# Weka

- Universidade de Waikato - Nova Zelândia
- Coleção de algoritmos de aprendizado de máquina para resolução de problemas de Data Mining
- implementado em Java
- open source software
- http://www.cs.waikato.ac.nz/ml/weka/

# Métodos de aprendizado suportados

- decision tree inducers
- rule learners
- naive Bayes
- decision tables
- locally weighted regression
- support vector machines
- instance-based learners
- logistic regression
- voted perceptrons

# Preparando os dados

- O *weka* lê os dados no formato .arff
  - Uma lista de todas as instâncias, onde o valor dos atributos são separados por vírgula mais um cabeçalho
- Ex(weather.arff):

```
@relation  weather                %Nome do arquivo
@attribute outlook {sunny, overcast, rainy}
@attribute temperature real %Atributo e tipo
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute play {yes, no}
@data                             %Início dos dados
sunny,85,85,FALSE,no
overcast,83,86,FALSE,yes
```

# Instalando o software

1. Crie um diretório chamada ´weka´ na raiz da sua área
   - c:\weka
2. Copiando os arquivos:
   - `http://www.weka`
   - `weka.jar`

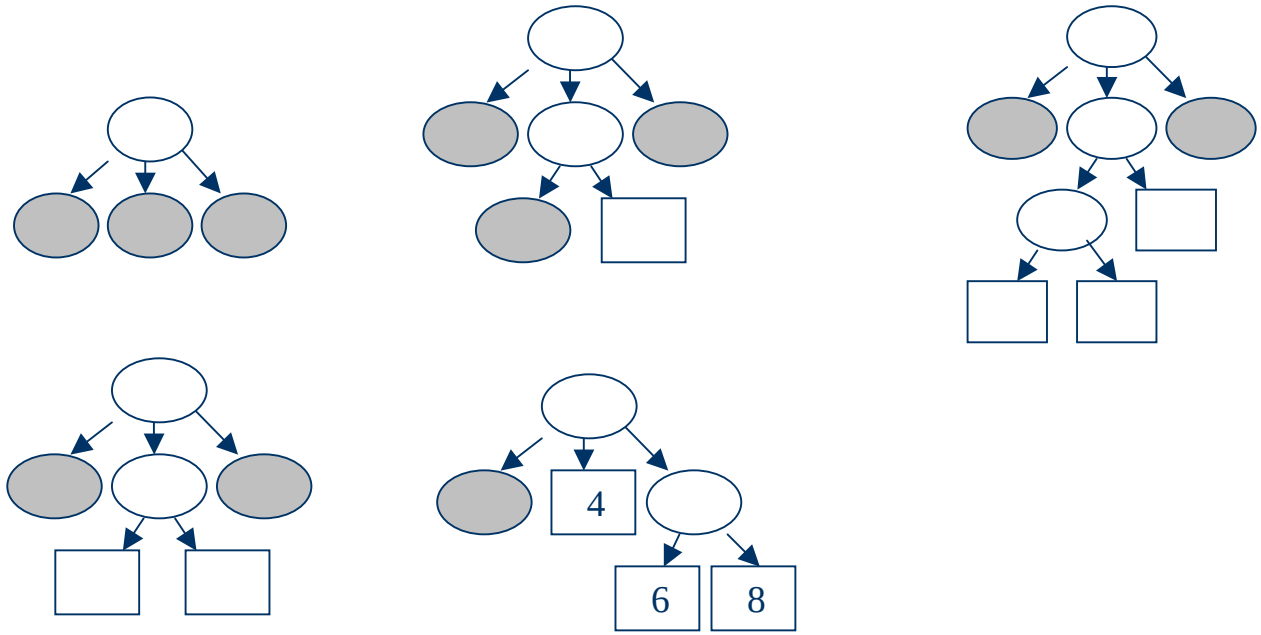
3. Certifique-se de que os dados foram salvos no formato correto.

