Trabalho Final: Projeto Completo de Ciência de Dados

Modalidade: Avaliação em dupla ou trio. Entrega via MS Teams.

Data de Apresentação: 01 de dezembro de 2025, até as 23h59min. Atraso será punido com 1,0 ponto de desconto a cada hora.

Visão Geral

O objetivo deste trabalho é consolidar seu aprendizado aplicando o ciclo completo de um projeto de dados, desde a concepção do problema até a "entrega" de um modelo funcional.

Para isso, vocês **retomarão o projeto original** que iniciaram no começo do semestre (referente à avaliação N1), onde cada grupo definiu um domínio de problema e escolheu um dataset. O foco desta avaliação será demonstrar a construção de um modelo preditivo (de classificação ou regressão) para aquele problema, justificando cada etapa do processo.

Compare os trabalhos que vocês fizeram para a N1 com o trabalho guiado que vocês fizeram para a N2. É possível que vários grupos entre vocês perceberão que a pergunta de negócio original não é adequada para uma atividade de Ciência de Dados e, portanto, deverá ser ajustada. Se vocês acharem necessário, podem até mudar completamente o tema original.

O trabalho deverá ser entregue até o dia 01/12/2025 as 23h59min. A avaliação será realizada pela análise do trabalho (explicações + código).

O trabalho final deve ser entregue como um link para um repositório no GitHub. O repositório deve conter um `README.md` detalhado que sirva como o relatório do projeto, além de todos os notebooks Jupyter, scripts e arquivos necessários para a sua execução. O repositório deve ser apresentado com a seguinte estrutura:

- `README.md` : O "rosto" do projeto. Deve explicar o problema, a estrutura do projeto e como rodá-lo.
- `/notebooks` : Pasta com os Jupyter Notebooks de exploração e modelagem.
- `/data` : Pasta com o dataset.
- `/scripts` : Pasta para scripts de deploy ou funções auxiliares.
- `requirements.txt` : O arquivo de dependências.
- `modelo_final.pkl` : O arquivo do modelo salvo.

Estrutura do Trabalho e Critérios de Avaliação

Seu projeto e apresentação devem ser estruturados em quatro partes principais, detalhadas abaixo.

Parte 1: A Fundação do Projeto - O Problema de Negócio (1,0 ponto)

O que fazer: Inicie sua apresentação contextualizando o projeto. Conte a história que motivou seu trabalho.

- **1.1. Apresente o Domínio do Problema:** Descreva o cenário e o contexto do problema que vocês escolheram. Por que ele é relevante? (Ex: "Nosso projeto se insere no contexto do mercado imobiliário, onde a precificação de imóveis é um desafio complexo...").
- **1.2. Apresente a Pergunta de Negócio:** Declare de forma clara e específica a pergunta que guiou toda a sua análise. (Ex: "A pergunta central que buscamos responder foi: 'Quais características de um imóvel (como área, número de quartos e localização) têm o maior impacto em seu preço de venda?'").
- **1.3. Defina o Objetivo do Modelo:** Explique o que o modelo preditivo (ou de classificação) que vocês construíram se propõe a fazer. (Ex: "O objetivo foi construir um modelo de regressão capaz de estimar o preço de um imóvel com base em suas características, fornecendo uma ferramenta de apoio para corretores e proprietários.").

Parte 2: A Jornada dos Dados - Pipeline e Arquitetura (1,0 ponto)

O que fazer: Descreva o caminho completo que os dados percorreram, desde sua fonte original até estarem prontos para a modelagem. O uso de um fluxograma ou diagrama visual é fortemente recomendado.

- **2.1. Origem e Repositório de Dados:**
  - Identifique a fonte original dos seus dados (ex: API do Twitter, dataset do Kaggle, portal de dados abertos do governo).
  - Descreva a arquitetura de armazenamento que vocês definiram para esses dados (ex: Data Lake para dados brutos, Data Warehouse para dados tratados, ou um Data Lakehouse). Justifique por que essa arquitetura foi escolhida para o seu projeto.
- **2.2. Apresente o Pipeline de Dados:** Explique, passo a passo, o fluxo de processamento.
  - **Ingestão:** Como os dados foram coletados e armazenados inicialmente no seu repositório?
  - **Limpeza e Transformação (ETL/ELT):** Quais foram as principais etapas de limpeza e preparação que vocês realizaram, e por quê? (Ex: tratamento de valores ausentes, padronização de formatos, remoção de duplicatas).

- **Análise Exploratória (EDA):** Explique como a EDA (realizada em etapas anteriores) ajudou a entender os dados e a selecionar as variáveis para o modelo.
- **Preparação para Modelagem:** Detalhe a etapa final de preparação, incluindo a seleção de *features*, a transformação de variáveis categóricas em numéricas (`One-Hot Encoding` / `get_dummies`) e a divisão dos dados em conjuntos de treino e teste.

Parte 3: O Coração do Projeto - Modelagem e Avaliação Comparativa (6,0 pontos)

O que fazer: Esta é a parte central do trabalho. Demonstre sua capacidade de treinar, comparar e avaliar criticamente diferentes modelos de Machine Learning.

- **3.1. Treinamento de Três Modelos:**
  - Escolha e treine **pelo menos três algoritmos diferentes** que sejam apropriados para o seu problema (classificação ou regressão).
  - *Sugestões para Classificação:* Árvore de Decisão, Regressão Logística, Random Forest, KNN, SVM.
  - *Sugestões para Regressão:* Regressão Linear, Ridge, Lasso, Árvore de Decisão para Regressão.
- **3.2. Avaliação com Três Métricas:**
  - Escolha **pelo menos três métricas de desempenho** para avaliar e comparar seus modelos.
  - *Sugestões para Classificação:* Acurácia, Precisão, Recall, F1-Score.
  - *Sugestões para Regressão:* RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), R^2 (R-squared).
  - **Explique cada métrica escolhida:** Antes de apresentar os resultados, dedique um momento para explicar o que cada métrica significa e por que ela é relevante para o seu problema. (Ex: "Para nosso problema de fraude, o Recall é crucial, pois mede a capacidade do modelo de encontrar todas as fraudes reais, mesmo que isso gere alguns alarmes falsos...").
- **3.3. Análise Comparativa dos Resultados:**
  - Apresente os resultados de forma clara, preferencialmente em uma **tabela comparativa**.
  - Discuta os resultados: Qual modelo obteve a melhor performance geral? Houve algum modelo que se destacou em uma métrica específica? Com base na sua análise e no objetivo do seu problema de negócio, **qual modelo você escolheria e por quê?**

Parte 4: Tornando o Modelo Útil - Deploy (2,0 pontos)

O que fazer: Demonstre que seu modelo não é apenas um exercício acadêmico, mas que pode ser reutilizado para fazer novas previsões.

- **4.1. Salvando o Modelo Treinado:**
  - Após escolher o seu melhor modelo na Parte 3, mostre o código utilizado para salvá-lo em um arquivo usando uma biblioteca como `pickle` ou `joblib`.
  - **Exemplo:** `joblib.dump(meu_melhor_modelo, 'modelo_final.pkl')`
- **4.2. Carregando e Utilizando o Modelo:**
  - Em um novo script ou célula de notebook, demonstre como carregar este arquivo de modelo salvo.
  - Crie um **exemplo de um novo dado** (uma nova entrada que o modelo nunca viu) e use o modelo carregado para fazer uma previsão sobre ele.
  - Apresente o resultado da previsão e explique o que ele significa. (Ex: "Carregamos nosso modelo de preços e, para um novo imóvel com estas características, ele previu um preço de R$ X.").