Universidade Federal da Grande Dourados

PRÉ-PROCESSAMENTO

Bibliografia:

Cap 1, 2, 3 – Data Mining, um Guia Prático

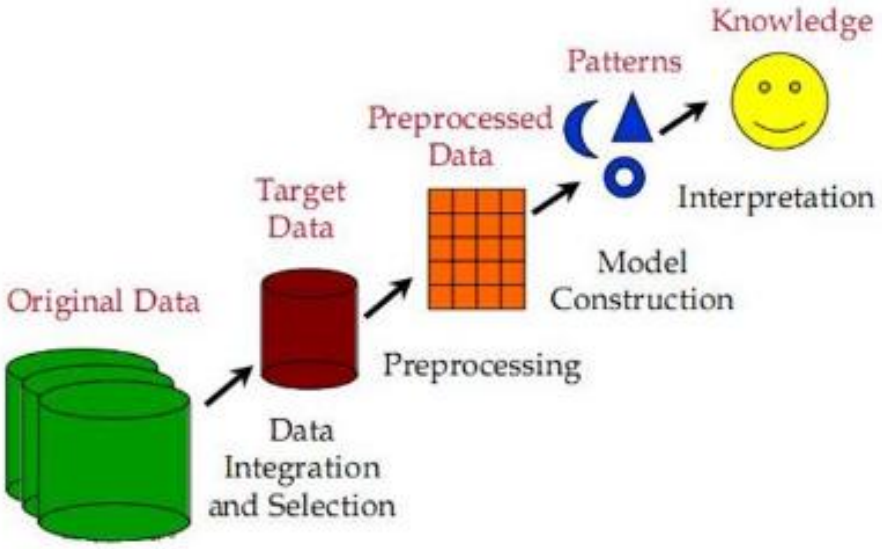
R. Goldschmidt e E. Passos, Ed. Campos

Agenda:

DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

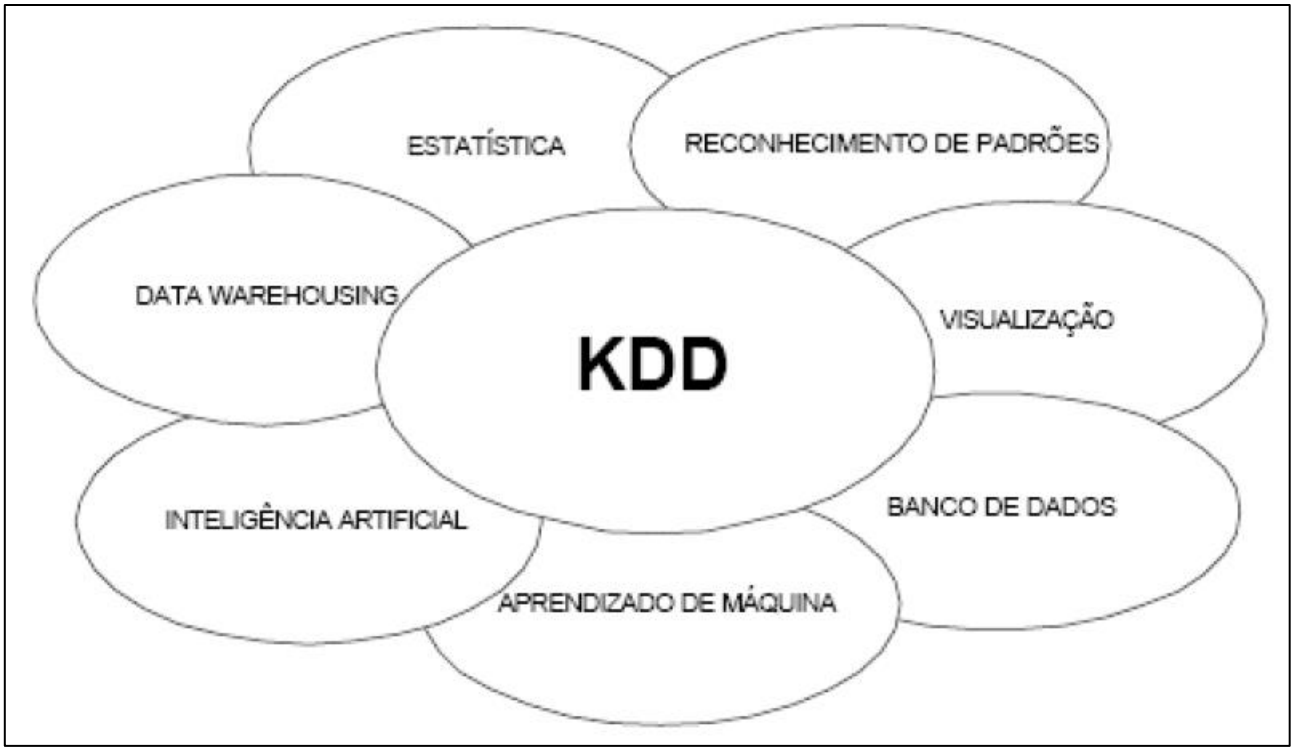
- São três grandes etapas principais:

