CURSO SUPERIOR DE ADS

Mineração de Dados Introdução e Algoritmos de Associação

Prof. Fernando Marlon Soares Figueiredo

Disciplina: Ciência de Dados e Bigdata

SECTRAS

ADS

Mineração de dados

KDD


Inteligência Artificial
Aprendizado de Máquina
Aprendizado Profundo
Ciência de Dados
KDD
Mineração de Dados

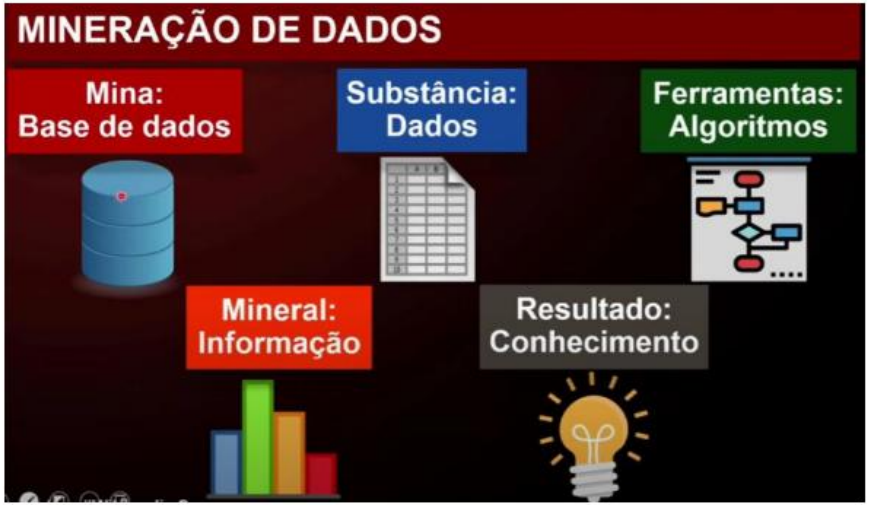
Posicionamento

DATA MINING – MINERAÇÃO DE DADOS

- Selecionar os métodos adequados para identificar padrões nos dados.
- Realizar a busca efetiva por padrões relevantes, utilizando uma forma específica de representação ou um conjunto de representações.
- Ajustar os parâmetros do algoritmo para obter o melhor desempenho na tarefa em questão.

DATA MINING

- **Importância:** Com o crescimento exponencial dos bancos de dados, a mineração de dados se torna essencial para transformar grandes volumes de dados em informações valiosas que podem ajudar na tomada de decisões, seja em negócios, ciência ou engenharia.



Atividades de Mineração de Dados

- **Tarefas de Previsão**
- Prever valor de um atributo baseado em valores de outros;
- Variável dependente (alvo - target) ⟶ Variáveis independentes (explicativas);
- **Tarefas Descritivas**
- Estatísticas;
- Correlações;
- Tendências;
- Grupos;
- Trajetórias;
- Distorções e anomalias;

Atributos e Medidas

- **Atributo**: Uma característica ou propriedade de um objeto que pode variar, seja de um objeto para outro ou de tempo para outro.
  - **Exemplo**:
    - Cor dos olhos: (objeto para objeto)
    - Temperatura, idade: (tempo para tempo)
- **Escala de medição**:
  - Regra ou função que associa um valor numérico ou categórico a um atributo (objeto).
  - Ex.: Peso físico, Sexo, Altura (categórico ou numérico), Temperatura.

Tipos de Atributos

- **Propriedades de um atributo** não precisam ser as mesmas dos valores usados para medi-los.
- **Valores usados** para representar um atributo podem ter propriedades que não sejam propriedades do próprio atributo e vice-versa.
    - **Exemplos**:
        - Idade: faz sentido obter estatísticas.
        - Nº de ID: serve apenas para distinção.
- **4 Tipos de atributos**:
    - Nominais
    - Ordinal

