Desafio DS Lighthouse - Classificação
Relatório de Modelagem

Hugo Silveira Sousa

Descrição do Problema

Os dados se referem a indivíduos que estavam na embarcação Titanic, sendo estes classificados entre os que sobreviveram e os que não sobreviveram. O objetivo do projeto é conseguir classificar as amostras nestas duas categorias, apenas com informações do indivíduo, como idade, onde embarcou, em qual classe estava, dentre outras informações. A base de treinamento é composta por 891 amostras, e a de teste por 418 amostras. Cada indivíduo é composto por 8 atributos, incluindo a informação se sobreviveu ou não.

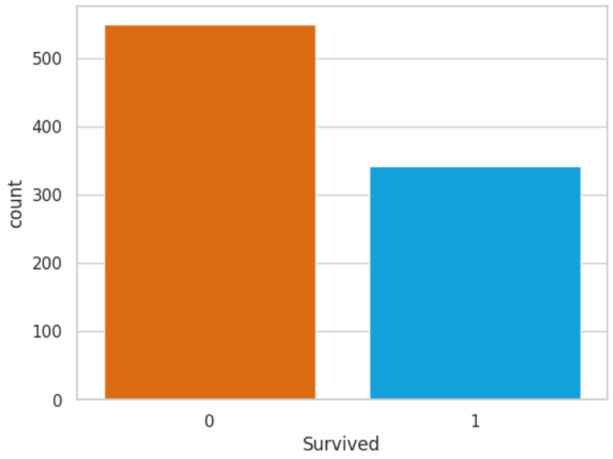
Análise exploratória dos dados (EDA)

Os 8 atributos de cada amostra são:

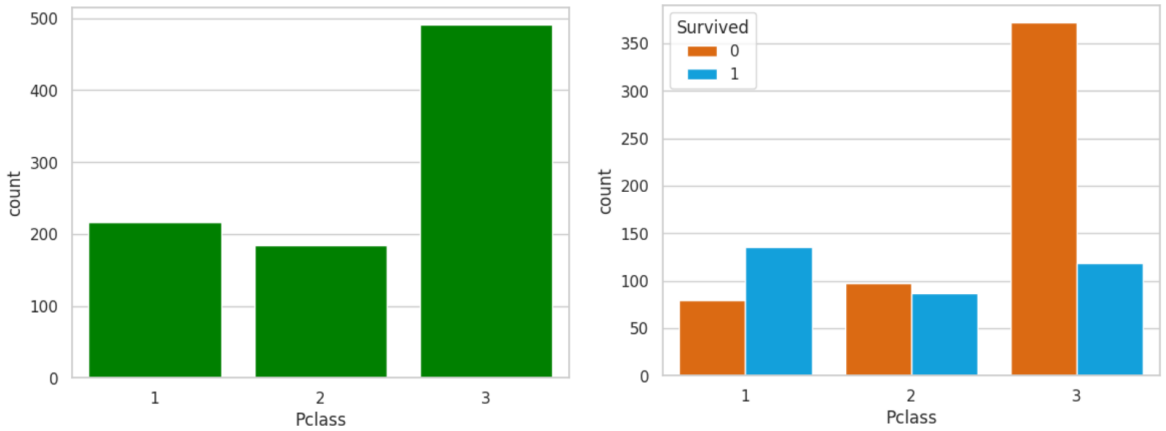
- *PassengerId*: ID do passageiro
- *Survived*: Se o passageiro sobreviveu (1) ou não (0)
- *Pclass*: Classe da passagem (1, 2 ou 3)
- *Name*: Nome do passageiro
- *Sex*: Sexo do passageiro
- *Age*: Idade
- *SibSp*: Número de irmãos/cônjuge a bordo do Titanic

- *Parch*: Número de pais/filhos a bordo do Titanic
- *Ticket*: Número da passagem
- *Fare*: Valor da passagem
- *Cabin*: Cabine do passageiro
- *Embarked*: Porto de embarque, Cherbourg (C), Queenstown (Q) ou Southampton (S)

Sobre o atributo *Survived*, temos que mais de 500 amostras da base são de indivíduos que não sobreviveram, e um pouco mais de 300 amostras sobreviveram.