1. Pré-Processamento
2. Mineração de Dados
3. Pós-Processamento



DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

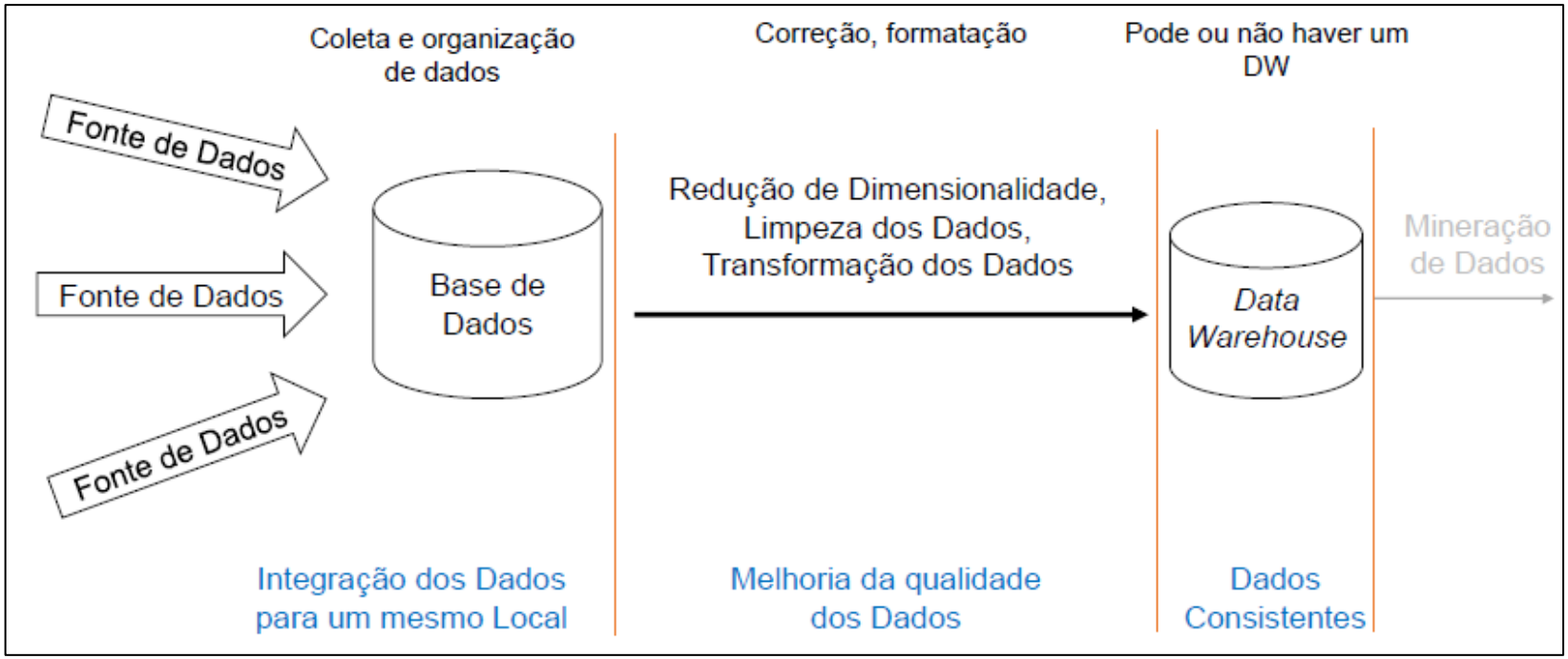
- Área Multidisciplinar



PRÉ-PROCESSAMENTO

- ✓Integração dos Dados
- ✓Redução de Dimensionalidade
- ✓Limpeza dos Dados
- ✓Transformação

