Data Mining com o programa Weka

Alípio Jorge

U.PORTO
FEP FACULDADE DE ECONOMIA UNIVERSIDADE DO PORTO

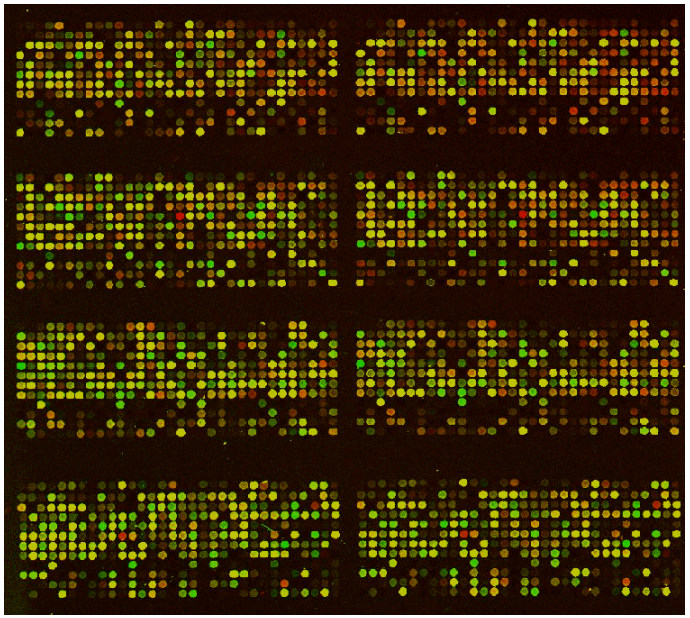
U.PORTO
LIAAD LABORATÓRIO DE INTELIGÊNCIA ARTIFICIAL E APOIO À DECISÃO
INESC PORTO LABORATÓRIO ASSOCIADO

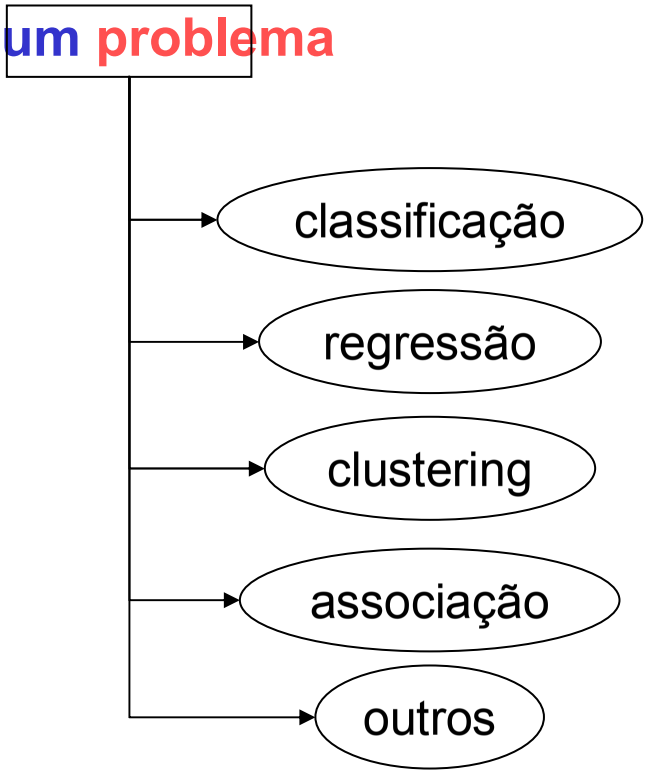
Sumário

- O que é Data Mining?
- Dados
- Classificação
- Regressão
- Clustering
- Associação
- Processo de Data Mining

O que é Data Mining?

- **Procura de padrões úteis em grandes quantidades de dados**
  - padrão: motivo que se repete com alguma frequência
  - útil: o padrão deve servir para resolver um problema





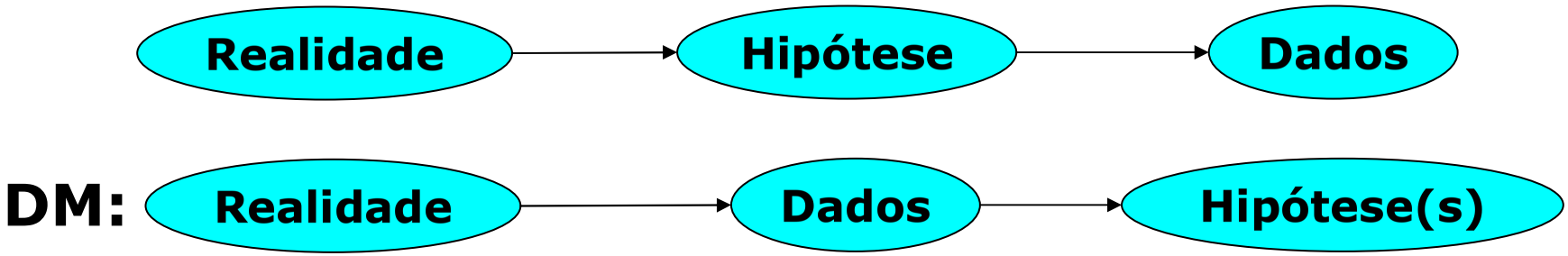
Dados

- **Tipicamente**
  - **tabela com N indivíduos e M atributos (depois de processados)**

sexo	IDA	CIV	ESCOL	PROF	HDORM	ACTIV	DESP	TAB	ALC	CAF	Peso	ALT	IMC	Colest
m	40anos	c	EstSuperiores	sup	8ha10h	pouca	nao	nao	bebe	sim	70a60Kg	m160	normal	alto
f	40anos	c	12ano	int	6ha8h	pouca	sim	nao	bebe	sim	70a60Kg	m150	excessopeso	baixo
f	50anos	s	9Classe	sup	6ha8h	pouca	nao	nao	nao	sim	50a60Kg	m150	normal	baixo
m	60anos	c	4Classe	semi-qual	6ha8h	pouca	nao	ex	bebe	nao	mais80	m160	excessopeso	medio
f	60anos	c	4Classe	sem-prof	menos6h	alguma	nao	nao	nao	sim	50a60Kg	m150	excessopeso	medio
f	50anos	c	EstSuperiores	sup	8ha10h	pouca	sim	nao	ocas	sim	50a60Kg	m150	normal	medio
m	40anos	c	4Classe	esp-man	mais10h	alguma	nao	ex	bebe	sim	mais80	m170	excessopeso	baixo
m	40anos	c	EstSuperiores	sup	6ha8h	nenhuma	sim	ex	bebe	sim	70a60Kg	m170	normal	baixo
m	40anos	c	4Classe	esp-n-man	6ha8h	pouca	sim	nao	bebe	sim	80a70kg	m160	excessopeso	medio
m	60anos	c	EstSuperiores	sup	8ha10h	nenhuma	sim	ex	bebe	sim	mais80	m170	excessopeso	medio
m	60anos	c	4Classe	semi-qual	8ha10h	pouca	nao	ex	ex	sim	70a60Kg	m180	normal	alto
m	50anos	c	9Classe	esp-n-man	8ha10h	pouca	nao	nao	bebe	sim	70a60Kg	m150	excessopeso	medio
f	40anos	v	4Classe	esp-n-man	8ha10h	nenhuma	nao	nao	nao	sim	50a60Kg	m160	normal	baixo
m	40anos	c	9Classe	esp-n-man	6ha8h	nenhuma	sim	fuma	bebe	sim	mais80	m160	obesidade	alto
f	50anos	c	12ano	int	6ha8h	alguma	sim	ex	bebe	sim	70a60Kg	m150	excessopeso	medio
m	50anos	c	12ano	int	8ha8h	pouca	sim	nao			80a70kg	m170	norm	medio
	40anos	d	EstSuperio			pouca	sim					m160		

Dados em DM

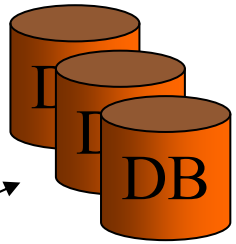
- dados observados (não experimentais)
  - recolhidos para outros fins
  - os dados não são recolhidos em função de uma hipótese
    - experiência de Galileu



- grandes quantidades de dados
  - scale-up dos métodos
  - preparação dos dados

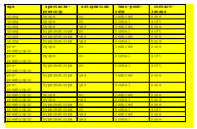
Data mining: o ciclo

action

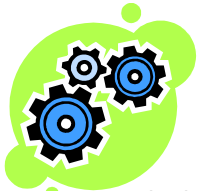


data bases

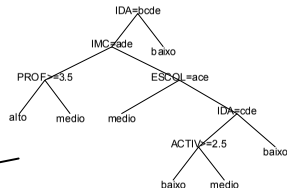
data warehouse



flat dataset



Data Mining



model

evaluation

decision support

Fontes de dados

- Actividade clínica (transaccional)
  - consultas
  - exames
  - tratamentos
  - diagnósticos
- Laboratório
  - genética, microarrays
  - proteínas, compostos químicos
- Textos
  - abstracts
  - receitas médicas
  - ...

O output: Padrões, Hipóteses, Modelos,...

