Atividades Integradas em Data Science

Seu Nome ou Instituição

4 de fevereiro de 2025

Sumário

1 Atividade 1 – Introdução à Data Science com o Dataset Titanic

Objetivo: Familiarizar os alunos com o ambiente Python e a leitura de dados.

Dataset: Titanic Dataset (Kaggle)

Descrição:

- Importar o dataset utilizando `pandas` com `pd.read_csv()`.
- Usar `numpy` para verificar os tipos de dados e contagem de registros (ex.: `np.unique()`).
- Exibir as 5 primeiras linhas com `DataFrame.head()` e gerar um sumário descritivo com `DataFrame.describe()`.
- Criar um gráfico simples com `matplotlib` (por exemplo, gráfico de barras para a distribuição de sobreviventes).

2 Atividade 2 – Análise Exploratória Básica com o Dataset Wine Quality

Objetivo: Desenvolver uma análise inicial para entender as variáveis e suas distribuições.

Dataset: Wine Quality Dataset (Kaggle)

Descrição:

- Carregar os dados utilizando `pandas`.
- Calcular estatísticas básicas (média, mediana, variância) para cada variável utilizando `numpy`.
- Gerar histogramas e gráficos de dispersão com `matplotlib` para visualizar a distribuição das notas de qualidade dos vinhos e relacioná-las com variáveis como pH e acidez.

3 Atividade 3 – Detecção e Visualização de Outliers

Objetivo: Ensinar técnicas para identificar e analisar outliers.

Dataset: Mall Customers Dataset (Kaggle)

Descrição:

- Carregar os dados com `pandas` e utilizar descrições estatísticas para detectar possíveis outliers.
- Criar boxplots com `matplotlib` para visualizar outliers em variáveis numéricas.
- Aplicar filtros com `numpy` (por exemplo, valores acima de um determinado desvio padrão) e discutir estratégias de tratamento.

4 Atividade 4 – Manipulação e Transformação de Dados com `pandas` e `numpy`

Objetivo: Trabalhar com operações básicas e avançadas de manipulação de dados.

Dataset: Iris Dataset (Kaggle)

Descrição:

- Criar e manipular `DataFrames` com `pandas`: reorganização de colunas, filtragem e agrupamento de dados.
- Realizar operações matemáticas com `numpy` (operações vetoriais, cálculos de médias) aplicadas a arrays extraídos do DataFrame.
- Converter dados entre `DataFrame` e arrays do `numpy` e discutir as vantagens de cada estrutura.

5 Atividade 5 – Limpeza e Tratamento de Dados

Objetivo: Desenvolver habilidades para a preparação dos dados para análise.

Dataset: Heart Disease Dataset (Kaggle)

Descrição:

- Importar o dataset com `pandas` e identificar dados faltantes, duplicados e inconsistências utilizando funções como `isnull()`, `drop_duplicates()` e `fillna()`.
- Usar `numpy` para converter colunas em arrays e aplicar normalizações ou padronizações.
- Documentar as alterações e justificar as escolhas de tratamento dos dados.

6 Atividade 6 – Estatística I: Medidas de Tendência Central e Dispersão

Objetivo: Calcular e interpretar medidas estatísticas dos dados.

Dataset: Titanic Dataset (Kaggle)

Descrição:

- Calcular média, mediana, moda e desvio padrão utilizando métodos do `pandas` e funções do `numpy`.
- Apresentar os resultados em tabelas e discutir a relevância de cada medida.
- Criar gráficos com `matplotlib` para ilustrar a distribuição dos dados e as medidas calculadas.

7 Atividade 7 – Estatística II: Correlação e Visualização com Heatmaps

Objetivo: Explorar relações entre variáveis utilizando medidas de correlação.

Dataset: House Prices Dataset (Kaggle)

Descrição:

- Calcular a matriz de correlação com o método `DataFrame.corr()` do `pandas`.
- Gerar um heatmap com `matplotlib` para visualizar as correlações, destacando variáveis com relações fortes ou fracas.
- Analisar quais variáveis podem impactar a variável alvo (ex.: preço) e discutir implicações para a modelagem.

8 Atividade 8 – Introdução à Machine Learning: Divisão de Dados

Objetivo: Introduzir a separação de dados para treinamento e teste.

