Projeto Final

1. Visão Geral e Objetivo

O objetivo deste projeto é simular um desafio real de ciência de dados dentro de uma empresa. Cada grupo deverá selecionar um conjunto de dados, definir um problema de negócio relevante e conduzir todo o fluxo de trabalho para desenvolver, avaliar e justificar a escolha de um modelo de Machine Learning. O foco está não apenas na precisão técnica, mas na **capacidade de comunicação, justificativa de decisões e alinhamento com os objetivos de negócio**.

2. Requisitos Detalhados do Projeto

2.1. Seleção do Conjunto de Dados (Dataset)

- Fonte: **Kaggle**.
- Tamanho Mínimo: O dataset deve conter pelo menos 10.000 linhas (10k) e um número suficiente de colunas para permitir uma análise significativa e a criação de *features*.
- Justificativa: O grupo deve justificar brevemente por que o dataset escolhido é interessante e adequado para um problema de ML.

2.2. Definição do Problema de Negócio

- O grupo deve formular claramente uma pergunta de negócio que será respondida pelo modelo.
- Exemplos:
    - "Prever a rotatividade de clientes (*churn*) para priorizar ações de retenção."
    - "Classificar transações financeiras como fraudulentas ou legítimas."
    - "Estimar o preço de vendas de imóveis com base em suas características."
- Este problema será o fio condutor de todas as decisões tomadas ao longo do projeto.

2.3. Pré-processamento e Engenharia de *Features*

- Realizar a limpeza dos dados (tratamento de valores missing, outliers, etc.).

- **Criação de *Features*:** O grupo deve criar **pelo menos 3 novas *features*** (variáveis) que não existiam originalmente no dataset. Estas devem ser justificadas com base no domínio do problema.
    - **Exemplos:** Criar uma *feature* "Idade do Imóvel" a partir do "Ano de Construção"; agregar transações anteriores para criar "Média de Gasto dos Últimos 3 Meses"; extrair o dia da semana de um *timestamp*.

2.4. Modelagem e Avaliação

- **Seleção de Modelos:** Avaliar **pelo menos 5 algoritmos de Machine Learning** diferentes. Sugestões:
    - `LogisticRegression (Regressão Logística) / LinearRegression (Regressão Linear)`
    - `KNeighborsClassifier (K-Vizinhos) / KNeighborsRegressor`
    - `DecisionTreeClassifier (Árvore de Decisão) / DecisionTreeRegressor`
    - `RandomForestClassifier (Floresta Aleatória) / RandomForestRegressor`
    - `GradientBoostingClassifier (Gradient Boosting) / GradientBoostingRegressor`
    - `SVC (Máquinas de Vetor de Suporte)`
- **Métrica Principal:**
    - O grupo deve **escolher e justificar** a métrica principal de avaliação (ex: Acurácia, Precisão, *Recall*, F1-Score, RMSE, R^2).
    - A justificativa deve estar diretamente ligada ao **problema de negócio**. Por exemplo: "Para detecção de fraude, otimizamos o *Recall* porque é mais custoso deixar uma fraude passar do que investigar um falso positivo."
- **Otimização de Hiperparâmetros:** Realizar a busca pelos melhores hiperparâmetros para **todos os modelos** usando métodos como *GridSearchCV* ou *RandomizedSearchCV*.

2.5. Análise Comparativa e Teorização

- Apresentar uma tabela/resumo comparativo do desempenho de todos os modelos nas métricas escolhidas.
- **Escolha do Modelo Final:** Explicar qual modelo foi escolhido como final, baseando-se **nas métricas e na pergunta de negócio**.

- **Teorização do Desempenho:** O grupo deve **criar e apresentar teorias** para explicar **por que um modelo (o vencedor) performou melhor do que outro (um dos perdedores)**. Isso demonstra compreensão do funcionamento dos algoritmos.
    - **Exemplo:** "A Floresta Aleatória performou melhor que a Árvore de Decisão simples porque o *ensemble* (conjunto) de árvores reduz o overfitting, que era um problema claro no modelo de árvore única, que tinha acurácia altíssima no treino mas baixa no teste."

2.6. Conclusão e Próximos Passos

- Responder às perguntas finais:
    1. **"Meu modelo resolve adequadamente o problema proposto?"** (Com base nas métricas e no contexto).
    2. **"Meu modelo pode ser colocado em produção?"** Discutir viabilidade, custos computacionais, manutenção e possíveis vieses.

3. Produtos Entregáveis

3.1. Apresentação de Slides (Formato Livre - 15 a 20 minutos)
A apresentação para a turma e professor deve ser clara e conter, no mínimo, os seguintes tópicos:

1. **Introdução:** Apresentação do grupo e do problema de negócio.
2. **O Dataset:** Breve descrição dos dados (fonte, tamanho, variáveis principais).
3. **Análise Exploratória & Engenharia de *Features*:** Insights iniciais e explicação das 3 novas *features* criadas.
4. **Metodologia:** Explicação **didática** de como funcionam pelo menos 2 dos modelos usados (escolher os mais relevantes). Justificativa para a métrica principal escolhida.
5. **Resultados e Análise:** Tabela comparativa dos 5 modelos, teoria sobre o desempenho e justificativa para a escolha do modelo final.
6. **Conclusão:** Resposta às perguntas sobre adequação do modelo e produção. Lições aprendidas.

3.2. Código Fonte

- Um *script* ou *notebook* (Jupyter/Colab) bem documentado, contendo todo o fluxo de trabalho, desde a carga de dados até a geração do modelo final.

3.3. Relatório Técnico Resumido

- Um documento de 1-2 páginas resumindo o projeto, focado na tomada de decisão e nos resultados, simulando um documento para a gestão.

4. Critérios de Avaliação

- **Clareza do Problema de Negócio (10%)**
- **Qualidade do Pré-processamento e Engenharia de *Features* (20%)**
- **Rigor na Modelagem e Avaliação (25%)** (Diversidade de modelos, otimização, métricas)
- **Análise Crítica e Profundidade de Explicação (25%)** (Justificativa de métricas, teorização do desempenho, escolha do modelo)
- **Qualidade da Apresentação e Comunicação (20%)** (Organização, clareza dos slides, capacidade de responder perguntas)

Instruções Finais:

Este projeto vai além da codificação; é sobre contar uma história com dados. A comunicação das suas escolhas é tão importante quanto a implementação técnica. Dividam as tarefas de forma equilibrada (análise exploratória, engenharia de *features*, codificação de modelos, criação de slides) e assegurem-se de que todos os membros compreendem o projeto como um todo.

Para dúvidas, entrem em contato com o instrutor do módulo.

Bons estudos e bom trabalho