

Data Mining com Software Livre

Marcos Vinicius Fidelis

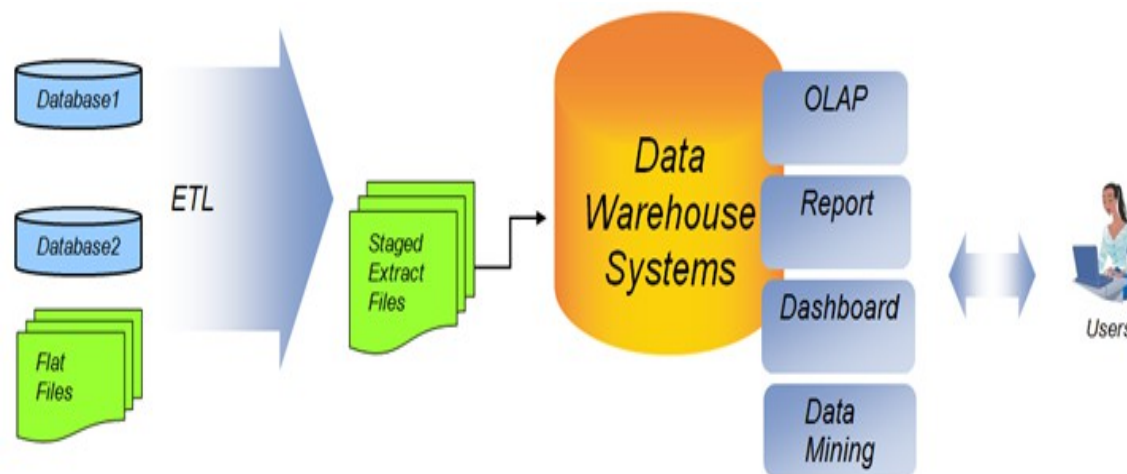
Analista de Informática – Universidade Estadual de Ponta Grossa
Professor – Universidade Tecnológica Federal do Paraná

fidelis@utfpr.edu.br - mvfidelis@uepg.br

Pentaho

- Pentaho é um software de código aberto para inteligência empresarial, desenvolvido em Java. A solução cobre as áreas de ETL, reporting, OLAP e mineração de dados (data-mining). Desenvolvido desde 2004 pela Pentaho Corporation.
- Comunidade Pentaho Brasil
 - O Pentaho Open Source Business Intelligence oferece poderosas ferramentas de análise de informações, monitoramento de indicadores e data mining para que as organizações revolucionem o uso da informação gerencial, atingindo ganhos significativos de eficiência e eficácia.
 - O software - uma plataforma completa de BI desenvolvida, distribuída e implantada como Open Source — apresenta grande flexibilidade e independência de plataformas, alta confiabilidade e segurança a um custo mínimo de implantação e manutenção.

Módulos Pentaho

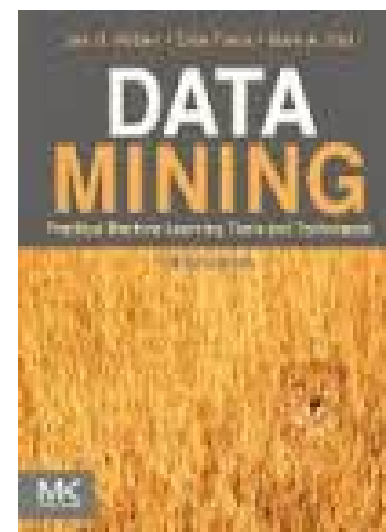
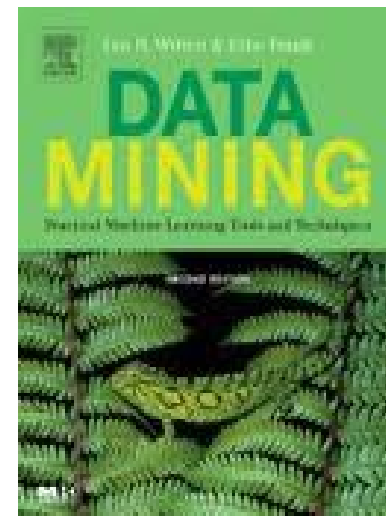


A Suite Pentaho OSBI é composta pelo Pentaho BI Server, Pentaho Data Integration, Pentaho Analysis, Pentaho Reporting, Pentaho Dashboards e Pentaho Data Mining.

- **Pentaho Data Integration:** Também conhecido como Kettle é uma solução robusta para integração de dados, recomendada para processos de ETL (do inglês Extract, Transformation and Load) responsáveis por popular um Data Warehouse, Migração de base de dados e Integração entre Aplicações. Não deixa nada a desejar para os principais players do mercado.
- **Pentaho Analysis:** Também conhecido como Mondrian é um poderoso motor olap, baseado em uma arquitetura ROLAP, onde pode-se utilizar os principais SGBD's do mercado. Possui diversas funcionalidades, como, camada de metadados, linguagem MDX, cache em memória, tabelas agregadas e muito mais.
- **Pentaho Reporting:** Este módulo da suite contempla duas ferramentas, uma ferramenta de geração de relatórios, também conhecida como JFreeReport e outra para geração de metadados, a qual permite a criação Ad-Hoc de relatórios via web browser.
- **Pentaho Dashboards:** Este módulo da suite permite a criação de painéis de controle, mais conhecidos como Dashboards e através dele é possível reunir em uma mesma tela, os principais indicadores de um departamento ou de toda a empresa.
- **Pentaho Data Mining:** Também conhecido como Weka é o módulo mais antigo da suite e possui poderosos recursos para mineração de dados.