Tipos de Atributos

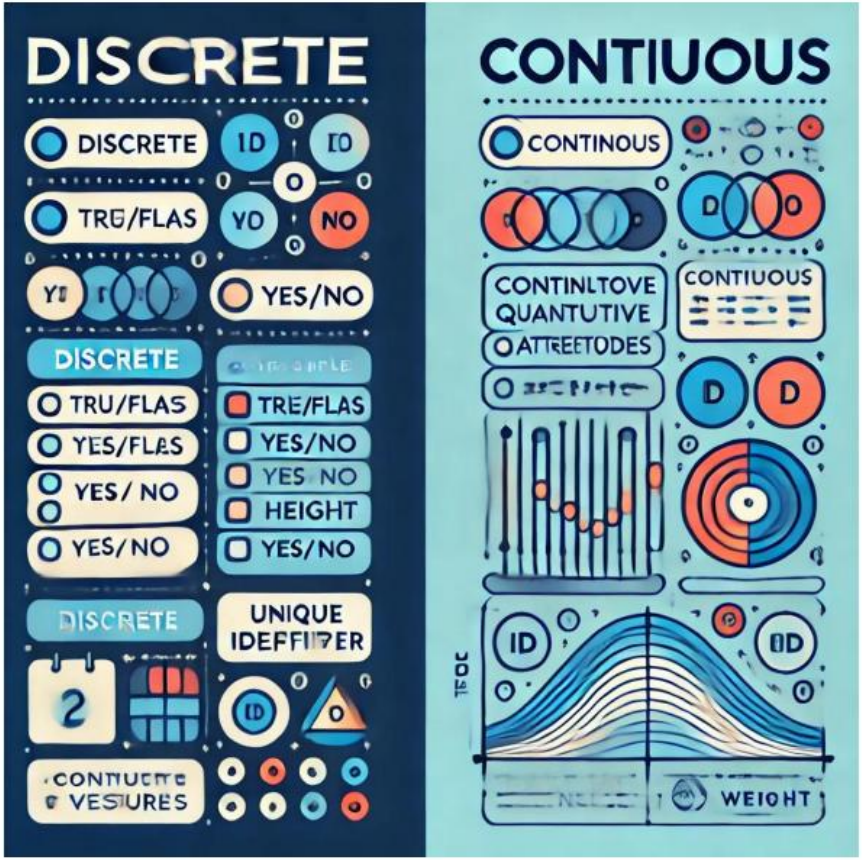
  - Intervalo
  - Proporção
- **Podemos descrever atributos:**
  - **Distinção**: = ou != ou ~
  - **Ordenação**: <, <=, >, >=
  - **Adição**: + e -
  - **Multiplicação**: *, ×

Atributos: Qualitativos X Quantitativos

- **Qualitativos (categorizados)**:
  - **Nominal**: ID
  - **Ordinal**: {regular, bom, ótimo}{1,2,3}
- **Quantitativo (numéricos)**:
  - **Intervalar**: temperatura (zero até x graus)
  - **Proporcional**: comprimento (metros, pés...)

Atributos Quantitativos

- **Discretos**:
  - ID
  - Verdadeiro, Falso
  - Sim/Não
- **Contínuos**:
  - Números reais: altura, peso



Conjuntos de Dados

- **Principais são 3:**
  - **Dados de registros**
  - **Dados em grafos**
  - **Dados ordenados**

Conjuntos de Dados

- **Características gerais:**
  - **Dimensão:**
    - Número de atributos
    - Maldição da dimensionalidade (dificuldade em análise)
    - Redução de dimensionalidade (IA - PCA)
  - **Dispersão:**
    - Variabilidades
    - Importantes para a Mineração
  - **Resolução:**
    - Proximidade (terra vista de um prédio ou avião) - fractais

Sobre a Qualidade dos dados

- **Dados possuem:**
  - Imprecisões
  - Falhas em medições
  - Ruídos
  - Incertezas
  - Erros
  - Duplicidade
  - Ausência
  - Quantidade
  - Viés

Erros de Medição e Coleta de Dados

- **Erro**: Valor medido ≠ Valor Real
  - Sistemáticos
  - Aleatórios
- **Ruídos**:
  - Componente aleatório de um erro de medição
  - Medidas espaciais e temporais
- **Artefatos**:
  - Erros determinísticos - consistentes

Diferença entre: Precisão, Foco e Exatidão

- **Qualidade dos experimentos:**
  - **Precisão:**
    - Refere-se à proximidade de medições repetidas da mesma quantidade entre si. Ou seja, se você medir a mesma coisa várias vezes, os resultados devem ser muito próximos uns dos outros.
    - É exemplificado pelo desvio padrão, que mede a dispersão dos valores em relação à média.

Diferença entre: Precisão, Foco e Exatidão

- **Foco**:
  - Refere-se à variação sistemática das medições em relação à quantidade real que está sendo medida. Isso significa o quão consistentemente as medições se desviam de um valor verdadeiro conhecido.
  - Indica a diferença entre a média do conjunto de medições e o valor real conhecido da quantidade medida.