- O sumário derivado pelos algoritmos de data mining:

- Padrão
  - afirmação sobre as variáveis envolvidas, válida para uma parte do espaço das variáveis (local)
- Modelo
  - afirmação sobre as variáveis, válida para todo o espaço (global)
- Hipótese
  - Modelo ou padrão candidato

Tipos de problemas de DM

- Classificação
- Regressão

} data mining dirigido

- Clustering
- Associação

} data mining exploratório

- Outros
  - o conjunto de tipos de problemas não é fechado

Problemas de Classificação

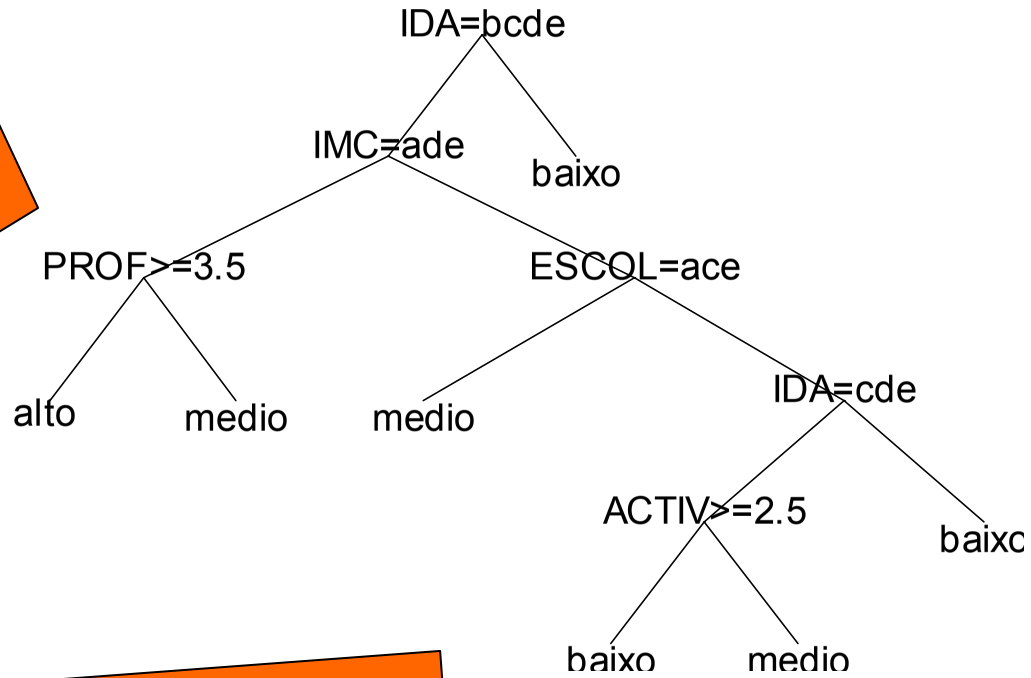
- **Dados**
  - N casos de classes conhecidas
- **Descobrir**
  - descobrir uma relação funcional f(X) =y
  - que a um novo caso X, faça corresponder uma classe y

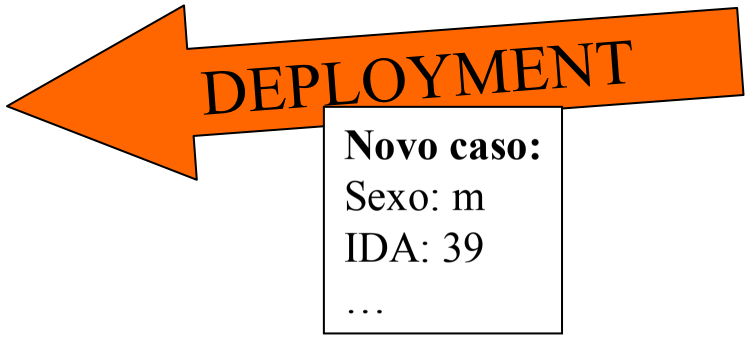
3 classes:
- alto
- médio
- baixo

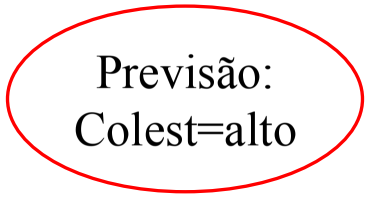
sexo	IDA	CIV	ESCOL	PROF	HDORM	ACTIV	DESP	TAB	ALC	CAF	Peso	ALT	IMC	Colest
m	40anos	c	EstSuperiores	sup	8ha10h	pouca	nao	nao	bebe	sim	70a60Kg	m160	normal	alto
f	40anos	c	12ano	int	6ha8h	pouca	sim	nao	bebe	sim	70a60Kg	m150	excessopeso	baixo
f	50anos	s	9Classe	sup	6ha8h	pouca	nao	nao	nao	sim	50a60Kg	m150	normal	baixo
m	60anos	c	4Classe	semi-qual	6ha8h	pouca	nao	ex	bebe	nao	mais80	m160	excessopeso	medio
f	60anos	c	4Classe	sem-prof	menos6h	alguma	nao	nao	nao	sim	50a60Kg	m150	excessopeso	medio
f	50anos	c	EstSuperiores	sup	8ha10h	pouca	sim	nao	ocas	sim	50a60Kg	m150	normal	medio
m	40anos	c	4Classe	esp-man	mais10h	alguma	nao	ex	bebe	sim	mais80	m170	excessopeso	baixo
m	40anos	c	EstSuperiores	sup	6ha8h	nenhuma	sim	ex	bebe	sim	70a60Kg	m170	normal	baixo
m	40anos	c	4Classe	esp-n-man	6ha8h	pouca	sim	nao	bebe	sim	80a70kg	m160	excessopeso	medio
m	60anos	c	EstSuperiores	sup	8ha10h	nenhuma	sim	ex	bebe	sim	mais80	m170	excessopeso	medio
m	60anos	c	4Classe	semi-qual	8ha10h	pouca	nao	ex	ex	sim	70a60Kg	m180	normal	alto
m	50anos	c	9Classe	esp-n-man	8ha10h	pouca	nao	nao	bebe	sim	70a60Kg	m150	excessopeso	medio
f	40anos	v	4Classe	esp-n-man	8ha10h	nenhuma	nao	nao	nao	sim	50a60Kg	m160	normal	baixo
m	40anos	c	9Classe	esp-n-man	6ha8h	nenhuma	sim	fuma	bebe	sim	mais80	m160	obesidade	alto
f	50anos	c	12ano	int	6ha8h	alguma	sim	ex	bebe	sim	70a60Kg	m150	excessopeso	medio
m	50anos	c	12ano	int	6ha8h	pouca	sim	nao	bebe	sim	80a70kg	m170	normal	medio
m	40anos	d	EstSuperiores	sup	6ha8h	pouca	sim	fuma	bebe	sim	70a60Kg	m160	normal	alto

Classificação

sexo	IDA	CIV	ESCOL	PROF	HDORM	ACTIV	DESP	TAB	ALC	CAF	Peso	ALT	IMC	Colest
m	40anos	c	EstSuperiores	sup	8ha10h	pouca	nao	nao	bebe	sim	70a60Kg	m160	normal	alto
f	40anos	c	12ano	int	6ha8h	pouca	sim	nao	bebe	sim	70a60Kg	m150	excessopeso	baixo
f	50anos	s	9Classe	sup	6ha8h	pouca	nao	nao	nao	sim	50a60Kg	m150	normal	baixo
m	60anos	c	4Classe	semi-qual	6ha8h	pouca	nao	ex	bebe	nao	mais80	m160	excessopeso	medio
f	60anos	c	4Classe	sem-prof	menos6h	alguma	nao	nao	nao	sim	50a60Kg	m150	excessopeso	medio
f	50anos	c	EstSuperiores	sup	8ha10h	pouca	sim	nao	ocas	sim	50a60Kg	m150	normal	medio
m	40anos	c	4Classe	esp-man	mais10h	alguma	nao	ex	bebe	sim	mais80	m170	excessopeso	baixo
m	40anos	c	EstSuperiores	sup	6ha8h	nenhuma	sim	ex	bebe	sim	70a60Kg	m170	normal	baixo
m	40anos	c	4Classe	esp-n-man	6ha8h	pouca	sim	nao	bebe	sim	80a70kg	m160	excessopeso	medio
m	60anos	c	EstSuperiores	sup	8ha10h	nenhuma	sim	ex	bebe	sim	mais80	m170	excessopeso	medio
m	60anos	c	4Classe	semi-qual	8ha10h	pouca	nao	ex	ex	sim	70a60Kg	m180	normal	alto
m	50anos	c	9Classe	esp-n-man	8ha10h	pouca	nao	nao	bebe	sim	70a60Kg	m150	excessopeso	medio
f	40anos	v	4Classe	esp-n-man	8ha10h	nenhuma	nao	nao	nao	sim	50a60Kg	m160	normal	baixo
m	40anos	c	9Classe	esp-n-man	6ha8h	nenhuma	sim	fuma	bebe	sim	mais80	m160	obesidade	alto
f	50anos	c	12ano	int	6ha8h	alguma	sim	ex	bebe	sim	70a60Kg	m150	excessopeso	medio
m	50anos	c	12ano	int	6ha8h	pouca	sim	nao	bebe	sim	80a70kg	m170	normal	medio
m	40anos	d	EstSuperiores	sup	6ha8h	pouca	sim	fuma	bebe	sim	70a60Kg	m160	normal	alto



