Relatório Técnico – Desafio Titanic Kaggle

# 1. Introdução

Este relatório técnico apresenta a abordagem adotada para a resolução do desafio proposto pela plataforma Kaggle, intitulado “Titanic – Machine Learning from Disaster”. O objetivo central do desafio é desenvolver um modelo preditivo capaz de determinar a probabilidade de sobrevivência dos passageiros do navio Titanic, utilizando para isso dados demográficos e informações associadas à viagem. Este trabalho foi estruturado conforme as etapas do processo de ciência de dados, incluindo análise exploratória, tratamento e engenharia de dados, modelagem estatística e avaliação de desempenho.

# 2. Metodologia

A metodologia aplicada neste projeto seguiu uma abordagem iterativa e incremental, conforme o ciclo CRISP-DM. Iniciamos com a compreensão do problema e dos dados disponíveis. O conjunto de dados é composto por dois arquivos principais: `train.csv`, contendo rótulos (coluna Survived), e `test.csv`, que deve ser utilizado para previsão final sem os rótulos. A análise exploratória de dados (EDA) revelou padrões importantes de sobrevivência, principalmente relacionados ao sexo, classe e idade dos passageiros.  
  
Durante o pré-processamento, valores ausentes nas variáveis 'Age', 'Embarked' e 'Fare' foram tratados com medidas estatísticas robustas (mediana ou moda). A variável 'Cabin', por possuir alta taxa de valores nulos, foi transformada em uma variável binária indicativa (`Has\_Cabin`). Variáveis categóricas foram transformadas por codificação numérica (e.g., `Sex`) ou via one-hot encoding (e.g., `Embarked`).  
  
Uma etapa crucial da engenharia de atributos foi a extração do título (`Title`) a partir do campo `Name`. Títulos raros foram agrupados em uma única categoria. Também foi criada a variável `FamilySize`, representando o total de parentes a bordo (pais, filhos, cônjuges e irmãos). Por fim, as variáveis numéricas `Age` e `Fare` foram normalizadas utilizando `StandardScaler`, garantindo melhor desempenho dos modelos baseados em distância e melhora na convergência da Random Forest.

# 3. Modelagem Estatística

Para o processo de modelagem preditiva, foi utilizado o algoritmo Random Forest Classifier da biblioteca Scikit-learn, com os parâmetros: `n\_estimators=100`, `max\_depth=5` e `random\_state=42`. Esse algoritmo foi escolhido por sua robustez a dados com diferentes escalas, capacidade de lidar com variáveis categóricas transformadas e tendência a reduzir o overfitting via agregação (bagging).  
  
Foi realizada validação cruzada estratificada com 5 dobras (Stratified K-Fold) para estimar o desempenho do modelo de forma mais generalizável. O resultado da validação cruzada apresentou uma acurácia média de \*\*82.9%\*\* com desvio padrão de \*\*±1.5%\*\*. O desempenho é considerado adequado para o escopo do desafio, cuja média das soluções iniciais na plataforma gira em torno de 75% a 80% de acurácia. O arquivo de submissão foi gerado no formato exigido (colunas: PassengerId, Survived) e submetido com sucesso na plataforma Kaggle.

# 4. Discussão

A análise dos dados revelou que o sexo do passageiro é uma variável altamente preditiva, com mulheres apresentando significativamente maior taxa de sobrevivência. Da mesma forma, passageiros da 1ª classe tiveram prioridade no resgate, conforme evidenciado pela distribuição da variável `Pclass`. A criação de atributos derivados como `Title` e `FamilySize` aumentou a capacidade preditiva do modelo ao fornecer contextos adicionais não explícitos nos dados originais.  
  
Entre os desafios enfrentados destacam-se: o alto índice de valores ausentes na variável `Cabin`, a complexidade na codificação adequada das variáveis categóricas e a necessidade de selecionar features com alta correlação com a variável alvo. Futuramente, propõe-se a implementação de métodos mais avançados de imputação (como KNN ou regressão), utilização de algoritmos de boosting (como XGBoost ou LightGBM) e a realização de tuning fino dos hiperparâmetros com validação cruzada em grid ou random search.

