Projeto 2 - Machine Learning

Brazilian E-Commerce Public Dataset by Olist

1. Introdução

Este projeto tem como finalidade aplicar os conhecimentos adquiridos na disciplina de Ciência de Dados em um conjunto de dados real, por meio da construção de **dois problemas de Machine Learning: um de aprendizado supervisionado e outro de aprendizado não supervisionado.** O dataset utilizado será o Brazilian E-Commerce Public Dataset by Olist, disponível no Kaggle.

Cada equipe deverá propor, implementar e justificar os dois problemas, explorando as variáveis disponíveis, tratando os dados de forma adequada, selecionando as features e avaliando os modelos treinados.

2. Dados Utilizados

- Fonte: Olist
- Conjunto de dados com aproximadamente 100 mil pedidos realizados entre 2016 e 2018.
- Abrange informações sobre pedidos, pagamentos, entregas, avaliações de clientes, produtos, vendedores e localização.

🔗 Acesse aqui: Brazilian E-Commerce Public Dataset by Olist

Para auxiliar no início do projeto, foi disponibilizado um **notebook base no Google Colab** contendo:

- Download e extração do dataset
- Leitura dos arquivos .csv
- Análise exploratória básica de todas as tabelas
- Visualizações com matplotlib e seaborn

Este material serve como ponto de partida para que as equipes possam entender a estrutura dos dados e iniciar a construção dos seus modelos.

🔗 Acesse aqui: Notebook Inicial no Colab

3. Estrutura do Projeto

As equipes devem desenvolver o projeto com base nas seguintes etapas:

1. **Exploração Inicial dos Dados**
    - Análise exploratória com visualizações e estatísticas descritivas.
    - Limpeza de dados e tratamento de valores ausentes e inconsistências.
2. **Problemas de Machine Learning**
    - Definição de dois problemas: Um de aprendizado supervisionado (classificação ou regressão). Um de aprendizado não supervisionado (clustering ou detecção de anomalias).
    - Definição clara do objetivo de cada modelo.
    - Justificativa para a escolha das features e, no caso supervisionado, do target.
3. **Modelagem**
    - Utilização de pelo menos 3 algoritmos para cada abordagem.
    - Obrigatório usar como baseline: Regressão Linear para problemas de regressão e Regressão Logística para problemas de classificação.
    - Aplicação de técnicas de redução de dimensionalidade em algum problema para pré-modelagem ou para visualização (ex: PCA, t-SNE, UMAP).
4. **Avaliação dos modelos**
    - Para regressão: mínimo de 3 métricas, incluindo R^2.
    - Para classificação: utilizar acurácia, precisão, recall, F1-score e matriz de confusão.
    - Apresentar a curva de aprendizado e análise de desempenho dos modelos.
    - Explicação sobre overfitting e underfitting, com base nos resultados.

4. Entrega

Cada equipe deverá realizar duas entregas obrigatórias via AVA, até a data estipulada:

Vídeo Explicativo

- O vídeo deve ter no máximo **10 minutos** de duração.
- O link de acesso ao vídeo deverá ser enviado no campo de texto da atividade no AVA. O vídeo pode estar hospedado no YouTube (não listado) ou no Google Drive (com link compartilhável).
- No vídeo, a equipe deve apresentar:
    - O problema de classificação/regressão e o de clusterização.
    - A justificativa para escolha das variáveis (features e target).
    - As decisões de modelagem e interpretação dos resultados.
    - Explicações sobre overfitting/underfitting, avaliação dos modelos e visualizações relevantes.

Penalização por tempo excedente: Vídeos com mais de 10 minutos sofrerão desconto de 0,5 ponto por minuto extra, arredondado para cima.

Notebook

- O **arquivo do notebook (.ipynb)** deverá ser enviado diretamente na tarefa do AVA (upload do arquivo).
- O notebook deve conter:
    - Todo o código executado no projeto
    - Comentários explicativos
    - Visualizações e análise exploratória
    - Modelagem supervisionada e não supervisionada
    - Avaliação dos modelos com as métricas obrigatórias
    - Aplicação da técnica de redução de dimensionalidade

5. Avaliação (0 a 10 pontos)

Critério	Peso
Definição e justificativa dos problemas (supervisionado e não sup.)	1,5
Escolha e justificativa das features e do target	1,0
Qualidade do pré-processamento e limpeza dos dados	1,0
Aplicação de pelo menos 3 algoritmos por abordagem	1,5
Aplicação de técnica de redução de dimensionalidade	1,0
Avaliação dos modelos com métricas obrigatórias e curva de aprendizado	2,0
Análise crítica (overfitting, underfitting, desempenho comparado)	1,0
Clareza e objetividade na apresentação em vídeo	1,0
Total	**10**

6. Sugestões de Problemas

A formulação dos problemas deve, preferencialmente, considerar **situações reais que possam ser úteis para a empresa Olist ou para seus fornecedores ou para seus clientes**, como estratégias para melhorar o atendimento ao cliente, otimizar a logística ou compreender melhor o comportamento dos consumidores. Isso estimula uma abordagem mais **prática, aplicada e com valor de negócio**.

As sugestões a seguir têm caráter **inspirador**, cabendo às equipes a formulação criativa e justificável dos seus próprios problemas:

Aprendizado Supervisionado:

- Prever a nota de avaliação (*review_score*) com base em dados do pedido, produto e pagamento.
- Prever se um pedido será entregue no prazo com base em suas características logísticas.
- Estimar o valor total pago em um pedido.
- Classificar um pedido como potencialmente problemático ou não com base nas características e avaliações anteriores.

Aprendizado Não Supervisionado:

- Agrupar clientes por comportamento de compra (valores, categorias, frequência).
- Clusterizar produtos por características como categoria, preço e dimensões.
- Identificar regiões com comportamentos de compra semelhantes com base na geolocalização.

6. Observações Finais

- A turma deve se dividir em 6 equipes, as quais devem conter no máximo 5 integrantes.
- O projeto é individual por equipe. Trabalhos idênticos ou plagiados serão desconsiderados.
- A avaliação considerará a qualidade da análise, a clareza da comunicação e a profundidade da interpretação dos resultados.