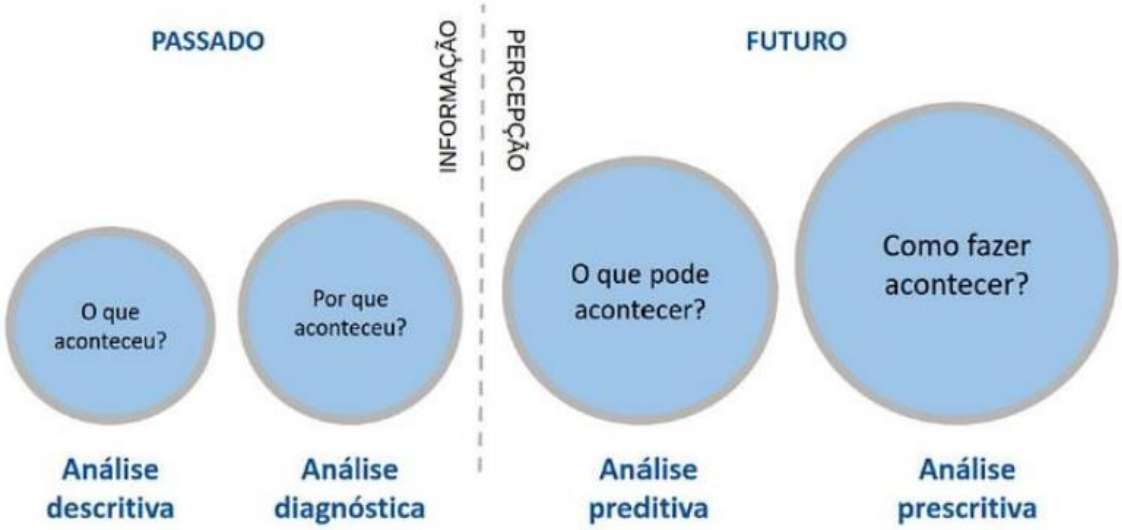
Diferença entre: Precisão, Foco e Exatidão

- **Exatidão**:
  - Refere-se ao grau de erro na medição dos dados.
  - É uma medida de quão perto as medições estão do valor real ou verdadeiro da quantidade sendo medida.
  - Uma medição é exata quando o valor medido é muito próximo do valor real, com pouco erro.

Em síntese

- Essas definições ajudam a entender a qualidade das medições em experimentos:
- **Precisão** é sobre a consistência entre as medições.
- **Foco** é sobre a precisão central ou média das medições em relação ao valor verdadeiro.
- **Exatidão** é sobre quão correta é uma medição em comparação ao valor real.

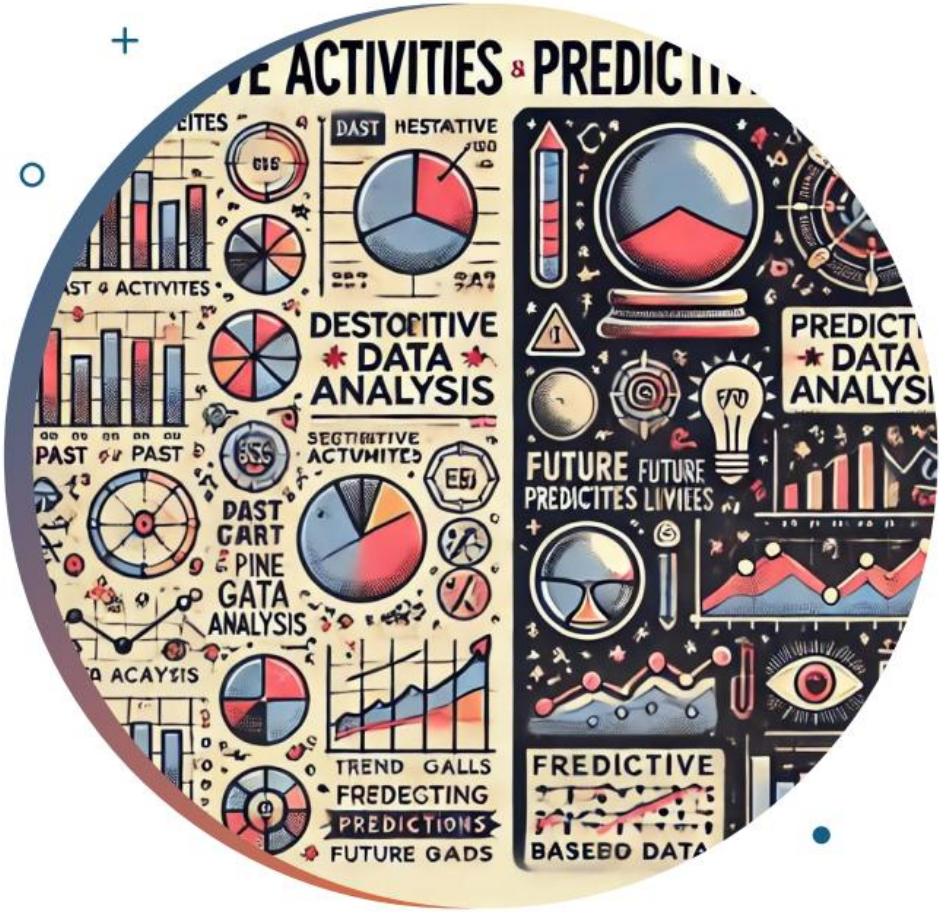
Mineração de Dados
Atividades descritivas vs preditivas



