Mineração de Dados

GRIMALDO OLIVEIRA

Sobre Grimaldo

- Grimaldo Oliveira
  - grimaldo_lopes@hotmail.com
- Formação
  - Mestre em Tecnologias Aplicadas a Educação Universidade do Estado da Bahia.
  - Especialização em Análise de Sistemas pela Faculdade Visconde de Cairu.
  - Estatístico pela Universidade Federal da Bahia.
- Atividades
  - Mais de 10 anos atuando como Consultor de Business Intelligence.
  - Projetos Governos Maranhão, Mato Grosso e Bahia.
  - Idealizador do Blog : BI com Vatapá – bicomvatapa.blogspot.com.
  - Livro: BI Como Deve Ser – bicomodeveser.com.br

Agenda

- Tarefas de Mineração de Dados
  - Classificação
  - Análise de Clusters (agrupamentos) – Segmentação
  - Análise de Outliers (exceções)
  - Estimativa (ou regressão)
  - Sumarização

Classificação

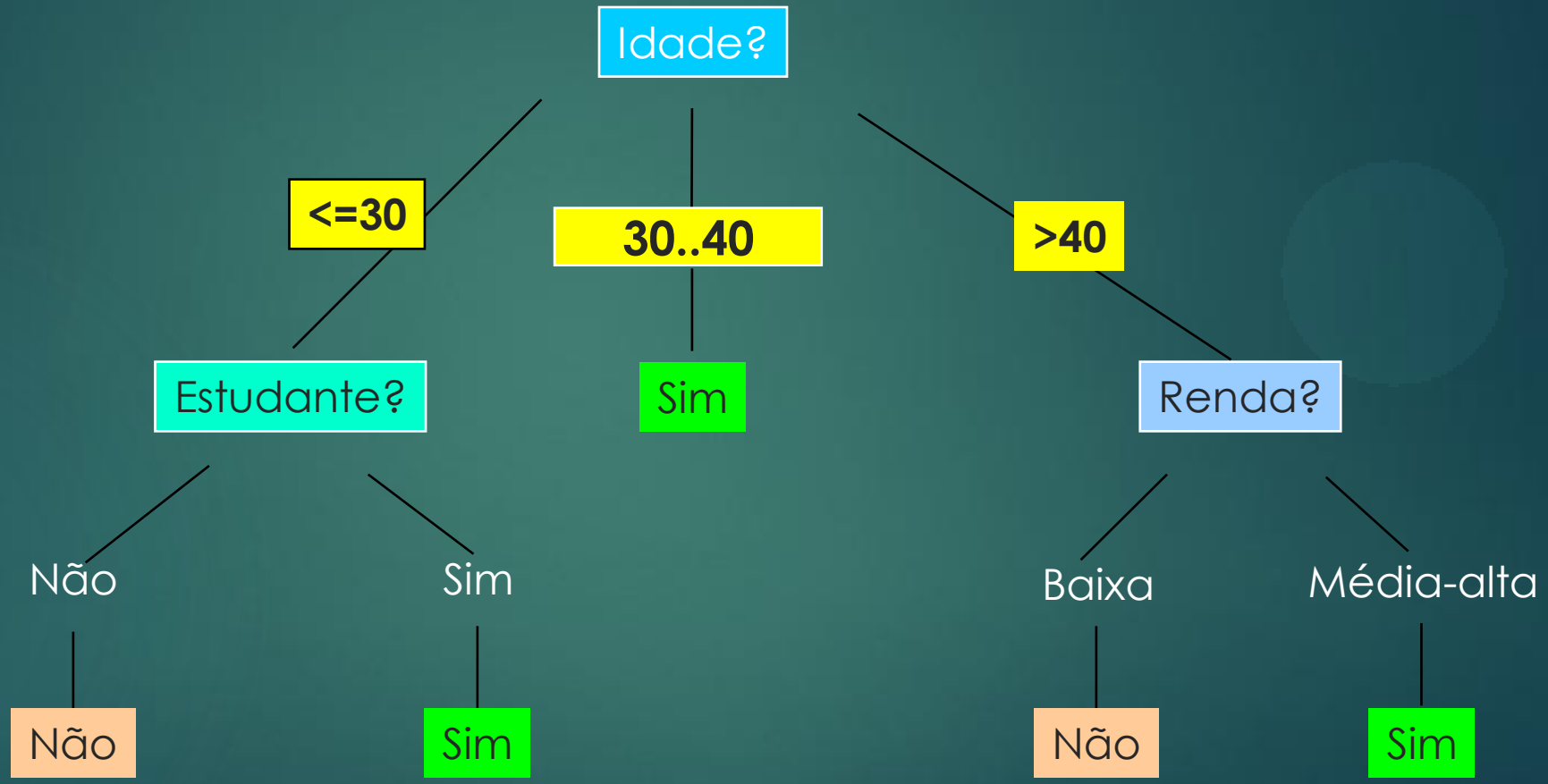
- Classificação
  - Predição dos nomes (rótulos) das classes;
  - Classifica os dados (constrói um modelo) com base no conjunto de treinamento e nos valores (rótulos) do atributo classificador, de forma a determinar a classe dos novos dados;

