Contexto do desafio

Você trabalha como analista de dados em uma empresa de educação online e recebeu dados sobre alunos.

O objetivo é prever a nota final de um aluno com base em hábitos de estudo e características observáveis.

Etapa 1 – Exploração do Dataset (Excel + Python)

Objetivos

- Familiarização com dados
- Uso de Excel e Python juntos

Tarefas

- Abrir o dataset no Excel
- Identificar:
    - Quantidade de linhas e colunas
    - Tipos de variáveis (numéricas e categóricas)
- Criar no Excel:
    - Média, mediana e moda da nota final
    - Quartis (Q1, Q2, Q3)
- Exportar ou carregar o CSV no Python

Etapa 2 – Estatística Descritiva com Python + NumPy

Objetivos

- Aprender NumPy
- Reforçar estatística básica

Tarefas

- Calcular usando NumPy:
    - Média
    - Mediana
    - Desvio padrão
    - Quartis
    - Comparar resultados do Python vs Excel
- Interpretar os resultados obtidos.

Etapa 3 – Análise de Correlação

Objetivos

- Entender relação entre variáveis

Tarefas

- Calcular correlação entre:
    - Horas de estudo × Nota final
    - Aulas assistidas × Nota final

Etapa 4 – Teste de Hipótese (Introdução)

Hipótese exemplo

- Alunos que estudam mais de 10 horas por semana têm nota média maior.

Tarefas

- Separar alunos em dois grupos:
    - Até 10h
    - Mais de 10h

Obs.: Não é preciso realizar o teste de hipótese, apenas façam a escolha das amostras, levando em consideração o que foi ensinado na trilha e expliquem o porque escolheram dessa forma.

Etapa 5 – Regressão Linear (Machine Learning)

Objetivo

- Prever a nota final

Tarefas

- Implementar regressão linear:
    - Manualmente com NumPy
    - OU usando sklearn (linear_model - LinearRegression)
- Avaliar o modelo (erro médio)(MSE)

Obs.: Essa etapa é opcional, somente para quem tem interesse na área, abaixo segue um passo a passo do que deve ser feito e sugestões de como realizar o processo.

- Carreguem o dataset no Python e o convertam para pandas.
- Inspecionem o dataset de vocês, utilizando funções como:
    - Describe
    - Info

    - Columns
    - Dtypes
    - Value_Counts
- Dividam o dataset em dados de treino e dados de teste, podem utilizar a função train_test_split da biblioteca scikit-learn.
    - O import é de sklearn.model_selection.
    - Podem usar a proporção 80% pra dados de treino e 20% para dados de teste.
- Importem um modelo de regressão linear (recomendo caso tenham interesse na área de modelos de IA).
    - Se não quiserem se aprofundar, podem apenas importar o modelo LinearRegression, também da scikit-learn.
    - O import é de sklearn.linear_model
- Instanciem o modelo:
    - model = LinearRegression()
- Obs.: A maioria das coisas feitas no campo de modelagem é feita com classes, então se tiverem interesse e ainda não souberem, estudem Programação Orientada a Objetos principalmente os conceitos:
    - Herança de Classes
    - Polimorfismo

- Utilizem os dados de treino para treinar o modelo:
    - model.fit(x_train, y_train)
- Façam as predições:
    - y_pred = model.predict(x_test)
- Importem as métricas MSE (mean square error), MAE (mean absolute error) e R2 Score da biblioteca scikit-learn.
    - O import é de sklearn.metrics
    - O que deve ser importado é mean_squared_error, mean_absolute_error, r2_score
- Avaliem a performance do modelo:
    - mse = mean_squared_error(y_test, y_pred)
    - mae = mean_absolute_error(y_test, y_pred)
    - r2 = r2_score(y_test, y_pred)
- Salvem os resultados.
- Responda essas perguntas:
    - O que é o Mean Square Error, para que ele é utilizado, e como é calculado?
    - O que é o Mean Absolute Error, para que ele é utilizado, e como é calculado?
    - O que é o R2 Score, para que ele é utilizado, e como é calculado?
    - Como essas métricas avaliam o desempenho do modelo?

- O que é o x_train, y_train, x_test, y_test? Porque a função train_test_split realiza essa divisão?

OBSERVAÇÕES FINAIS:

O que deve ser entregue é o código utilizado para executar cada uma das tarefas e a interpretação solicitada em texto. No caso da Etapa 5, caso realizar, envie somente o código e as respostas das perguntas em texto.