Pentaho Data Mining

- Pentaho Data Mining, é baseado no projeto WEKA, é um conjunto de ferramentas para aprendizado de máquina e mineração de dados.
- Seu amplo conjunto de classificação, regressão, regras de associação, e algoritmos de agrupamento pode ser usado para ajudá-lo a entender o negócio melhor e também ser explorado para melhorar o desempenho futuro através de análises preditivas.
- Possui 3 versões principais
 - WEKA 3.4 – versão estável criada em 2003 – para corresponder com o que está descrito na 2ª edição do livro de Witten e Frank livro Mineração de Dados (publicado 2005). Esta versão está congelada.
 - WEKA 3.6 – versão estável criada em 2008 – referente a 3ª edição
 - WEKA 3.7 – versão de desenvolvimento

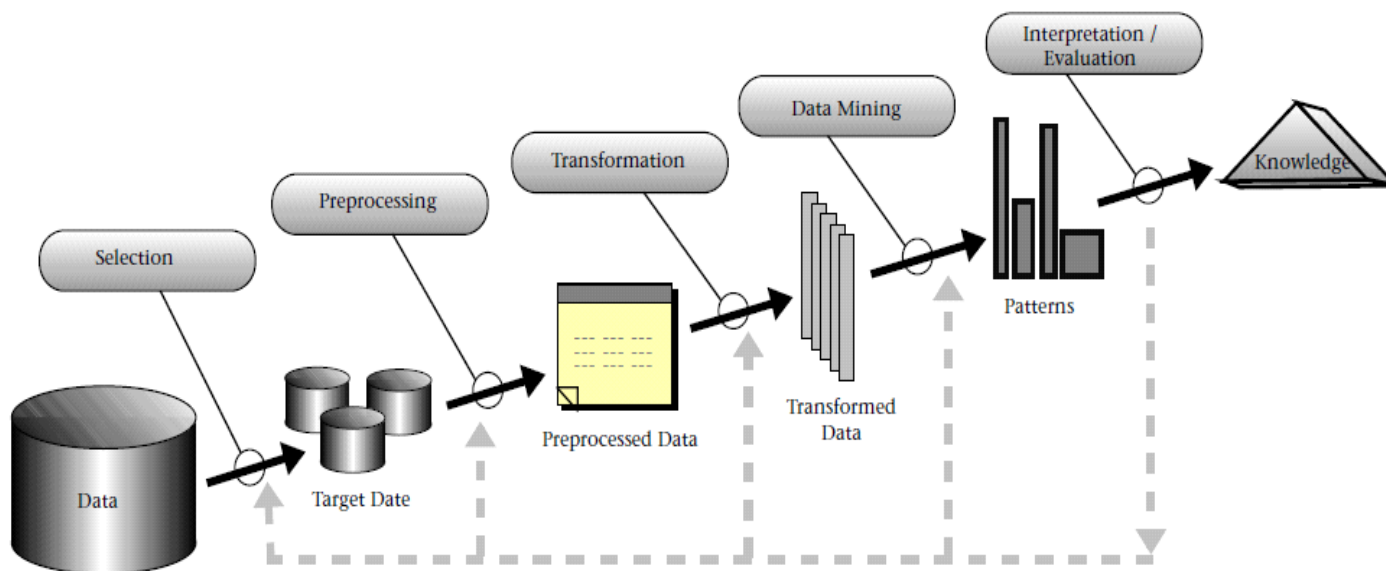


Knowledge Discovery in Databases

- KDD utiliza algoritmos de *data mining* para extrair padrões classificados como “conhecimento”. Incorpora também tarefas como escolha do algoritmo adequado, processamento e amostragem de dados e interpretação de resultados;
- “Torture os dados até eles confessarem”;

Fases do KDD

- Definição do problema
- Seleção dos dados
- Pré-processamento dos dados
- Transformação
- Mineração dos dados
- Interpretação/Avaliação



Fases do KDD

- **Seleção dos dados**

Selecionar ou segmentar dados de acordo com critérios definidos.

Ex: Todas as pessoas que são proprietárias de carros é um subconjunto de dados determinados.

- **Pré-processamento**

Estágio de limpeza dos dados, onde informações julgadas desnecessárias são removidas.

Ex: O sexo de um paciente gestante.

Reconfiguração dos dados para assegurar formatos consistentes (identificação).

Ex: sexo = "M" ou "F"

sexo = "F" ou "M"

- **Transformação**

Transforma-se os dados em formatos utilizáveis. Esta depende da técnica *Data Mining* usada.

Ex: rede neural -> converter valor literal em valor numérico.

Disponibilizar os dados de maneira usável e navegável.

- **Mineração de dados**

É a verdadeira extração dos padrões de comportamento dos dados.

Utilizando a definição de fatos, medidas de padrões, estados e o relacionamento entre eles.

- **Interpretação e Avaliação**

Identificado os padrões pelo sistema, estes são interpretados em conhecimentos, os quais darão suporte a tomada de decisões humanas.

Ex: Tarefas de previsões e classificações.

KDD x Data Mining

- Mineração de Dados é um passo no processo de KDD (Knowledge Discovery in Database) que consiste na aplicação de análise de dados e algoritmos de descobrimento que produzem uma enumeração de padrões (ou modelos) particular sobre os dados.
Fayyad - 1996.
- KDD utiliza algoritmos de data mining para extrair padrões classificados como “conhecimento”. Incorpora também tarefas como escolha do algoritmo adequado, processamento e amostragem de dados e interpretação de resultados.

Data Mining

É o processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados.

WIKIPEDIA

DATA MINING é um processo de busca de dados por PADRÕES anteriormente desconhecidos e uso frequente destes padrões para PREDIZER CONSEQUENCIAS futuras.”

Jeff Jonas e Jim Harper

Esse é um tópico recente em ciência da computação, mas utiliza várias técnicas da estatística, recuperação de informação, inteligência artificial e reconhecimento de padrões.

E daí? Quem vai se interessar?



- Uma empresa utilizando data mining é capaz de:
 - criar parâmetros para entender o comportamento do consumidor;
 - identificar afinidades entre as escolhas de produtos e serviços;
 - prever hábitos de compras
 - analisar comportamentos habituais para detectar fraudes.



