

Universidade do Minho
Escola de Engenharia

Sistemas de Aprendizagem e Extração de Conhecimento

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



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 Apply Stop

Current relation

Relation: None Attributes: None
Instances: None Sum of weights: None

Selected attribute

Name: None Weight: None Type: None
Missing: None Distinct: None Unique: None

Visualize All

Attributes

All None Invert Pattern

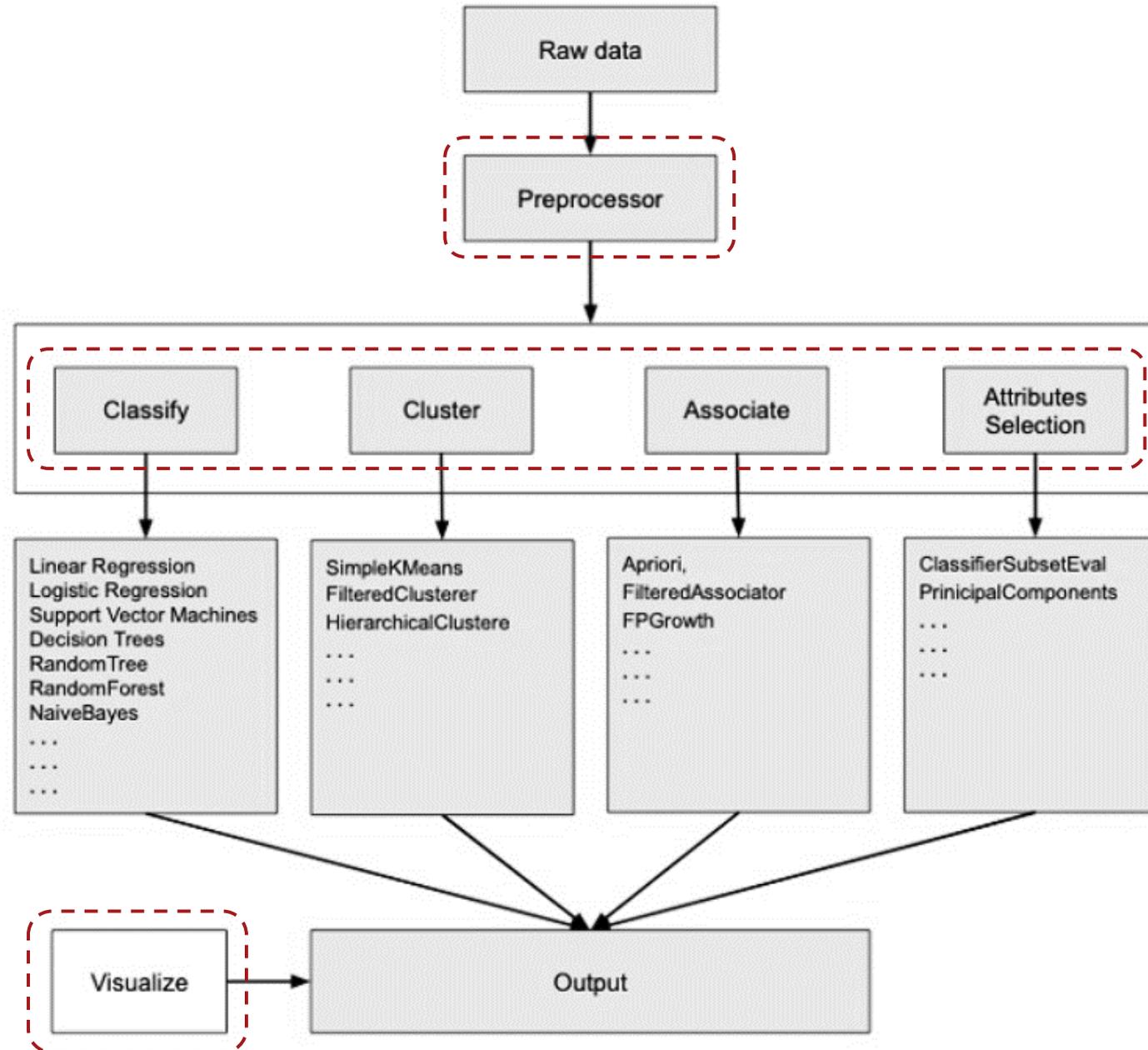
Remove

Status

