# Relatório de ciência de dados do Incidente do Titanic

Mateus P. Genaro, genaro.mateus@gmail.com
Setembro de 2020

# Sumário



## 1. Introdução

O incidente do navio "Titanic", em 1912, é um dos grandes eventos históricos mais trágicos na história da humanidade, devido ao imenso porte do navio, o maior já construído em sua época, e a quantidade massiva de pessoas que embarcou no mesmo. Com os dados coletados dos passageiros abordo, foi possível construir um grande banco de dados com diversas informações sobre seus tripulantes, como gênero, idade, renda, entre outros, sobrevivente ou não, entre outras informações. Baseado nesses dados, investigou-se e gerou-se, através da análise exploratória, novas informações que correlacionam as características dos passageiros, de modo que se pode identificar padrões sobre os resultados obtidos para, por exemplo, avaliar as chances de sobrevivência com base nas próprias características de um indivíduo. Além disso, para a análise preditiva, se utilizou de aprendizado de máquina para calcular as chances de sobrevivência através de diversos modelos de classificação, em aprendizado supervisionado, como máquina de vetores de suporte, florestas aleatórias, regressão logística, entre outras. Por fim, foi feita a hiper parametrização dos modelos com método ensemble para se aprimorar a acurácia obtida, na qual obteve-se como resultado aproximadamente 84% de acurácia para o modelo de máquina de vetores de suporte.

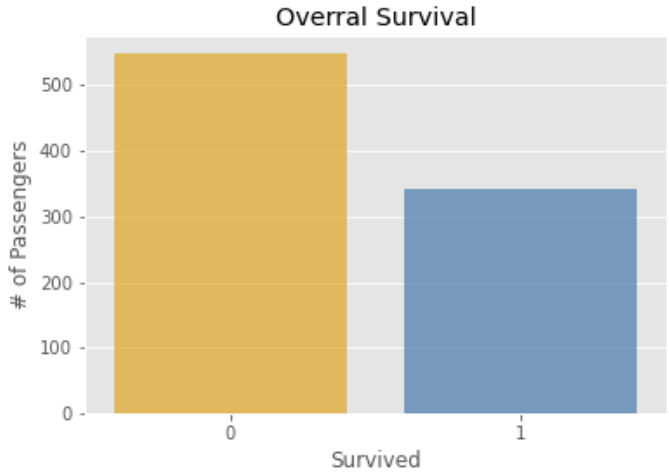
## 2. Metodologia

Para o desenvolvimento desse projeto, motivado pela familiaridade, praticidade e desempenho, como principal recurso, foi utilizado a linguagem de programação Python, juntamente às suas bibliotecas relevantes (como, pandas, numpy, sklearn, entre outras, as quais podem ser conferidas no próprio algoritmo), e os softwares de framework Anaconda Spyder.
Toda a base de dados utilizada neste trabalho foi obtida por disponibilidade do Kaggle, através de uma de suas competições nomeada "Titanic: Machine Learning from Disaster".