Motivação

- A informatização dos meios produtivos permitiu a geração de grandes volumes de dados
 - Transações eletrônicas, novos equipamentos e sensores.
- A sombra digital deixada por cada indivíduo aumenta significativamente com o avanço da tecnologia
 - Posts facebook, twitter, instagram, foursquare, blogs, smartphones, logs diversos, etc.
- A capacidade de analisar os dados é infinitamente menor que a velocidade de geração dos mesmos dados

Todos os padrões são interessantes?

Um padrão é interessante se:

- Facilmente pode ser compreendido por humanos
- É válido em dados de teste com um certo grau de certeza
- Potencialmente útil
- Novo
- Se valida uma hipótese do usuário

A internet em um minuto



Tarefas de Data Mining

- **Classificação:** aprendizado de uma função que mapeia um dado em uma de várias classes conhecidas.
- **Regressão** (predição): aprendizado de uma função que mapeia um dado em um valor real.
- **Agrupamento** (*clustering*): identificação de grupos de dados onde os dados tem características semelhantes entre si e os grupos tem características diferentes.
- **Sumarização:** descrição do que caracteriza um conjunto de dados (ex. conjunto de regras).
- **Detecção de desvios ou outliers:** identificação de dados que deveriam seguir um padrão mas não o fazem.

New York Times: Data Mining na Saúde

- De tempos em tempos o jornal New York Times publica uma notícia relevante sobre Data Mining. No último dia 06 de março uma reportagem revelou que pela primeira vez, os efeitos colaterais não declaradas de medicamentos foram capazes de ser detectados e classificados antes de isto ser feito pelo Food and Drug Administration(FDA – órgão governamental responsável pelo controle de alimentos, medicamentos, cosméticos e afins).
- Cientistas da Microsoft e das Universidades de Stanford e Columbia vasculharam milhões de consultas no Google, Yahoo e Bing . O estudo encontrou evidências de que o uso combinado de um antidepressivo (paroxetina) e um remédio para baixar o colesterol (pravastatina) causava elevação do açúcar no sangue. Antes deste estudo, a única maneira de tal fato ser notado seria se um médico o detectasse e o reportasse para o sistema da FDA, conhecido como o Sistema de Notificação de Eventos Adversos.
- A pesquisa iniciou “manualmente” por um grupo de trabalho do departamento de bioengenharia de Stanford. Após identificar os primeiros indícios a Microsoft entrou no projeto e criou software para digitalização de dados anônimos recolhidos a partir um plugin instalado em navegadores da Web. Graças a isto 82 milhões buscas individuais estavam disponíveis para análise.
- Os pesquisadores inicialmente identificaram pesquisas individuais para os termosparoxetina e pravastatina, e depois pesquisas conjuntas. Eles então calcularam a probabilidade de que os usuários de cada grupo também procurar por hiperglicemia, bem como cerca de 80 de seus sintomas (palavras ou frases como “açúcar alto no sangue” e “visão embaçada”).
- Resultados: pessoas que procuravam por ambas as drogas durante o período de 12 meses foram significativamente mais propensos a procurar por termos relacionados a hiperglicemia do que aqueles que procuram por apenas uma delas (algo em torno de 10%, em comparação com 5% e 4% para apenas um medicamento).
- Eles também descobriram que as pessoas que fizeram as buscas por sintomas de ambas as drogas eram susceptíveis de fazer as buscas em um curto período de tempo: 30% fizeram a busca no mesmo dia, 40% durante a mesma semana e 50% durante o mesmo mês. Algo difícil de acontecer, se não houvesse relação entre elas.

- **How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did**
- Charles Duhigg apresenta no New York Times como a Target pesca os "propensos pais", naquele momento crucial antes de se transformar em um comprador leal de todas as coisas relativas a crianças. Para descobrir os futuros papais, a Target atribui a cada cliente um número de ID do cliente , ligada ao seu cartão de crédito , nome ou endereço de e-mail que se torna um cesto que armazena um histórico de tudo o que eles compraram e qualquer informação demográfica que a Target recolheu deles ou de produtos comprados a partir de outras fontes. Para isto foram analisados dados históricos de compra para todas as senhoras que se inscreveram para se o bebê Target no passado.
- Após executar o teste, analisando os dados, em pouco tempo alguns padrões úteis surgiram. Loções, por exemplo. Muita gente compra loção, mas um dos colegas do responsável pelo projeto, notou que as mulheres que se inscreveram no bebê Target estavam comprando grandes quantidades de loção sem perfume por volta do início do segundo trimestre. Outro analista observou que em algum momento nas primeiras 20 semanas, as mulheres grávidas reforçavam as compras em suplementos como cálcio, magnésio e zinco. Muitos compradores compram sabão e bolas de algodão, mas quando alguém de repente começa a comprar lotes de sabão sem perfume e sacos extra-grandes de bolas de algodão, além de desinfetantes para as mãos e panos, sinaliza que pode ser estar perto de sua data de parto.

