Preparação de Dados

O principal objetivo da **preparação de dados** consiste em **transformar** os *data sets* por forma à **informação** neles contida seja adequadamente exposta à ferramenta de análise de dados. Para além disso, a preparação de dados também prepara o preparador por forma selecionar os modelos de AD mais adequados.

Tarefas na preparação de dados

Discretização/Enumeração

- Utiliza-se a discretização para reduzir o número de valores de um atributo contínuo, dividindo-o em intervalos;
  - Os métodos mais utilizados (Naive Bayes, CHAID, etc) requerem valores discretos;
  - Para além disso permitem reduzir o tamanho dos dados;
  - Sinónimo de *binning*;

Equall-Width binning: - Divide o intervalo dos valores em `k` intervalos de igual largura, resultando numa grelha uniforme; - Vantagens: - Simples e fácil de implementar; - Produz abstrações de dados razoáveis; - Desvantagens: - Pode não ser eficaz se os dados não estiverem distribuídos uniformemente;

Equall-Height binning: - Divide a gama de valores em `k` intervalos de **aproximadamente** igual frequência; - Vantagens: - Cada intervalo contém o mesmo número de valores sendo vantajoso pois os intervalos são mais equilibrados; - Desvantagens: - Pode não ser eficaz se os dados não estiverem distribuídos uniformemente;

Limpeza

- Preenchimento de valores de atributos;
- Remoção de lixo dos dados;
- Remoção de valores impossíveis;
- Resolução de inconsistências;

Integração

- Integração de múltiplas fontes de dados (BD's, ficheiros, papel, web, etc);
- Abrange também a detecção e resolução de conflitos entre os dados;

- Integração exige **conhecimento do negócio**;

Transformação **Normalização e agregação dos dados**: - **Alisamento** (*smoothing*): remover lixo/ruído dos dados; - Por exemplo: substituir valores nulos pela média - **Agregação**: Pressupões que o resultado sumaria os dados iniciais; - Por exemplo: somar valores de vendas diárias para obter vendas mensais; - **Generalização**: Hierarquização de conceitos; - Por exemplo: transformar valores de idade em faixas etárias; - **Construção de atributos**: Construção de novos atributos a partir de outros; - Por exemplo: calcular o IMC a partir do peso e altura; - **Uniformização**: Pretende evitar que atributos com uma grande amplitude de valores sobressaiam em relação a outros atributos com menor quantidade de valores; - **Normalization**: normalizar valores de salários para uma escala de 0 a 1; - **Standardization**: padronizar valores de salários para uma média de 0 e desvio padrão de 1; - **Detecção de valores atípicos**: Por visualização recorrendo a box plots ou histogramas; - Por exemplo: detetar valores de salários muito elevados ou muito baixos;

Redução

- Obtenção de representações dos dados menos volumosas, mas com capacidade de produzir idênticos resultados analíticos;
- Redução de dimensões;
- Compressão de dados;

Seleção de atributos A seleção de atributos é uma técnica de redução de dados que consiste em selecionar um subconjunto de atributos relevantes para a análise. A seleção de atributos é importante porque: - Reduz a complexidade do modelo; - Melhora a precisão do modelo; - Reduz o tempo de treino do modelo; - Melhora a interpretabilidade do modelo;

Algumas técnicas de seleção de atributos são: - **Filtro**: Seleciona atributos com base em métricas estatísticas (correlação de Pearson) - **Wrapper**: Utiliza um algoritmo de aprendizagem para avaliar a qualidade dos atributos; - **Embeded methods**: Seleciona atributos durante o treino do modelo;

Tipos de dados

Os tipos de dados diferem na sua natureza e na informação que proporcionam.

Qualitativos

Nominais

- Atribui nomes únicos a objetos:
  - Não existe outra informação que se possa deduzir;
  - Nomes de pessoas;
  - Código de identificação;

Categorias

- Atribui categorias a objetos:
  - Podem ser valores numéricos, mas **não ordenados**;
  - Código postal;
  - Sexo;
  - Cor dos olhos;

Ordinais

- Os valores podem ser ordenados naturalmente:
  - Classificação: excelente, bom, suficiente, etc;
  - Temperatura: frio, morno, quente;

Quantitativos (discretos e contínuos)

Intervalos

- É possível calcular a distância entre dois valores:
  - Temperatura;
  - Humidade;

Rácio

- Os valores podem ser usados para determinar um rácio significativo entre eles:
  - Salário;
  - Balanço bancário;