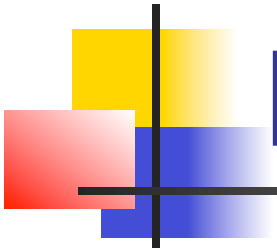
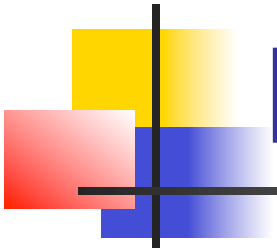


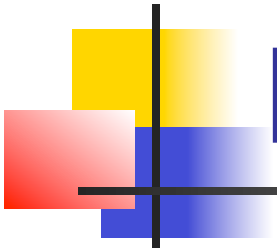
# Mineração de Dados

Arthur Emanuel de O. Carosia
Cristina Dutra de Aguiar Ciferri



# Motivação

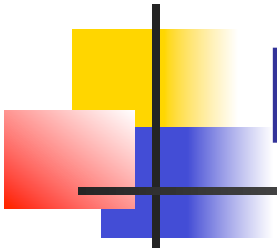
- Aumento da capacidade de processamento e de armazenamento de dados;
- Baixo custo;
- Grande quantidade de dados armazenados;
- Inviabilidade de análise manual dos dados.



# Mineração de Dados

- Extração de conhecimento de grandes volumes de dados.

- Consistem em um dos processos de KDD (*Knowledge Discovery in Databases*).

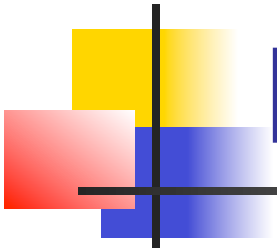


# KDD

***Knowledge Discovery in Databases***

Processo de descobrimento de conhecimento em bancos de dados. Composto das seguintes etapas:

1. *Limpeza dos dados:* etapa em que os ruídos e dados inconsistentes são eliminados;
2. *Integração dos dados:* etapa em que diferentes fontes de dados são integradas, produzindo um único repositório de dados;
3. *Seleção:* etapa em que são selecionados apenas os atributos de interesse ao usuário;



# KDD

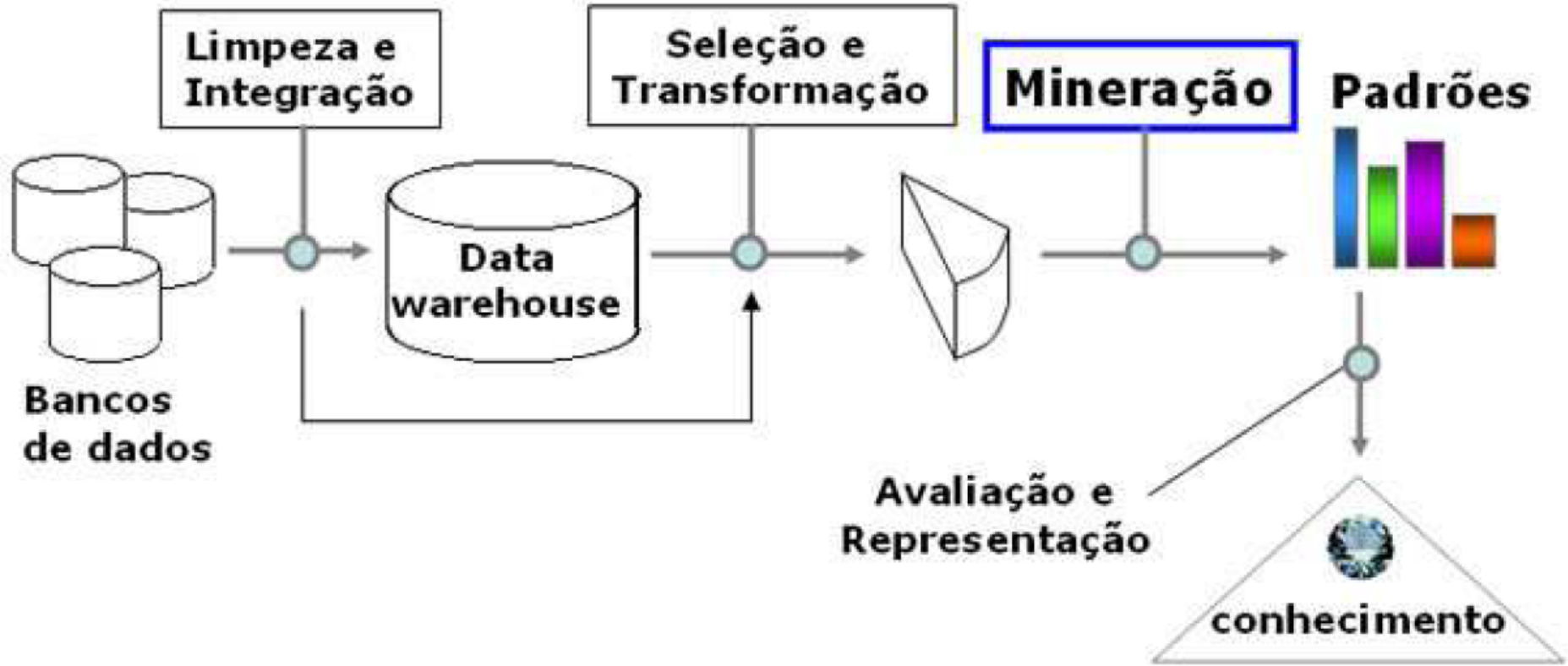
4. *Transformação dos dados:* etapa em que os dados são transformados em um formato correto para a aplicação de algoritmos de mineração de dados;