# Executando o software

- Inicialize o ambiente jdk1.2.2
  - menu: programs\Linguagens\Java\JDK1.2.2
- Entre no diretório ´weka’
  - `cd weka`
- Execute o aplicativo
  - `java weka.gui.explorer.Explorer`

# Regras de classificação

- PART
- Forma lista de regras a partir de árvores parciais podadas



# Saída do algoritmo

```
PART decision list
------------------

outlook = overcast: yes (4.0)
humidity = high: no (5.0/1.0)
: yes (5.0/1.0)

Number of Rules  :        3

=== Confusion Matrix ===
 a b   <-- classified as
 8 1 | a = yes
 1 4 | b = no
```

# Gerando regras de associação

- APRIORI
- Algoritmo para minerar regras de associação.

| | |
|---|---|
| **IF** umidade = normal **AND** vento = não **THEN** jogar = sim | 4/4 |
| **IF** umidade = normal **AND** jogar = sim **THEN** vento = não | 4/6 |
| **IF** vento = não **AND** jogar = sim **THEN** umidade = normal | 4/6 |
| **IF** umidade = normal **THEN** vento = não **AND** jogar = sim | 4/7 |
| **IF** vento = não **THEN** umidade = normal **AND** jogar = sim | 4/8 |
| **IF** jogar = sim **THEN** vento = não **AND** umidade = normal | 4/9 |
| **IF** ? **THEN** vento = não **AND** umidade = normal **AND** jogar = sim | 4/12 |

# Saída do algoritmo

Best rules found:

1. temperature=cool humidity=normal windy=FALSE **2** ==> play=yes **2** conf:(1)
2. temperature=cool windy=FALSE play=yes **2** ==> humidity=normal **2** conf:(1)
3. outlook=overcast temperature=hot windy=FALSE **2** ==> play=yes **2** conf:(1)
4. temperature=cool windy=FALSE **2** ==> humidity=normal play=yes **2** conf:(1)
5. outlook=rainy temperature=mild windy=FALSE **2** ==> play=yes **2** conf:(1)

# Regras de associação

- PredictiveApriori
- Algoritmo para minerar regras de associação.

PredictiveApriori
==========

Best rules found:

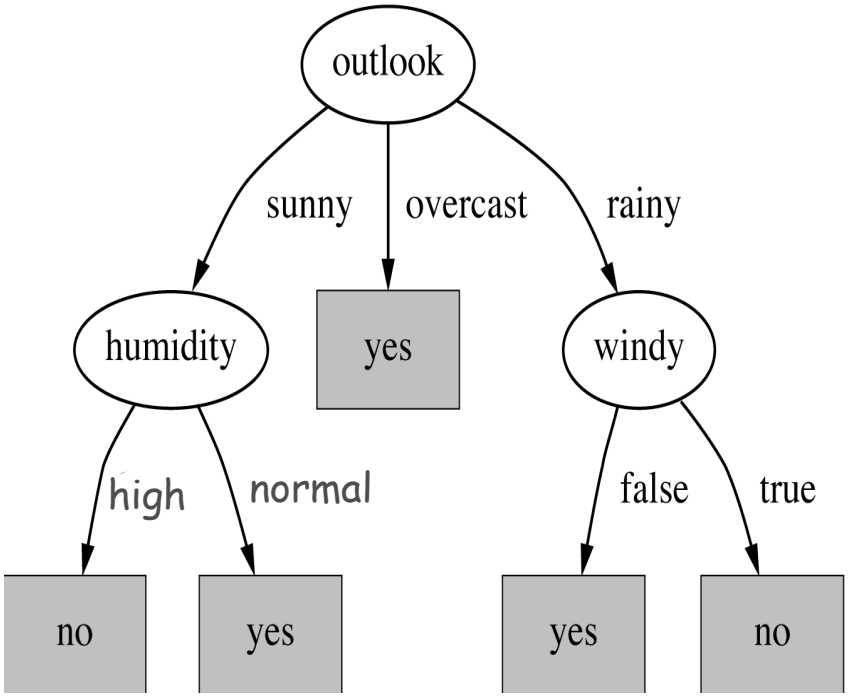
1. taxa-lacrimal=reduzido 12 ==> classe-lente-contato=nenhuma 12  acc:(0.99371)
2. classe-lente-contato=leve 5 ==> astigmatismo=nao taxa-lacrimal=normal 5  acc:(0.9664)
3. classe-lente-contato=pesada 4 ==> astigmatismo=sim taxa-lacrimal=normal 4  acc:(0.94354)
4. idade-paciente=jovem classe-lente-contato=nenhuma 4 ==> taxa-lacrimal=reduzido 4  acc:(0.94354)
5. quadro-clinico=miopia classe-lente-contato=nenhuma 7 ==> taxa-lacrimal=reduzido 6  acc:(0.67059)
6. astigmatismo=nao classe-lente-contato=nenhuma 7 ==> taxa-lacrimal=reduzido 6  acc:(0.67059)