Welcome to the Weka Explorer

Log  x 0

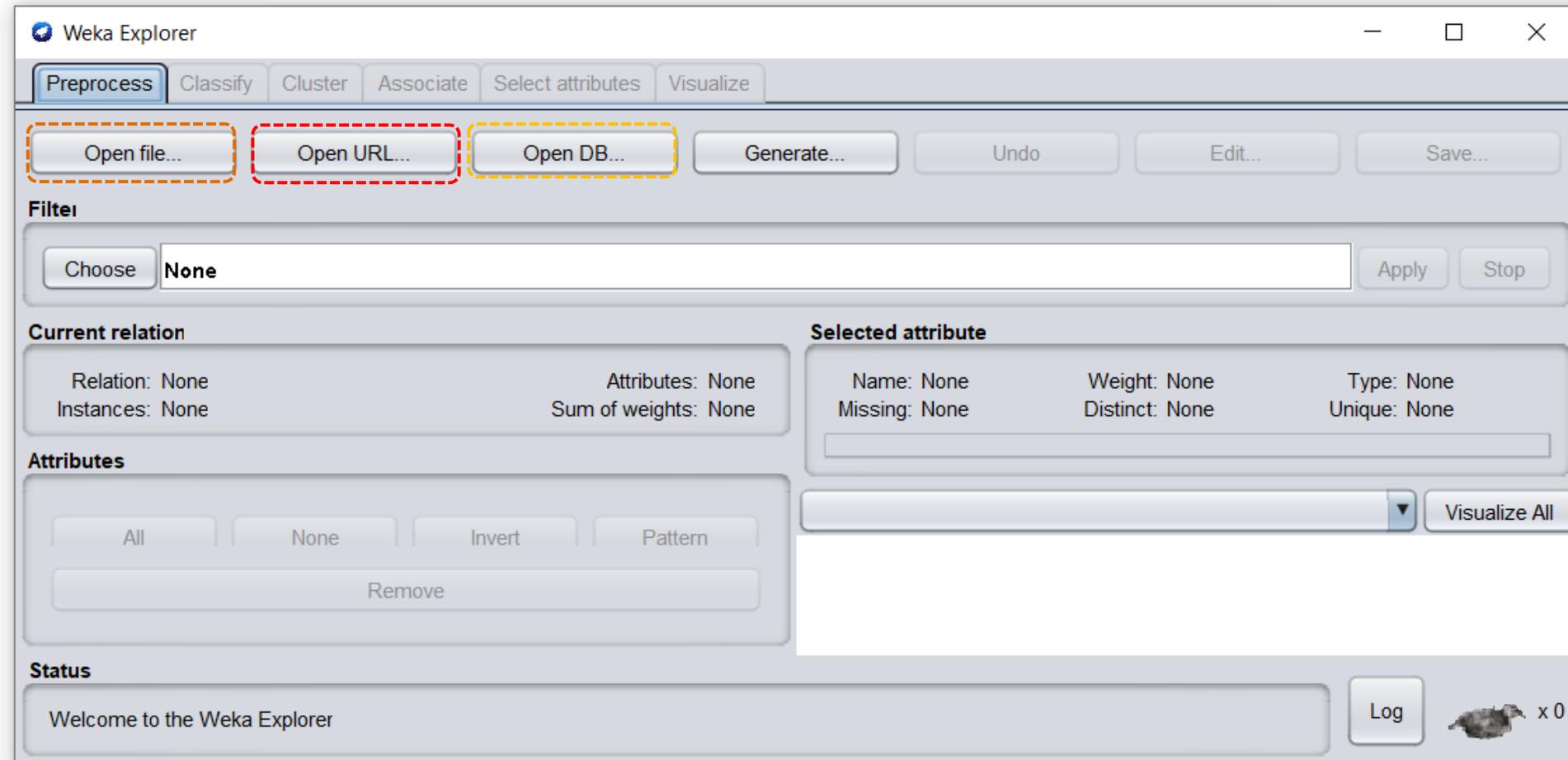
WEKA



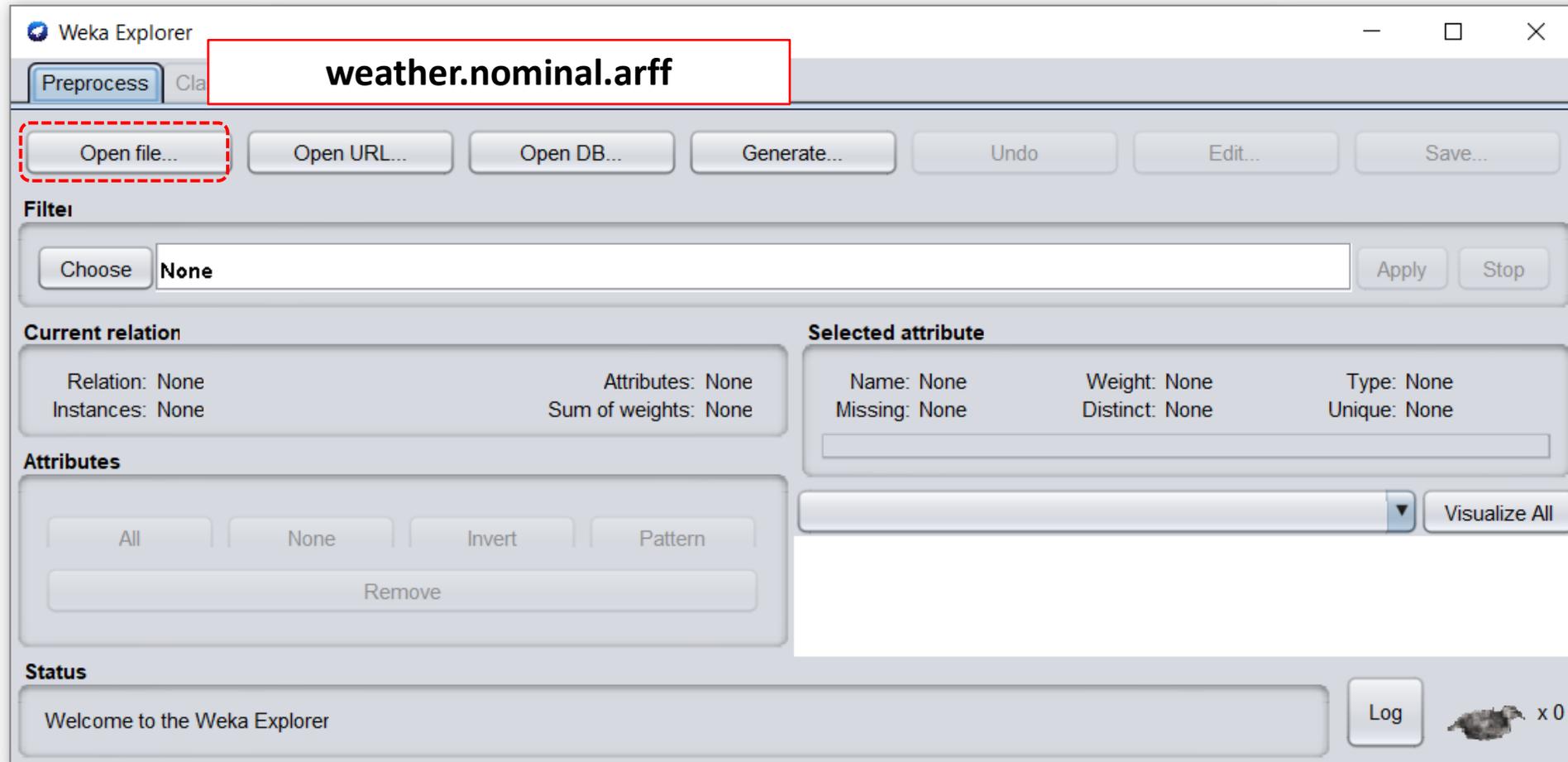
WEKA – carregar dados



- 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 – cargar datos

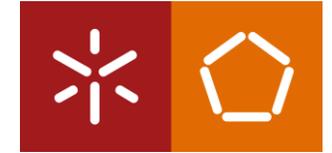


Weather.arff

		attributes				
		Outlook	Temp	Humidity	Windy	Play
instances	1	Sunny	Hot	High	False	No
	2	Sunny	Hot	High	True	No
	3	Overcast			False	Yes
	4	Rainy			False	Yes
	5	Partly sunny		Normal	False	Yes
	6	Partly cloudy		Normal	True	No