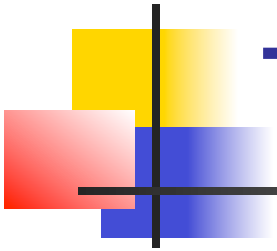
5. *Mineração:* etapa em que são aplicadas técnicas inteligentes a fim de se extrair padrões de interesse;

6. *Avaliação:* etapa em que são identificados padrões de interesse de acordo com algum critério;

7. *Visualização:* etapa em que são utilizadas técnicas de representação de conhecimento para apresentar ao usuário o conhecimento minerado.

# KDD


Limpeza e Integração
Seleção e Transformação
Mineração
Padrões
Data warehouse
Bancos de dados
Avaliação e Representação
conhecimento



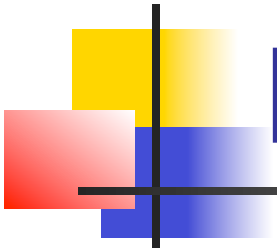
# Tarefas x Técnicas

**Tarefa de mineração de dados:**

*o que* se quer buscar nos dados.

**Técnica de mineração de dados:**

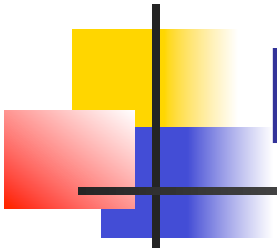
métodos que garantem *como* descobrir padrões de interesse nos dados.



# Regras de Associação

- Busca por padrões associativos que indiquem relacionamentos entre conjunto de itens

- Estes padrões têm a forma X -> Y: a ocorrência de um conjunto de itens (*k-itemsets*) X implica na ocorrência de um um conjunto de itens Y.



# Regras de Associação

Exemplo:

Análise de cesta de compras: identificação das associações entre itens tal que a presença de alguns itens na cesta implique frequentemente a presença de outros.

# Regras de Associação

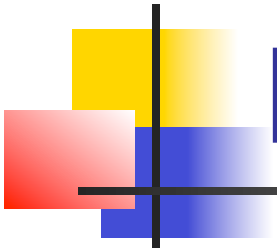
Para obter e mensurar as regras de associação são utilizadas duas medidas de interesse:

- *Suporte:* indica a porcentagem de ocorrência dos conjuntos X e Y na base de dados

$$suporte(X \rightarrow Y) = \frac{\text{nº de registros contendo X e Y}}{\text{total de registros}}$$

- *Confiança:* indica a frequência em que a ocorrência do conjunto de itens X implica na ocorrência do conjunto Y.

$$confiança(X \rightarrow Y) = P(Y \mid X) = \frac{\text{nº de registros contendo X e Y}}{\text{nº de registros contendo X}}$$