Dataset: Breast Cancer Wisconsin Dataset (Kaggle)

Descrição:

- Carregar os dados utilizando `pandas`.
- Utilizar o método `train_test_split` do `scikit-learn` para dividir o dataset (por exemplo, 70% treinamento e 30% teste).
- Utilizar `numpy` para validar a integridade dos dados após a divisão e discutir a importância da separação para evitar overfitting.

9 Atividade 9 – Aprendizado Supervisionado: Classificação com Regressão Logística

Objetivo: Implementar um modelo de classificação supervisionada.

Dataset: Iris Dataset (Kaggle)

Descrição:

- Realizar o pré-processamento dos dados (codificação de labels, normalização se necessário) com `pandas` e `numpy`.
- Dividir os dados com `train_test_split` do `scikit-learn`.
- Implementar uma regressão logística utilizando a classe `LogisticRegression` do `scikit-learn` e treinar o modelo.
- Plotar a acurácia do modelo ao longo de diferentes iterações utilizando `matplotlib` e discutir os resultados.

10 Atividade 10 – Aprendizado Não Supervisionado: Clusterização com K-Means

Objetivo: Aplicar técnicas de clusterização para identificar padrões.

Dataset: Mall Customers Dataset (Kaggle)

Descrição:

- Carregar e explorar os dados com `pandas`.
- Pré-processar os dados (normalização de variáveis numéricas) utilizando `numpy`.
- Aplicar o algoritmo K-Means do `scikit-learn` para criar clusters.
- Gerar gráficos com `matplotlib` para visualizar os clusters e discutir a escolha do número ótimo de clusters (por exemplo, método do cotovelo).

11 Atividade 11 – Separação de Bases e Pré-processamento para Modelos

Objetivo: Reforçar a importância de uma correta divisão dos dados.

Dataset: House Prices Dataset (Kaggle)

Descrição:

- Importar os dados com `pandas` e verificar dados ausentes com `isnull()`.
- Dividir o dataset em treino e teste com `train_test_split` do `scikit-learn`.
- Realizar normalização ou padronização dos dados utilizando funções do `numpy` ou o `StandardScaler` do `scikit-learn`.
- Discutir a importância de evitar o vazamento de informações (data leakage).

12 Atividade 12 – Cross Validation: Validação Cruzada de Modelos

Objetivo: Implementar a técnica de validação cruzada para uma avaliação robusta.

Dataset: Wine Quality Dataset (Kaggle)

Descrição:

- Dividir os dados (mantendo uma reserva para teste final) e aplicar `train_test_split`.
- Utilizar `cross_val_score` do `scikit-learn` para realizar k-fold cross validation (por exemplo, k=5) em um modelo (como regressão linear ou árvore de decisão).
- Calcular médias e desvios das métricas com `numpy` e gerar gráficos comparativos com `matplotlib`.

13 Atividade 13 – Análise Exploratória Avançada com Dataset Financeiro

Objetivo: Aprofundar a análise exploratória em um dataset mais complexo.

Dataset: Stock Market Data (Kaggle)

Descrição:

- Importar o dataset com `pandas` e realizar a limpeza dos dados (remoção de outliers, tratamento de datas).
- Calcular indicadores financeiros simples (médias móveis, variações percentuais) utilizando `numpy`.
- Criar gráficos de linhas, scatter plots e barras com `matplotlib` para visualizar tendências e sazonalidades.
- Analisar correlações entre os indicadores e discutir possíveis relações.

14 Atividade 14 – Feature Engineering e Criação de Novas Variáveis

Objetivo: Aprimorar o dataset por meio da criação de novas features para melhorar a modelagem.

Dataset: Airbnb New User Bookings (Kaggle)

Descrição:

- Analisar o dataset com `pandas` e identificar colunas que podem ser transformadas ou combinadas (ex.: extração de mês e dia a partir de datas).
- Utilizar `numpy` para aplicar transformações matemáticas ou estatísticas que resultem em novas variáveis.
- Documentar o processo de engenharia de features e discutir como elas podem impactar a performance dos modelos.