- Aplicações típicas
  - Aprovação de crédito, marketing dirigido, diagnóstico médico ...

Classificação

Nome	Idade	Renda	Profissão	Bom Pagador
Daniel	≤ 30	Média	Estudante	Sim
João	31..50	Média-Alta	Professor	Sim
Carlos	31..50	Média-Alta	Engenheiro	Sim
Maria	41..50	Baixa	Vendedora	Não
Paulo	≤ 30	Baixa	Porteiro	Não
Otavio	> 60	Baixa	Aposentado	Não

Classificação : Árvore de Decisão



É Bom Pagador?

Exemplo : Árvore de Decisão

- Representação por regras IF-THEN:
  - Cada par (atributo, valor) forma uma conjunção;
- Regras são de mais fácil compreensão aos usuários:
  - IF Idade = "<=30" AND Estudante = "Não"
    THEN Bom_Pagador = "Não"
  - IF Idade = ">40" AND Renda = "Média-Alta"
    THEN Bom_Pagador = "Sim"

Classificação

Classificador Bayesiano

Classificador Bayesiano

- Aprendizagem probabilista: cálculo da probabilidade explícita da hipótese, de ampla aplicação em vários domínios;

- Incremental:
  - cada exemplo de treinamento pode aumentar / diminuir a probabilidade da hipótese;
  - Conhecimento a priori pode ser combinado com os dados observados;

- Previsão probabilista:
  - Várias hipótese podem ser previstas, ponderadas por suas probabilidades;
  - Fornece uma referência a ser comparada a outros métodos.

Fundamento: Teorema de Bayes

- Dado um conjunto de dados D, a probabilidade a posteriori de uma hipótese h, P(h|D) é dada por:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- A probabilidade máxima a posteriori MAP é:

$$h_{MAP} \equiv \arg\max_{h \in H} P(h|D) = \arg\max_{h \in H} P(D|h)P(h).$$

- Dificuldade prática: requer conhecimento inicial de muitas probabilidades, custo computacional elevado;

Exemplo: Jogar ou não Tênis

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

P(p) = 9/14
P(n) = 5/14

outlook	
P(sunny|p) = 2/9	P(sunny|n) = 3/5
P(overcast|p) = 4/9	P(overcast|n) = 0
P(rain|p) = 3/9	P(rain|n) = 2/5
temperature	
P(hot|p) = 2/9	P(hot|n) = 2/5
P(mild|p) = 4/9	P(mild|n) = 2/5
P(cool|p) = 3/9	P(cool|n) = 1/5
humidity	
P(high|p) = 3/9	P(high|n) = 4/5
P(normal|p) = 6/9	P(normal|n) = 2/5
windy	
P(true|p) = 3/9	P(true|n) = 3/5
P(false|p) = 6/9	P(false|n) = 2/5

Exemplo: Jogar ou não Tênis

- Um novo exemplo: X = <rain, hot, high, false>

$P(X \mid p) \cdot P(p) =$

$P(rain \mid p) \cdot P(hot \mid p) \cdot P(high \mid p) \cdot P(false \mid p) \cdot P(p) = 3/9 \cdot 2/9 \cdot 3/9 \cdot 6/9 \cdot 9/14 = 0.010582$