# Regras de Associação

## Algoritmo Apriori

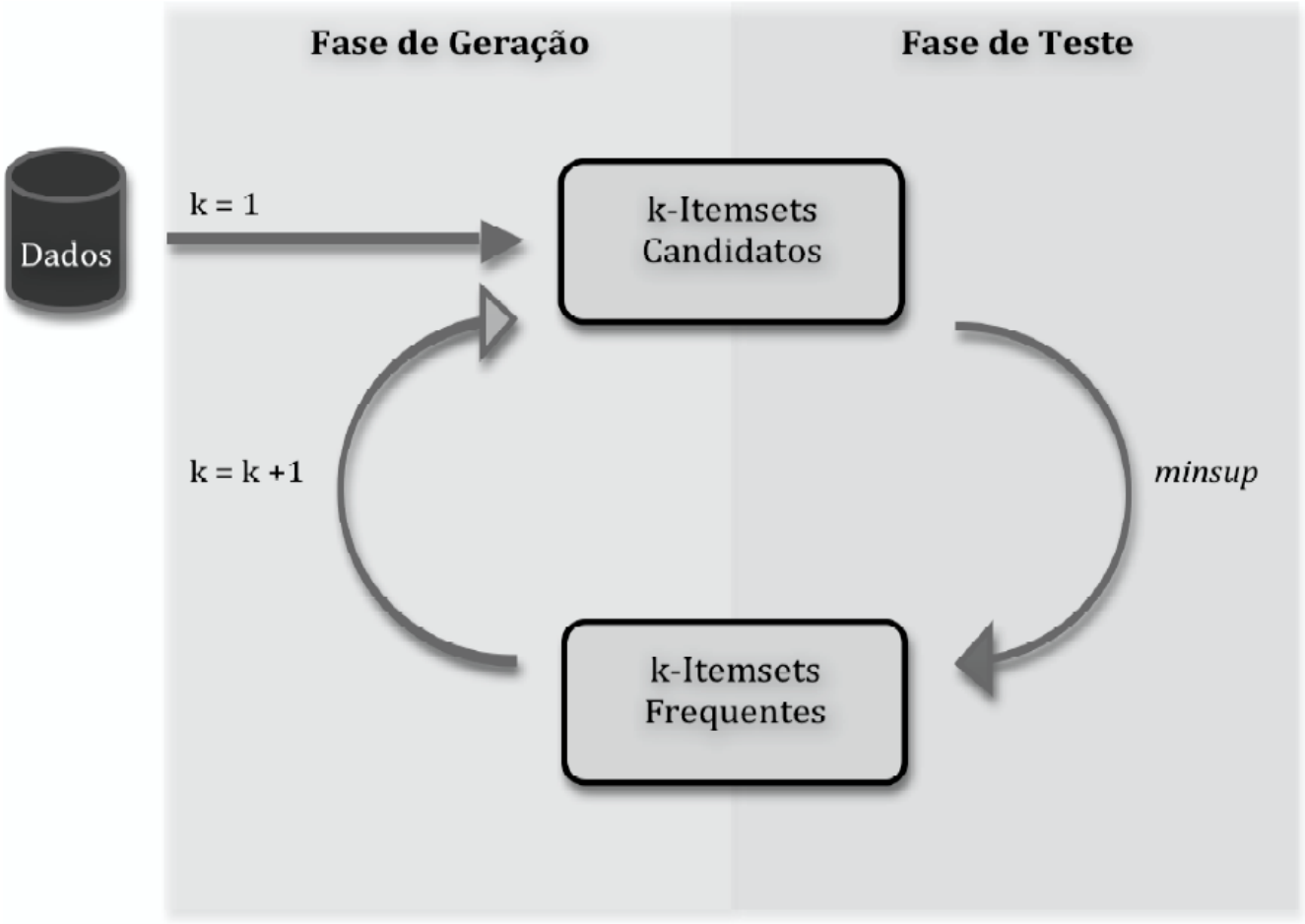
- Conceito de geração-e-teste de candidatos:

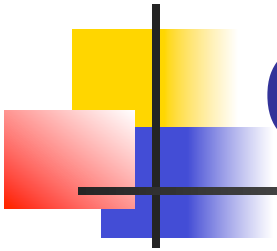
1. **Obtenção dos k-*itemsets* candidatos.** Na primeira iteração, os 1-itemsets candidatos são obtidos por meio da varredura no conjunto de dados. A partir da segunda iteração, o conjunto de k-*itemsets* candidatos é gerado por meio de combinações dos (k-1)-*itemsets* frequentes.

