Atividade Prática Final – Projeto Integrado em Data Science
Contexto: Vendas no E-commerce Brasileiro (Olist)

Objetivo
Desenvolver um projeto completo de Ciência de Dados utilizando microdados reais de vendas do e-commerce brasileiro (Olist), passando pelas etapas de pré-processamento, análise exploratória em formato de relatório, visualização em Power BI e aplicação de algoritmos de regressão e classificação.

Contexto do Projeto
O dataset da Olist reúne informações de pedidos realizados em marketplaces brasileiros, incluindo dados de clientes, produtos, vendedores, entregas, pagamentos e avaliações. O objetivo do projeto é analisar padrões de compra, identificar fatores associados a atrasos/cancelamentos e prever valores de vendas, aplicando técnicas de Ciência de Dados utilizadas no mercado.

Link do conjunto de dados
https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Estrutura do Trabalho
O projeto será realizado individualmente.

• Utilizar as seguintes chaves para integração das tabelas:
o order_id para unir pedidos, pagamentos, reviews e itens.
o customer_id para integrar clientes.
o product_id para integrar produtos.
o seller_id para integrar vendedores.

• Construir uma tabela fato unificada contendo, no mínimo:
o Dados do cliente (cidade, estado, tipo de cliente).
o Dados do produto (categoria).
o Dados do pedido (valor, frete, datas).
o Dados do pagamento (tipo, número de parcelas).
o Avaliação do cliente (review_score).

• Criar a variável PROBLEMA:
o 1 = Pedido com problema → atraso na entrega OU avaliação ≤ 2.
o 0 = Pedido normal → entregue no prazo e avaliação > 2.

• Tratar valores faltantes de forma justificada (remoção, imputação por média, mediana ou zero, dependendo do caso).

• Criar dois dataframes finais:
o df_classificacao: prever PROBLEMA.
o df_regressao: apenas pedidos sem problema, prever VALOR_DO_PEDIDO.

• Tratamento de variáveis categóricas:
o LabelEncoder para variáveis ordinais (ex.: rating).
o OneHotEncoder para variáveis nominais (ex.: forma de pagamento, categoria do produto, estado).

• Normalizar variáveis numéricas utilizando Min-Max Scaling:
o valor do pedido, frete, número de parcelas, tempo de entrega, etc.

Entregável: código com os dois dataframes finais prontos para modelagem.

---

2. Análise Exploratória dos Dados (EDA)

A EDA será entregue como um relatório no formato de perguntas e respostas.
Cada pergunta deve conter:
• Um gráfico (histograma, boxplot, barras, dispersão ou heatmap).
• Uma explicação curta (2 a 3 frases) interpretando o gráfico.

EDA – Parte 1 (Distribuições Básicas – mínimo 6 gráficos)

1. Qual a proporção de pedidos com problema vs. pedidos normais?
2. Como se distribui o valor dos pedidos?
3. Como se distribui o tempo de entrega?
4. Quais são as categorias de produtos mais vendidas?
5. Quais são os estados com maior número de pedidos?
6. Quais são as formas de pagamento mais utilizadas?

EDA – Parte 2 (Análises Avançadas – mínimo 8 gráficos)

1. Pedidos com problema variam por estado?
2. O valor do frete influencia a chance de problema?
3. O número de parcelas influencia a chance de atraso?
4. Existe relação entre valor do pedido e avaliação do cliente?
5. Clientes de determinados estados gastam mais em média?
6. Há diferença no valor médio por categoria de produto?
7. Quais variáveis têm maior correlação com o valor do pedido (heatmap)?
8. Quais variáveis têm maior correlação com PROBLEMA (heatmap)?

Entregável: relatório em Word com prints dos gráficos e explicações.

---

3. Power BI – Modelo Estrela e Dashboard

Modelo Estrela Sugerido:
• dim_cliente: cidade, estado, tipo de cliente.
• dim_produto: categoria do produto.
• dim_vendedor: cidade, estado do vendedor.
• dim_tempo: data do pedido, mês, ano, dia da semana.
• dim_pagamento: tipo de pagamento, número de parcelas.
• fact_pedidos: valor do pedido, frete, tempo de entrega, avaliação, PROBLEMA.

Dashboard obrigatório (mínimo 3 páginas):

1. Visão Geral
   - o Faturamento total.
   - o Ticket médio.
   - o Percentual de pedidos com problema.
   - o Quantidade de pedidos por mês.
2. Perfil de Consumo
   - o Categoria × faturamento.
   - o Forma de pagamento × valor médio.
   - o Número de parcelas × problemas.
3. Mapa Geográfico
   - o Faturamento por estado.
   - o Percentual de pedidos com problema por estado.

Publicação:
• Publicar o dashboard no Power BI Service.
• Entregar o link compartilhável ou apresentação local do arquivo .pbix.

Entregável: arquivo .pbix com no mínimo 3 páginas de dashboard.

---

4. Modelagem de Machine Learning

Classificação (obrigatória):
• Prever PROBLEMA (0 = normal, 1 = com problema).
• Algoritmos obrigatórios:
o Rede Neural (MLP).
o SVM.
o Um terceiro à escolha (Random Forest, KNN, etc.).
• Avaliação com:
o Matriz de confusão.
o Acurácia.

o F1-Score.
o AUC-ROC (opcional).
• Discussão obrigatória sobre desbalanceamento das classes.

Regressão (obrigatória):
• Prever VALOR_DO_PEDIDO (somente pedidos normais).
• Algoritmos obrigatórios:
o Rede Neural (MLP).
o SVM Regressor.
o Um terceiro à escolha (Regressão Linear, Random Forest Regressor, etc.).
• Avaliação com:
o RMSE.
o MAE.
o R^2.

Opcional (extra):
• Criar a variável PERFIL_CLIENTE (Alto, Médio, Baixo gasto).
• Prever o PERFIL_CLIENTE com classificação multiclasses.

Entregável: código com todos os experimentos + tabela comparativa de métricas.

---

5. GitHub

O gerenciamento do projeto deve ocorrer desde a primeira semana.

Estrutura mínima do repositório:
• data/ → dados tratados
• codigos/ → pré-processamento, EDA, modelagem
• bi/ → arquivo Power BI (.pbix)
• README.md → documentação do projeto

O README deve conter:
• Objetivo do projeto.
• Descrição do dataset.
• Etapas realizadas.
• Instruções básicas de reprodução.

Requisito mínimo: commits semanais mostrando a evolução do projeto.

Desafios Opcionais (Crédito Extra)

• Banco de Dados (MySQL ou PostgreSQL):
o Carregar os dados tratados no banco.

o Consultar via Python.
o Conectar ao Power BI.

• Interface:
o Criar aplicação em Streamlit ou Flask para consulta de pedidos, previsão de problema ou valor da venda.

Cronograma – 6 Semanas

Semana 1 – Pré-processamento + Estruturação do GitHub
Entregável: integração das tabelas, criação da variável PROBLEMA, separação dos dataframes, encoding, normalização e repositório criado.

Semana 2 – EDA Parte 1
Entregável: relatório com gráficos de distribuição e interpretações.

Semana 3 – EDA Parte 2
Entregável: análises avançadas, correlações e comparações entre grupos.

Semana 4 – Power BI
Entregável: arquivo .pbix com modelo estrela e no mínimo 3 páginas de dashboards.

Semana 5 – Machine Learning
Entregável: códigos de classificação e regressão com métricas comparativas.

Semana 6 – Revisão e Entrega Final
Entregável: repositório completo, dashboards finalizados e documentação pronta.