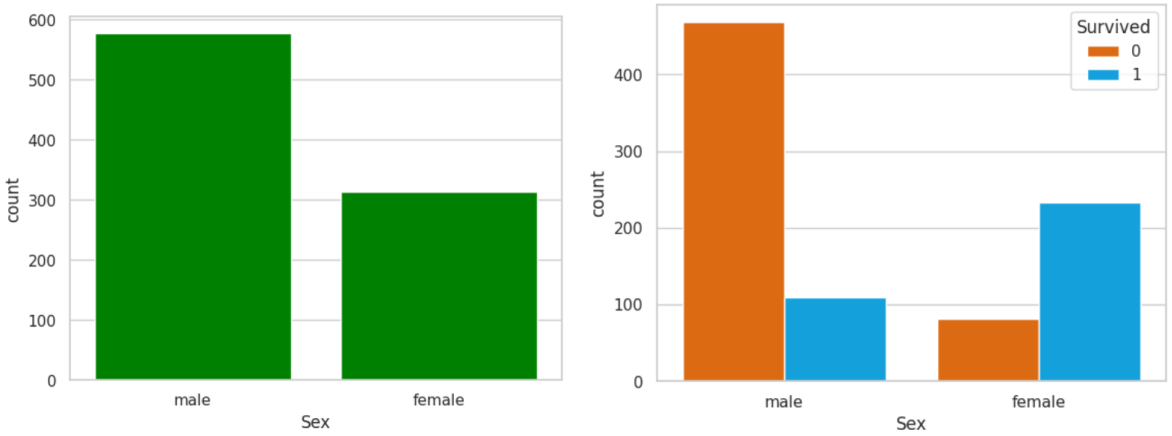


Sobre o atributo *Pclass*, a maior parte das amostras estão na 3º classe, quando analisado por sobreviventes, observou-se que a maioria das amostras da 3º classe não sobreviveram, e a maioria da 1º classe sobreviveu, indicando uma relação entre esses 2 atributos.

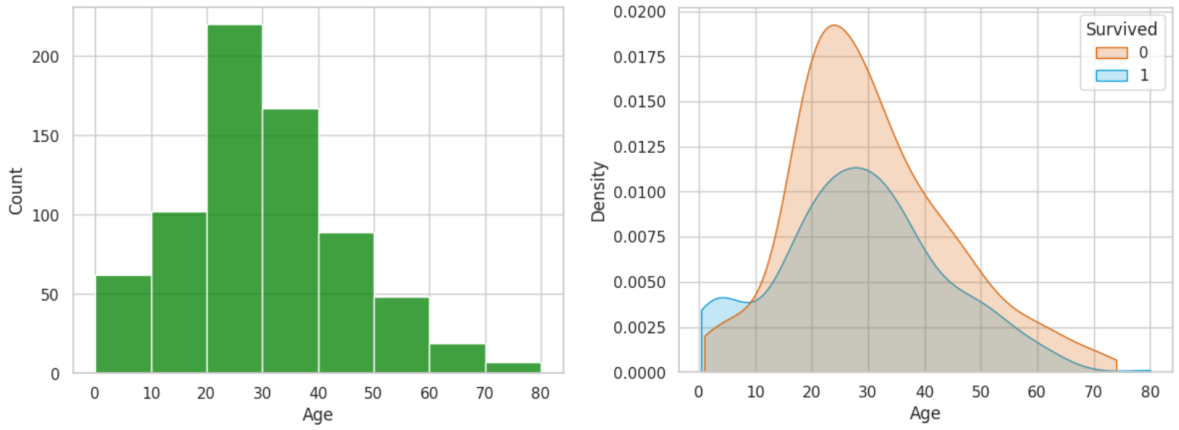


Sobre o atributo *Sex*, a maior parte das amostras eram de homens, quando analisado por sobreviventes, percebe-se que a maior parte dos homens não