2. **Teste dos *itemsets* candidatos.** Etapa que Filtra aqueles que são de interesse, ou seja, que atendem à frequência mínima pré-estabelecida.

# Regras de Associação

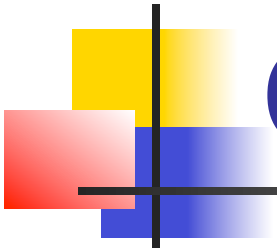
## Algoritmo Apriori





# Classificação

- Encontra modelos que descrevem e distinguem classes de objetos que ainda não foram classificados.
- Baseado na análise de um conjunto de dados já classificados.
- Aprendizado supervisionado.

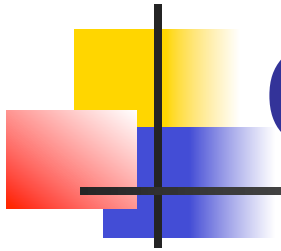


# Classificação

Processo realizado em 2 etapas:

1. Criação do modelo de classificação a partir de dados de treinamento.

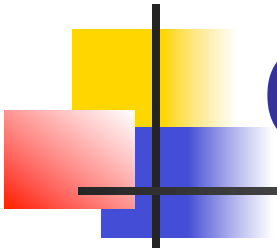
2. Verificação do modelo a partir do testes das regras com dados diferentes dos utilizados para a sua criação.



# Classificação

Árvore de decisão

1. Cada nó interno é um atributo do banco de dados de amostras diferente do atributo-classe.
2. Folhas são valores do atributo-classe.
3. Cada ramo ligado de um nó filho ao nó pai é etiquetado com o valor do atributo contido no nó pai.
4. Um atributo que aparece em um nó não pode aparecer nos seus descendentes.