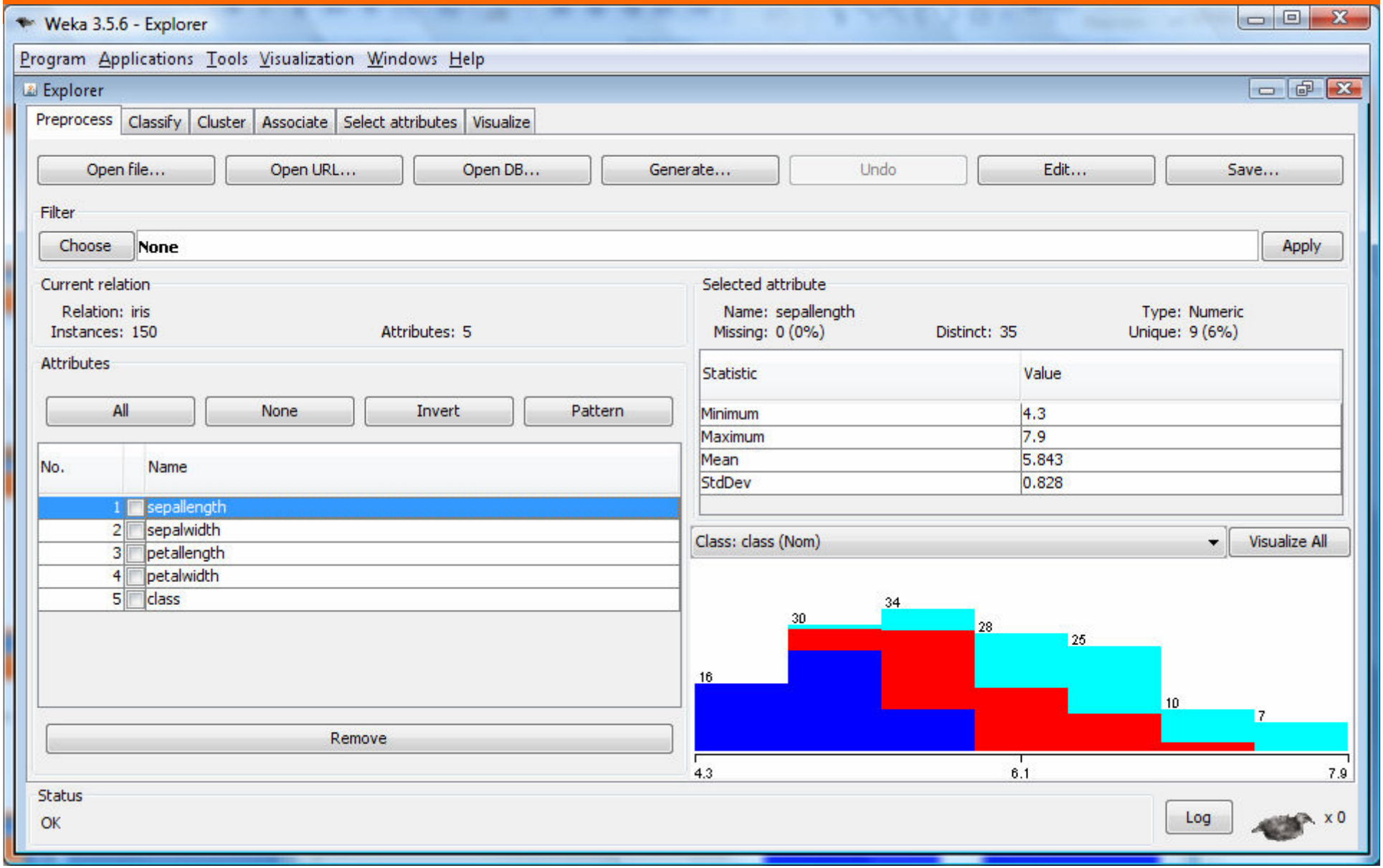




Weka

- **WEKA (da Universidade de Waikato, Nova Zelândia)**
  - http://www.cs.waikato.ac.nz/ml/weka/
  - é um suite de DM
  - tem uma vasta colecção de métodos disponíveis
  - faz pré-processamento
  - faz (alguma) visualização
  - permite organizar experiências
  - é gratuito
  - é open source (java)

Weka


Weka 3.5.6 - Explorer
Program Applications Tools Visualization Windows Help
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 sepallength
2 sepalwidth
3 petallength
4 petalwidth
5 class
Remove
Selected attribute
Name: sepallength Type: Numeric
Missing: 0 (0%) Distinct: 35 Unique: 9 (6%)
Statistic Value
Minimum 4.3
Maximum 7.9
Mean 5.843
StdDev 0.828
Class: class (Nom) Visualize All
34
30
28
25
16
10
7
4.3
6.1
7.9
Status
OK
Log
x 0

Weka: dados

- **ARFF (Attribute-Relation File Format)**
  - o formato nativo do Weka

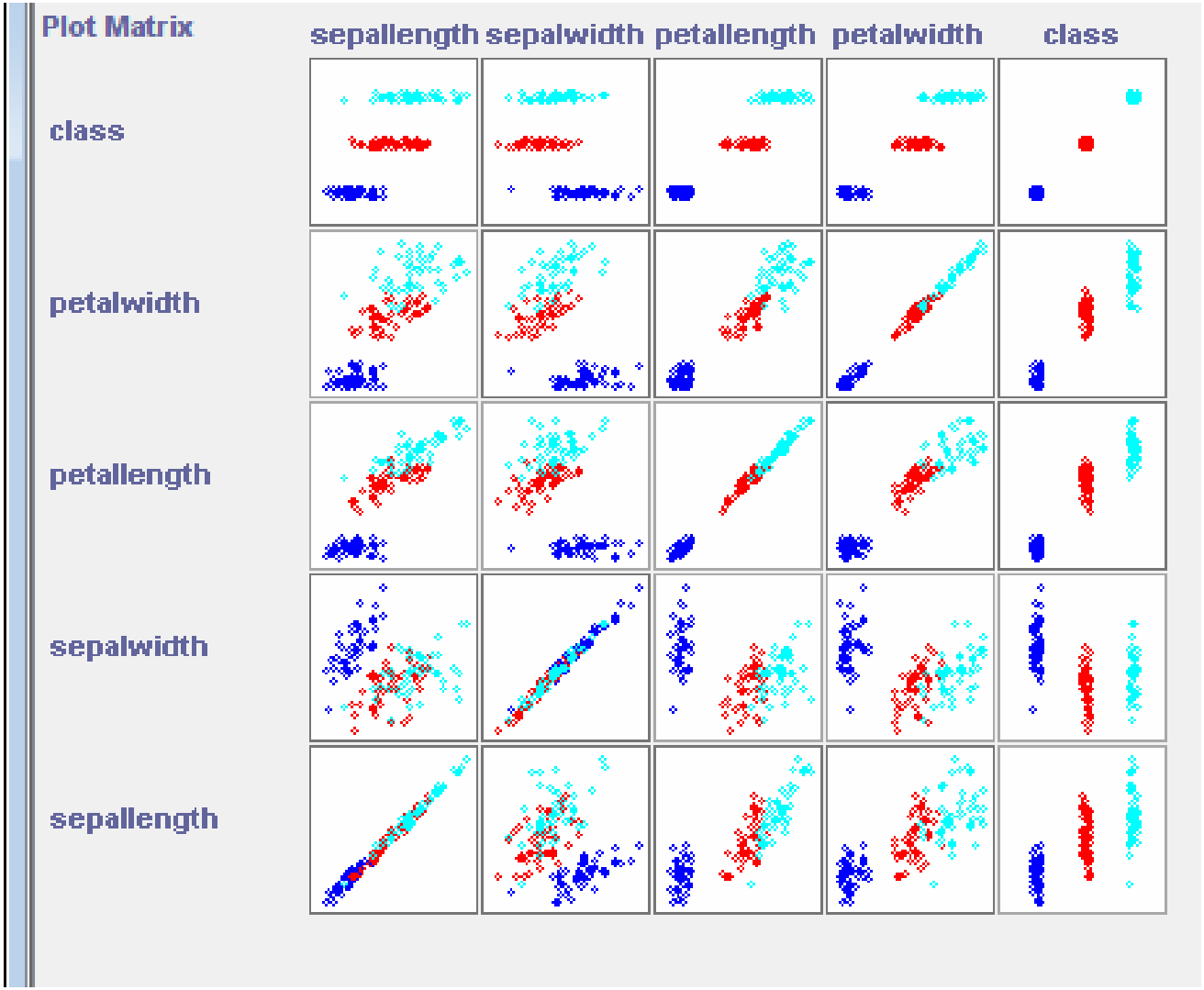
```
@RELATION iris

@ATTRIBUTE sepallength          REAL
@ATTRIBUTE sepalwidth           REAL
@ATTRIBUTE petallength          REAL
@ATTRIBUTE petalwidth REAL
@ATTRIBUTE class       {Iris-setosa,Iris-versicolor,Iris-virginica}

@DATA
5.1,3.5,1.4,0.2,Iris-setosa
4.9,3.0,1.4,0.2,Iris-setosa
4.7,3.2,1.3,0.2,Iris-setosa
…
```

- **CSV (Comma Sepparated Value)**
  - tem conversor para este formato e outros formatos populares

Actividade : Iris : compreensão dos dados

- O conjunto de dados Iris
  - descreve 150 plantas de três tipos diferentes. O que distingue esses 3 tipos?
- Carregar os dados Iris para o Weka
- Observar as distribuições dos valores
- procurar visualmente atributos correlacionados com o target (class / Espécie)
- interpretar o plot em Visualize
  - jitter: espalha os pontos
  - class colour
  - plot size: escala do gráfico