# Indução de árvores de decisão

- Dados do tempo: ‘weather.nominal.arff’
- Algoritmo Id3: weka.classifiers.Id3

```
Id3
------------------

outlook = sunny
|   humidity = high: no
|   humidity = normal: yes
outlook = overcast: yes
outlook = rainy
|   windy = TRUE: no
|   windy = FALSE: yes
```



# Indução de árvores de decisão

- Medidas de desempenho do classificador gerado.
- Id3: avaliação por `Stratified cross-validation`

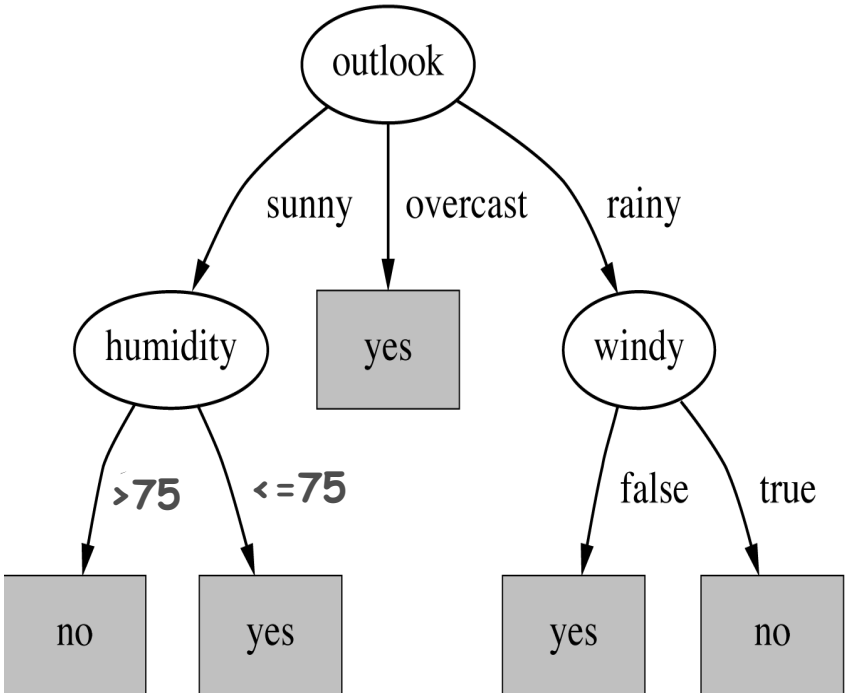
```
=== Summary ===
Correctly Classified Instances    11      78.5714 %
Incorrectly Classified Instances   3      21.4286 %
=== Detailed Accuracy By Class ===
TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.889     0.4        0.8       0.889     0.842    yes
  0.6       0.111      0.75      0.6       0.667    no
=== Confusion Matrix ===
 a b   <-- classified as
 8 1 | a = yes
 2 3 | b = no
```

# Indução de árvores de decisão

- Dados do tempo: ‘weather. arff’
- Algoritmo C4.5: weka.classifiers.j48.J48

```
J48 pruned tree
------------------
outlook = sunny
|   humidity <= 75: yes (2.0)
|   humidity > 75: no (3.0)
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
```



# Indução de árvores de decisão

- Medidas de desempenho do classificador gerado.
- C4.5: avaliação por `Stratified cross-validation`

```
=== Summary ===
Correctly Classified Instances    8     57.1429 %
Incorrectly Classified Instances  6     42.8571 %
=== Detailed Accuracy By Class ===
TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.778     0.8        0.636     0.778     0.7      yes
  0.2       0.222      0.333     0.2       0.25     no
=== Confusion Matrix ===
 a b   <-- classified as
 7 2 | a = yes
 4 1 | b = no
```

# Classificador Naive Bayes

- Dados do tempo: ‘weather. arff’
- Algoritmo weka.classifiers.NaiveBayesSimple

```
Class yes: P(C) = 0.625
Attribute outlook
sunny   overcast        rainy
0.25    0.4166667       0.3333333
Attribute temperature
Mean: 73 Std: 6.164414
Attribute humidity
Mean: 79.11 Std: 10.215728
Attribute windy
TRUE        FALSE
0.363636  0.6363=P(false/yes)
```

```
Class no: P(C) = 0.375
Attribute outlook
sunny   overcast        rainy
0.5     0.125           0.375
Attribute temperature
Mean: 74.6      Std: 7.8930349
Attribute humidity
Mean: 86.2      Std: 9.7313925
Attribute windy
TRUE    FALSE
0.57142857      0.42857143
```