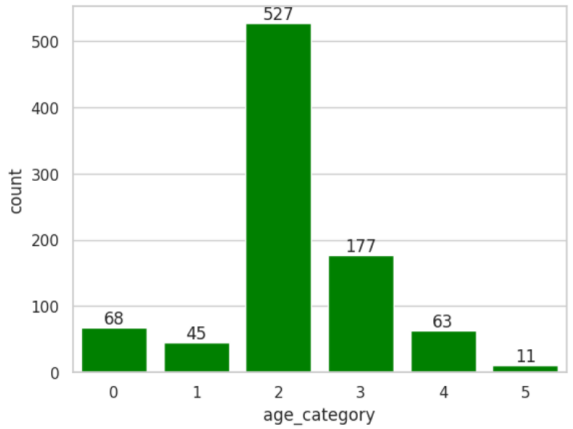
sobreviveu, e esse comportamento é contrário na classe das mulheres.

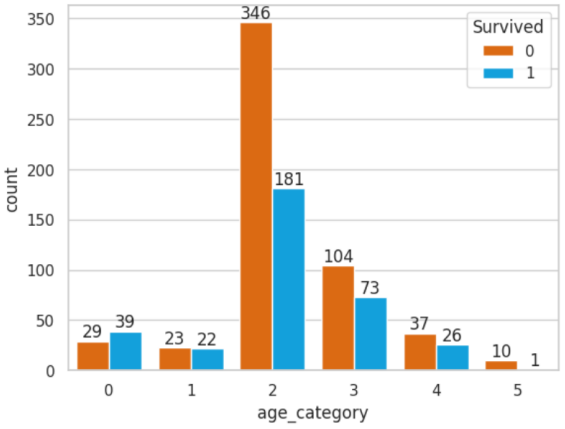


Sobre o atributo *Age*, o menor valor é de 0.42 e o maior é 80, a média das idades é de 30 anos. Com o histograma, pode-se perceber que a maioria dos indivíduos estavam entre os 20 e 40 anos, no gráfico de densidade por sobrevivência, observa-se que os não sobreviveram têm sua maioria entre os 20 e 30 anos, e que sobreviveram têm um pequeno pico entre os 0 e 10 anos, o maior pico entre os 20 e 30 anos.