Actividade : Iris : visualize



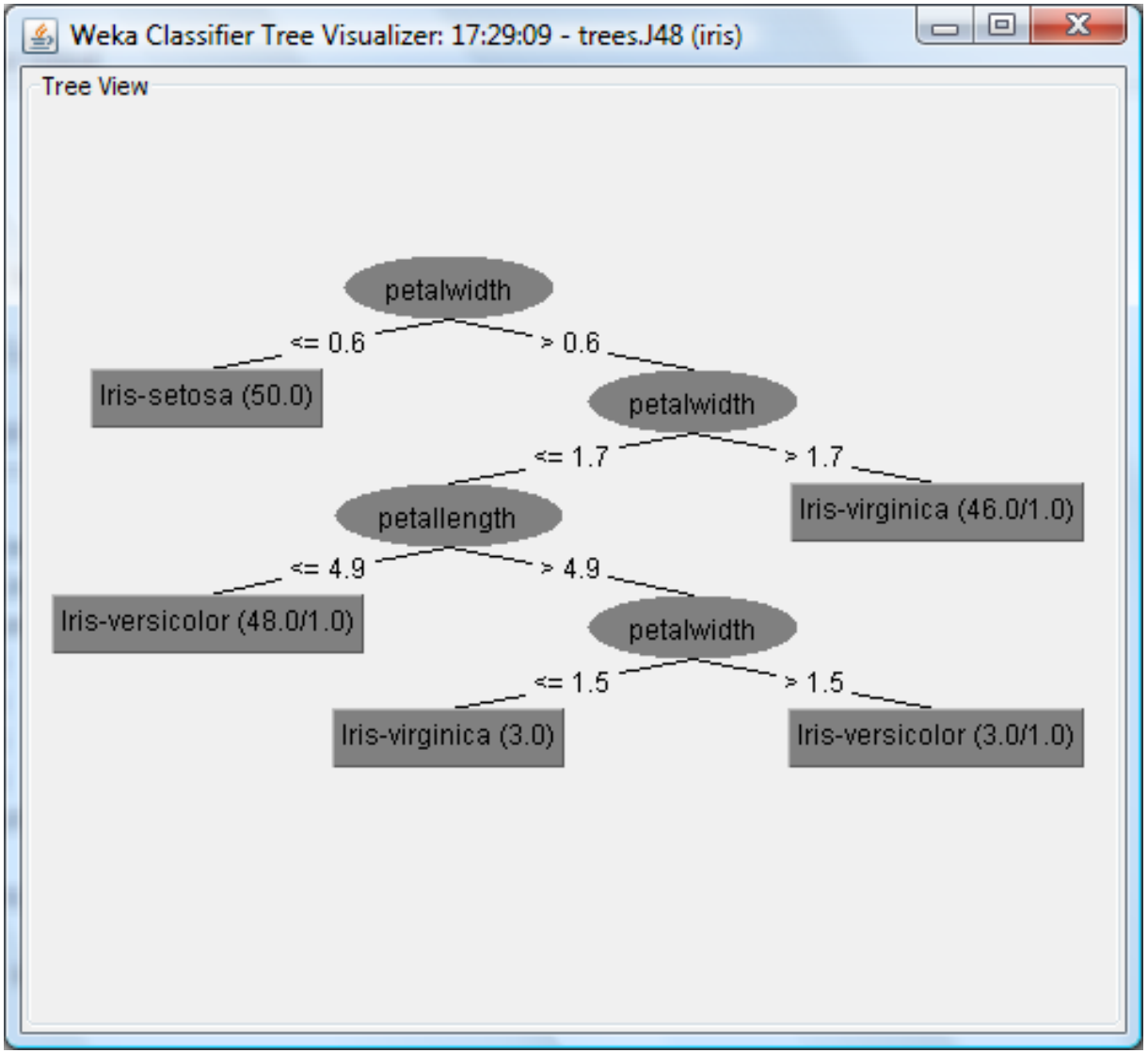
Actividade : Iris : Classificação : OneR

- **Método OneR (um classificador muito simples)**
  - correr o método com treino e teste: split de 66%
  - qual o modelo obtido?
  - quantos respostas erradas no teste?
  - qual a percentagem de erro?
  - qual a classe mais fácil?
  - quais as mais difíceis?
  - quais as classes que se confundiram no teste?
  - relacione a matriz de confusão com a visualização
  - correr com cross-validation: mudou a sua opinião sobre a qualidade do modelo?

Actividade : Iris : Classificação : J48

- **Método J48 (árvore de decisão)**
  - usar cross validation:
  - qual a árvore obtida? Como se lê?
  - visualize a árvore (botão direito sobre Result list)
  - qual a percentagem de erro?
  - como se compara com o OneR?

Visualizar a árvore



Interpretar o output

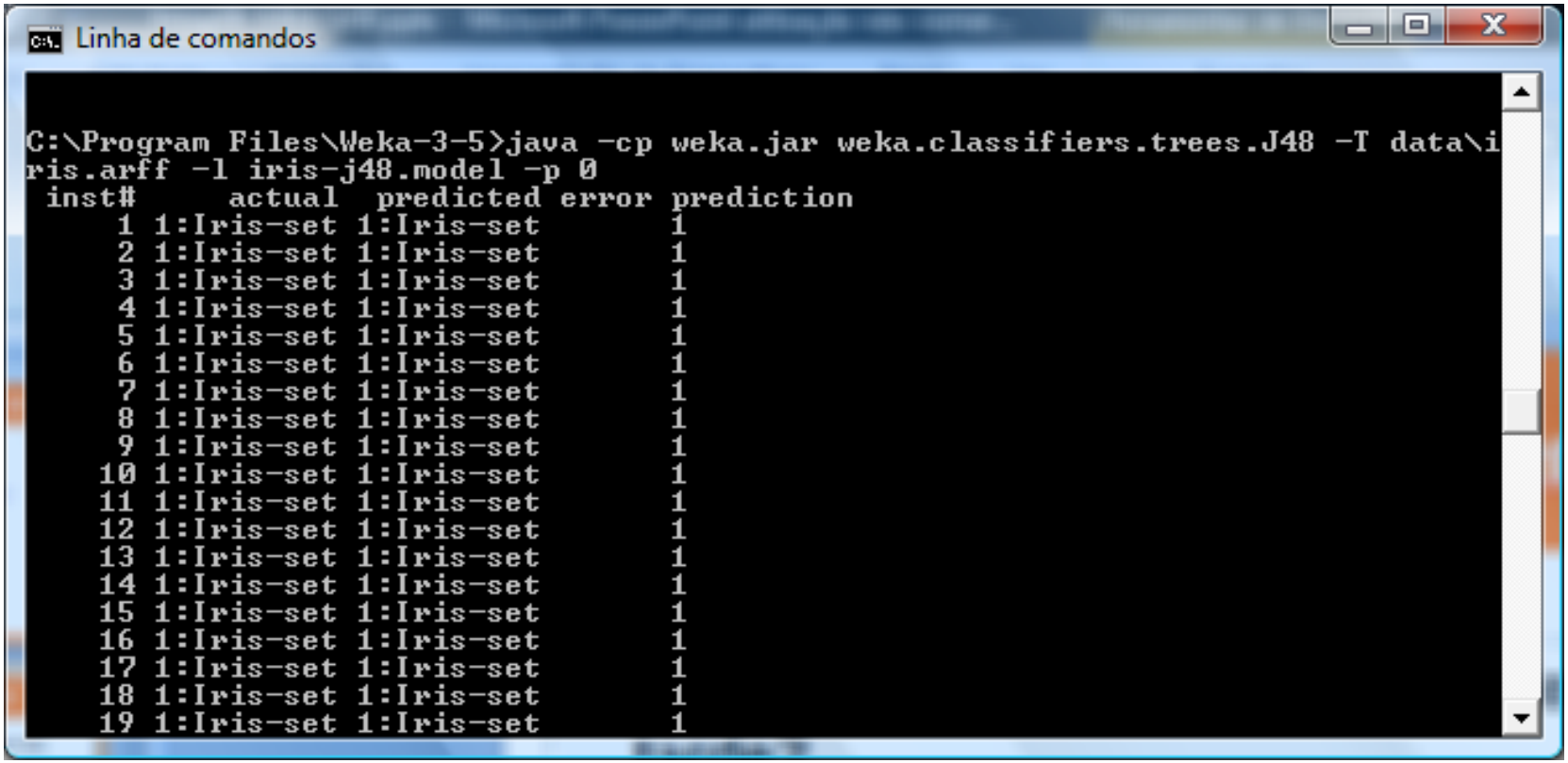
- mean absolute error
  - SOMA (|prob(prev) – verdade| / N)
- TP rate (true positives) da classe C
  - #previstos correctamente da classe C / #exemplos da classe C
- FP rate (false positives) da classe C
  - #previstos como da classe C / #exemplos de outras classes
- Precision da classe C
  - #previstos correctamente da classe C / #previstos na classe C
- Recal da classe C
  - #previstos correctamente da classe C /#exemplos da classe C
- F-measure
  - 2*Prec*Rec / (Prec+Rec)

Actividade : Iris : Deployment : J48

- Como se faz para guardar o modelo e o usar mais tarde?
  - com o botão direito do rato sobre o item da result list
    - save model (ex. iris-j48.model)
  - o modelo é agora um ficheiro autónomo
- Novos dados
  - são colocados no formato ARFF
- Usar o modelo
  - na linha de comando