$P(X \mid n) \cdot P(n) =$

$P(rain \mid n) \cdot P(hot \mid n) \cdot P(high \mid n) \cdot P(false \mid n) \cdot P(n) = 2/5 \cdot 2/5 \cdot 4/5 \cdot 2/5 \cdot 5/14 =$ **0.018286**

O exemplo X é classificado como da classe **n (não jogar).**

Redes Neurais

Redes Neurais

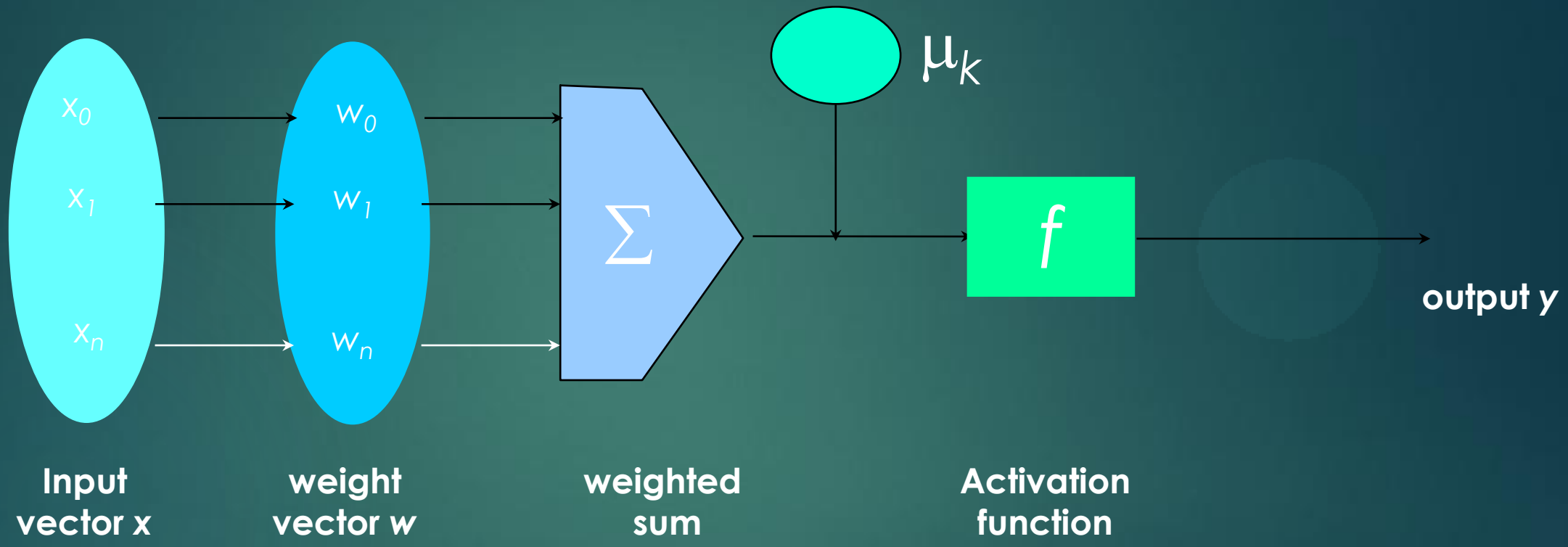
Vantagens:

- Correção de predição em geral elevada;
- Robustez, bom funcionamento na presença de ruídos;
- Saídas discretas, reais, ou mistas;
- Avaliação rápida da função de aprendizagem.

Desvantagens / crítica:

- Tempo de treinamento lento;
- Dificuldade no entendimento da função de aprendizagem (pesos);
- Difícil incorporação de conhecimento de domínio.

Um neurônio



- Um vetor n-dimensional x de entrada é mapeado em uma variável y por meio de um produto escalar e de um mapeamento não-linear.

Agrupamento(Clustering)

Agrupamento

Cluster: uma coleção de objetos de dados;

- Similares entre si no mesmo cluster;
- Não similares aos objetos fora do respectivo cluster;

Análise de clusters:

- Agrupamento de dados em clusters;

Agrupamento (*clustering*) é uma classificação não-supervisionada: não há classes pré-definidas.

