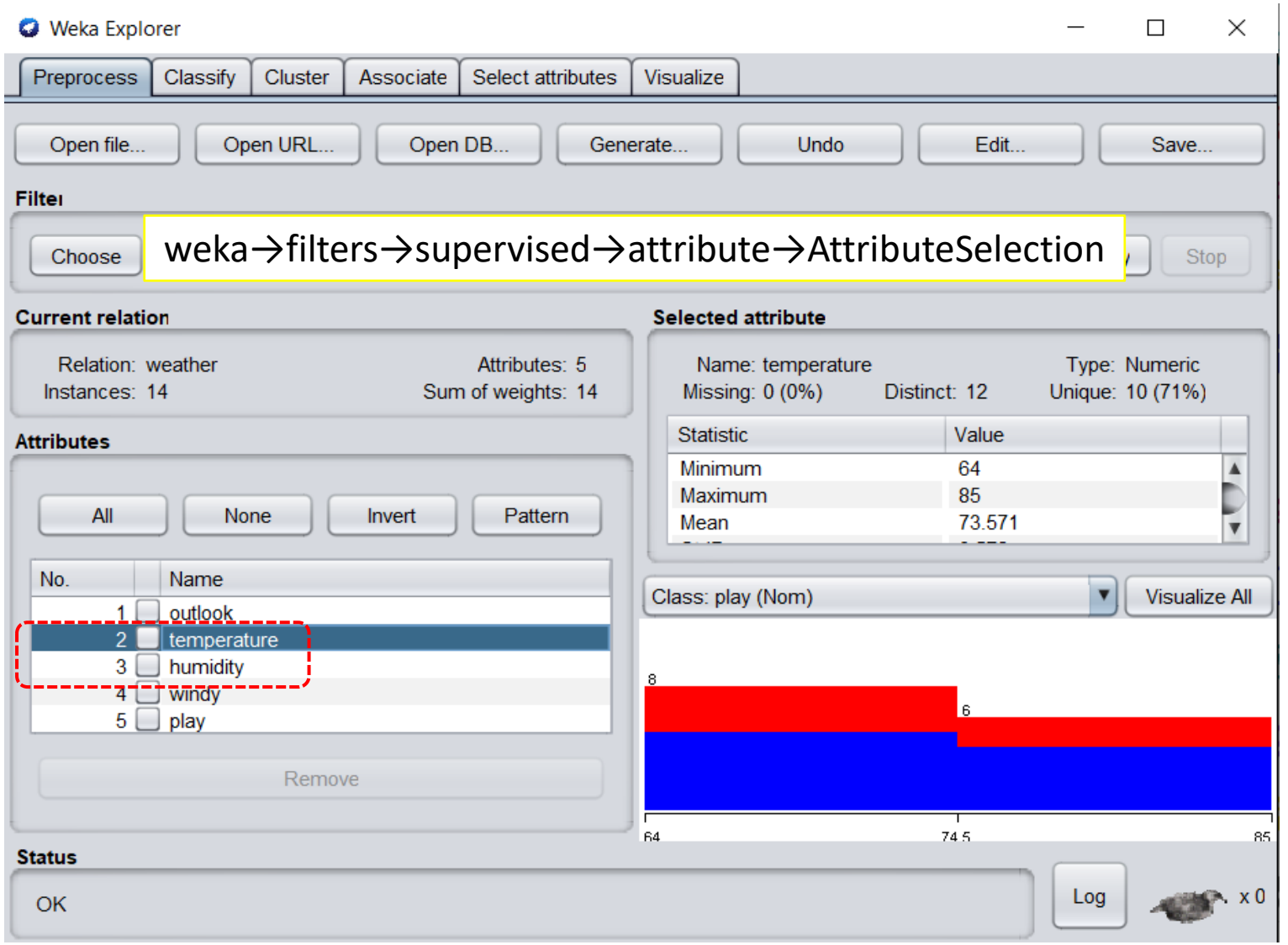
*Os *datasets* estão guardados na pasta Data que está dentro da pasta de instalação do *software*
C:\Program Files\Weka-3-8-4\data