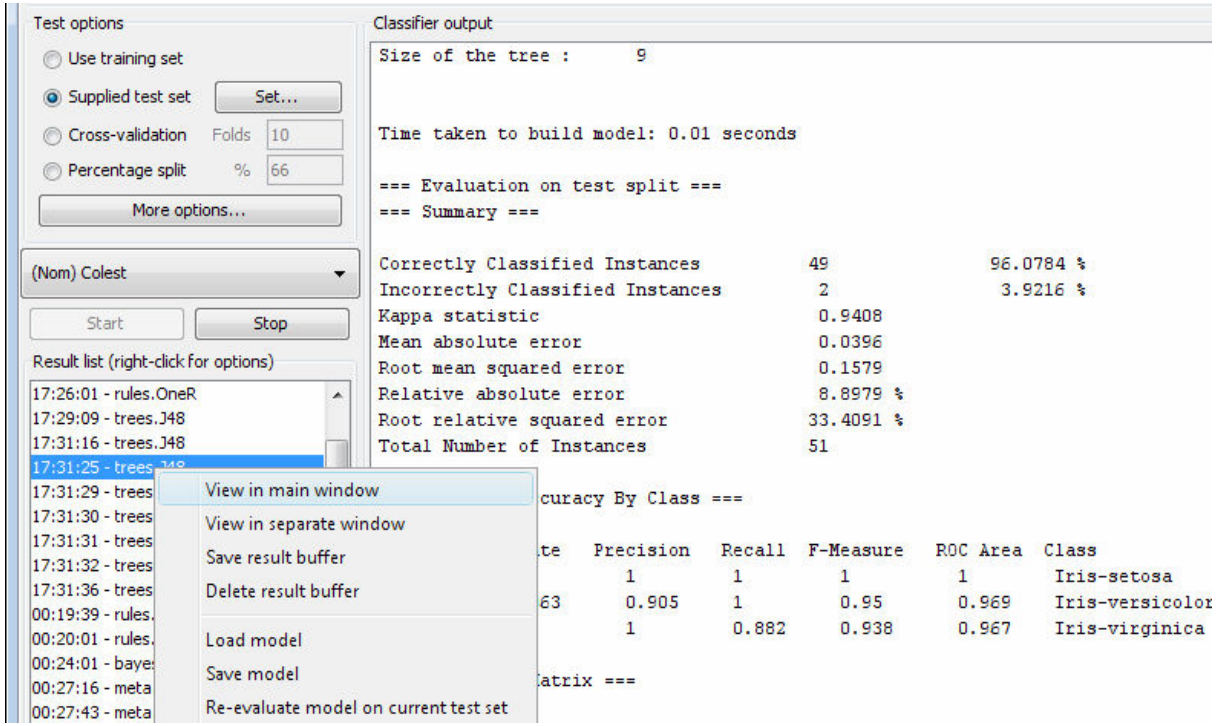
```
java -cp weka.jar weka.classifiers.trees.J48 -T
iris-novos.arff -l iris-j48.model -p 0
```

Actividade : Iris : Deployment : J48



Actividade : Iris : Deployment : J48

- De uma forma mais rápida (mas menos flexível)
  - usar Supplied test set, com novos exemplos, classe "?"
  - carregar o modelo (ou aprendê-lo)
  - com o botão direito sobre a Result list, Re-evaluate model on current test set.



Actividade : Iris : Classificação : PART

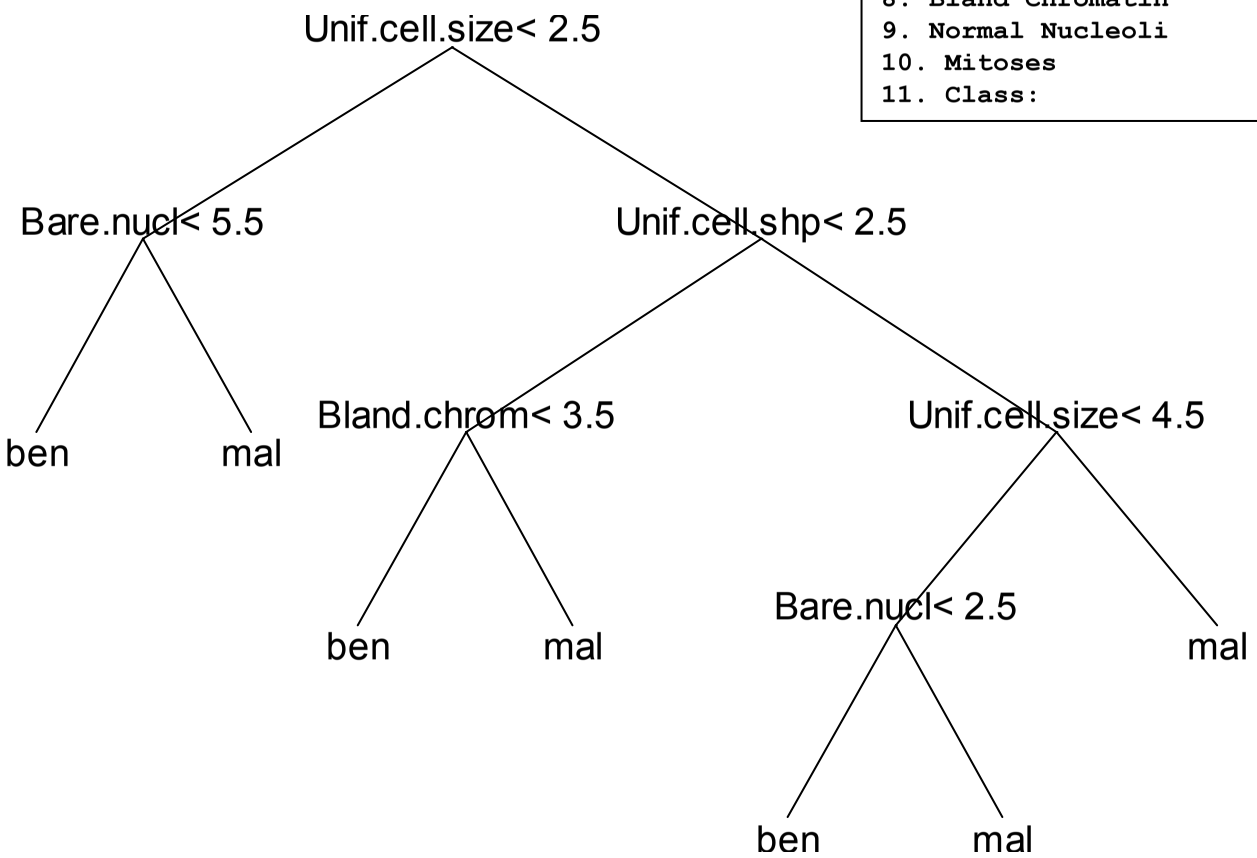
- **PART**
  - obtém regras a partir de árvores de decisão parciais

- **regras**
  - petalwidth <= 0.6: Iris-setosa (50.0)
  - petalwidth <= 1.7 AND
  - petallength <= 4.9: Iris-versicolor (48.0/1.0)
  - : Iris-virginica (52.0/3.0)

Classificação: breast cancer>diagnóstico

- **Breast cytology**
  - **Análise de 699 exames**

Attributes	
1. Sample code number	id number
2. Clump Thickness	1 - 10
3. Uniformity of Cell Size	1 - 10
4. Uniformity of Cell Shape	1 - 10
5. Marginal Adhesion	1 - 10
6. Single Epithelial Cell Size	1 - 10
7. Bare Nuclei	1 - 10
8. Bland Chromatin	1 - 10
9. Normal Nucleoli	1 - 10
10. Mitoses	1 - 10
11. Class:	benign, malignant



Actividade : breast cancer

- Obtenha e avalie classificadores para os dados
  - como são os dados? distribuição das classe
  - OneR
  - J48 (veja a árvore)
  - erro?
  - clas
- Como se compara com o conjunto de dados Iris?

Classificação: Caso

- **Fertilização "in vitro"**
  - recolha de óvulos na mulher
  - fertilização externa -> produção de vários embriões
  - selecção de embriões
  - colocação de embriões no útero

- **O problema**
  - seleccionar os embriões mais promissores.
  - 60 variáveis descritivas
  - dados históricos
  - limites cognitivos do embriologista

Classificação: Caso

- **Análise:**
  - **Situação operacional:**
    - recolha de vários óvulos de uma mulher, para fertilização externa e posterior reimplantação no útero da mulher.
  - **Problema de decisão:**
    - Quais os embriões a seleccionar (quais os que têm mais hipóteses de sobreviver?)
  - **Falta de conhecimento sobre o processo de decisão:**
    - não há teoria (ou é incompleta) sobre como seleccionar os embriões.
  - **Necessidade de obtenção do conhecimento explícito:**
    - Validação dos critérios de decisão obtidos automaticamente
  - **Necessidade de automação do processo:**
    - demasiadas variáveis e conhecimento relacionado, limites cognitivos do embriologista.
  - **Existência de dados:**
    - Registos históricos.
  - **Utilização de uma técnica de DM:**
    - classificação.

Classificação

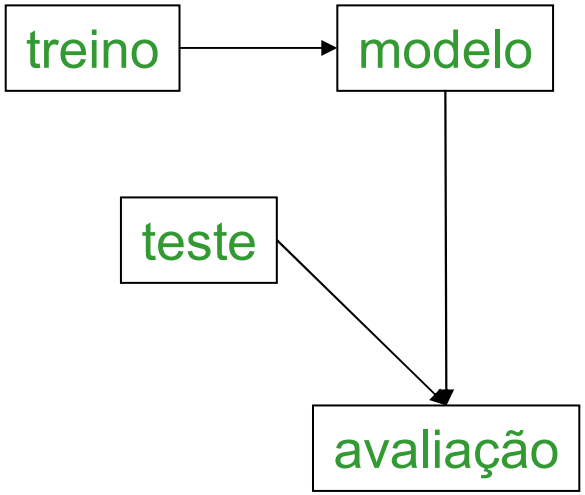
- **Aplicações**
  - **Microarrays**
    - quais os genes importantes para esta doença?
  - **Farmacologia**
    - qual o padrão que corresponde a um químico carcinogénico?
  - **Diagnóstico a partir de imagens (mamografia)**
    - benigno ou maligno?
  - **…**

Classificação

- **Avaliação**
  - **Que confiança podemos ter na resposta do modelo de classificação?**
  - **Importa estimar o erro cometido pelo modelo**
    - assumindo que dados futuros mantêm a distribuição presente
  - **O que medir?**
    - Error rate: percentagem de respostas erradas
    - Accuracy / acerto: percentagem de respostas certas
    - Recall: percentagem de casos de uma classe identificados
    - Precision: acerto numa classe
  - **Esse estimador deve ser fiável**
    - pode ser medido de várias formas

Classificação: Avaliação

- **Exemplos de treino**
  - para construir o modelo
  - *training-set*
- **Exemplos de teste**
  - para avaliar o modelo
  - *test-set*



