

Análise Exploratória de Dados

Tecnologias de Informação - Data & Analytics

- **Nível:**
 - **Duração:** 18h
-

Sobre o curso

A Análise Exploratória de Dados é uma das principais fases da estrutura de um projecto de Data Science. Contribui para uma diminuição das dimensões do conjunto de dados, tornando o problema menos complexo, e ainda a levar menos tempo a efectuar o treino dos modelos de machine learning, num passo subsequente. Além disso, a Análise Exploratória de Dados é uma actividade fundamental para termos uma primeira noção do tipo de dados que temos à nossa disposição.

Assim, os principais objectivos desta análise são:

- Maximizar os insights no conjunto de dados
- Relevar estruturas e padrões características do conjunto de dados
- Detectar anomalias e outliers
- Extrair e seleccionar variáveis mais importantes
- Avaliar premissas existentes (e.g., intuições de negócio actuais)
- Preparar todo o conjunto de dados para um modelo mais simples e com capacidade preditiva e explicativa elevada

No final da ação de formação os participantes deverão estar aptos a:

- Perceber o que é a Análise Exploratória de Dados e como se encaixa no workflow de Data Science
 - Criar visualizações objectivas e eficazes de dados que resultem em acções concretas
 - Criar sumário estatísticos de dados
 - Seleccionar e justificar variáveis mais relevantes para utilizar num modelo de machine learning
 - Reduzir a dimensão do conjunto de dados através de selecção e extração de variáveis
 - Construir uma *pipeline* composto por diferentes módulos de pré-processamento de dados
-

Destinatários

- Data Scientists / Data Analysts / Business Analysts

- Managers
 - IT
-

Pré-requisitos

- Programação Básica em Python (com conhecimentos de pandas e numpy)
 - Noções de Estatística
-

Metodologia

- A avaliação deste curso é feita através de um projecto final prático (programação) individual sobre um dataset real.
-

Programa

- Introduction to EDA
- Data Visualization
 - Introduction to Matplotlib and Seaborn
 - Line chart, bar plot, histograms, heatmaps, box-plot, distribution plots
- Descriptive Statistics
 - Numerical and Categorical Variables
- Dimensionality Reduction
 - Feature Selection: Pearson Correlation, Chi-square test, Mutual Information, etc.
 - Feature Extraction: Principal Component Analysis
- Pipelines
 - Scikit-learn pipelines
 - How to build your own Transformer