Business Intelligence

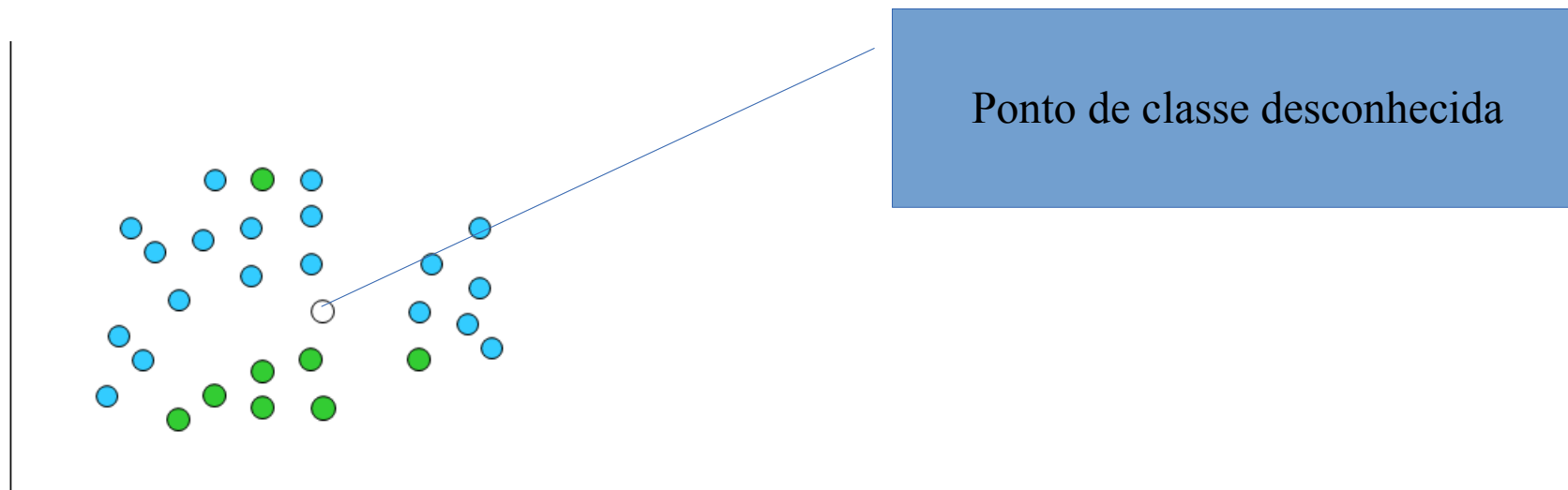
O conceito de inteligência de Negócios (*Business Intelligence - BI*) é entendido por técnicas e ferramentas que possibilitam ao usuário analisar dados e com base nestas análises emitir respostas que possam subsidiar objetiva e confiavelmente os processos de decisão numa empresa.

Em um mercado cada vez mais competitivo e a busca cada vez maior por soluções para proporcionar vantagens competitivas, as empresas buscam cada vez mais:

- Entender melhor o nicho de atuação no mercado;
- Promover melhoramentos na competência essencial da empresa;
- Identificar oportunidades;
- Responder adequada e eficientemente às mudanças de mercado;
- Melhorar o relacionamento com clientes e fornecedores;
- Reduzir custos operacionais.

Classificação

Encontrar um método para prever a classe de uma instância a partir de instâncias pré-classificadas



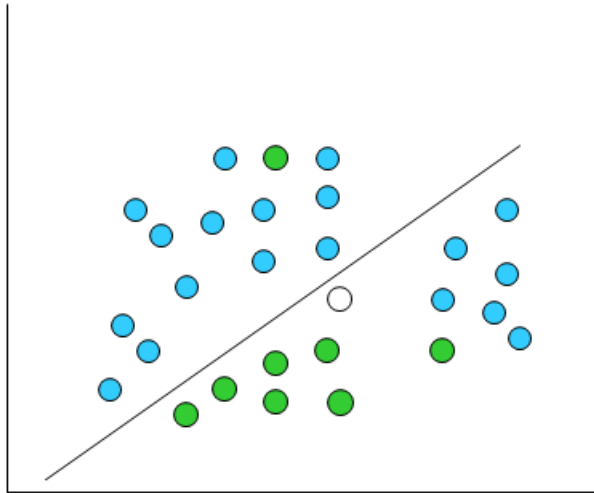
Dado um conjunto de pontos das classes (V)erde e (A)zul

Qual é a classe para o novo ponto (D)esconhecido ?

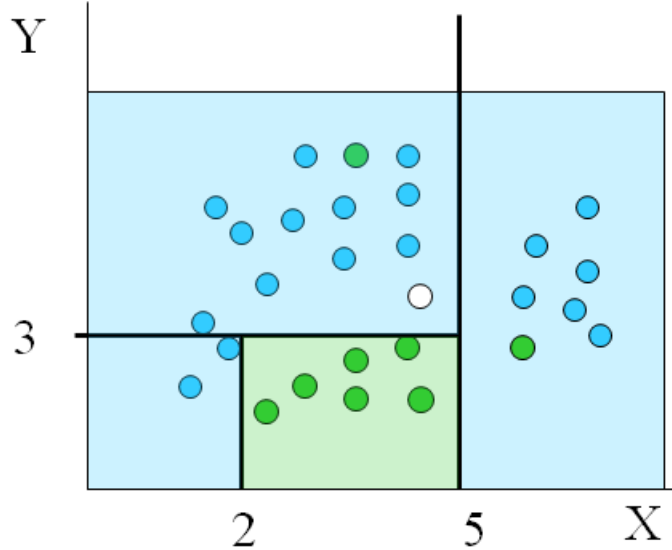
Regressão Linear

$$w_0 + w_1 x + w_2 y \geq 0$$

Regressão calcula w_i a partir dos dados para minimizar o erro



Árvore de decisão



```
IF X > 5
  THEN A
ELSE IF Y > 3
  THEN A
ELSE IF X > 2
  THEN V
ELSE A
```

- Cada nó interno é um teste em um atributo
 - Cada ramo representa um valor de teste
 - Cada folha representa uma classe
-
- Novas instâncias são classificadas seguindo o caminho que leva da raiz até a folha.

Weather: Jogar ou não jogar?

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No

```
% Arq file for the weather data with some
numeric features
%
%relation weather

%attribute outlook { sunny, overcast, rainy }
%attribute temperature numeric
%attribute humidity numeric
%attribute windy { true, false }
%attribute play? { yes, no }

%data
%
% 14 instances
%
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
rainy, 70, 96, false, yes
rainy, 68, 80, false, yes
rainy, 65, 70, true, no
overcast, 64, 65, true, yes
sunny, 72, 95, false, no
sunny, 69, 70, false, yes
rainy, 75, 80, false, yes
sunny, 75, 70, true, yes
overcast, 72, 90, true, yes
overcast, 81, 75, false, yes
rainy, 71, 91, true, no
```