https://medium.com/@edselferri/an%C3%A1lise-preditiva-ou-prescritiva-sua-empresa-precisa-de-ambos-63b49caf09cf

Nicole Fallon. businessnewsdaily.com

O Big Data recebe muita agitação no mundo dos negócios. É verdade que a análise de dados pode fornecer insights úteis e profundos sobre sua empresa e seus clientes, mas somente se você usar esses insights em todo o seu potencial.



Mineração de Dados

1. **Mineração de Dados**
    1. **Atividades Descritivas**
        1. Regras de Associação
        2. Agrupamento
        3. Sumarização
    2. **Atividades Preditivas**
        1. Classificação
        2. Regressão
        3. Séries Temporais

Atividades descritivas

Regras de Associação

- **Definição:** Descobrir relações entre variáveis em um grande conjunto de dados.
- **Exemplo:** Em um supermercado, se alguém compra pão, é provável que compre leite também (análise de cestas de compras).

Agrupamento

- **Definição:** Agrupamento de dados semelhantes em clusters.
- **Exemplo:** Agrupar clientes de um banco com base em comportamento similar de gastos.

Exemplo algoritmo de agrupamento

- O **K-means** é um algoritmo de agrupamento (clustering) amplamente utilizado em mineração de dados e aprendizado de máquina.
- Ele é usado para particionar um conjunto de dados em um número pré-definido de grupos chamados **clusters**.
- O objetivo do K-means é agrupar os dados de tal maneira que os itens dentro de um mesmo cluster sejam o mais similares possível entre si, enquanto os clusters diferentes sejam o mais distintos possível uns dos outros.

Como Funciona o K-means?

1. Escolha do Número de Clusters (K):

1. Primeiro, você deve escolher o número de clusters K que deseja encontrar nos dados.

2. Inicialização dos Centróides:

1. O algoritmo começa escolhendo aleatoriamente K pontos nos dados para serem os centróides iniciais dos clusters. Um centróide é basicamente o ponto central de um cluster.

3. Atribuição dos Pontos aos Clusters:

Como Funciona o K-means?

1. Cada ponto do conjunto de dados é atribuído ao cluster cujo centróide está mais próximo. Isso cria K clusters com base na proximidade dos pontos ao centróide.

1.Atualização dos Centróides:

1. Depois de atribuir todos os pontos a um cluster, os centróides de cada cluster são recalculados como a média de todos os pontos que pertencem a esse cluster.

2.Repetição:

1. Os passos 3 e 4 são repetidos até que os centróides não mudem mais ou mudem muito pouco entre as iterações, o que indica que os clusters se estabilizaram.

Exemplo Agrupamento

- Imagine que você tem um conjunto de dados de pessoas com suas idades e rendas anuais.
- Você quer agrupar essas pessoas em 3 grupos com base em suas semelhanças.
- O K-means pode ser utilizado para agrupar as pessoas de maneira que, por exemplo, um cluster agrupe jovens com baixa renda, outro cluster agrupe adultos com renda média, e o último cluster agrupe pessoas mais velhas com alta renda.