PRÉ-PROCESSAMENTO



ETAPAS DO PRÉ-PROCESSAMENTO

- **Integração** dos Dados

✓Dados de diferentes origens para um mesmo local

- **Redução** de Dimensionalidade (vertical e horizontal)

✓Seleção de atributos: **melhora a precisão** do algoritmo

✓Remoção de objetos: **melhora a eficiência** do algoritmo

- **Limpeza** dos Dados

✓Correção de inconsistências(divergências), dados incompletos (ou ausentes)

✓Remoção/Correção de ruídos (erros ou *outliers*), duplicidade

ETAPAS DO PRÉ-PROCESSAMENTO

- **Transformação** dos Dados (formato adequado para aplicá-los à MD)

✓Discretização de atributos (redução do domínio); de contínuo para discreto: temperatura, peso, pressão...

✓Codificação: categórica (nominal) para numérica

Os atributos podem ser:

- Numéricos: idade, peso, tamanho, temperatura, pressão
- Categóricos: sexo, conceito, religião, grau de satisfação

✓Normalização dos dados (mesma faixa de valores)

ETAPAS DO PRÉ-PROCESSAMENTO

- **Normalização dos dados**

✓Atributos com grandes domínios não devem dominar atributos com domínios menores

✓Evita a influência de atributos de forma tendenciosa em certos algoritmos de AM

✓Alguns tipos de normalização

- ❑Normalização Linear
- ❑Normalização pelo Valor Máximo
- ❑Normalização Max-Min
- ❑Normalização pelo Escore-z

ETAPAS DO PRÉ-PROCESSAMENTO

- **Normalização dos dados**

✓Normalização Linear

$$f(X) = \frac{X - Min}{Max - Min}$$

Atributo X	Atributo X Normalizado (f(X))
1000	0,14
2000	0,43
3000	0,71
1500	0,29
1500	0,29
1000	0,14
3000	0,71
500	0
4000	1
1000	0,14

ETAPAS DO PRÉ-PROCESSAMENTO

- **Normalização dos dados**

✓Normalização pelo Valor Máximo

$$f(X) = X / máximo$$

Atributo X	Atributo X Normalizado (f(X))
1000	0,25
2000	0,50
3000	0,75
1500	0,38
1500	0,29
1000	0,25
3000	0,75
500	13
4000	1
1000	0,25

ETAPAS DO PRÉ-PROCESSAMENTO

- **Normalização dos dados**

✓Normalização Max-Min

- Mapeia um valor X em um valor X' no domínio [$novo\_min_X, novo\_max_X$]:

$$f(X) = \frac{X - minx}{max_x - minx}(novo\_maxx - novo\_minx) + novo\_minx$$

ETAPAS DO PRÉ-PROCESSAMENTO

- **Normalização dos dados**

✓Normalização pelo Score-z

- Os valores do atributo X são normalizados com base na média e no desvio padrão

$$f(X) = (X - \overline{X}) / \sigma_X$$

- $\bar{X}$ é a média de X, e σ_X é o desvio padrão de X

➢ ETAPAS DO PRÉ-PROCESSAMENTO

- **Motivações para a tarefa de pré-processamento**

➢ Preparação dos dados antes de aplicá-los às tarefas de mineração
  - ➢ Melhora a **qualidade** dos dados
  - ➢ **Modelo consistente** que retrate a realidade
  - ➢ Promove a **eficácia** do modelo de dados e eficiência da aprendizagem

➢ Geralmente esta etapa consome 70-80% do tempo e esforço

ATIVIDADE

Realizar a atividade disponível no Moodle