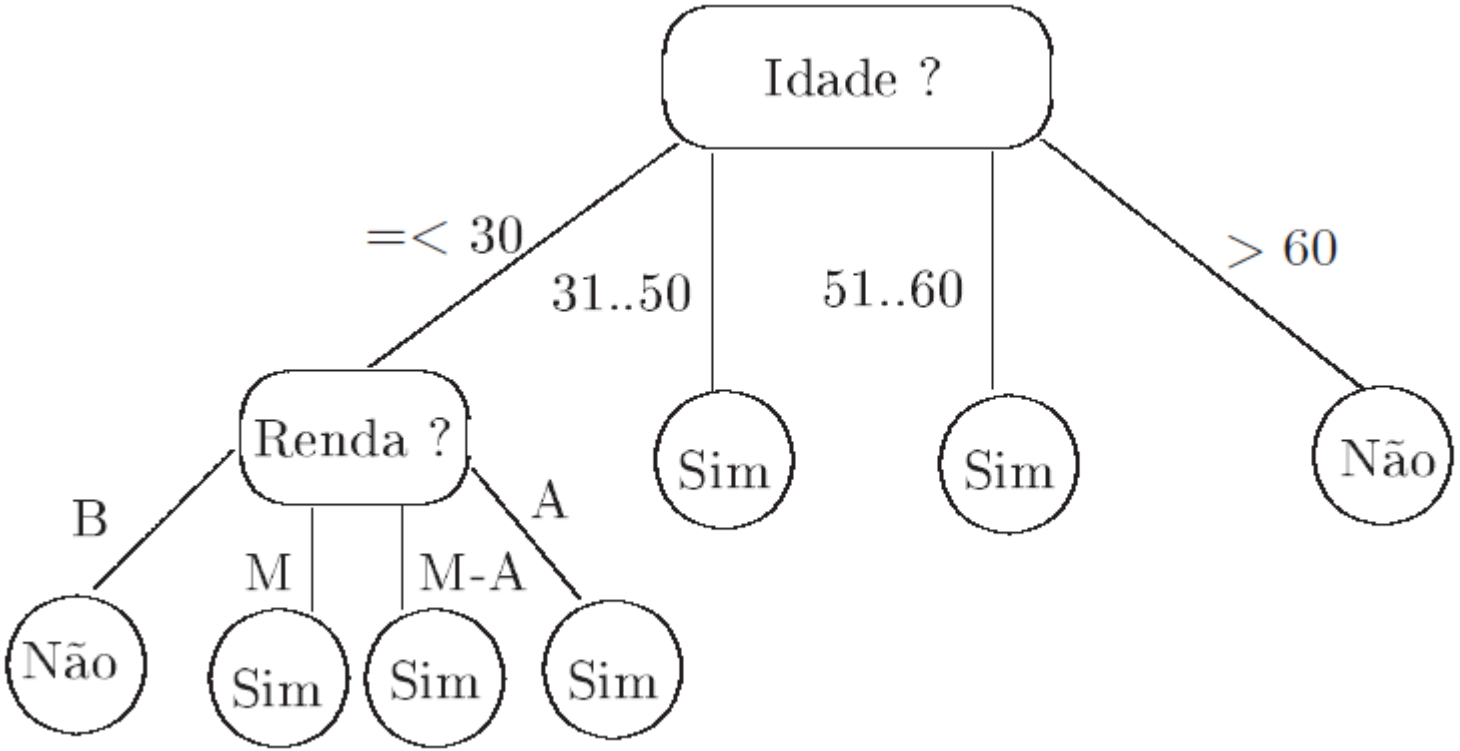
# Classificação

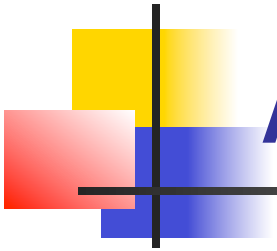
Exemplo: Descobrir se o cliente é um potencial comprador de produtos eletrônicos

| Nome | Idade | Renda | Profissão | ClasseProdEletr |
|---|---|---|---|---|
| Daniel | =< 30 | Média | Estudante | Sim |
| João | 31..50 | Média-Alta | Professor | Sim |
| Carlos | 31..50 | Média-Alta | Engenheiro | Sim |
| Maria | 31..50 | Baixa | Vendedora | Não |
| Paulo | =< 30 | Baixa | Porteiro | Não |
| Otávio | > 60 | Média-Alta | Aposentado | Não |

# Classificação

## Árvore de Decisão





# Agrupamento ou Clusterização

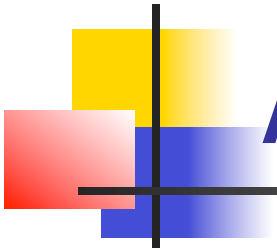
- Identifica agrupamentos de objetos através de algum critério de similaridade.
- Um *cluster* se trata de uma coleção de objetos similares entre si e diferentes de objetos que pertençam a outros clusters
- Aprendizado não supervisionado.

# Agrupamento ou Clusterização

## Algoritmo K-Means

Banco de dados é representado como matriz de dissimilaridade (distância) entre os objetos.

$$\begin{vmatrix} 0 & \dots & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ d(n,1) & d(n,2) & d(n,3) & \dots & 0 \end{vmatrix}$$



# Agrupamento ou Clusterização

## Algoritmo K-Means

1. Escolhe-se arbitrariamente k objetos do banco de dados que serão o centro de cada cluster.

2. Inserção de outros objetos no cluster: considera-se a distância entre o objeto do centro e cada um dos demais e insere no cluster aquele elemento cuja distância é mínima.