Weka – Explorer

Weka Explorer

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

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

Filter: Choose **None** Apply

Current relation
Relation: weather
Instances: 14
Attributes: 5

Attributes: All None Invert Pattern

No.	Name
1	outlook
2	temperature
3	humidity
4	windy
5	play

Remove

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

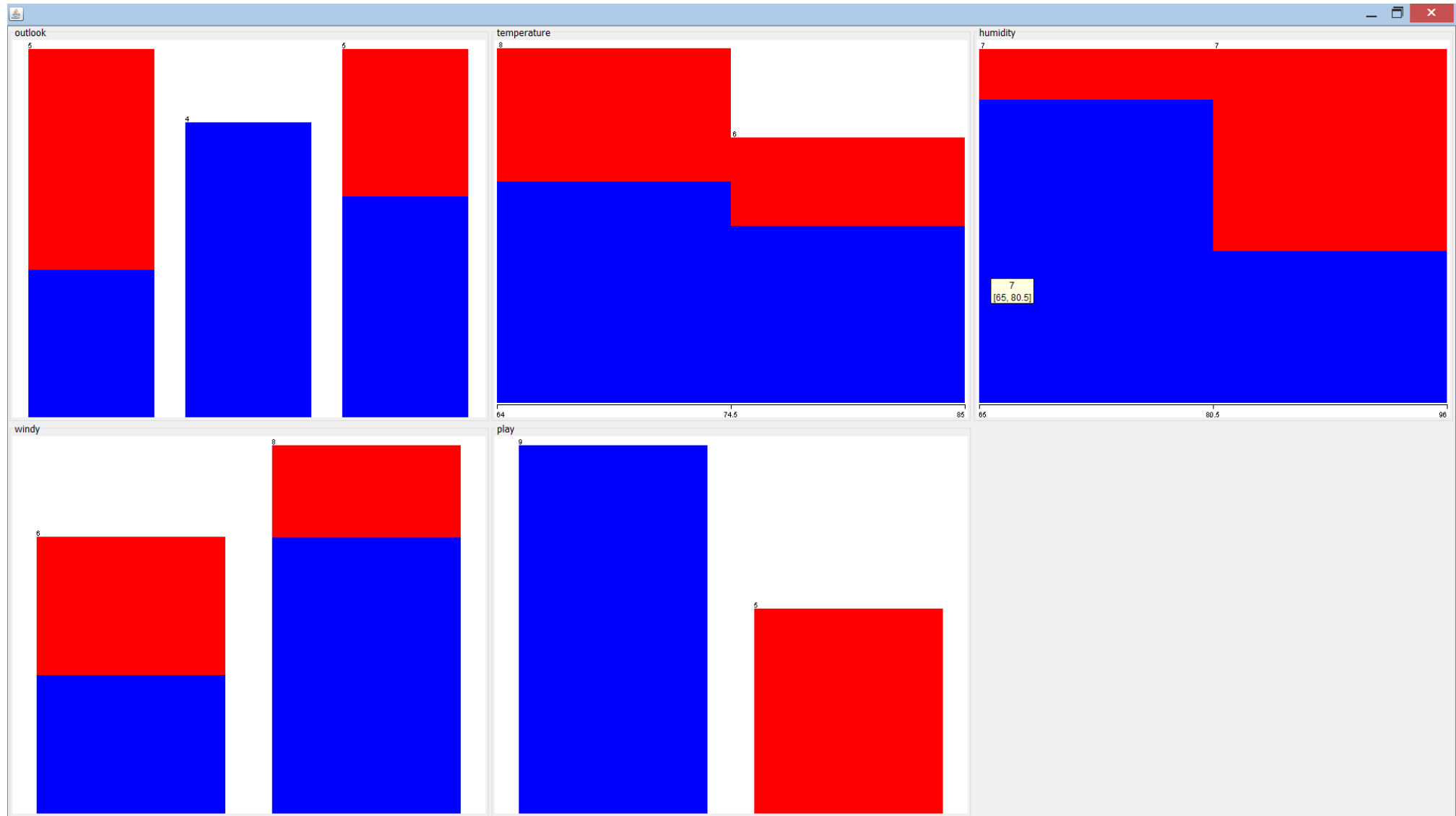
No.	Label	Count
1	sunny	5
2	overcast	4
3	rainy	5

Class: play (Nom) Visualize All

Outlook	play = no (blue)	play = yes (red)
sunny	5	5
overcast	4	0
rainy	5	5

Status: OK Log x 0

Visualização Gráfica



Avaliando cada atributo

Selected attribute		
Name: outlook		Type: Nominal
Missing: 0 (0%)		Unique: 0 (0%)
Distinct: 3		
No.	Label	Count
1	sunny	5
2	overcast	4
3	rainy	5

