# Aula 01

I.A E BIG DATA

## Introdução

Este livro foi elaborado com o objetivo de fornecer uma base sólida e didática sobre os principais conceitos e práticas em **Inteligência Artificial (IA)** e **Big Data**, conforme as competências e objetivos de aprendizagem propostos em cursos técnicos e de graduação tecnológica. O material segue uma abordagem prática e acessível e é ideal para estudantes, docentes e profissionais iniciantes na área de ciência de dados.

## Capítulo 1 - Mineração de Dados (Data Mining)

### 1.1 Histórico e Definições

A Mineração de Dados é o processo de explorar grandes conjuntos de dados à procura de padrões e informações úteis. O termo surgiu nos anos 1990 com a disseminação de grandes bancos de dados e a necessidade de extrair valor deles. Com raízes na estatística, aprendizado de máquina e bancos de dados, a mineração de dados é usada em diversas áreas como saúde, segurança pública, varejo, indústria financeira e telecomunicações.

### 1.2 Processos de Descoberta do Conhecimento

* **KDD (Knowledge Discovery in Databases)**: é um processo completo que envolve seleção, pré-processamento, transformação, mineração de dados e interpretação/avaliação.
* **CRISP-DM (Cross-Industry Standard Process for Data Mining)**: é um modelo padrão composto pelas seguintes etapas:
  1. Compreensão do negócio
  2. Compreensão dos dados
  3. Preparação dos dados
  4. Modelagem
  5. Avaliação
  6. Implantação

### 1.3 PCA - Análise dos Componentes Principais

A Análise dos Componentes Principais é uma técnica estatística usada para reduzir a dimensionalidade dos dados ao mesmo tempo em que mantém a maior variabilidade possível. Isso é feito transformando o conjunto original de variáveis em um novo conjunto de variáveis ortogonais (componentes principais).

### 1.4 Tratamento de Dados Faltantes

* **Remoção de linhas ou colunas com dados faltantes**
* **Imputação estatística** (média, mediana, moda)
* **Modelagem preditiva** para preencher os valores com algoritmos de regressão

### 1.5 Engenharia de Atributos

* **Label Encoding**: transforma rótulos de texto em números inteiros.
* **One Hot Encoding**: cria variáveis binárias para cada categoria.
* **Normalização**: escala os dados entre 0 e 1.
* **Padronização**: escala os dados com média 0 e desvio padrão 1.
* Separação entre:
  + **Features (X)**: atributos de entrada
  + **Target (y)**: atributo de saída

### Exercícios de Fixação

1. Explique a diferença entre KDD e CRISP-DM.
2. Dê um exemplo prático de como o PCA pode ser aplicado.
3. Crie um conjunto de dados com 3 atributos categóricos e aplique One Hot Encoding.
4. Qual a diferença entre normalização e padronização?
5. Aplique um método de tratamento de dados faltantes em uma base com valores ausentes.

## Capítulo 2 - Inteligência Artificial

### 2.1 Histórico e Aplicações

A IA surgiu oficialmente em 1956 na conferência de Dartmouth. Desde então, passou por fases de grande investimento e recessão (invernos da IA), até o recente avanço com o aprendizado profundo (deep learning). É aplicada em sistemas de recomendação, carros autônomos, detecção de fraudes, medicina diagnóstica, robótica e assistentes virtuais como Alexa, Siri e Google Assistant.

### 2.2 Fundamentos da Aprendizagem de Máquina

Machine Learning é uma subárea da IA focada em criar algoritmos que aprendem com dados. A ideia é permitir que o sistema encontre padrões e tome decisões sem ser explicitamente programado para cada tarefa.

### 2.3 Modelos Supervisionados e Não Supervisionados

* **Supervisionados**: algoritmos treinados com entradas e saídas conhecidas (regressão, classificação)
* **Não supervisionados**: não possuem rótulos, apenas entradas (agrupamento, associação)

### 2.4 Algoritmos

* **K-means**: agrupa dados em K clusters. Exige definição de K previamente.
* **Agrupamento Hierárquico**: constrói uma árvore de agrupamento (dendrograma).
* **DBScan**: baseado em densidade, bom para dados com ruído.
* **KNN (K-Nearest Neighbors)**: classifica com base na maioria dos vizinhos mais próximos.
* **Regressão Linear**: modela uma relação linear entre variáveis.
* **Regressão Logística**: usada para problemas de classificação binária.
* **Árvores de Decisão**: dividem os dados com base em regras.
* **Random Forest**: conjunto de árvores de decisão.
* **SVM**: separa dados com margens máximas.
* **Gradient Boosting / XGBoost**: otimizam previsões combinando modelos fracos.

### 2.5 Métricas de Desempenho

* **Erro Quadrático Médio (MSE)**
* **Erro Absoluto Médio (MAE)**
* **R² (Coeficiente de Determinação)**
* **Matriz de Confusão**: avalia classificações verdadeiras e falsas
* **Acurácia, Precisão, Recall, F1-score**: métricas de desempenho de classificadores

### Exercícios de Fixação

1. Compare modelos supervisionados e não supervisionados.
2. Aplique o algoritmo K-means a um conjunto de dados simples.
3. Use uma árvore de decisão para classificar dados binários.
4. Calcule acurácia e F1-score com base em uma matriz de confusão.
5. Modele um problema real usando regressão logística.