	7	Cloudy	Cool	Normal	True	Yes
	8	Cloudy	Mild	High	False	No
	9	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

Classification problem:
predict the "class" value

WEKA – preprocess



Atributos

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: weather.symbolic Instances: 14 Attributes: 5 Sum of weights: 14

Attributes: All | **None** | Invert | Pattern

No.	Name
1	<input checked="" type="checkbox"/> outlook
2	<input type="checkbox"/> temperature
3	<input type="checkbox"/> humidity
4	<input type="checkbox"/> windy
5	<input type="checkbox"/> play

Remove

Selected attribute

Name: outlook Missing: 0 (0%) Distinct: 3 Type: Nominal Unique: 0 (0%)

No.	Label	Count	Weight
1	sunny	5	5.0
2	overcast	4	4.0
3	rainy	5	5.0

Class: play (Nom) Visualize All

Status: OK Log x 0

Classe

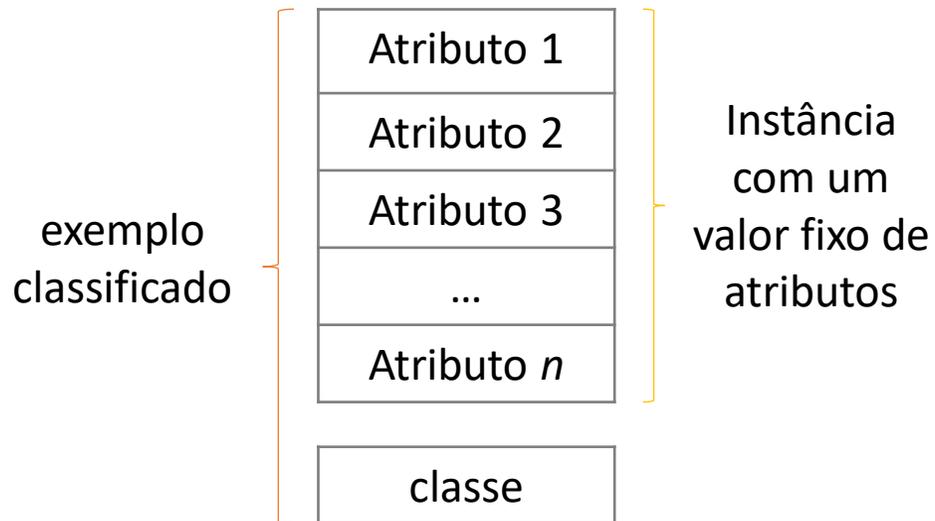
WEKA – preprocess



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 – cargar datos



Atributos
Numéricos

The screenshot shows the Weka Explorer interface with the following components:

- Preprocess** tab selected.
- Filter:** A text box contains the path `weka → filters → supervised → attribute → Discretize`, highlighted in yellow. Buttons for **Choose**, **Apply**, and **Stop** are visible.
- Current relation:** Relation: weather, Instances: 14, Attributes: 5, Sum of weights: 14.
- Selected attribute:** Name: temperature, Type: Numeric, Missing: 0 (0%), Distinct: 12, Unique: 10 (71%).
- Attributes:** A table with columns 'No.' and 'Name'. The 'temperature' attribute (No. 2) is selected and highlighted in blue. A red dashed box highlights the 'temperature' row. A red arrow points from the text 'Atributos Numéricos' to this row. Buttons for 'All', 'None', 'Invert', and 'Pattern' are above the table. A 'Remove' button is below.
- Class:** play (Nom). A 'Visualize All' button is next to it.
- Visualize:** A horizontal bar chart showing the distribution of the 'temperature' attribute. The x-axis ranges from 64 to 85. The y-axis shows counts of 8 and 6. The bars are colored red and blue.
- Status:** OK. A 'Log' button and a small icon with 'x 0' are at the bottom right.

WEKA – cargar datos



Atributos
Numéricos

Weka Explorer

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

Open file... | Open URL... | Open DB... | Generate... | Undo | Edit... | Save...

Filter: **weka → filters → supervised → attribute → AttributeSelection** Stop

Current relation
Relation: weather | Instances: 14 | Attributes: 5 | Sum of weights: 14

Selected attribute
Name: temperature | Type: Numeric | Missing: 0 (0%) | Distinct: 12 | Unique: 10 (71%)

Statistic	Value
Minimum	64
Maximum	85
Mean	73.571

Class: play (Nom) Visualize All

Attributes
All | None | Invert | Pattern

No.	Name
1	<input type="checkbox"/> outlook
2	<input checked="" type="checkbox"/> temperature
3	<input type="checkbox"/> humidity
4	<input type="checkbox"/> windy
5	<input type="checkbox"/> play

Remove

Status
OK | Log | x 0

WEKA – classificação



OPÇÕES DE TESTE

CLASSE

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose ZeroR

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 10

Percentage split % 66

More options...

Classifier output

(Nom) play

Start Stop

Result list (right-click for options)

OK Log x 0

WEKA – classificação



OPÇÕES DE TESTE

CLASSE

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose weka→classifiers>trees>J48

Test options

Use training set

Supplied test set Set...

Cross-validation Folds 10

Percentage split % 66

More options...

Classifier output

(Nom) play

Start Stop

Result list (right-click for options)

Status

OK

Classifier

- weka
 - classifiers
 - bayes
 - functions
 - lazy
 - meta
 - misc
 - rules
 - trees
 - DecisionStump
 - HoeffdingTree
 - J48
 - LMT
 - MSP
 - RandomForest
 - RandomTree
 - REPTree

Close

WEKA – classificação



Classifier

Choose **J48 -C 0.25 -M 2**

Test options

Use training set

Supplied test set

Cross-validation Folds **10**

Percentage split % **66**

(Nom) play

Result list (right-click for options)

02:25:18 - trees.J48

Classifier output

```
=== Run information ===
Scheme:      weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:    weather-weka.filters.supervised.attribute.Discrete
Instances:   14
Attributes:  3
              outlook
              windy
              play
Test mode:   10-fold cross validation

=== Classifier model (full training set) ===
J48 pruned tree
-----
outlook = sunny: no (5)
outlook = overcast: yes (2)
outlook = rainy
| windy = TRUE: no (2.0)
| windy = FALSE: yes (3.0)
```

Confusion Matrix

```
=== Confusion Matrix ===
a b  <-- classified as
6 3  | a = yes
5 0  | b = no
```

