Relatório Técnico — Desafio Titanic (Kaggle)

1. Introdução

O Desafio Titanic da plataforma Kaggle é um dos projetos mais clássicos e introdutórios da área de Ciência de Dados. O objetivo é prever quais passageiros sobreviveram ao naufrágio do Titanic, com base em variáveis como idade, sexo, classe, entre outras. Essa tarefa de classificação binária utiliza um conjunto de dados histórico que permite praticar análise exploratória, pré-processamento de dados e modelagem supervisionada.

2. Metodologia

2.1 Análise Exploratória dos Dados (EDA)

Na etapa de EDA, foram analisadas as principais variáveis disponíveis no conjunto de treino, como Sex, Pclass, Age, Fare e Embarked. Os seguintes insights foram destacados:
- A taxa de sobrevivência foi mais alta entre mulheres (≈ 74%) do que entre homens (≈ 19%).
- Passageiros da 1ª classe apresentaram maior taxa de sobrevivência (≈ 63%) do que os da 3ª classe (≈ 24%).
- Houve muitos valores ausentes na variável Cabin e alguns valores ausentes em Age e Embarked.

Foram gerados gráficos de contagem e tabelas cruzadas que evidenciaram correlações entre as variáveis preditoras e a variável-alvo Survived.

2.2 Pré-processamento

As principais etapas de pré-processamento incluíram:
- Preenchimento de valores ausentes em Age com a média e exclusão da coluna Cabin.
- Conversão de variáveis categóricas (Sex, Embarked) para valores numéricos por meio de codificação.
- Normalização de variáveis contínuas usando StandardScaler.

2.3 Modelos Utilizados

Vários modelos supervisionados foram testados utilizando cross_val_score com validação cruzada k-fold. Os modelos incluíram:
- Logistic Regression
- Decision Tree
- K-Nearest Neighbors
- Naive Bayes
- Support Vector Machines (SVM)

- Random Forest
- XGBoost

2.4 Submissão

O modelo final escolhido foi treinado com o conjunto de treino completo, e as previsões foram geradas para o conjunto de teste do Kaggle. Os resultados foram salvos em um arquivo .csv de submissão.

3. Resultados

3.1 Acurácia de Validação Cruzada

Modelo	Acurácia Média (CV)
Logistic Regression	≈ 78%
Random Forest	≈ 81%
XGBoost	≈ 83%
SVM	≈ 79%
KNN	≈ 77%
Decision Tree	≈ 76%