- **Distribuição dos exemplos**
  - geralmente assume-se dist. treino = dist. teste
  - *concept drift* (deriva conceptual)
- **Avaliação mais robusta**
  - validação cruzada
  - leave-one-out

Classificação

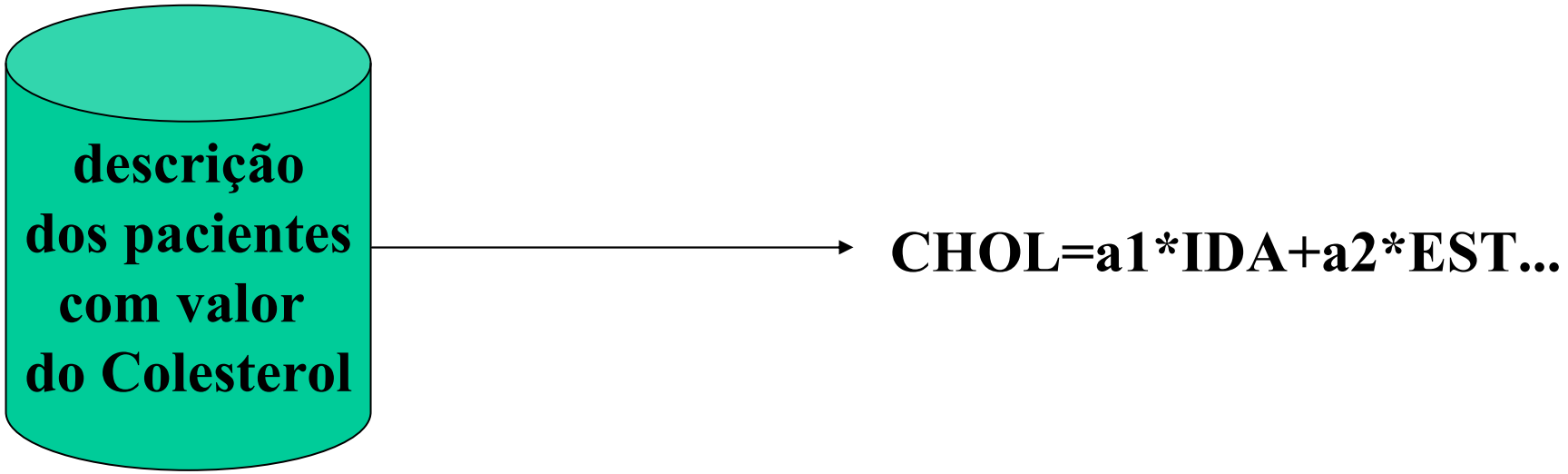
- **Técnicas**
  - Árvores de decisão
  - Sistemas de regras
  - Support Vector Machines (SVM)
  - Redes Neuronais (NN)
  - Discriminante linear
  - Regressão logística
  - Abordagens Bayesianas
  - Baseadas em instâncias
  - …

Actividade : Comparação

- **Com o dataset Colestlab**
  - **comparar métodos**
    - OneR
    - J48
    - Bagging com OneR
    - Bagging com J48
  - **Sugestão: usar o experimenter do Weka**

Problemas de Regressão

- **Dados**
  - N casos descritos por M variáveis, sendo uma delas a var. objectivo
- **Descobrir**
  - uma relação funcional f(X) =y
  - que a um novo caso X, faça corresponder um novo valor y



Problemas de Regressão

- **Aplicações**
  - Previsão numérica
  - Séries temporais
  - Planeamento logístico
- **Avaliação**
  - erro: distância "média" entre previsão e verdade
  - treino e teste, validação cruzada, etc.
- **Técnicas**
  - Regressão linear
  - Árvores de regressão
  - Support vector machines
  - …

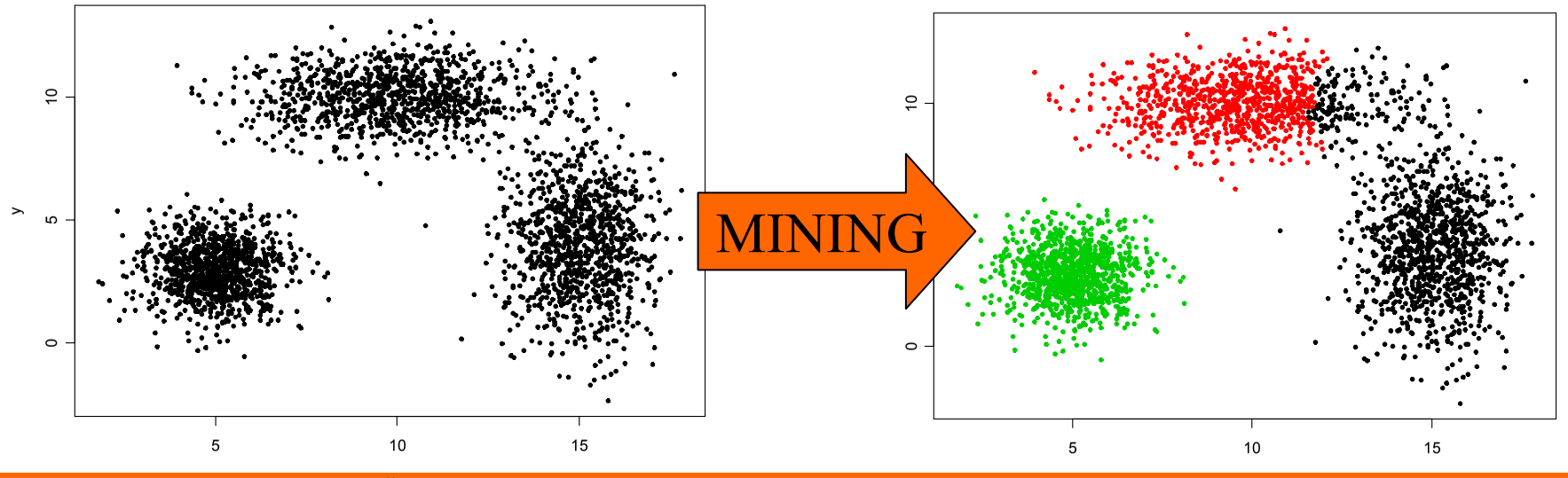
DM: dirigido vs. exploratório

- **Dirigido (Classificação, Regressão,...)**
  - há um objectivo bem definido
  - é facilmente avaliável
  - presta-se à automação das decisões

- **Exploratório (Clustering, Associação,Visual DM,...)**
  - procuram-se padrões interessantes, accionáveis
  - é difícil de avaliar
  - o perito é muito importante no ciclo de descoberta
  - a visualização é importante

Problemas de Clustering

- **Dados**
  - N casos, sem classe conhecida
- **Descobrir**
  - Uma partição dos casos que maximize a semelhança intra-grupo e a dissemelhança entre grupos distintos
- **Aplicações**
  - descobrir grupos de genes / pacientes / compostos químicos



Clustering: breast cancer

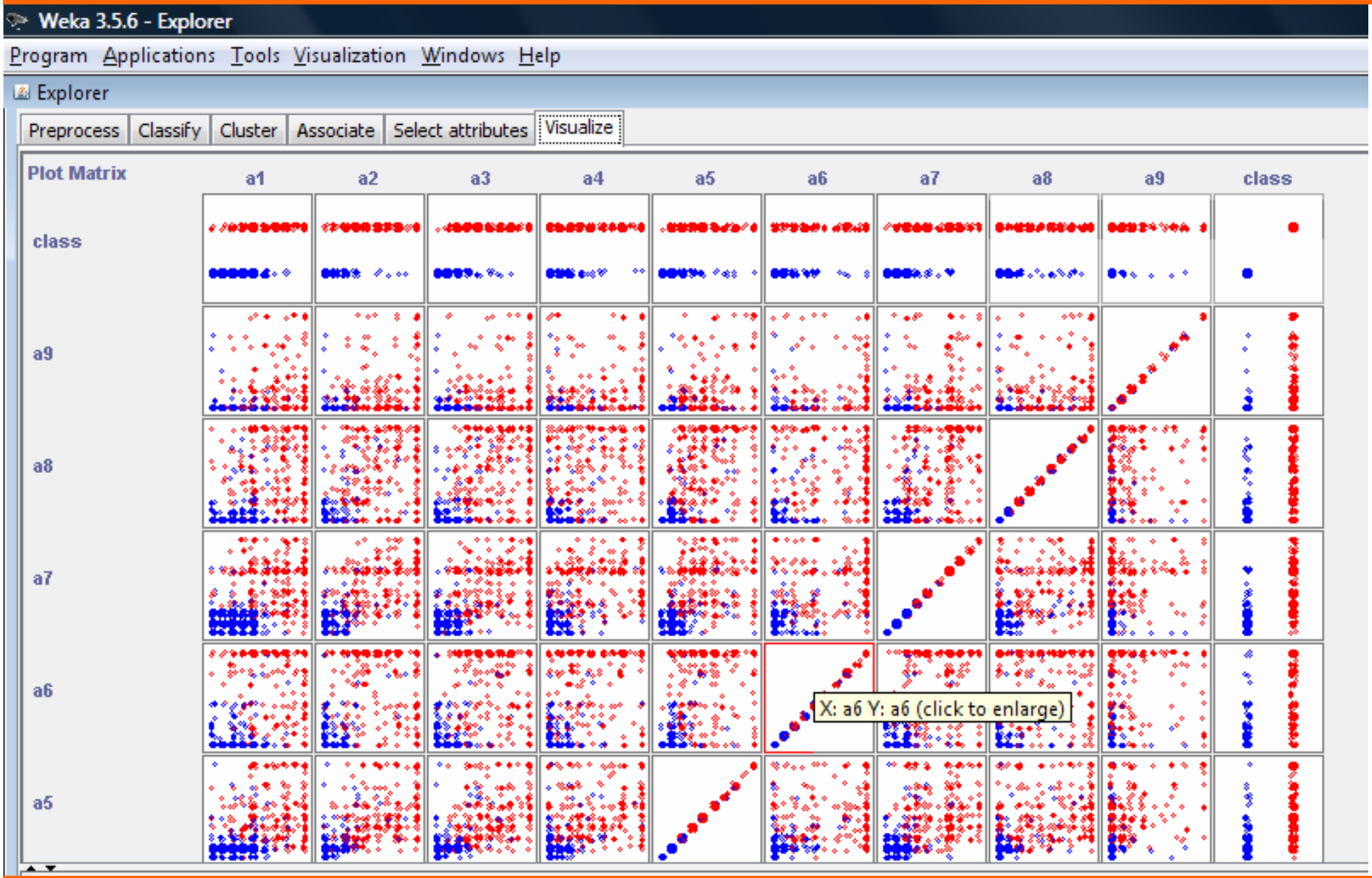
- Vamos tentar redescobrir a classe
  - Usamos clustering hierárquico
  - Pedimos 2 clusters
  - Fornecemos a tabela com todos os atributos menos Diag
  - Confrontamos com os verdadeiros valores de classe
  - Sobreposição de 91% (avaliação?)