Visualização

- **Visualização**
- Visualmente, o K-means pode ser entendido como encontrar os "centros" de grupos de pontos em um gráfico, onde cada grupo (cluster) contém pontos que estão próximos uns dos outros.

Limitações do K-means

- **Número de Clusters:** É necessário definir o número de clusters K antes de rodar o algoritmo, o que nem sempre é trivial.
- **Sensibilidade a Outliers:** O K-means pode ser fortemente influenciado por outliers, ou seja, pontos de dados que estão longe dos outros e podem distorcer os centróides.
- **Clusters Esféricos:** O K-means assume que os clusters são aproximadamente esféricos, o que pode não ser adequado para todos os tipos de dados.

Exercício Prático:

Exemplo K-Means

```
1  from sklearn.cluster import KMeans
2  from sklearn.datasets import load_iris
3  import matplotlib.pyplot as plt
4
```

✓ 0.0s

- `sklearn.cluster import KMeans`: Importa a classe `KMeans` da biblioteca `scikit-learn`, que é usada para executar o algoritmo K-means.
- `sklearn.datasets import load_iris`: Importa a função `load_iris`, que carrega o conjunto de dados Iris, um dos datasets de exemplo mais comuns em aprendizado de máquina.
- `matplotlib.pyplot as plt`: Importa a biblioteca `matplotlib.pyplot`, comumente usada para criar visualizações gráficas.

- **Tarefa:** Usar o algoritmo K-means em Python para agrupar um conjunto de dados simples, como as características de flores (dataset Iris).

Conjunto de Dados Iris

- O **conjunto de dados Iris** é um dos datasets mais conhecidos e amplamente utilizados em aprendizado de máquina e estatística. Foi introduzido pelo botânico e estatístico britânico Ronald A. Fisher em 1936 em seu artigo "The use of multiple measurements in taxonomic problems" como um exemplo de análise discriminante.

- **Estrutura do Conjunto de Dados Iris**
- O conjunto de dados Iris contém informações sobre três diferentes espécies de íris (um tipo de flor) - *Iris setosa*, *Iris versicolor*, e *Iris virginica*. Cada espécie tem 50 exemplos, totalizando 150 amostras no conjunto de dados.

- **Atributos**
- Cada amostra no conjunto de dados Iris tem quatro características (ou atributos) que foram medidas em centímetros:

1. **Comprimento da sépala** (*Sepal Length*)
2. **Largura da sépala** (*Sepal Width*)
3. **Comprimento da pétala** (*Petal Length*)
4. **Largura da pétala** (*Petal Width*)

- Essas características foram escolhidas porque, na época, eram facilmente mensuráveis e acreditava-se que poderiam ser usadas para diferenciar as espécies.
- **Classes (Rótulos)**
- Além dessas características, cada amostra é rotulada com uma das três espécies de íris:
- **Iris setosa**
- **Iris versicolor**
- **Iris virginica**

- **Por que o Conjunto de Dados Iris é Popular?**
- **Simplicidade:** Com apenas 150 amostras e 4 características, o conjunto de dados Iris é pequeno e simples, o que o torna ideal para introduzir técnicas de análise de dados, estatística e aprendizado de máquina.
- **Classificação de Três Classes:** Ele contém exemplos de três classes distintas, o que permite testar e comparar algoritmos de classificação.
- **Facilidade de Visualização:** Como o conjunto de dados tem apenas quatro características, é possível visualizá-lo de forma simples, usando gráficos de dispersão e outras técnicas de visualização.

Usos Comuns do Conjunto de Dados Iris

- **Classificação:** Testar algoritmos de classificação, como K-nearest neighbors, árvores de decisão, e, claro, K-means para agrupamento.
- **Visualização:** Praticar técnicas de visualização de dados, como gráficos de dispersão 2D ou gráficos 3D.
- **Exploração de Dados:** Usado em exercícios de pré-processamento e exploração de dados para entender a distribuição das características e a separabilidade das classes.

Carregando

```
from sklearn.datasets import load_iris

# Carregar o conjunto de dados
data = load_iris()

# Visualizar as primeiras 5 amostras
print(data.data[:5])
print(data.target[:5])
```

Aqui, data.data contém as características das flores, e data.target contém os rótulos correspondentes às espécies.

O conjunto de dados Iris é um excelente ponto de partida para aprender sobre técnicas de classificação e agrupamento, bem como para praticar a aplicação de algoritmos de aprendizado de máquina.