# Classificador Naive Bayes

- Medidas de desempenho do classificador gerado.
- NB: avaliação por `Stratified cross-validation`

```
=== Summary ===
Correctly Classified Instances    8     57.1429 %
Incorrectly Classified Instances  6     42.8571 %
=== Detailed Accuracy By Class ===
TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.778     0.8        0.636     0.778     0.7      yes
  0.2       0.222      0.333     0.2       0.25     no
=== Confusion Matrix ===
 a b   <-- classified as
 7 2 | a = yes
 4 1 | b = no
```

# Classificador Prism

- Medidas de desempenho do classificador gerado.
- Algoritmo weka.classifiers.rules.Prism

```
=== Classifier model (full training set) ===
Prism rules
----------
If aparencia = encoberto then sim
If umidade = normal
   and vento = falso then sim
If temperatura = agradável
   and umidade = normal then sim
If aparencia = chuvoso
   and vento = falso then sim
If aparencia = sol
   and umidade = alta then nao
```

# Classificador Decision Stump

- Dados Lentes: ‘Lentes. arff’
- Algoritmo weka.classifiers.trees.DecisionStump

```
=== Classifier model (full training set) ===
Decision Stump
Classifications
taxa-lacrimal = reduzido : nenhuma
taxa-lacrimal != reduzido : leve
taxa-lacrimal is missing : nenhuma
Class distributions
taxa-lacrimal = reduzido
leve      pesada    nenhuma
0.0 0.0   1.0
taxa-lacrimal != reduzido
leve      pesada    nenhuma
0.4166666666666667  0.3333333333333333  0.25
taxa-lacrimal is missing
leve      pesada    nenhuma
0.20833333333333334 0.16666666666666666 0.625
```

# Classificador M5Rules

- Dados CPU: ‘CPU. arff’
- Algoritmo weka.classifiers.rules.M5Rules

```
=== Classifier model (full training set) ===
M5 pruned model rules
(using smoothed linear models) :
Number of Rules : 2
Rule: 1
IF
    MMAX <= 14000
THEN
class =
    0.0057 * MYCT + 0.0063 * MMIN + 0.0032 * MMAX + 0.6394 * CHMAX + 1.3823
    [141/17.502%]
Rule: 2
IF
    MMAX <= 22485
THEN
class = -0.1389 * MYCT + 0.0092 * MMIN  + 0.0026 * MMAX + 0.9361 * CHMAX + 16.8095
    [37/13.703%]
```

# Classificador Simple Linear Regression

- Dados Flores: ‘Flores. arff’
- Algoritmo weka.classifiers.functions.SimpleLinearRegression
- Class=tamanho_petala

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

Linear regression on Comprimento_petala

0.42 * Comprimento_petala - 0.37


Time taken to build model: 0 seconds
```

# Cluster K-means

- Algoritmo weka.clusterers.**SimpleKMeans.Kmeans**
- Parâmetros:-N Número Grupo;-S semente Aleatória
- Atributos: Numéricos , Nominais , Em falta

```
=== Clustering model (full training set) ===
kMeans
======
Cluster centroids:
Cluster 0
sunny mild high FALSE
Cluster 1
overcast hot high TRUE
```

```
Cluster centroids:
Cluster 0
                6.26 2.87 4.90 1.67
Cluster 1
                5.01 3.41 1.46 0.24
```

# Cluster COWEB

- Algoritmo weka.clusterers.Cobweb
- Parâmetros:-A 1.0 (acuity);
  -C 0.0028209479177387815 (cutoff)

```
=== Clustering model (full training set) ===
Number of merges: 1
Number of splits: 1
Number of clusters: 4
node 0 [150]
     | leaf 1 [50]
node 0 [150]
     | leaf 2 [50]
node 0 [150]
     | leaf 3 [50]
```

# Cluster EM

- Algoritmo weka.clusterers.EM

```
EM
==
Number of clusters selected by cross validation: 5

Cluster: 0 Prior probability: 0.2539

Attribute: MYCT
Normal Distribution. Mean = 497.3459 StdDev = 359.5158
Clustered Instances
0       54 ( 26%)    1       45 ( 22%)    2       15 (  7%)   3       17 (  8%)
4       78 ( 37%)
Log likelihood: -40.75772
```