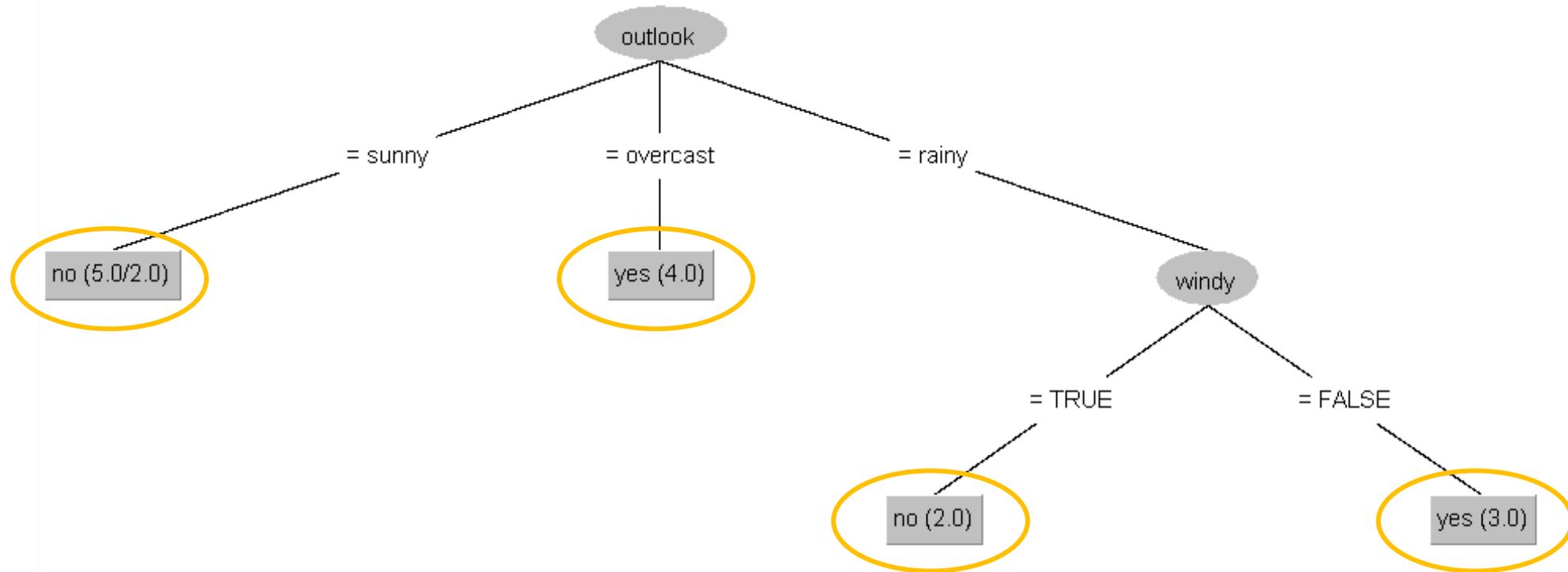
Summary

Correctly Classified Instances	6	42.8571 %
Incorrectly Classified Instances	8	57.1429 %
Kappa statistic	-0.3659	
Mean absolute error	0.4571	
Root mean squared error	0.5589	
Relative absolute error	95.9918 %	
Root relative squared error	113.2761 %	
Total Number of Instances	14	

Number of Leaves : 4

Size of the tree : 6

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 - ‘unpruned 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



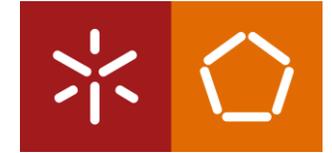
[1] Abrir o Weka / Explorer e carregar o data set “*contact-lens.arff*”.

- [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.

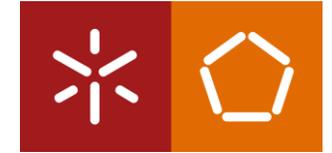
[c] Qual o número de “*headlamps*” que foram classificadas como “*build wind float*”?

[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?