Aplicações típicas:

- Como ferramenta para análise da distribuição dos dados;
- Como pré-processamento para outros métodos.

Aplicações gerais do agrupamento

- Reconhecimento de padrões;
- Análise de dados espaciais:
  - Criação de mapas temáticos em GIS por agrupamento de espaços de características;
  - Detecção de clusters espaciais e sua explanação em data mining;
- Processamento de imagens;
- Pesquisas de mercado;
- WWW:
  - Classificação de documentos;
  - Agrupamento de dados de weblogs para descobrir padrões similares de acesso;

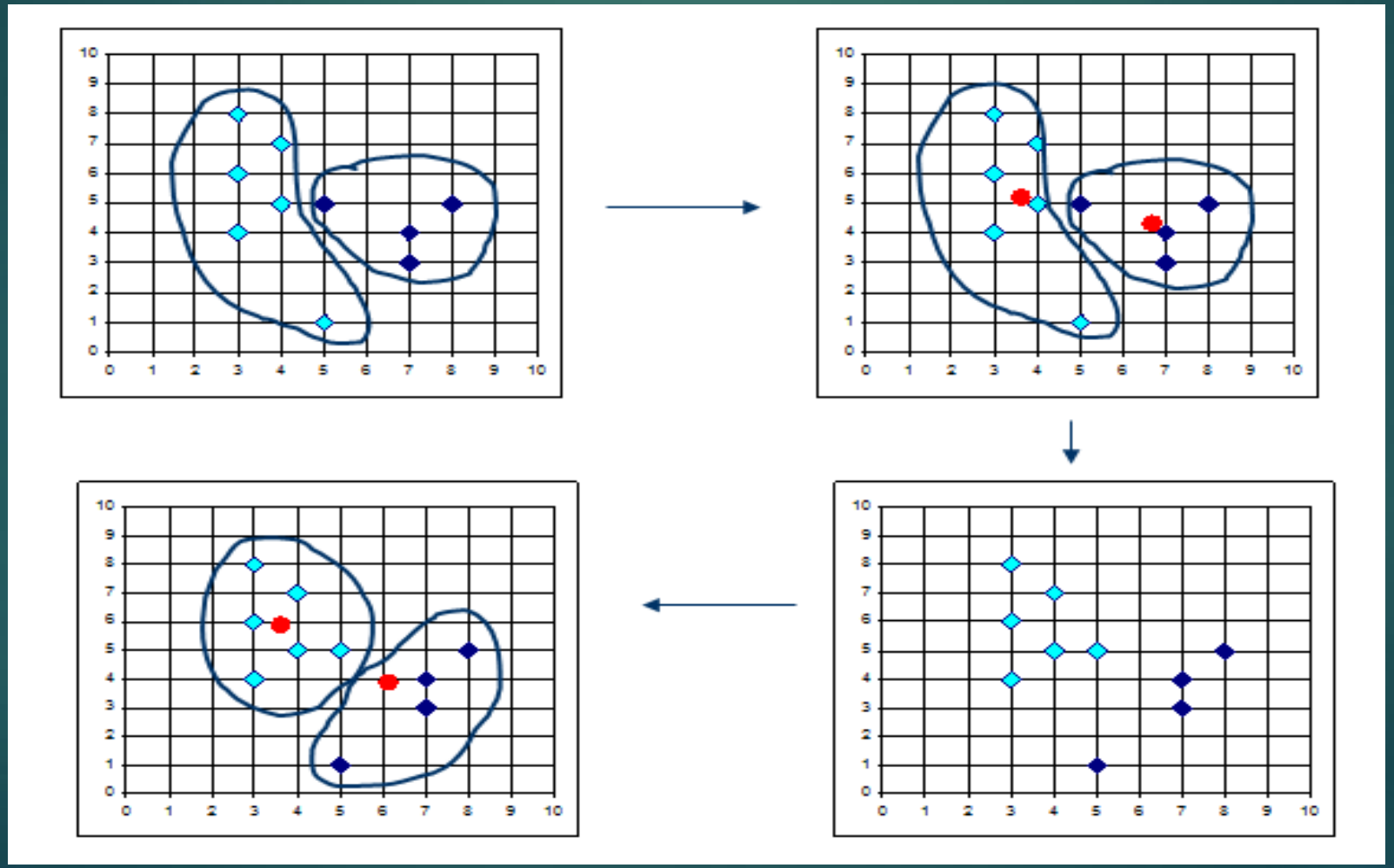
O método k-means (k-médias)