WEKA – carregar dados



Atributos Numéricos

WEKA – carregar dados

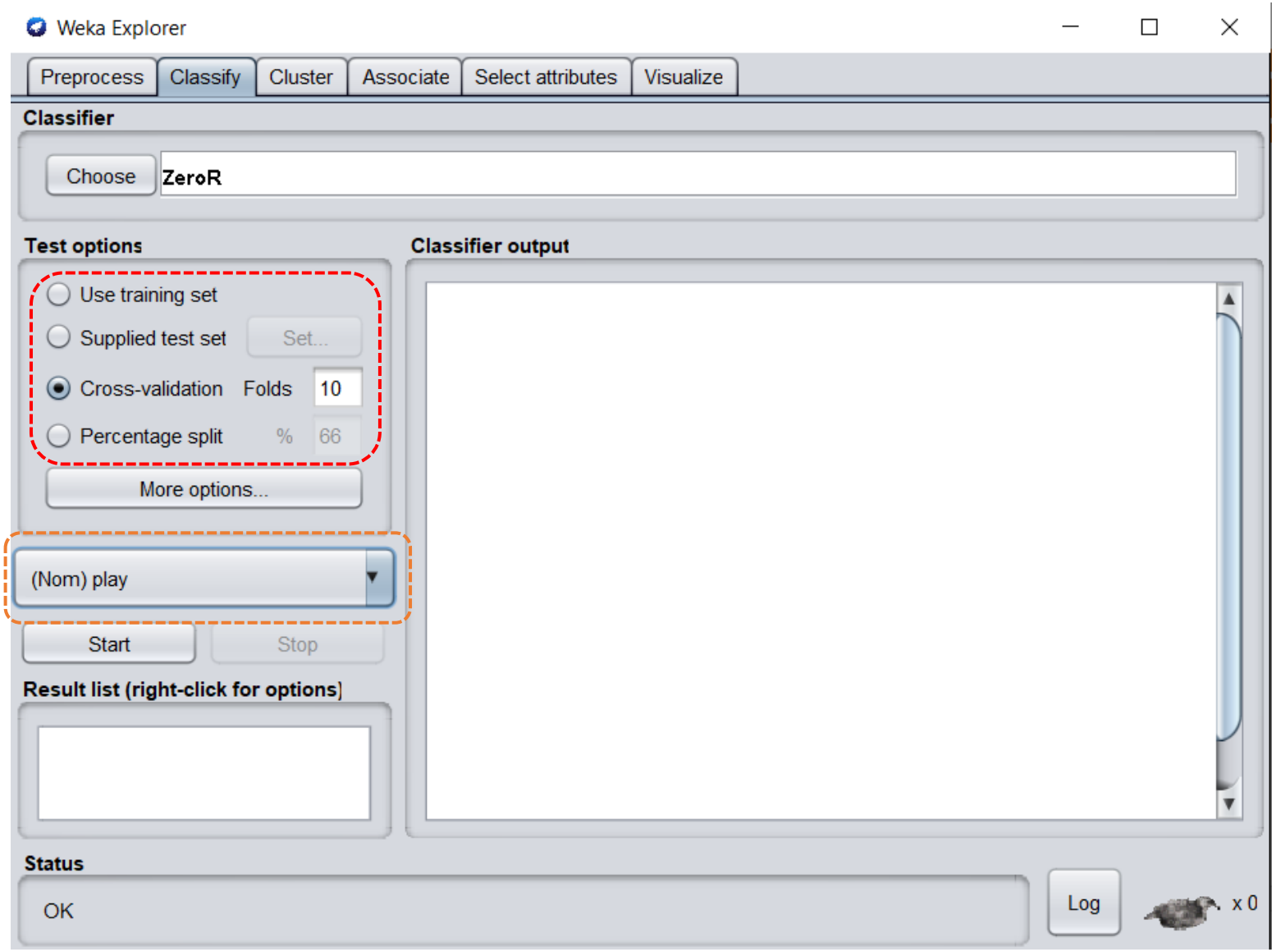


Atributos Numéricos

WEKA – classificação

OPÇÕES DE TESTE

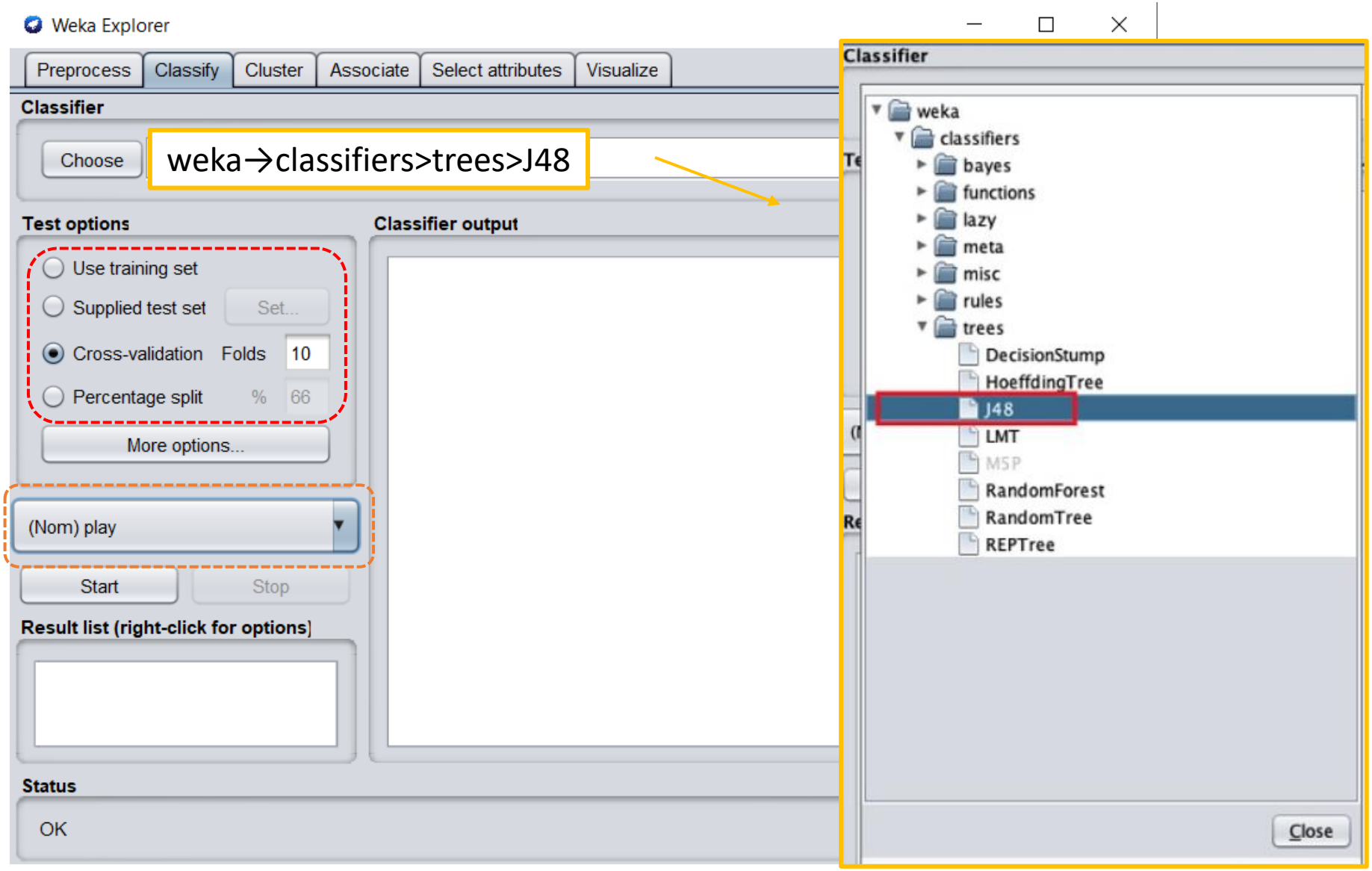
CLASSE



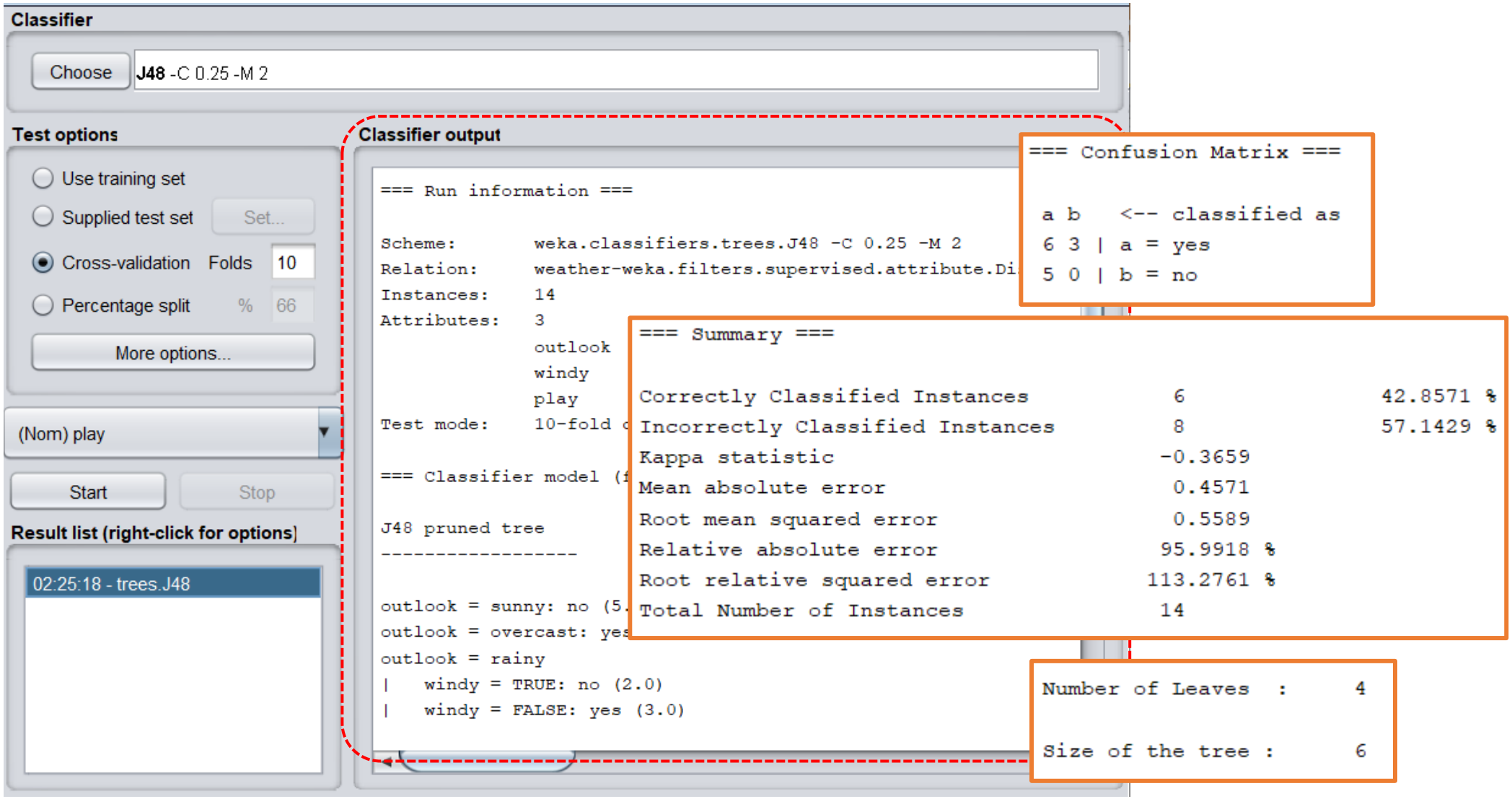
WEKA – classificação

OPÇÕES DE TESTE

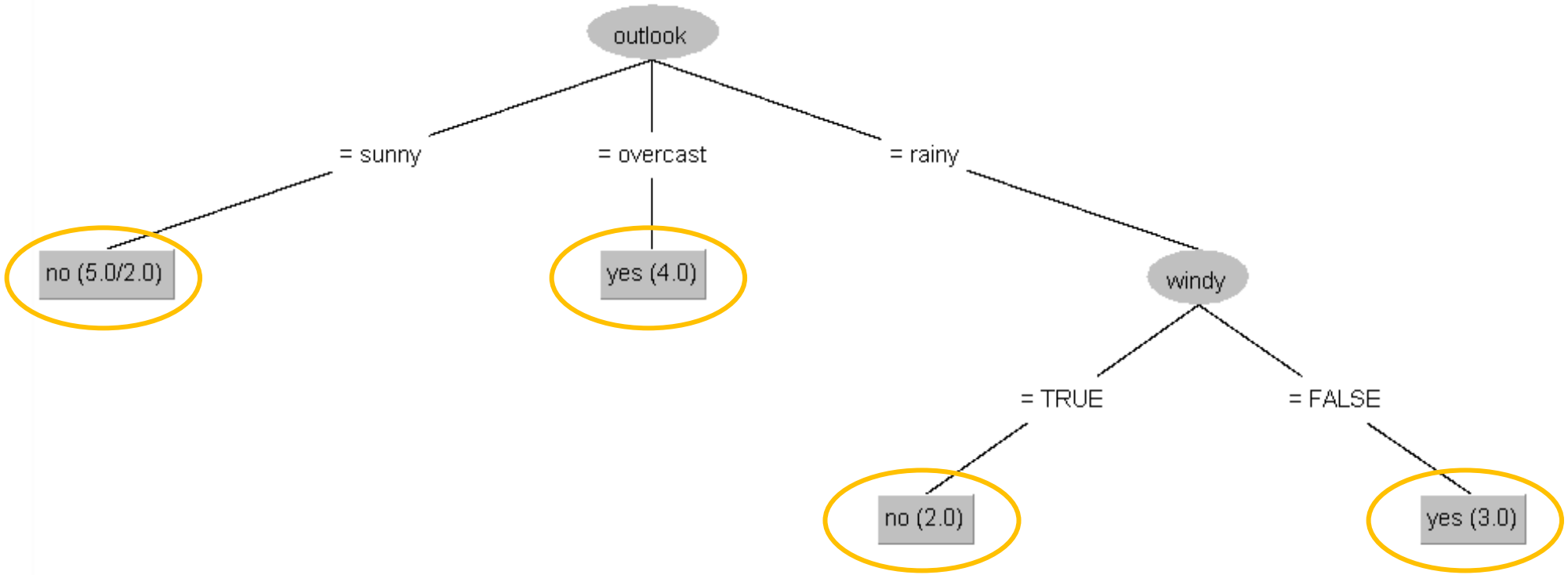
CLASSE



WEKA – classificação



WEKA – classificação



WEKA – classificação

- Abrir o *dataset* glass.arff;
- Escolher o algoritmo J48;
- Analisar os resultados e visualizar a árvore;
- Carregar em cima do algoritmo J48;
- Examinar as diferentes opções;
- Usar uma árvore não podada - ‘unproned tree’;
- Colocar a propriedade ‘minNumObj’ igual a 15 para evitar folhas pequenas;
- Comparar com os resultados obtidos anteriormente.

WEKA – classificação

PRUNNING DECISION TREES