## Capítulo 3 - Big Data & Analytics

### 3.1 Fundamentos e Importância

Big Data se refere a conjuntos de dados que são tão grandes e complexos que não podem ser tratados com técnicas tradicionais de processamento de dados. Sua importância cresce à medida que sensores, smartphones e redes sociais geram quantidades massivas de dados.

### 3.2 Os 5 V’s do Big Data

* **Volume**: Petabytes e Exabytes
* **Variedade**: dados estruturados, semiestruturados e não estruturados
* **Velocidade**: fluxo de dados em tempo real
* **Veracidade**: qualidade e confiabilidade dos dados
* **Valor**: utilidade que pode ser extraída desses dados

### Aplicações Reais

* Monitoramento de redes sociais
* Previsão de falhas em equipamentos
* Recomendação de produtos em e-commerce
* Análise preditiva no setor de saúde

### Exercícios de Fixação

1. Descreva os 5 V's do Big Data com exemplos.
2. Dê três aplicações reais do Big Data.
3. Identifique as principais fontes de dados em uma organização.
4. Quais os desafios para processar Big Data?

## Capítulo 4 - Plataforma de Big Data: Hadoop

### 4.1 Características

O Hadoop é um framework que permite o armazenamento distribuído de dados e seu processamento paralelo em clusters. É open-source e escalável.

### 4.2 Armazenamento e Processamento Distribuído

* Utiliza o HDFS (Hadoop Distributed File System)
* Executa tarefas com MapReduce
* Ideal para análise de dados em larga escala

### 4.3 Conceitos de Cluster

* **NameNode**: gerencia metadados do sistema de arquivos
* **DataNode**: armazena os dados efetivos
* **JobTracker** e **TaskTracker**: gerenciam e executam tarefas (em MapReduce)

### 4.4 Ecossistema Hadoop

* **HDFS**: sistema de arquivos distribuído
* **YARN**: gerenciador de recursos
* **MapReduce**: modelo de programação para processamento paralelo
* **Hive**: SQL para Hadoop
* **Pig**: linguagem de script
* **HBase**: banco de dados NoSQL
* **Spark**: engine de processamento rápido

### Exercícios de Fixação

1. O que é um cluster Hadoop? Quais seus componentes?
2. Explique a diferença entre NameNode e DataNode.
3. O que é MapReduce e como ele funciona?
4. Compare Hive, Pig e HBase.
5. Instale e configure um cluster Hadoop local com 2 nós.

## Capítulo 5 - Metodologia e Avaliação

### 5.1 Metodologia

* Aulas expositivas e práticas
* Projetos com dados reais
* Simulações com ferramentas como Weka, Orange, Jupyter, Spark
* Debates sobre privacidade e ética

### 5.2 Avaliação Formativa

* Participação ativa
* Estudos de caso
* Relatórios semanais

### 5.3 Avaliação Somativa

* Provas teóricas e práticas
* Projeto integrador com IA e Big Data
* Apresentações técnicas

## Referências Bibliográficas

### Bibliografia Básica

* MEDEIROS, Luciano F. Inteligência artificial aplicada. Intersaberes, 2018.
* VALDATI, Aline de B. Inteligência artificial - IA. Contentus, 2020.
* TAURION, Cezar. Big Data. Brasport, 2013.

### Bibliografia Complementar

* KAUFMAN, Dora. Desmistificando a inteligência artificial. Autêntica, 2022.
* LUGER, G. F. Inteligência artificial. Pearson, 2013.
* ARAÚJO, R. C. A. Urban data analytics e IOT. Contentus, 2020.
* QUINSLER, Aline P. Probabilidade e estatística. Intersaberes, 2022.
* MORETTIN, Luiz G. Estatística básica. Pearson, 2010.

## Conclusão

Este livro foi desenvolvido para auxiliar estudantes e profissionais na compreensão e aplicação dos conceitos fundamentais de IA e Big Data. Com uma abordagem prática e aprofundada, incentiva a inovação, o pensamento crítico e a responsabilidade ética. Os capítulos incluem conteúdo teórico, exemplos práticos e exercícios para fixação, resultando em um aprendizado mais completo e eficaz.

ATIVIDADE 01:

Quais são as colunas ou variáveis do dataseat:  
voiceNo  
StockCode  
Quantity  
InvoiceDate  
UnitPrice  
CustomerID  
Country  
  
Existem dados faltantes? Em quais linhas? Como eles afetam a análise?  
Descreption – impossibilita saber qual foi o produto vendido  
CustomerID – impossibilita saber qual foi o cliente que comprou  
UnitPrice – impede saber o valor do produto e quantidade que foi comprado  
Country – impede saber o pais que o produto foi comprado  
  
  
Quais variáveis parecem ter mais impacto para prever o comportamento de compra?  
  
  
Existem colunas com dados categóricos que precisem ser convertidos para análise?   
  
  
Quais tipos de variáveis temos? Numéricas, Categóricas, Temporais?