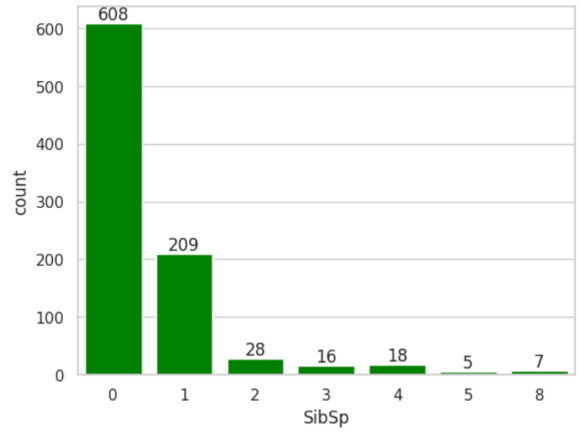


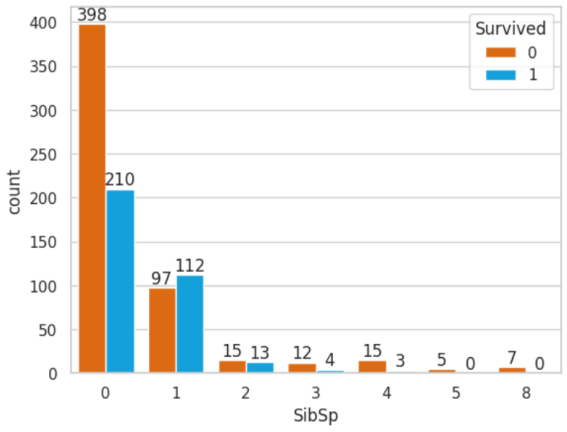
Para o projeto, as idades foram divididas em 6 categorias, 0 para menores que 12 anos, 1 para entre 12 e 18, 2 para entre 18 e 34, 3 para entre 34 e 50, 4 para entre 50 e 65 e 5 para os maiores de 65 anos. Nota-se que a maioria das amostras estão na classe 2. Quando analisado por sobreviventes, nota-se que a maioria das amostras da classe 0 sobreviveram.



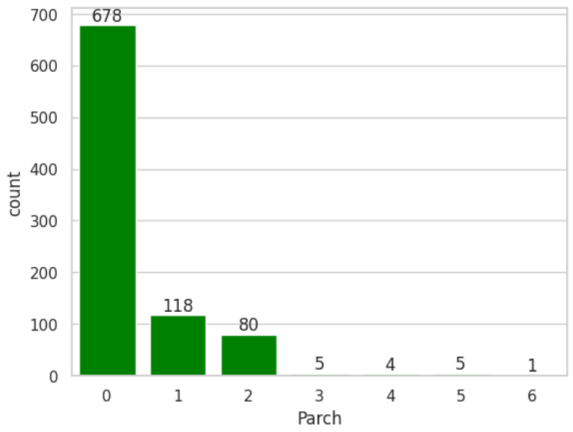


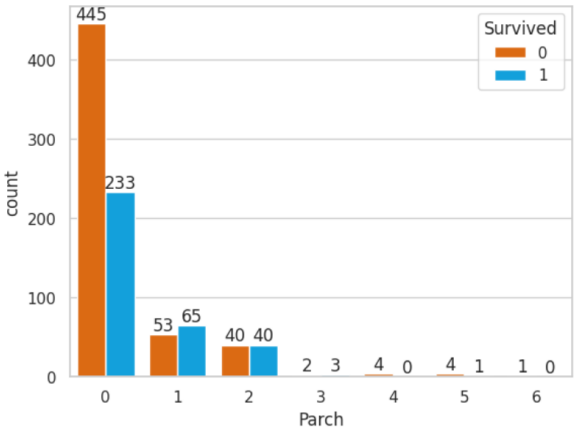
Sobre o atributo *SibSp*, a maior parte das amostras tem o valor 0, quando analisado por sobreviventes, nota-se que aqueles com o valor 1, a maioria sobreviveu.





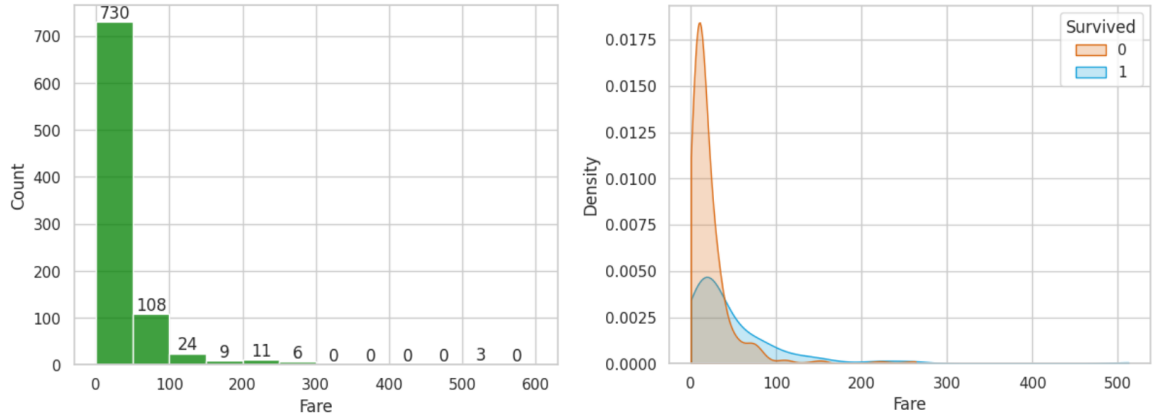
Sobre o atributo *Parch*, a maior parte das amostras tem o valor 0, quando analisado por sobreviventes, nota-se que aqueles com o valor 1, a maioria sobreviveu, mesmo comportamento do atributo *SibSp*.