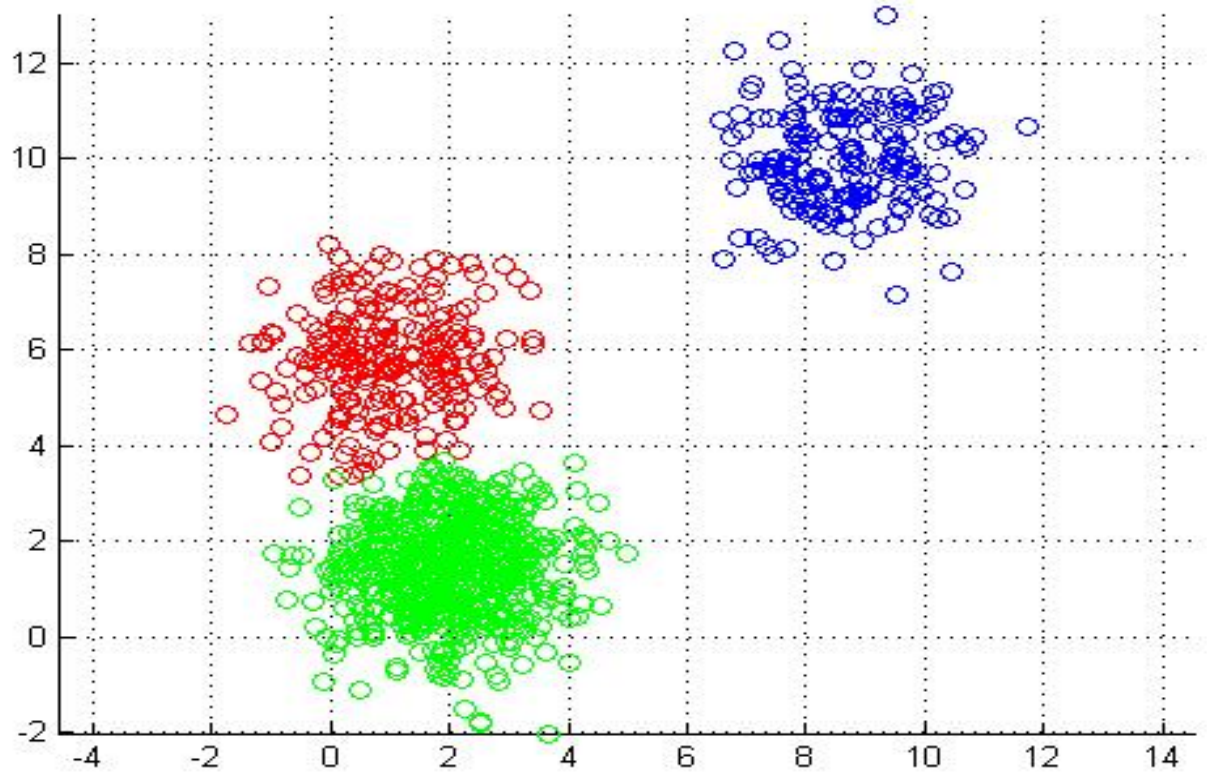
3. A média dos elementos do cluster será o seu novo representante.

4. Repete 2 atualizando os clusters e calcula-se os novos centros.

5. O processo pára quando nenhum objeto for realocado para outro cluster distinto do qual ele pertence.

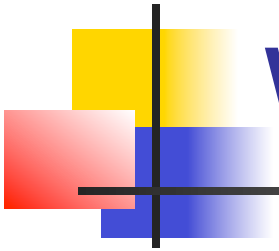
# Agrupamento ou Clusterização

Exemplo: identificação de grupos homogêneos de clientes em um supermercado.



# Weka

- Software que contém um conjunto de implementações de algoritmos de diversas tarefas de Mineração de Dados.