cluster	ben	mal
1	452	50
2	6	191

Actividade : Weka : Clustering : Breast Cancer

- Carregue os dados Breast Cancer
  - use o SimpleKMeans
  - use class para avaliar
  - veja os resultados

Clustering: breast cancer > visualização



Problemas de Associação

- **Dados**
  - N casos descritos por M variáveis
- **Descobrir**
  - Co-ocorrências frequentes das variáveis
- **Aplicações**
  - Prognóstico / prevenção / planeamento
  - Descoberta de interacção entre factores
- **Técnicas**
  - Regras de associação
  - ...

Regras de Associação

Rules	Support	Confidence
IMC=obesidade, DESP=nao, ALC=bebe -> Colest=alto	0,03714	0,5
IMC=obesidade, DESP=nao, CIV=c, ALC=bebe -> Colest=alto	0,03429	0,52174
IMC=obesidade, DESP=nao, CIV=c, CAF=sim, ALC=bebe -> Colest=alto	0,03333	0,51471
DESP=nao, CAF=sim, IDA=60anos, TAB=nao -> Colest=alto	0,02952	0,51667
CAF=sim, IDA=60anos, TAB=nao, ALC=bebe -> Colest=alto	0,02476	0,53061
Peso=80a70kg, ALT=m150, DESP=nao, CIV=c -> Colest=alto	0,02381	0,5102
HDORM=8ha10h, ESCOL=4Classe, DESP=nao, CIV=c, CAF=sim, TAB=nao -> Colest=alto	0,02381	0,5
HDORM=8ha10h, ESCOL=4Classe, ALT=m150, CIV=c, CAF=sim -> Colest=alto	0,02286	0,6
HDORM=8ha10h, ESCOL=4Classe, ALT=m150, CIV=c -> Colest=alto	0,02286	0,55814
ACTIV=pouca, DESP=nao, IDA=60anos -> Colest=alto	0,02286	0,53333
CIV=c, IDA=60anos, TAB=nao, ALC=bebe -> Colest=alto	0,02286	0,53333
IMC=obesidade, DESP=nao, TAB=nao, ALC=bebe -> Colest=alto	0,02286	0,51064
HDORM=8ha10h, IMC=excessopeso, ALT=m150 -> Colest=alto	0,02286	0,51064
CIV=c, CAF=sim, IDA=60anos, TAB=nao, ALC=bebe -> Colest=alto	0,0219	0,62162
HDORM=8ha10h, CAF=sim, IDA=60anos, TAB=nao -> Colest=alto	0,0219	0,52273
Peso=80a70kg, sexo=f, ALT=m150, DESP=nao, CIV=c -> Colest=alto	0,0219	0,52273
HDORM=8ha10h, sexo=f, ESCOL=4Classe, CIV=c, CAF=sim -> Colest=alto	0,0219	0,52273
sexo=f, DESP=nao, CAF=sim, IDA=60anos, TAB=nao -> Colest=alto	0,0219	0,51111

Regras de Associação

- **Simbolicamente**

  { A1, A2 } => { A3 }

  – **Se observarmos um conjunto de** itens **A1 e A2 devemos também observar o item A3.**

  – suporte
    - percentagem dos cestos onde a co-ocorrência se observa
    - Estima Prob(A1 & A2 & A3)
  – confiança
    - percentagem dos casos onde a ocorrência de {A1,A2} "prevê" correctamente a ocorrência de {A3}
    - Estima Prob(A3 | A1 & A2)

Em geral

- dado um conjunto de transacções D
- D = {t | t é um conjunto de items i}

$$D = \{t \mid t \text{ é um conjunto de items } i\}$$

- uma regra de associação tem a forma A→B
- A e B são conjuntos de items

$$\text{suporte}(A \rightarrow B) = \frac{\#(A \cup B)}{\#D} \qquad \text{confiança}(A \rightarrow B) = \frac{\#(A \cup B)}{\#A}$$

Interesse de uma regra

- **Regra interessante [Silberschatz & Tuzhilin]**
  - **Inesperada: surpreendente para o utilizador**
    - medida de interesse: *desvio do esperado ou do acreditado*
  - **Útil (accionável)**
    - medida de interesse: *benefício estimado*
- **Subjectivo**
  - **interesse depende do conhecimento do utilizador**
    - um utilizador pode identificar uma oportunidade de aplicação
  - **envolve itens de interesse**
    - um utilizador pode estar mais interessado num determinado domínio
- **Objectivo**
  - **desvio da independência estatística**
  - **valores que se destacam**
    - e.g.: regras sintactimente singulares

Regras interessantes

- **Tipicamente**
  - ***A*→*B* é interessante se *A* e *B* não são estatisticamente independentes**
  - **se *A* e *B* são estatisticamente independentes**

    $sup(A \cup B) \approx sup(A) \times sup(B)$

    $conf(A \rightarrow B) \approx conf(\emptyset \rightarrow B)$
  - ***A*→*B* pode ter suporte elevado, confiança elevada e não ser interessante.**

    { jornal } → { combustível }

    sup = 5 %

    conf = 95 %

    não é *inesperada*

    não é *útil*

Regras interessantes

- **O desvio da independência**
  - **Regra inesperada ≈ que se desvia do pressuposto de independência**
- **Medidas de interesse**
  - **quociente**

$$\frac{probabilidade\,a\,posteriori}{probabilidade\,a\,priori} > 1$$

$$lift(A \rightarrow B) = \frac{confiança(A \rightarrow B)}{suporte(B)}$$

$$conviction(A \rightarrow B) = \frac{suporte(\neg B)}{lift(A \rightarrow \neg B)}$$

  - **diferença**

$$probabilidade\,a\,posteriori - probabilidade\,a\,priori > 0 \quad leverage$$

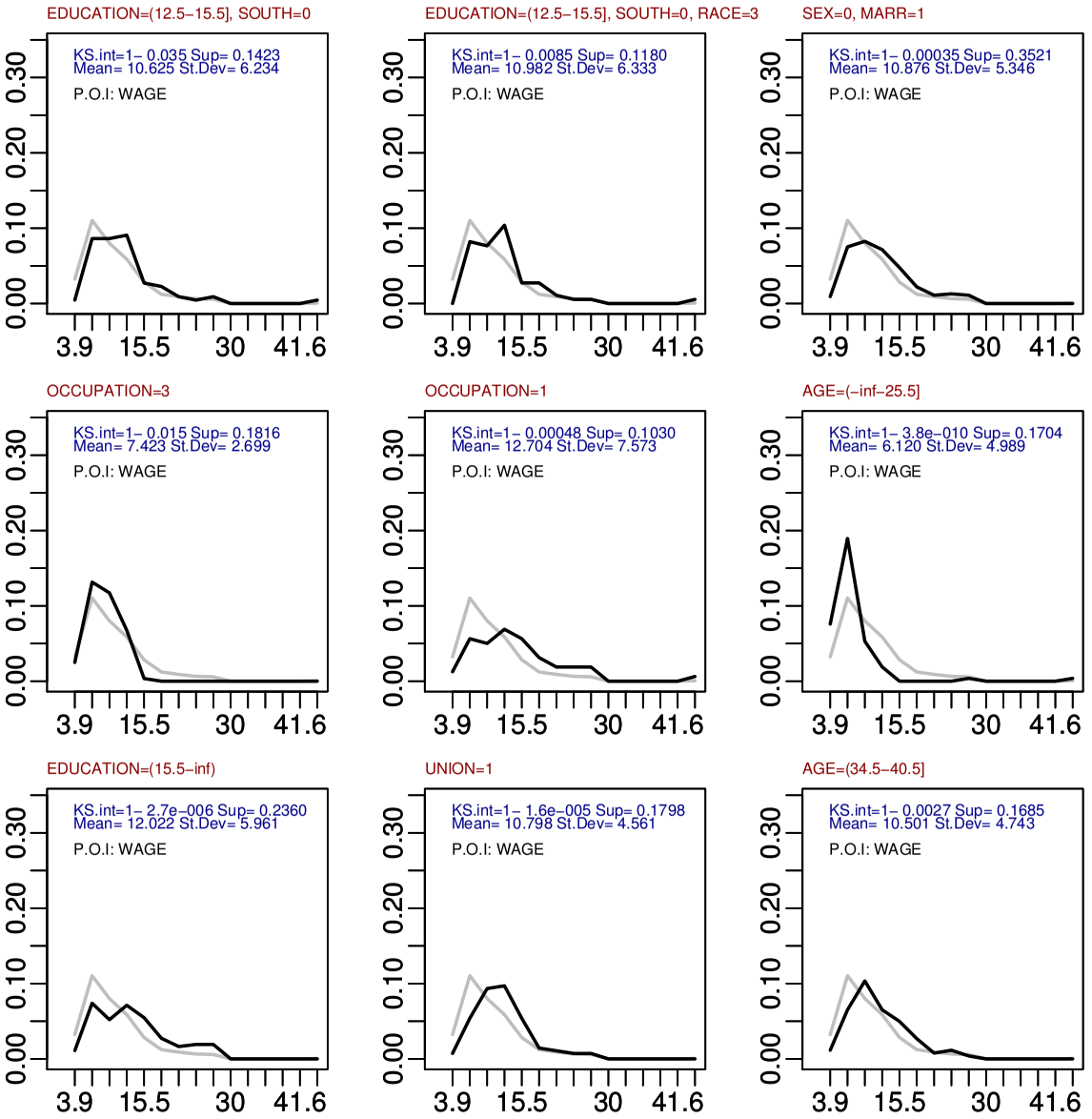
  - **teste de hipóteses**

$$probabilidade\,a\,posteriori >> probabilidade\,a\,priori\,? \quad \chi 2$$

Actividade : Colestlab : Associação

- Use o Weka para obter regras de associação
  - quais as regras mais interessantes?