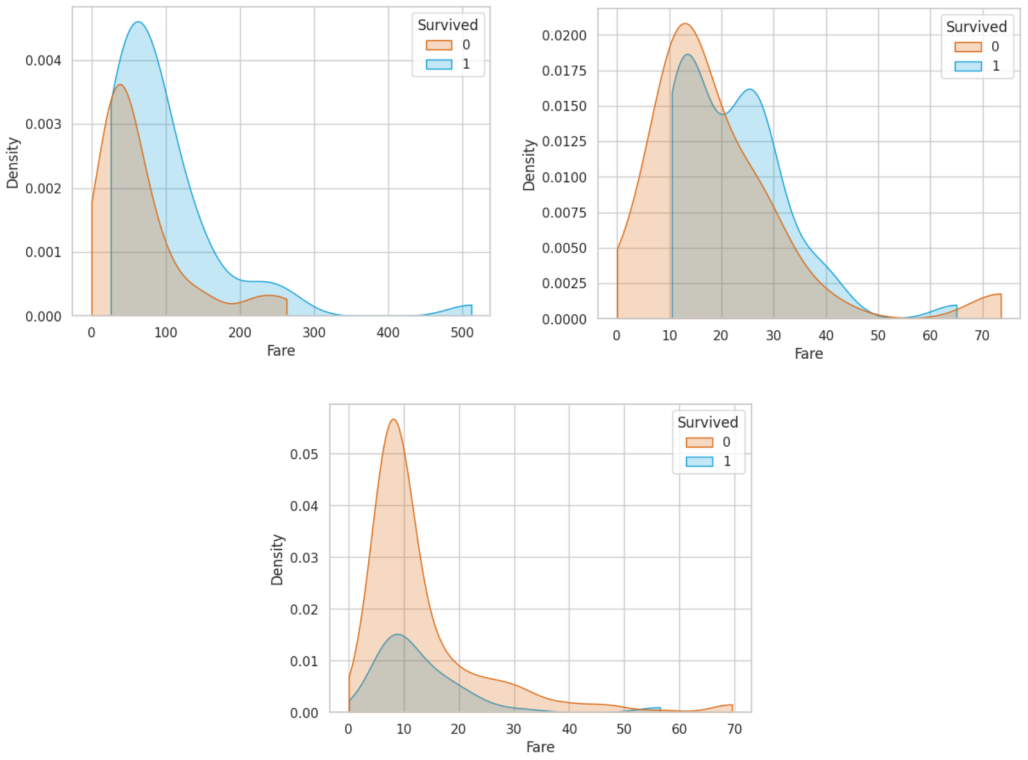


Sobre o atributo *Fare*, a grande maioria das amostras pagaram valores entre 0 e 100. Também tiveram 3 passageiros que pagaram valores maiores que 500. Quando