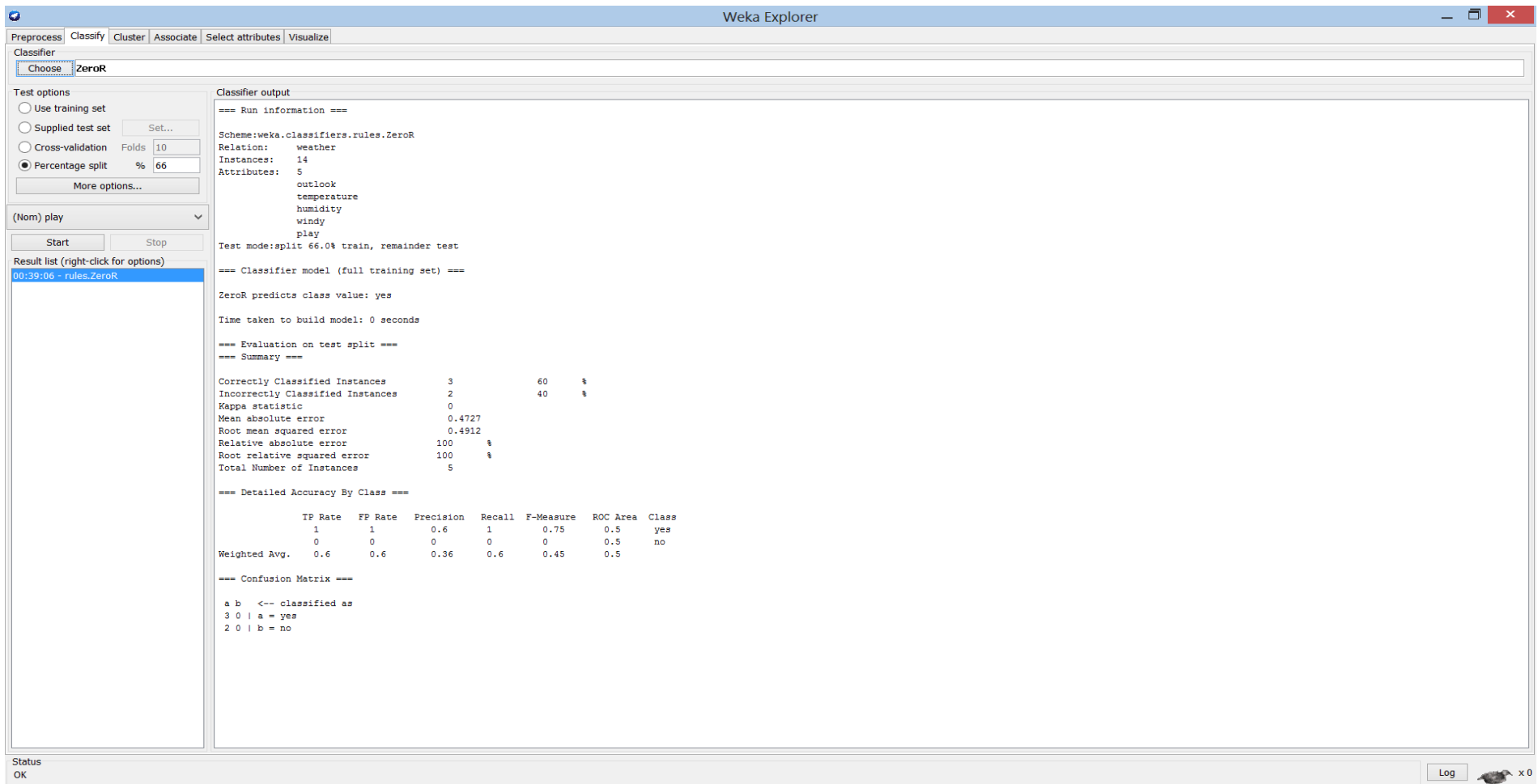
Selected attribute	
Name: temperature	
Missing: 0 (0%)	
Distinct: 12	
Type: Numeric	
Unique: 10 (71%)	
Statistic	Value
Minimum	64
Maximum	85
Mean	73.571
StdDev	6.572

Selected attribute	
Name: humidity	
Missing: 0 (0%)	
Distinct: 10	
Type: Numeric	
Unique: 7 (50%)	
Statistic	Value
Minimum	65
Maximum	96
Mean	81.643
StdDev	10.285

Selected attribute		
Name: windy		Type: Nominal
Missing: 0 (0%)		Unique: 0 (0%)
Distinct: 2		
No.	Label	Count
1	TRUE	6
2	FALSE	8

Classificador ZeroR

O classificador ZeroR prevê a classe mais frequente para atributos categóricos e a média para Atributos numéricos. Útil para servir de “baseline” para avaliação de outros classificadores.



The screenshot shows the Weka Explorer interface with the ZeroR classifier selected. The 'Test options' section on the left shows 'Percentage split' at 66%. The 'Classifier output' section on the right displays the following information:

```
=== Run information ===  
Scheme: weka.classifiers.rules.ZeroR  
Relation: weather  
Instances: 14  
Attributes: 5  
  outlook  
  temperature  
  humidity  
  windy  
  play  
Test mode: split 66.0% train, remainder test  
=== Classifier model (full training set) ===  
ZeroR predicts class value: yes  
Time taken to build model: 0 seconds  
=== Evaluation on test split ===  
=== Summary ===  
Correctly Classified Instances      3      60 %  
Incorrectly Classified Instances    2      40 %  
Kappa statistic                     0  
Mean absolute error                 0.4727  
Root mean squared error             0.4912  
Relative absolute error             100 %  
Root relative squared error         100 %  
Total Number of Instances          5  
=== Detailed Accuracy By Class ===  
      TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class  
      1      1      0.6      1      0.75      0.5      yes  
      0      0      0      0      0      0.5      no  
Weighted Avg.  0.6      0.6      0.36      0.6      0.45      0.5  
=== Confusion Matrix ===  
a b  <-- classified as  
3 0 | a = yes  
2 0 | b = no
```

The 'Result list' on the left shows the result for '00:39:06 - rules.ZeroR'.

Classificador J48 (C4.5) – Árvore de Decisão

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)

01:09:22 - trees.J48

Classifier output

temperature
humidity
windy
play

Test mode: split 66.0% train, remainder test

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

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

Time taken to build model: 0 seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances	2	40	%
Incorrectly Classified Instances	3	60	%
Kappa statistic	-0.3636		
Mean absolute error	0.6		
Root mean squared error	0.7746		
Relative absolute error	126.9231	%	
Root relative squared error	157.6801	%	
Total Number of Instances	5		

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.667	1	0.5	0.667	0.571	0.333	yes
	0	0.333	0	0	0	0.333	no
Weighted Avg.	0.4	0.733	0.3	0.4	0.343	0.333	

=== Confusion Matrix ===

a b <-- classified as

2 1 | a = yes

2 0 | b = no

Weka Classifier Tree Visualizer: 01:09:22 - trees.J48 (weather)

Tree View

```
graph TD; outlook((outlook)) -- sunny --> humidity((humidity)); outlook -- overcast --> yes40[yes (4.0)]; outlook -- rainy --> windy((windy)); humidity -- "<= 75" --> yes20[yes (2.0)]; humidity -- "> 75" --> no30[no (3.0)]; windy -- TRUE --> no20[no (2.0)]; windy -- FALSE --> yes30[yes (3.0)];
```

Navalha de Occam

- Entidades não devem ser multiplicadas sem necessidade
- Entre todas as hipóteses consistentes com a evidência, a mais simples é a mais provável de ser verdadeira.