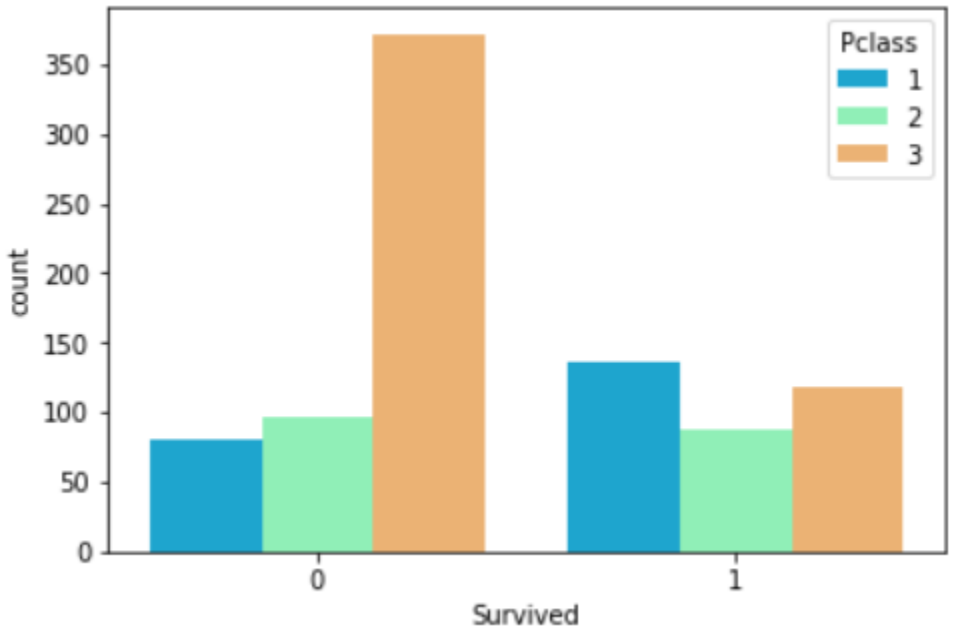
3.2 Pontuação no Leaderboard Kaggle

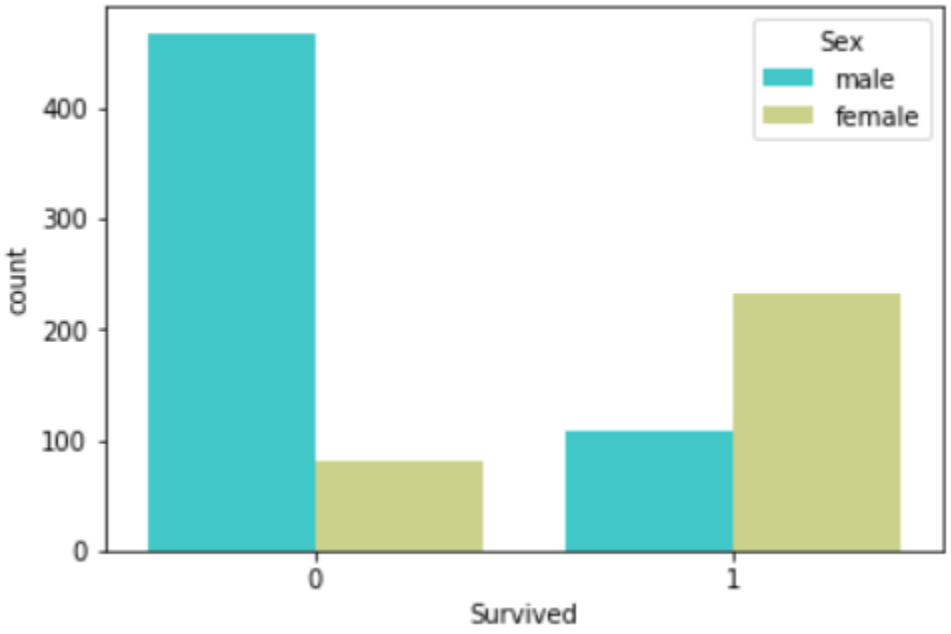
A pontuação da submissão gerada ficou em torno de 0.775 (77.5%) de acurácia pública, que é uma performance comum e respeitável para esse desafio com abordagem básica.

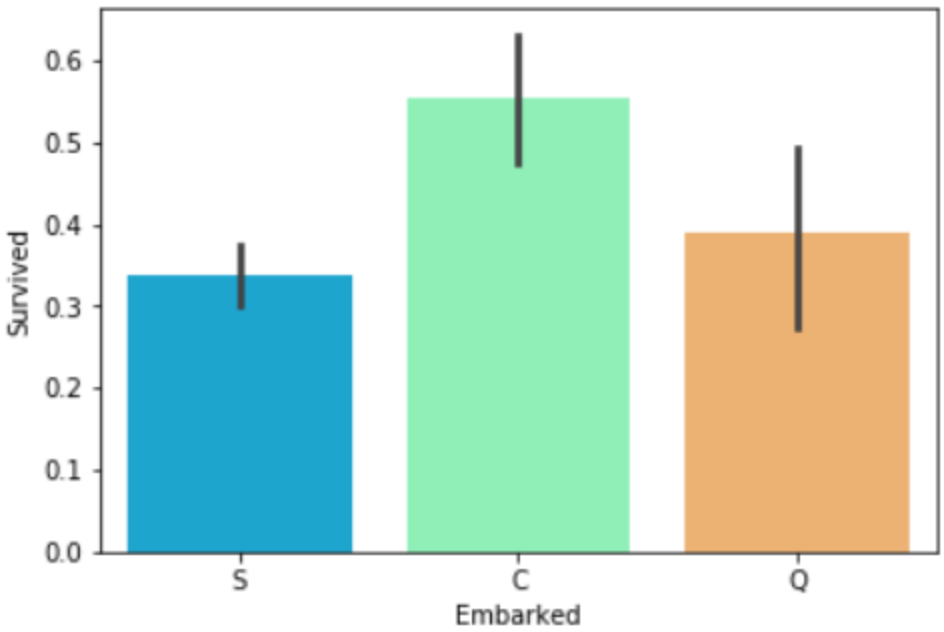
3.3 Insights

- O gênero e a classe dos passageiros foram os fatores com maior impacto na sobrevivência.
- Variáveis como idade e tarifa também tiveram correlação relevante.
- A ausência de dados completos (especialmente em Cabin) impôs limitações para uso total das informações disponíveis.

3.4 Gráficos


Pclass
1
2
3
count
350
300
250
200
150
100
50
0
0
1
Survived


Sex
male
female
count
400
300
200
100
0
0
1
Survived



4. Discussão

4.1 Desafios Enfrentados

- Lidar com valores ausentes, especialmente nas variáveis Age, Cabin e Embarked.
- Balanceamento dos dados — maior número de passageiros que não sobreviveram.
- Escolha do melhor modelo sem overfitting.

4.2 Limitações

- O uso de variáveis textuais como Name e Ticket não foi explorado.
- Cabin foi descartada por possuir muitos dados faltantes.

4.3 Possíveis Melhorias

- Aplicação de técnicas de imputação mais avançadas (como KNNImputer ou modelos baseados em regressão para Age).
- Engenharia de features com base nos títulos contidos nos nomes dos passageiros (Mr, Mrs, Miss).
- Uso de GridSearchCV para ajuste de hiperparâmetros dos modelos.

5. Conclusão

Este desafio permitiu praticar o ciclo completo de um projeto de ciência de dados supervisionado: desde a análise inicial dos dados, passando pelo pré-processamento e seleção de modelos, até a submissão no Kaggle. Foi possível compreender a importância da análise exploratória e da preparação dos dados na performance dos algoritmos de machine learning.

Além disso, o projeto reforçou conhecimentos sobre bibliotecas como pandas, seaborn, scikit-learn e XGBoost. A competição do Titanic é uma excelente porta de entrada para quem está começando na área, com desafios reais e aprendizado contínuo.