Exemplos Práticos de Mineração de Dados

- **Exemplo de Classificação:** Uma editora que deseja saber quais clientes são mais propensos a comprar um novo livro pode usar a classificação para direcionar campanhas de marketing.
- **Exemplo de Regras de Associação:** Supermercados utilizam regras de associação para entender quais produtos são frequentemente comprados juntos, melhorando o layout das prateleiras e as ofertas promocionais.

Sumarização

- **Definição:** A sumarização em mineração de dados é o processo de condensar grandes conjuntos de dados em informações mais simples e facilmente compreensíveis.
- Isso pode ser feito através de estatísticas descritivas, como médias, medianas, modas, desvio padrão, entre outros, ou por técnicas mais avançadas, como sumarização de textos ou resumos automáticos.

Exemplo

- Neste exemplo, vamos usar o conjunto de dados **Iris** para ilustrar como podemos sumarizar dados usando Python. Vamos focar em técnicas básicas de sumarização estatística, que são fundamentais para entender a distribuição dos dados antes de aplicar qualquer modelo de aprendizado de máquina.

1. **Importação das Bibliotecas**

Primeiro, importamos as bibliotecas necessárias:

python Copiar código

```python
import pandas as pd
from sklearn.datasets import load_iris
```

- `pandas as pd`: Pandas é uma biblioteca poderosa para manipulação e análise de dados em Python.
- `load_iris`: Função da scikit-learn para carregar o conjunto de dados Iris.

2. **Carregamento e Estruturação dos Dados**

Agora, vamos carregar o conjunto de dados Iris e organizá-lo em um DataFrame do Pandas para facilitar a análise.

python Copiar código

```python
# Carregar o conjunto de dados Iris
iris = load_iris()

# Criar um DataFrame
df = pd.DataFrame(data=iris.data, columns=iris.feature_names)
df['species'] = iris.target
df['species'] = df['species'].map({0: 'setosa', 1: 'versicolor', 2: 'virginica'})

# Exibir as primeiras linhas do DataFrame
df.head()
```

3. **Sumarização Estatística dos Dados**

Vamos sumarizar o conjunto de dados usando algumas funções estatísticas básicas, como `mean`, `median`, `std`, etc.

python Copiar código

```python
# Sumarização estatística básica
summary = df.describe()

# Exibir o resumo estatístico
print(summary)
```

Este código produzirá um resumo estatístico das quatro características das flores, fornecendo valores como média (`mean`), desvio padrão (`std`), valores mínimo (`min`), e máximo (`max`), bem como os percentis.

4. **Sumarização por Categoria**

Podemos também sumarizar os dados agrupando-os por espécie de íris. Isso é útil para comparar as características entre as diferentes espécies.

python Copiar código

```python
# Sumarização por espécie
species_summary = df.groupby('species').mean()

# Exibir a sumarização por espécie
print(species_summary)
```

Aqui, calculamos a média de cada característica (comprimento da sépala, largura da sépala, comprimento da pétala, largura da pétala) para cada espécie de íris.

5. **Visualização da Sumarização**

Para tornar a sumarização mais intuitiva, podemos usar gráficos:

python Copiar código

```python
import matplotlib.pyplot as plt

# Visualização das médias das características por espécie
species_summary.plot(kind='bar', figsize=(10, 6))
plt.title('Média das Características por Espécie')
plt.ylabel('Média')
plt.show()
```

Este gráfico de barras mostrará a média das características de cada espécie, facilitando a comparação visual entre elas.