Classificador OneR

The screenshot shows the Weka Explorer interface with the OneR classifier selected. The 'Test options' section on the left shows 'Percentage split' at 66%. The 'Result list' on the left shows the OneR model selected. The 'Classifier output' pane on the right displays the following information:

```
=== Run information ===
Scheme:weka.classifiers.rules.OneR -B 6
Relation: weather
Instances: 14
Attributes: 5
  outlook
  temperature
  humidity
  windy
  play
Test mode:split 66.0% train, remainder test

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

outlook:
  sunny   -> no
  overcast -> yes
  rainy   -> yes
(10/14 instances correct)

Time taken to build model: 0 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances      2      40 %
Incorrectly Classified Instances    3      60 %
Kappa statistic                    -0.3636
Mean absolute error                 0.6
Root mean squared error             0.7746
Relative absolute error             126.9231 %
Root relative squared error        157.6501 %
Total Number of Instances          5

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
      0.667    1      0.5      0.667  0.571    0.333   yes
      0      0.333    0      0      0      0.333   no
Weighted Avg.  0.4    0.733    0.3    0.4    0.343    0.333

=== Confusion Matrix ===

a b  <-- classified as
2 1 | a = yes
2 0 | b = no
```

The status bar at the bottom shows 'Status OK' and a 'Log' button.

Database Iris

Weka Explorer

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

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

Filter: Choose **None** Apply

Current relation:
Relation: iris
Instances: 150
Attributes: 5

Attributes: All None Invert Pattern

No.	Name
1	sepal.length
2	sepal.width
3	petal.length
4	petal.width
5	class

Remove

Selected attribute:
Name: class
Missing: 0 (0%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	Iris-setosa	50
2	Iris-versicolor	50
3	Iris-virginica	50

Class: class (Nom) Visualize All

Status: OK Log x 0

Iris - ZeroR

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose NaiveBayes

Test options

☐ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 10

☒ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

00:59:03 - rules.ZeroR

00:59:07 - rules.OneR

00:59:11 - trees.J48

00:59:18 - bayes.NaiveBayes

Classifier output

```
class
Test mode:split 66.0% train, remainder test

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

J48 pruned tree
-----

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
| petalwidth <= 1.7
| | petallength <= 4.9: Iris-versicolor (48.0/1.0)
| | petallength > 4.9
| | | petalwidth <= 1.5: Iris-virginica (3.0)
| | | petalwidth > 1.5: Iris-versicolor (3.0/1.0)
| petalwidth > 1.7: Iris-virginica (46.0/1.0)

Number of Leaves :    5
Size of the tree :    9

Time taken to build model: 0.02 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances      49           96.0784 %
Incorrectly Classified Instances     2           3.9216 %
Kappa statistic                    0.9408
Mean absolute error                 0.0396
Root mean squared error             0.1579
Relative absolute error             8.8979 %
Root relative squared error        33.4091 %
Total Number of Instances          51

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
      1      0      1      1      1      1      Iris-setosa
      1      0.063  0.905    1      0.95    0.969  Iris-versicolor
      0.882    0      1      0.882  0.938    0.967  Iris-virginica
Weighted Avg.   0.961    0.023  0.965    0.961    0.961    0.977

=== Confusion Matrix ===

 a  b  c  <-- classified as
15  0  0 | a = Iris-setosa
 0 19  0 | b = Iris-versicolor
 0  2 15 | c = Iris-virginica
```

Status
OK

Log x 0

Iris - J48

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose NaiveBayes

Test options

- ☐ Use training set
- ☐ Supplied test set Set...
- ☐ Cross-validation Folds 10
- ☒ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

- 00:59:03 - rules.ZeroR
- 00:59:07 - rules.OneR
- 00:59:11 - trees.J48
- 00:59:18 - bayes.NaiveBayes

Classifier output

Test mode: split 66.0% train, remainder test

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

J48 pruned tree

```
petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
| petalwidth <= 1.7
| | petalwidth <= 4.9: Iris-versicolor (48.0/1.0)
| | | petalwidth <= 1.5: Iris-virginica (3.0)
| | | petalwidth > 1.5: Iris-versicolor (3.0/1.0)
| | petalwidth > 4.9
| | | petalwidth <= 1.5: Iris-virginica (46.0/1.0)
| | | petalwidth > 1.5: Iris-versicolor (3.0/1.0)
```

Number of Leaves : 5

Size of the tree : 9

Time taken to build model: 0.02 seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances	49	96.0784 %
Incorrectly Classified Instances	2	3.9216 %
Kappa statistic	0.9408	
Mean absolute error	0.0396	
Root mean squared error	0.1579	
Relative absolute error	8.8979 %	
Root relative squared error	33.4091 %	
Total Number of Instances	51	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	1	0	1	1	1	1	Iris-setosa
1	0.882	0.063	0.905	1	0.95	0.969	Iris-versicolor
1	0.882	0	1	0.882	0.938	0.967	Iris-virginica
Weighted Avg.	0.961	0.023	0.965	0.961	0.961	0.977	

=== Confusion Matrix ===

a	b	c	-- classified as
15	0	0	a = Iris-setosa
0	19	0	b = Iris-versicolor
0	2	15	c = Iris-virginica

Weka Classifier Tree Visualizer: 00:59:11 - trees.J48 (iris)

Tree View

Status: OK

Log x 0

Iris - OneR

Weka Explorer

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier

Choose NaiveBayes

Test options

☐ Use training set

☐ Supplied test set Set...

☐ Cross-validation Folds 10

☒ Percentage split % 66

More options...