- Dado k, o algoritmo k-means é implementado em quatro passos:

1. Partição dos objetos em k conjuntos não vazios;
2. Cálculo de pontos "semente" como os centróides (médias) dos clusters das partições correntes;
3. Assinalação de cada objeto ao cluster (centróide) mais próximo de acordo com a função de distância;
4. Retorno ao passo 2 até que não haja mais alterações de assinalação.

O método k-means (k-médias)

- Exemplo



Técnicas de Mineração de Dados

Técnica	Tarefas	Exemplos
Descoberta de Regras de Associação	Associação	Apriori, AprioriTid, AprioriHybrid, AIS, SETM (Agrawal e Srikant, 1994) e DHP (Chen *et al.*., 1996).
Árvores de Decisão	Classificação Regressão	CART, CHAID, C5.0, Quest (Two Crows, 1999); ID-3 (Chen *et al.*., 1996); SLIQ (Metha *et al.*., 1996); SPRINT (Shafer *et al.*., 1996).
Raciocínio Baseado em Casos ou MBR	Classificação Segmentação	BIRCH (Zhang *et al.*., 1996); CLARANS (Chen *et al.*., 1996); CLIQUE (Agrawal *et al.*., 1998).
Algoritmos Genéticos	Classificação Segmentação	Algoritmo Genético Simples (Goldberg, 1989); Genitor, CHC (Whitley, 1993); Algoritmo de Hillis (Hillis, 1997); GA-Nuggets (Freitas, 1999); GA-PVMINER (Araújo *et al.*., 1999).
Redes Neurais Artificiais	Classificação Segmentação	Perceptron, Rede MLP, Redes de Kohonen, Rede Hopfield, Rede BAM, Redes ART, Rede IAC, Rede LVQ, Rede Counterpropagation, Rede RBF, Rede PNN, Rede Time Delay, Neocognitron, Rede BSB (Azevedo, 2000), (Braga *et al.*, 2000), (Haykin, 2001)

Próximos vídeos...

Finalidade: Coleta de dados com os gestores para a construção do BI.

Fatos / Dimensões		Diária
Hóspede		✓
Tipo Quarto		✓
Código Tipo Quarto		
Tipo Quarto	HISTÓRICO	
Classe Quarto		✓
Tempo (Data Registro Primeira Diária)		✓

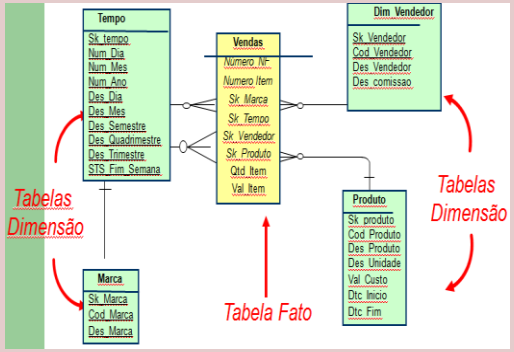
Tarefas de Mineração de Dados -Parte01

Finalidade: Levantamento dos relacionamentos e objetos que armazenam os dados da empresa.

DIMENSÕES	ORIGEM	
	TABELA/VISÃO	CAMPO
Hóspede		
Nome Hóspede	HOSPEDE	NOM_HOSPEDE
Cidade Hóspede	CIDADE_ORIGEM	NOM_CIDADE
País Hóspede	PAIS_ORIGEM	NOM_PAIS
Aeroporto Hóspede	AEROPORTO_SAIDA	DES_AEROPORTO
Local Aeroporto Saída	AEROPORTO_SAIDA	NOM_LOCALIDADE
Código Hóspede	HOSPEDE	COD_HOSPEDE

Tarefas de Mineração de Dados -Parte02

Finalidade: Modelo adequado para realizar as consultas nas bases que servirão ao BI



Mineração Visual

contato@bicomodeveser.com