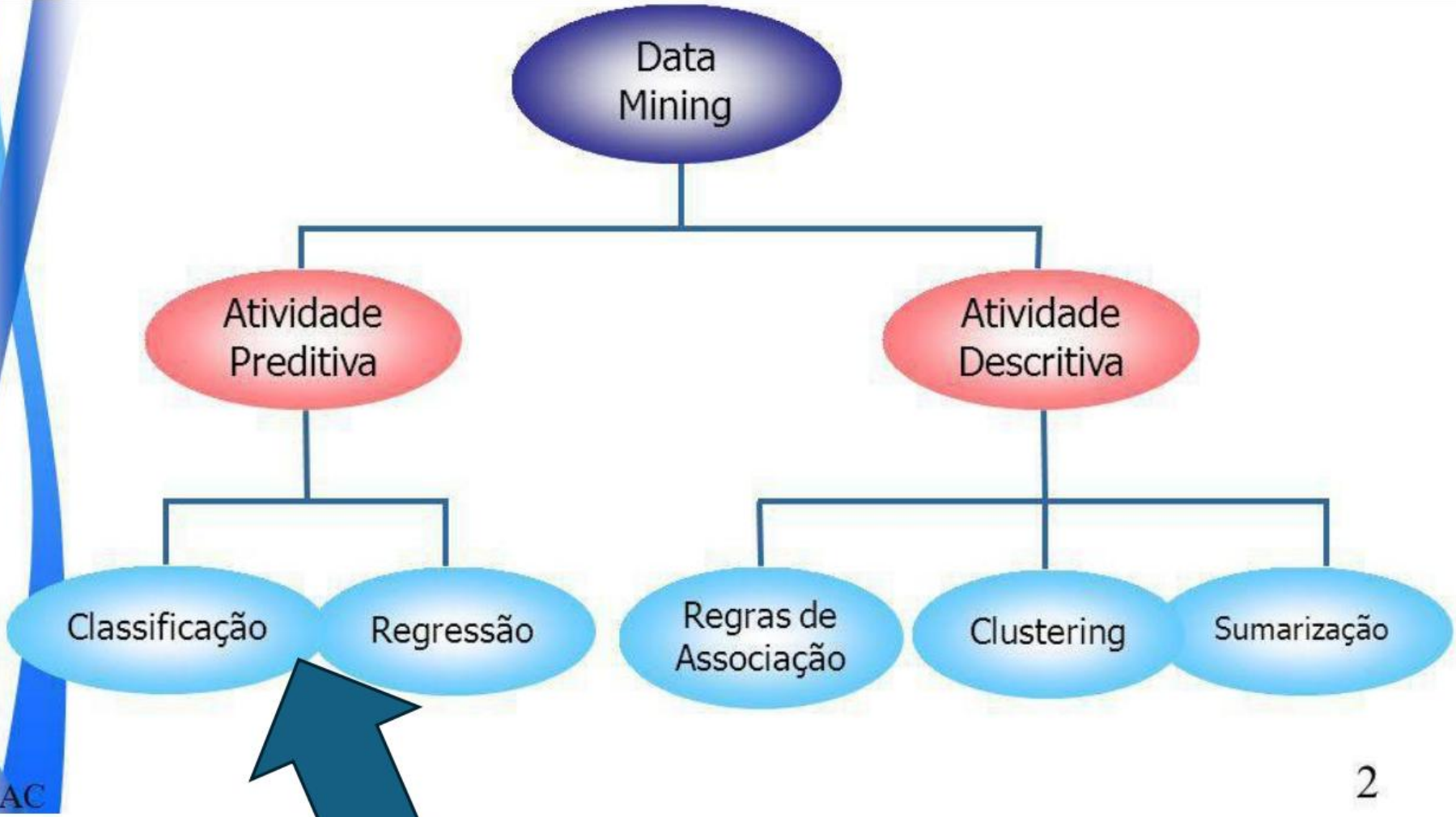
Entendendo a sumarização

- A sumarização é um passo essencial na análise de dados, pois permite:
- **Entender a distribuição dos dados:** Através de estatísticas descritivas, podemos ter uma noção do comportamento dos dados, como qual espécie tem sépalas ou pétalas maiores.
- **Identificar anomalias:** Analisando valores extremos ou outliers, podemos descobrir inconsistências nos dados que podem precisar de tratamento especial.
- **Facilitar a visualização e comunicação dos resultados:** Resumos e gráficos ajudam a condensar grandes volumes de dados em insights claros e comunicáveis.

Em síntese

- Sumarização é uma técnica fundamental em qualquer análise de dados, pois fornece uma visão geral do conjunto de dados e ajuda a identificar padrões e tendências.
- No contexto de aprendizado de máquina, uma boa compreensão dos dados através da sumarização pode guiar na escolha dos melhores modelos e métodos a serem aplicados.

Atividades Preditivas


Data Mining
Atividade Preditiva
Atividade Descritiva
Classificação
Regressão
Regras de Associação
Clustering
Sumarização

Atividades Preditivas

- **Classificação**

- **Definição:** Processo de atribuir uma etiqueta ou categoria a um dado com base em características. Classificação é uma técnica de aprendizado supervisionado onde o objetivo é prever a classe ou categoria de uma amostra com base em suas características. Em termos simples, a classificação ajuda a dividir os dados em categorias predefinidas.