(Nom) class

Start Stop

Result list (right-click for options)

- 00:59:03 - rules.ZeroR
- 00:59:07 - rules.OneR
- 00:59:11 - trees.J48
- 00:59:18 - bayes.NaiveBayes

Classifier output

```
=== Run information ===

Scheme:weka.classifiers.rules.OneR -S 6
Relation: iris
Instances: 150
Attributes: 5
  sepalength
  sepalwidth
  petallength
  petalwidth
  class
Test mode:split 66.0% train, remainder test

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

petallength:
  < 2.45 -> Iris-setosa
  < 4.85 -> Iris-versicolor
  >= 4.85 -> Iris-virginica
(143/150 instances correct)

Time taken to build model: 0 seconds

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances      49      96.0784 %
Incorrectly Classified Instances    2       3.9216 %
Kappa statistic                    0.9408
Mean absolute error                 0.0261
Root mean squared error             0.1617
Relative absolute error              5.8688 %
Root relative squared error         34.2012 %
Total Number of Instances          51

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
      1      0      1      1      1      1      Iris-setosa
      1      0.063  0.905    1      0.95      0.969  Iris-versicolor
      0.882    0      1      0.882  0.938      0.941  Iris-virginica
Weighted Avg.  0.961  0.023  0.965  0.961  0.961      0.969

=== Confusion Matrix ===

  a  b  c  <-- Classified as
15  0  0 | a = Iris-setosa
 0 19  0 | b = Iris-versicolor
 0  2 15 | c = Iris-virginica
```

Status
OK

Log x 0

Outras abordagens para classificadores

- Naive Bayes
- Rules
- Support Vector Machines
- Genetic Algorithms
- Neural Net
- E muitos outros.

Como avaliar classificadores?

- Acurácia
- Custo/benefício total – quando diferentes erros envolvem diferentes custos
- Curvas de Lift e ROC
- Erro em previsões numéricas
- Quanto confiável são os resultados previstos?

Taxa de erro do classificador

- Medida de desempenho natural para problemas de classificação
 - Sucesso: a classe das instâncias é prevista corretamente
 - Erro: a classe das instâncias é prevista incorretamente
- Precisão, compreensível e interessante
- Acurácia = classificados corretamente / total de exemplos
- Erro = $1 - \text{Acurácia}$

Matriz de Confusão para duas classes

Classe	predita C_+ predita C_-		Taxa de Erro da Classe	Taxa de Erro Total
	T_P	F_N	$\frac{F_N}{T_P + F_N}$	$\frac{F_P + F_N}{n}$
verdadeira C_+				
verdadeira C_-	F_P	T_N	$\frac{F_P}{F_P + T_N}$	

T_P = Verdadeiro Positivo (True Positive)
 F_N = Falso Negativo (False Negative)
 F_P = Falso Positivo (False Positive)
 T_N = Verdadeiro Negativo (True Negative)
 $n = (T_P + F_N + F_P + T_N)$

J48 pruned tree

```

petalwidth <= 0.6: Iris-setosa (50.0)
petalwidth > 0.6
| petalwidth <= 1.7
| | petalwidth <= 4.9: Iris-versicolor (48.0/1.0)
| | petalwidth > 4.9
| | | petalwidth <= 1.5: Iris-virginica (3.0)
| | | petalwidth > 1.5: Iris-versicolor (3.0/1.0)
| petalwidth > 1.7: Iris-virginica (46.0/1.0)
    
```

Number of Leaves : 5

Size of the tree : 9

Correctly Classified Instances	144	96	%
Incorrectly Classified Instances	6	4	%
Kappa statistic	0.94		

=== Confusion Matrix ===

```

a b c <-- classified as
49 1 0 | a = Iris-setosa
0 47 3 | b = Iris-versicolor
0 2 48 | c = Iris-virginica
    
```

Aplicações de Data Mining

- Advertising
- Bioinformatics
- Customer Relationship Management (CRM)
- Database Marketing
- Fraud Detection
- eCommerce
- Health Care
- Investment/Securities
- Manufacturing, Process Control
- Sports and Entertainment
- Telecommunications
- Web

Data Mining e privacidade

- Data Mining busca PADRÕES e não PESSOAS.
- Soluções técnicas podem limitar a invasão de privacidade
 - Substituir informações sigilosas com um id anônimo
 - Fornecer saídas aleatórias
 - Utilizar rótulos em instâncias que escondam o real significado.

7 passos para aprender DM

- Languages: Learn R, Python, and SQL
- Tools: Learn how to use data mining and visualization tools
- Textbooks: Read introductory textbooks to understand the fundamentals
- Education: watch webinars, take courses, and consider a certificate or a degree in data science
- Data: Check available data resources and find something there
- Competitions: Participate in data mining competitions
- Interact with other data scientists, via social networks, groups, and meetings

Conclusão

- DM é necessário para tratar conjuntos grandes de dados.
- DM é um processo que permite compreender o comportamento dos dados.
- DM é uma etapa dentro do processo de KDD, embora atualmente DM e KDD seja encarada como a mesma atividade.
- Evitar overfitting.
- Pode ser bem aplicado em diversas áreas de negócios como auxiliar no processo decisório.

Onde conseguir mais informações?

- <http://weka.pentaho.com/>
- <http://www.cs.waikato.ac.nz/ml/weka/>
- Mineração de Dados - Conceitos, Aplicações e Experimentos com Weka
 - <http://www.lbd.dcc.ufmg.br/colecoes/erirjes/2004/004.pdf>
- Sítio da IBM
 - Mineração de dados com WEKA, Parte 1: Introdução e regressão
 - <http://www.ibm.com/developerworks/br/opensource/library/os-weka1/>
 - Mineração de dados com o WEKA, Parte 2: Classificação e armazenamento em cluster
 - <http://www.ibm.com/developerworks/br/opensource/library/os-weka2/>
- Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems)
- KDnuggets
 - news, software, jobs, courses,...
 - www.KDnuggets.com
- ACM SIGKDD – data mining association
 - www.acm.org/sigkdd

Não perca nossos próximos eventos!



VI
Fórum de
Tecnologia em
Software
Livre

18 e 19 de setembro
www.ftsl.org.br



**Software
Freedom Day**

20 de setembro de 2014
www.curitibalivre.org.br/sfd

Curitiba - Campus central da UTFPR



Contato

Obrigado a todos!

Marcos Vinicius Fidelis

mvfidelis@uepg.br - fidelis@utfpr.edu.br