15 Atividade 15 – Estatística Aplicada à Modelagem e Interpretação

Objetivo: Integrar conceitos estatísticos na interpretação de modelos de machine learning.

Dataset: Titanic Dataset (Kaggle)

Descrição:

- Calcular estatísticas descritivas e de dispersão utilizando `pandas` e `numpy`.
- Criar gráficos (dispersão, histogramas) com `matplotlib` para visualizar os dados.
- Discutir como essas análises podem indicar quais variáveis possuem maior potencial preditivo.

16 Atividade 16 – Construção de um Pipeline de Machine Learning

Objetivo: Automatizar o fluxo de trabalho integrando pré-processamento, modelagem e validação.

Dataset: Breast Cancer Wisconsin Dataset (Kaggle)

Descrição:

- Construir um pipeline utilizando a classe `Pipeline` do `scikit-learn` que inclua:
  - Pré-processamento (tratamento de dados com `pandas` e normalização com `StandardScaler`).
  - Criação de novas features (usando transformações com `numpy`).
  - Treinamento de um modelo (por exemplo, SVM ou árvore de decisão).
- Executar validação cruzada integrada no pipeline e gerar gráficos comparativos de performance com `matplotlib`.

17 Atividade 17 – Avaliação Avançada de Modelos e Métricas

Objetivo: Implementar métricas avançadas para avaliação de modelos de classificação.

Dataset: Breast Cancer Wisconsin Dataset (Kaggle)

Descrição:

- Treinar um modelo de classificação (por exemplo, árvore de decisão ou regressão logística) utilizando o `scikit-learn`.
- Calcular métricas como precisão, recall, F1-score e ROC-AUC.
- Plotar a curva ROC e a matriz de confusão com `matplotlib` e discutir a interpretação de cada métrica.

18 Atividade 18 – Introdução ao Uso do Orange3 para Data Mining

Objetivo: Integrar uma ferramenta visual de análise de dados e comparar resultados com abordagens em código.

Dataset: Mall Customers Dataset (Kaggle)

Descrição:

- Importar o dataset no Orange3 e configurar um fluxo de trabalho visual para realizar análise exploratória, pré-processamento e clusterização.
- Comparar os resultados do clusteramento (por exemplo, K-Means) realizados no Orange3 com os obtidos via `scikit-learn`.
- Discutir vantagens e limitações do uso de ferramentas visuais versus a programação em Python.

19 Atividade 19 – Projeto Integrado: Mini Pipeline de Análise de Dados

Objetivo: Consolidar os conhecimentos integrando diversas etapas do processo de Data Science.

Dataset: House Prices Dataset (Kaggle)

Descrição:

- Importar os dados e realizar uma análise exploratória completa utilizando `pandas`, `numpy` e `matplotlib`.
- Executar o tratamento dos dados e aplicar técnicas de feature engineering.
- Construir um pipeline completo de machine learning com o `scikit-learn`, incluindo pré-processamento, treinamento e validação (utilizando cross validation).
- Comparar os resultados obtidos e discutir os desafios encontrados.

20 Atividade 20 – Desafio Final: Projeto Completo e Avançado de Data Science

Objetivo: Desenvolver um projeto final que integre todas as competências adquiridas.

Dataset: COVID-19 Dataset (Kaggle) ou outro dataset complexo e atual de sua escolha.

Descrição:

- **Importação e Limpeza:** Importar os dados com `pandas`, tratar valores ausentes, inconsistências e formatar datas. Utilizar `numpy` para normalização.
- **Análise Exploratória:** Realizar uma análise descritiva e explorar tendências com `pandas` e `numpy`. Gerar visualizações (histogramas, scatter plots, heatmaps) com `matplotlib`.
- **Engenharia de Features:** Desenvolver novas variáveis que possam potencializar a modelagem e documentar o processo.
- **Modelagem:** Implementar modelos supervisionados (regressão, classificação) e não supervisionados (clusterização) utilizando o `scikit-learn`. Dividir os dados em conjuntos de treinamento e teste e aplicar validação cruzada.
- **Comparação com Orange3:** Opcionalmente, reproduzir parte da análise utilizando o Orange3 para comparar os resultados.
- **Relatório Final:** Consolidar todos os resultados, insights e desafios em um relatório detalhado.