Requisitos para o Desenvolvimento do MVP

1. Escolha bases de dados que não tenham sido utilizadas em aula. Sugere-se usar bases de dados disponibilizada em algum dos repositórios a seguir:
   - UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets.php
   - Kaggle: https://www.kaggle.com/datasets
   - Google Datasets: https://datasetsearch.research.google.com/
   - Hugging Face: https://huggingface.co/datasets

Aproveite os filtros que oferecem os repositórios para encontrar com mais facilidade os datasets da sua preferência. Caso você prefira usar um dataset que reflita outro problema, será muito bem-vindo.

2. Você deverá escolher uma das duas alternativas abaixo:
Treinar modelos de machine learning para um problema de **classificação** ou **regressão**: Você deverá treinar modelos clássicos de machine learning, iniciando na carga e preparação dos dados, incluindo a separação entre treino e teste, a seleção de atributos, transformação de dados, modelagem, otimização de parâmetros, até a avaliação e comparação de resultados dos modelos treinados.

3. Produza um notebook no Google Colab, considerando o item 1 e o item 2, com as características a seguir:
   - O notebook servirá como relatório, descrevendo textualmente (utilizando as células de texto) o contexto do problema e as operações com os dados (veja o checklist sugerido abaixo).
   - Utilize a linguagem Python e bibliotecas que considere apropriadas para abordar o problema.
   - Crie o notebook seguindo as boas práticas de codificação.
   - Veja aqui um modelo de um projeto de Ciência de Dados no Google Colab.

Checklist Sugerido

Definição do problema

Objetivo: entender e descrever claramente o problema que está sendo resolvido.

- Qual é a descrição do problema?
- Este é um problema de classificação ou regressão?
- Que premissas ou hipóteses você tem sobre o problema?
- Que restrições ou condições foram impostas para selecionar os dados?
- Defina cada um dos atributos do *dataset*.

Análise exploratória dos dados

Objetivo: entender a informação disponível.

Estatísticas descritivas:

- Quantos atributos e instâncias existem?
- Quais são os tipos de dados dos atributos?
- Verifique as primeiras linhas do *dataset*. Algo chama a atenção?
- Há valores faltantes, discrepantes ou inconsistentes?
- Faça um resumo estatístico dos atributos com valor numérico (mínimo, máximo, mediana, moda, média, desvio-padrão e número de valores ausentes). O que você percebe?

Visualizações:

- Verifique a distribuição de cada atributo. O que você percebe? Dica: esta etapa pode dar ideias sobre a necessidade de transformações na etapa de preparação de dados (por exemplo, converter atributos de um tipo para outro, realizar operações de discretização, normalização, padronização, etc.).
- Se for um problema de classificação, verifique a distribuição de frequência das classes. O que você percebe? Dica: esta etapa pode indicar a possível necessidade futura de balanceamento de classes.
- Analise os atributos individualmente ou de forma combinada, usando os gráficos mais apropriados.

Preparação dos dados

Objetivo: realizar operações de limpeza, tratamento e preparação dos dados.

- Verifique quais operações de pré-processamento podem ser interessantes para o seu problema e salve versões diferentes do seu *dataset* (por exemplo, normalização, padronização e codificação de variáveis).
- Trate (removendo ou substituindo) os valores faltantes (se existentes).
- Separe o dataset entre treino e teste (e validação, se aplicável).
- Explique, passo a passo, as operações realizadas, justificando cada uma delas.

Modelagem e avaliação dos resultados

Objetivo: construir, treinar e avaliar os modelos para resolver o problema em questão.

- Selecione os algoritmos mais indicados para o problema e os dataset escolhidos, justificando as suas escolhas.
- Há algum ajuste inicial para os hiperparâmetros?
- O modelo foi devidamente treinado? Foi observado problema de underfitting?
- É possível otimizar os hiperparâmetros de algum dos modelos? Se sim, faça-o, justificando todas as escolhas.
- Selecione as métricas de avaliação condizentes com o problema, justificando.
- Treine o modelo escolhido com toda a base de treino, e teste-o com a base de teste.
- Os resultados fazem sentido?
- Foi observado algum problema de overfitting?
- Compare os resultados de diferentes modelos.
- Descreva a melhor solução encontrada, justificando.

Conclusão

Objetivo: resumir os principais achados, pontos de atenção e conclusões sobre esse projeto.

Requisitos e composição da nota

Execução sem erros (10%)
O código do notebook deve ser executável do início ao fim sem apresentar erros. Isso significa que todas as bibliotecas, funções e variáveis devem estar corretamente definidas e não deve haver falhas de execução durante o processo.

Análise exploratória e preparação dos dados (30%)
Aqui, é esperado que você faça uma análise inicial dos dados, identificando padrões, outliers, ou problemas de qualidade nos dados (como valores ausentes). Isso inclui a visualização e descrição de características dos dados, estatísticas descritivas e a exploração de correlações. Além disso, você precisa preparar os dados para modelagem, o que pode incluir a normalização, transformação de variáveis categóricas e tratamento de dados faltantes.

Modelagem e avaliação dos resultados (30%)
Após a preparação dos dados, você deve treinar modelos preditivos. A escolha do modelo adequado depende da natureza do problema. Durante essa fase, você deve também validar e avaliar o desempenho do modelo usando métricas apropriadas (ex.: acurácia, erro quadrático médio, AUC, etc.). Neste bloco, é necessário interpretar os resultados do modelo, levantar hipóteses sobre a performance, e apontar possíveis melhorias ou ajustes.

Qualidade e organização do trabalho (20%)
Esta parte avalia a clareza do código e da apresentação, a qualidade da documentação e a organização geral do notebook. Tudo deve estar bem documentado, com explicações claras dos passos tomados e uma apresentação visual limpa e fácil de seguir.

Conclusão (10%)
É essencial que a conclusão apresente um resumo claro e conciso dos principais achados de cada etapa do projeto. Isso inclui os insights da análise exploratória, o desempenho dos modelos, e os desafios enfrentados durante a preparação dos dados e a modelagem.

Sobre a entrega

Para a entrega:

- Você deverá disponibilizar UM ÚNICO notebook com o código em Python em um **repositório público** do GitHub.

- A utilização do dataset dentro do notebook deve ser feita através da URL do seu repositório do GitHub.

- O link do notebook no GitHub deve ser informado na tarefa de entrega no Google Classroom.