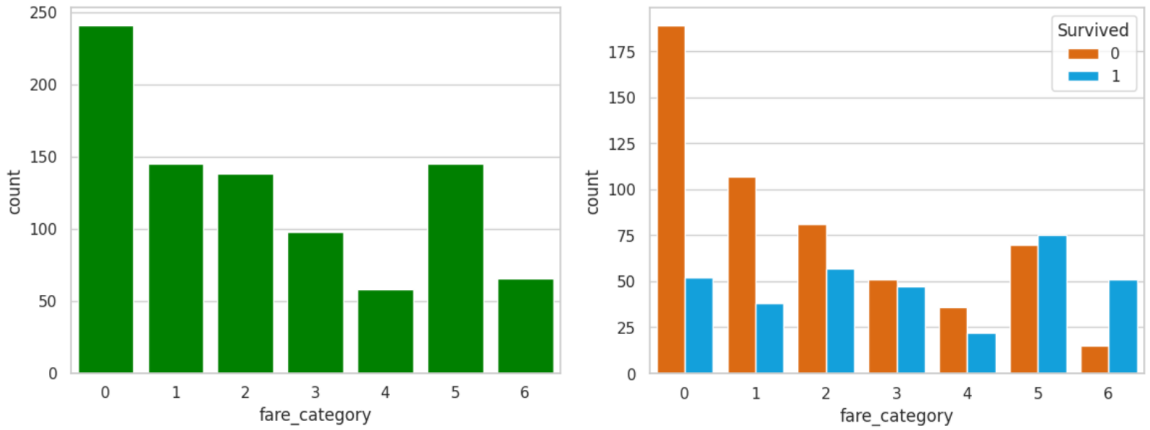
analisado o gráfico de densidade, as amostras dos não sobreviventes tem um pico bem acentuado para aqueles que pagaram os menores valores, já o dos sobreviventes é mais distribuído, mas ainda tem o seu pico nos menores valores, devido a maioria das amostras terem pagado valores mais baixos. A média geral desse atributo é 32, e a moda é 8 (valores de moda com arredondamentos).



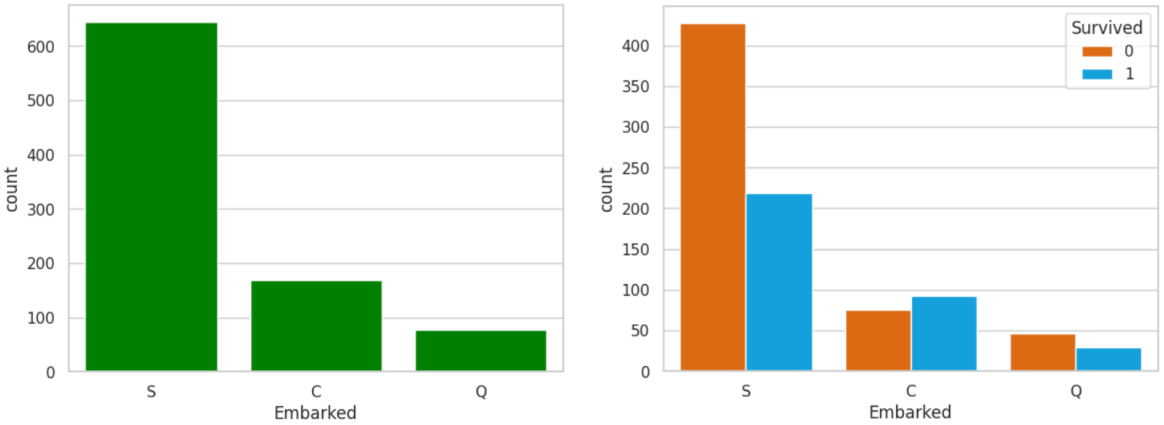
Para a 1º classe, a média é 84 e a moda é 27, 2º classe a média é 21 e a moda é 13, e para a 3º classe a média é 14 e a moda é 8. Nota-se no gráfico de densidade da 1º classe (superior esquerda) os sobreviventes tendem a pagar mais por suas passagens, mas esse comportamento não é tão evidente para as amostras da 2º classe (superior direita) e 3º classe (inferior).