Regras de distribuição (consequente numérico)



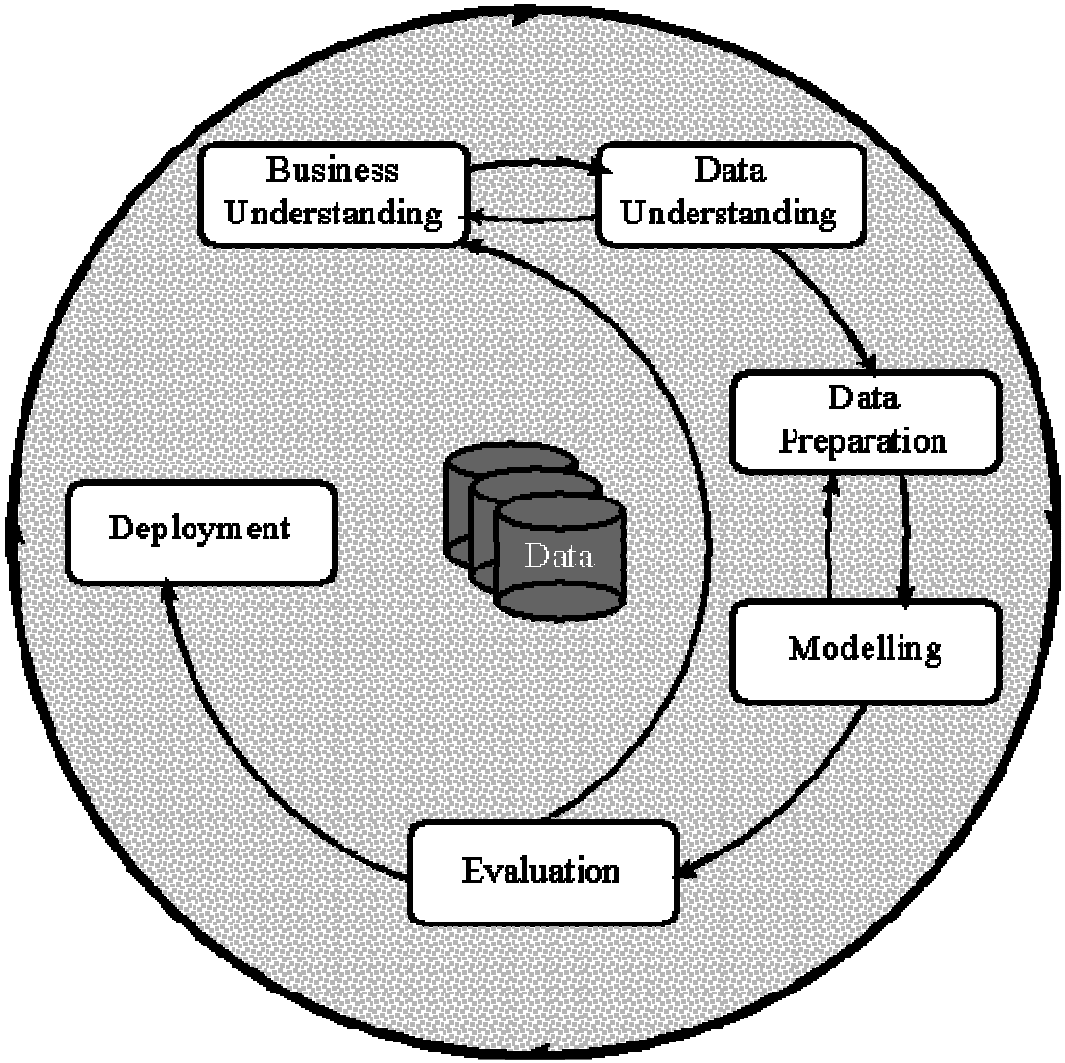
Outros problemas/técnicas de DM

- **Detecção de outliers**
  - **Previsão de fenómenos raros**
- **Text e Web mining**
  - **Sumariação automática de textos**
  - **Extracção de informação**
- **Data mining multi-relacional**
  - **Lidar com estruturas complexas (moléculas)**
  - **Descobrir padrões que explicam actividade**
- **Pré-processamento dos dados**
  - **Muito importante**
    - DNA micro array pode ter milhares de variáveis

O Processo de Data Mining

- Quais são os passos de um processo de DM?
  - como chegar do problema à acção?
- Existe alguma metodologia/ processo modelo de DM?
  - pronta a ser seguida ?

O processo de Data Mining (versão 4)
CRISP-DM (NCR, SPSS, Daimler-Chrysler, OHRA) (1997)



O (modelo de) processo CRISP-DM

- **Pretende ser um *standard***
  - **CRoss-Industry Standard Process for Data Mining**

- **Desenvolvido com base na experiência de muitos profissionais**

- **A partir do Modelo de Processo CRISP-DM podemos definir um Processo de DM**

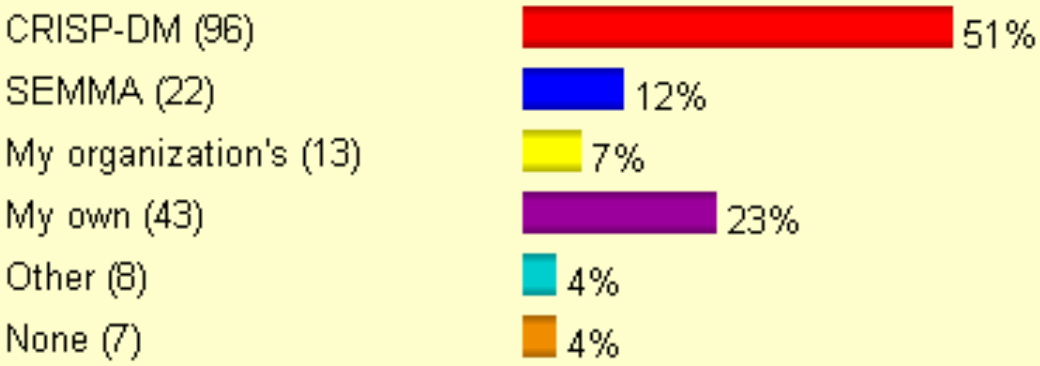
Fases do ciclo de vida de um projecto de DM

- **Compreensão da Actividade (*business understanding*)**
  - Os objectivos do projecto do ponto de vista da actividade
- **Compreensão dos dados**
  - Familiarização com os dados, qualidade dos dados
- **Preparação dos dados**
  - Construír o *data set* a utilizar
- **Modelação**
  - Experimentar e seleccionar técnicas de modelação
- **Avaliação**
  - Rever os passos seguidos para construir o modelo
- **Acção / Produção (Deployment)**
  - Utilização dos modelos criados em problemas reais

KDnuggets : Polls : What main methodology are you using for data mining?

Poll

What main methodology are you using for data mining? [189 votes total]



Comments

- Uta Winter, Subject: CRISP-DM
  As the ultimate reason of doing Data Mining is always to improve a business situation which is perceived as being not satisfactory it is only logical that the Crisp-DM methodology starts there: How do we tackle the business problem by transforming it into a DM-problem. It doesn´t start with sampling (which is not always necessary anyway)! This is a big improvement over more techically focused approaches. The Crisp-Manual is well-written and especially non-tekkies can benefit from it. After having done some projects myself I can feel that this process model is not written by people who have never been out in a company and only discussing DM in theory but who exacly know all the issues which occur when doing a real project.

Referências

- www.kdnuggets.com
- Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth: The KDD Process for Extracting Useful Knowledge from Volumes of Data. Commun. ACM 39(11): 27-34 (1996)
- David Hand, Heikki Mannila, Padhraic Smyth, Principles of Data Mining, MIT Press, 2001
- Igor Jurisica, John Mylopoulos, Janice I. Glasgow, Heather Shapiro, Robert F. Casper: Case-based reasoning in IVF: prediction and knowledge mining. Artificial Intelligence in Medicine 12(1): 1-24 (1998)
- Shearer, C., The CRISP-DM Model: The New Blueprint for Data Mining. In Journal Data Warehousing, Vol. 5, No. 4, (2000), 13—22
- Kim, J., Zhao, S., and Heber, S. 2007. Finding association rules of cis-regulatory elements involved in alternative splicing. In Proc. of the 45th Annual Southeast Regional Conference (2007). ACM.
- www.r-project-org
- www.liaad.up.pt/~amjorge/Projectos/Class/