Atividades Preditivas

- **Classificação**

- **Exemplos Comuns:**
- **Detecção de Spam:** Classificar e-mails como "spam" ou "não spam".
- **Diagnóstico Médico:** Determinar se um paciente tem uma doença com base em seus sintomas (ex.: "positivo" ou "negativo" para uma doença).

Classificação

- **Técnicas Comuns:**
- **Árvores de Decisão:** Estruturas em forma de árvore que dividem os dados com base em características até chegar a uma decisão final.
- **K-Nearest Neighbors (KNN):** Classifica os dados com base nas classes dos dados mais próximos (vizinhos).
- **Máquinas de Vetores de Suporte (SVM):** Encontra a melhor fronteira de decisão que separa as diferentes classes nos dados.

Atividades Preditivas

- **Regressão**

- **Definição:** Regressão é outra técnica de aprendizado supervisionado, mas ao invés de prever uma categoria, o objetivo é prever um valor contínuo. A regressão tenta modelar a relação entre uma variável dependente e uma ou mais variáveis independentes.
- **Exemplo:** Prever o preço de uma casa com base em tamanho, localização, etc.

Regressão

- **Exemplos Comuns:**
- **Previsão de Preços de Casas:** Prever o preço de uma casa com base em características como tamanho, localização e número de quartos.
- **Análise de Vendas:** Estimar as vendas futuras com base em dados históricos de vendas.

Regressão

- **Técnicas Comuns:**
- **Regressão Linear Simples:** Modela a relação linear entre duas variáveis (uma dependente e uma independente).
- **Regressão Linear Múltipla:** Extensão da regressão linear simples para várias variáveis independentes.
- **Regressão Polinomial:** Modela relações mais complexas, permitindo que a curva de ajuste seja polinomial (não-linear).

Séries Temporais

- **Definição:** Análise de dados que são coletados ao longo do tempo.
- **Exemplo:** Exemplos incluem preços de ações diários, leituras de temperatura horárias, vendas mensais de uma loja ou medições semanais de tráfego em um website.
- O objetivo da análise de séries temporais é compreender os padrões subjacentes nos dados para fazer previsões futuras.

Componentes de Séries Temporais

1. **Tendência (Trend):** Refere-se ao movimento de longo prazo nos dados. Por exemplo, o preço de uma ação pode ter uma tendência ascendente ao longo de vários anos.
2. **Sazonalidade (Seasonality):** Padrões repetitivos ou cíclicos que ocorrem em intervalos regulares, como vendas de varejo que aumentam durante o período de Natal.
3. **Ciclicidade (Cyclical):** Flutuações que ocorrem de forma irregular, frequentemente devido a fatores econômicos ou de negócios, como ciclos de mercado.
4. **Ruído (Noise):** Variações aleatórias nos dados que não seguem um padrão específico. Isso pode incluir erros de medição ou eventos imprevisíveis.

Exemplos de Uso de Séries Temporais

- **Previsão de Preços de Ações:** Usando o histórico de preços para prever os preços futuros.
- **Previsão de Vendas:** Estimando a demanda futura com base em dados de vendas passadas.
- **Análise Climática:** Estudando padrões de temperatura ao longo do tempo para prever condições futuras.

Técnicas Comuns em Séries Temporais

1. **Média Móvel (Moving Average):** Uma técnica simples que calcula a média dos valores em uma janela de tempo móvel. Usada para suavizar ruídos e destacar tendências.
2. **Decomposição de Séries Temporais:** Técnica que separa uma série temporal em seus componentes: tendência, sazonalidade e ruído.
3. **Modelos Autoregressivos (AR):** Modelos que utilizam valores passados da série temporal para prever valores futuros.
4. **Modelos de Alisamento Exponencial:** Usados para prever séries temporais com uma forte componente de tendência ou sazonalidade.
5. **Modelos ARIMA (AutoRegressive Integrated Moving Average):** Um dos modelos mais populares, combina autoregressão e média móvel para prever séries temporais.

Referências

- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Concepts and Techniques*. Elsevier.
- Larose, D. T. (2015). *Data Mining and Predictive Analytics*. Wiley.
- Shmueli, G., Bruce, P. C., & Patel, N. R. (2017). *Data Mining for Business Analytics: Concepts, Techniques, and Applications with XLMiner*. Wiley.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern Classification*. Wiley.
- Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.
- Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (2015). *Time Series Analysis: Forecasting and Control*. Wiley.