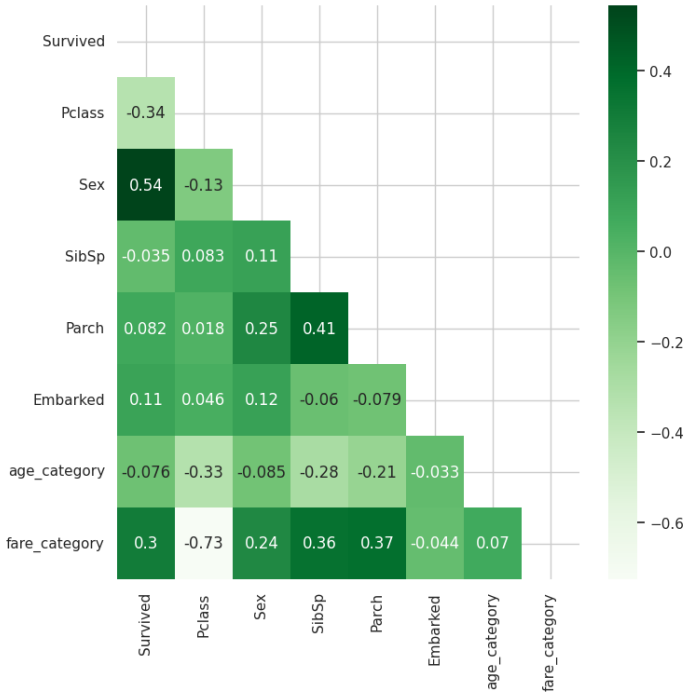
O atributo *Fare* também foi dividido em categorias, 0 para menores que 8, 1 para entre 8 e 13, 2 para entre 13 e 21, 3 para entre 21 e 27, 4 para entre 27 e 32, 5 para entre 32 e 84 e 6 para entre valores maiores que 84. Nota-se que a maioria está na categoria 0, e os de categorias menores tendem a não sobreviver e os de categorias maiores tem a sobreviver



Sobre o atributo *Embarked*, a maior parte das amostras tem o valor S, quando analisado por sobreviventes, nota-se que aqueles com o valor C, a maioria sobreviveu, e os de S a maioria não sobreviveu.



A matriz de correlações também foi gerada, percebe-se uma correlação mais elevada entre os atributos *Sex* e *Survived*, como tinha sido demonstrado nos gráficos, também há correlação entre os atributos *SibSp* e *Parch*, que se referem a parentes a bordo.



Descrição de Feature Engineering

Na fase de Feature Engineering foram eliminados os atributos:

- *PassengerId*: ID único por amostra
- *Name*: Único por amostra
- Ticket: ID único da passagem da amostra
- Cabin: Cabine do passageiro, por ter muitos valores nulos, e alguns passageiros com mais de uma cabine

Os valores nulos do atributo *Age,* foram preenchidos com a média das idades. Em seguida, o atributo foi substituído por *age_category*, com 6 categorias, sendo 0 para menores que 12 anos, 1 para entre 12 e 18, 2 para entre 18 e 34, 3 para entre 34 e 50, 4 para entre 50 e 65 e 5 para os maiores de 65 anos.

O atributo *Fare* também foi dividido em categorias, 0 para menores que 8, 1 para entre 8 e 13, 2 para entre 13 e 21, 3 para entre 21 e 27, 4 para entre 27 e 32, 5 para entre 32 e 84 e 6 para entre valores maiores que 84.

Os valores nulos do atributo *Embarked*, foram preenchidos com a moda (S).

O atributo categórico *Sex* teve seus valores substituídos por 0 para *male* e 1 para *female*.

O atributo categórico *Embarked, teve seus valores* substituídos por 0 para S, 1 para C e 2 para Q.