é uma técnica que reduz o tamanho das árvores de decisão ao remover secções da árvore que fornecem pouco poder para classificar as instâncias. A poda reduz a complexidade do classificador final e, portanto, melhora a precisão da previsão através da redução do excesso de ajustes - *overfitting*.

EXERCÍCIO – FE02

[1] Abrir o Weka / Explorer e carregar o data set *"contact-lens.arff"*.

[a] Quantas instâncias (registos) tem este data set?
[b] Quantos atributos (colunas) tem este data set?
[c] Quantos e quais os valores possíveis para o atributo "*age*"?
[d] Quais os valores possíveis para o atributo "*contact-lens*"?
[e] Qual o atributo que tem "*reduced*" como um dos valores?

[2] Abrir o Weka/Explorer e carregar o data set "*iris.arff*".

[a] Quantas instâncias registos tem este data set?
[b] Quantos atributos (colunas) tem este data set?
[c] A classe "*iris-setosa*" tende a ter maiores ou menores valores de "*sepal.length*"?
[d] A classe "*iris-viginica*" tende a ter maiores ou menores valores de "*petal.width*"?
[e] Qual destes atributos, sozinho, parece dar uma melhor indicação da "*class*"?

EXERCÍCIO – FE02

[3] Abrir o Weka/Explorer e carregar o data set "*weather.nominal.arff*".

[a] Identificar quais os atributos deste data set?

[b] A utilização de um algoritmo de classificação poderá trazer conhecimento específico através dos dados apresentados. Indique um objetivo que possa ser atingido com a aplicação de algoritmos de classificação, quando o mesmo for executado em dados semelhantes mas previamente desconhecidos.

[4] Abrir o Weka e carregar o data set *"glass.arff"*.

[a] Abrir o separador "Classify" e escolher o algoritmo J48 ("trees")

[b] Observar a *"Confusion Matrix"* e indicar quais as maiores falhas no processo de classificação.

[c] Qual o número de "*headlamps*" que foram classificadas como "*build wind float*"?

[d] Qual o número de instâncias classificadas corretamente como "*vehic wind non-float*"?

[e] Qual o número de instâncias classificadas corretamente como "*vehic wind float*"?

[f] Na lista de resultados obtidos clicar com o botão direito e selecionar "*Visualize tree*". Copiar os resultados para a ficha de solução e descrever sucintamente o processo de classificação do algoritmo.

EXERCÍCIO – FE02

[5] Abrir o Weka / Explorer e carregar o data set *"labor.arff"*.

[a] Correr o algoritmo de classificação J48 com os parâmetros por defeito. Indicar a percentagem de instâncias corretamente classificadas.

[b] Utilizando somente 2 casas decimais, abra a configuração do algoritmo J48 e coloque a opção "unpruned" a "True". Corra novamente a classificação e indique a percentagem de instâncias corretamente classificadas.

[6] Abrir o Weka / Explorer e carregar novamente o data set *"glass.arff"*.

[a] Retirar o atributo "Fe". Qual o resultado da classificação?

[b] Retirar todos excepto "Ri", "Mg". Qual o resultado da classificação?