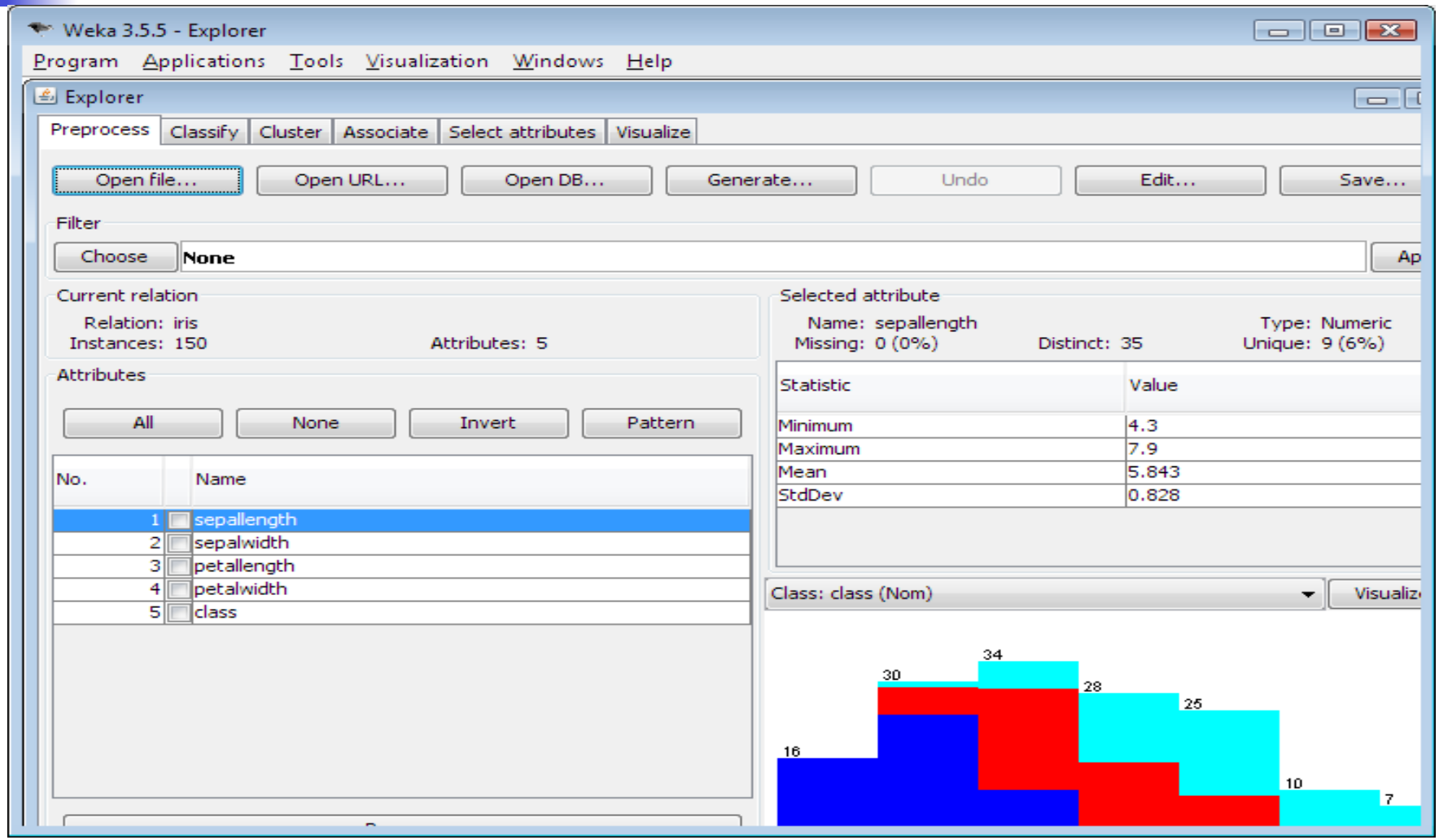


# Weka

- Formato de arquivo contendo dados: ARFF

```
@relation clima
@attribute céu {sol, nublado, chuva}
@attribute temperatura real
@attribute umidade real
@attribute vento {VERDADEIRO, FALSO}
@attribute jogar {sim, nao}
@data
sol,85,85,FALSO,nao
sol,80,90,VERDADEIRO,nao
nublado,83,86,FALSO,sim
chuva,70,96,FALSO,sim
```

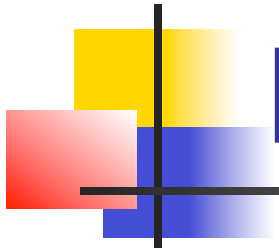
# Weka


Weka 3.5.5 - 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
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
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)
34
30
28
25
16
10
7

# Weka

Experimento:

Jogar ou não futebol de acordo com as condições climáticas?



# Referências


Sandra de Amo. Técnicas de Mineração de Dados. Disponível em <http://www.deamo.prof.ufu.br/arquivos/JAI-cap5.pdf>. Última visita em 24 de Novembro de 2010.

Weka: Data Mining Software in Java. ível em: <http:// www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 09 jul. 2010.

João Paulo Rodolfo de Siqueira. Mineração DE Regras DE Associação Multi-Rrelacional Transitivas- Aplicação na Área Biomédica. Dissertação (Mestrado), Universidade Metodista de Piracicaba, 2010.

Fernando Takehi Oyama. Mineração Multirrelacional de Regras de Associação em Grandes Bases de Dados. 2010. Dissertação (Mestrado em Ciência da Computação) - Universidade Estadual Paulista Júlio de Mesquita Filho.