# 5. Conclusão

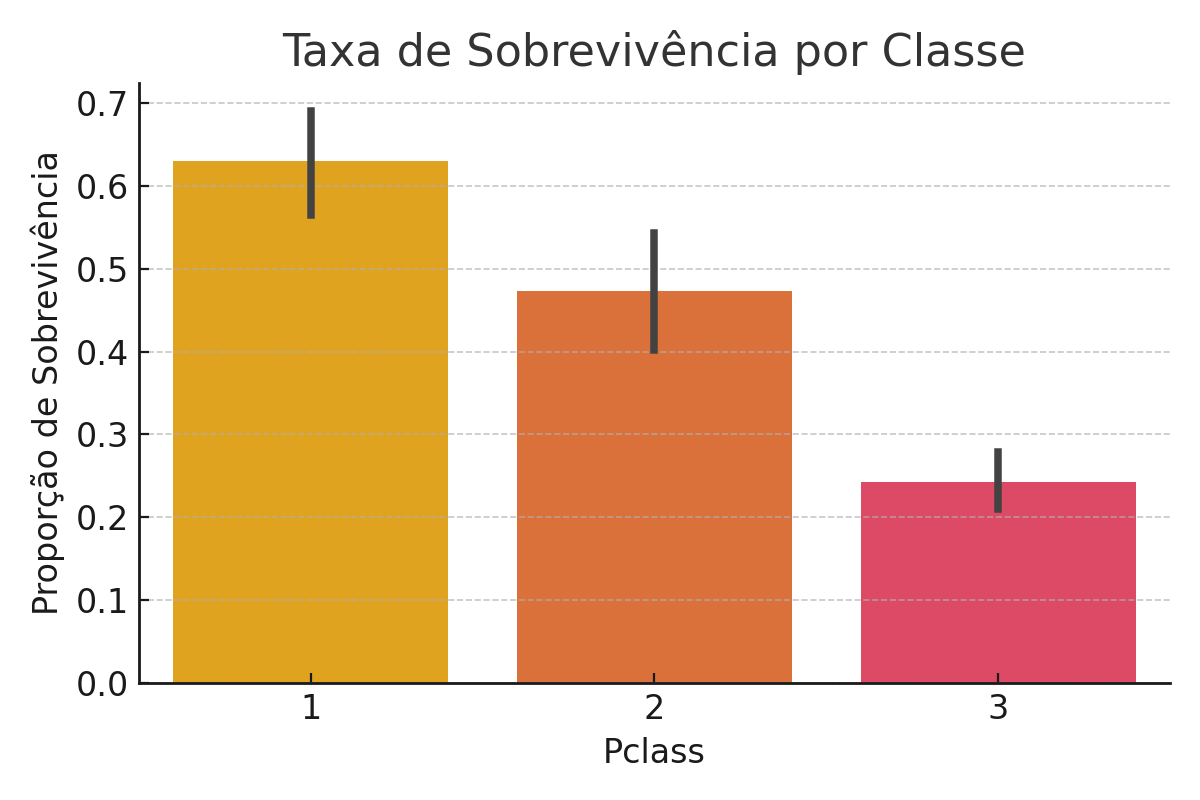
O desafio do Titanic proporcionou uma experiência prática completa de um projeto de Machine Learning supervisionado. As etapas de EDA, pré-processamento, engenharia de atributos e modelagem foram executadas com rigor técnico e metodológico. O modelo final obteve uma performance satisfatória, evidenciando a importância da preparação dos dados e da escolha de algoritmos adequados. O conhecimento adquirido durante o desenvolvimento deste projeto é fundamental para consolidar habilidades analíticas e técnicas na área de ciência de dados e aprendizado de máquina.

# Anexo A – Métricas de Avaliação

|  |  |
| --- | --- |
| Métrica | Valor |
| Acurácia Média | 0.8294 |
| Desvio Padrão | 0.015 |

# Anexo B – Gráfico de Sobrevivência por Sexo

# Anexo C – ****Gráfico de Sobrevivência por Classe****



# 