Critérios para escolha de um modelo de previsão

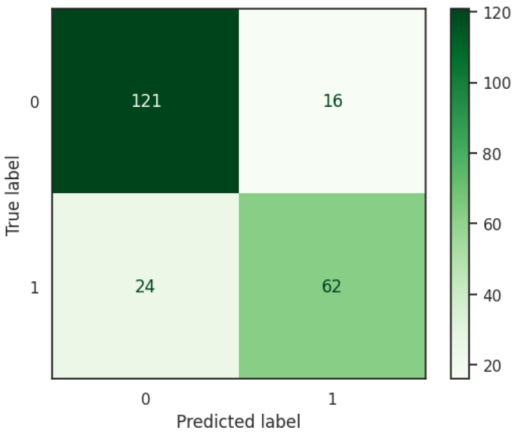
Para a escolha do modelo de classificação, foram testadas algumas abordagens clássicas, e avaliados os valores de métricas de desempenho. Aquele que trouxe maiores valores foi escolhido como o melhor modelo, e adicionado no projeto (Kedro).

Modelagem

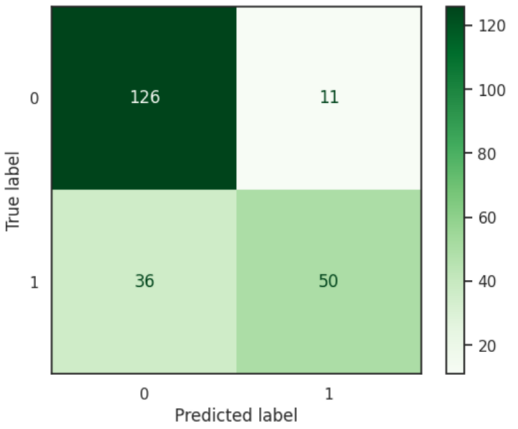
O processo de modelagem iniciou-se com a divisão da base de dados que deveria ser utilizada para treinamento. 75% da base foi usada efetivamente para treinar os dados e 25% foi usada para validar os modelos treinados.
Foram treinados modelos de 6 abordagens: Regressão Logística, Floresta Randômica, SVM, Naive Bayes, KNN e Árvore de Decisão. A biblioteca Scikit-Learn foi utilizada para usar a implementação desses modelos.
Os modelos foram treinados e validados com os mesmos dados.

Critérios para avaliação do desempenho do modelo

Para avaliar o desempenho dos modelos, foram utilizadas as métricas de acurácia, precisão, *recall* e f1-*score*, além de ser gerada a matriz de confusão para cada modelo.
O melhor modelo foi o de SVM, com 82% de acurácia, 81% de precisão, 80% de *recall* e 81% de f1-*score*. A matriz de confusão gerada foi:



O segundo melhor resultado, foi o do modelo de Floresta Randômica, com 79% de acurácia, 80% de precisão, 75% de *recall* e 76% de f1-*score*. A matriz de confusão gerada foi:



O modelo escolhido para ser implementado no Kedro foi o SVM.

As métricas de cada treinamento realizada via Kedro são salvas e podem ser acompanhadas via MLFlow, através do comando:

- kedro mlflow ui
    - http://localhost:5000

Análise dos resultados obtidos

Sobre os resultados com o modelo SVM, a matriz de confusão mostra que tiveram 121 casos que a amostra era da classe 0 e foi classificada como da classe 0 (Verdadeiro Positivo), 16 casos que era da classe 0 e foi avaliada como da classe 1 (Falso Positivo), 62 casos que era da classe 1 e foi classificado como da classe 1

(Verdadeiro Negativo) e 24 casos que era da classe 1 e foi classificado como da classe 0 (Falso Negativo).
O modelo tende a errar menos quando o indivíduo não sobreviveu, mas ainda assim quando o indivíduo sobreviveu, o modelo acertou a maioria dos testes.

Proposta de Deploy

Para deploy do modelo treinado, foi implementado junto ao Kedro uma API, utilizando o plugin kedro-fast-api, nela pode ser inserido valores para uma amostra e API leva essa amostra para o modelo que retorna a classe dele. A API pode ser gerada ao executar o projeto com o comando:

- kedro fast-api run
    - http://localhost:8000

Código do projeto

O código do projeto está disponível no link a seguir:

- https://github.com/hugosousa111/titanic-desafio-ds-lh