# 3. Análise Exploratória

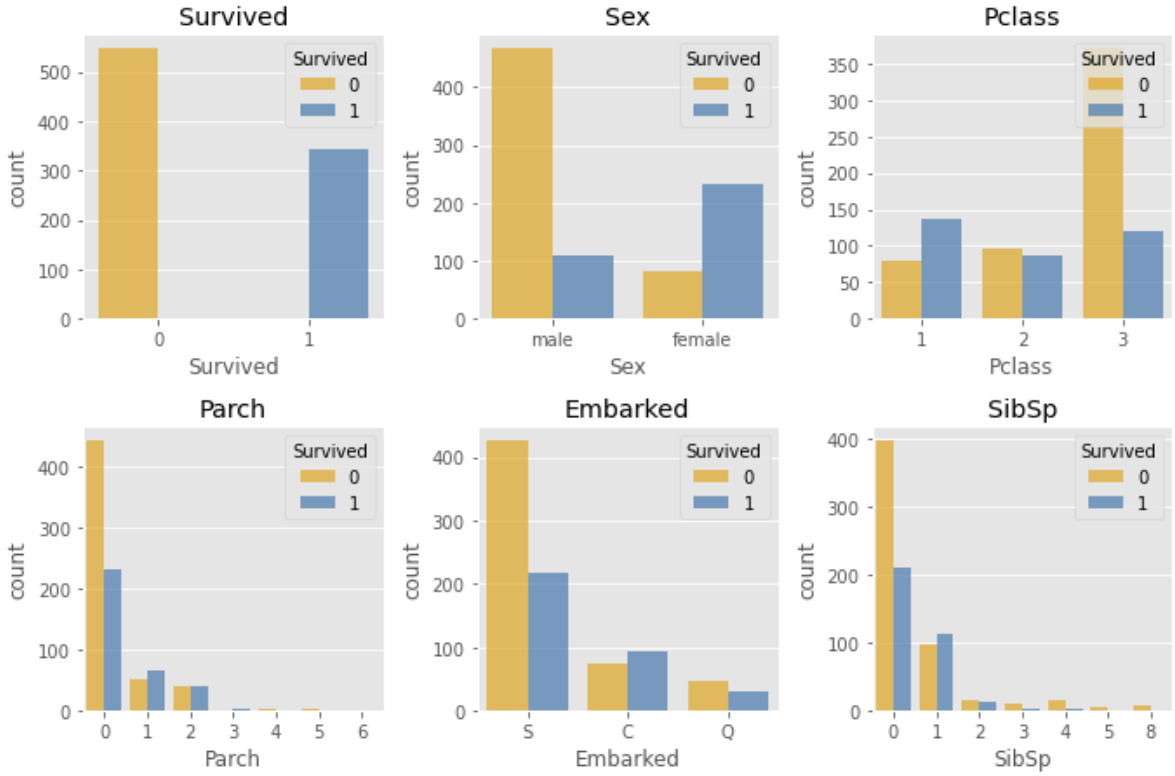
Iniciando a análise realizando a checagem de dados faltantes, nulos e inconsistentes, afim de aplicar o pré-processamento necessário e adquirir uma base de dados limpa e coerente, foram encontrados 177 valores faltantes no atributo “Age”, representando 19,9% da base de dados total, e 687 valores faltantes no atributo “Cabin”, representando 77,1% do total, e o atributo “Embarked” com 2 valores faltantes, sendo 0,2% do total. Como decisão para pré-processar os dados, afim de futuramente aplicar aprendizado de máquina, os valores faltantes foram lidados de maneira que fossem substituídos por suas respectivas médias aritméticas simples para o atributo “Age”, e por sua respectiva moda aritmética para o atributo “Embarked”. Para ao atributo “Cabin” pela grande quantidade percentual de valores faltantes, supõe-se que, além da falta de precisão na coleta de dados, grande parte dos passageiros realmente não tinham uma cabine para se alocar. Sendo assim, a decisão foi manter o mesmo atributo inalterado, afim de não criar algum tipo de viés, e verificar a intensidade do seu impacto para se determinar fatores importantes na análise.
Nenhum valor inconsistente foi encontrado para qualquer um dos atributos.
Para a detecção de outliers (pontos fora da curva), utilizou-se do método de Tukey e foram encontrados um número total de 25 outliers, representando 2,8% do número total de linhas da base de dados. A decisão para a solução foi remover os 25 outliers do dataframe, para, posteriormente, criar uma análise preditiva mais acurada, excluindo casos isolados e baseada em valores mais centrados na média e com menor variância. No entanto, para uma análise exploratória mais informativa, os outliers ainda foram considerados.

Prosseguindo para a sessão de análise gráfica, obteve-se resultados interessantes que correlacionam os atributos da base de dados de maneira a gerar novas informações sobre a própria base de dados. Começando pelo gráfico mais basal, temos, a seguir, o gráfico “Overral Survival” o qual é a contagem de número de sobreviventes:




, onde o valor ‘0’ representa a não sobrevivência e o valor ‘1’ representa os passageiros que sobreviveram. Como podemos observar, dada a base de dados que foi usada para esse

projeto, a tragédia do Titanic ocasionou em uma fatalidade de 549 vítimas (62%) e 342 sobreviventes (38%). Relacionando o atributo “Survived” com os demais atributos relevantes dentro da base de dados, temos o seguinte conjunto gráfico:



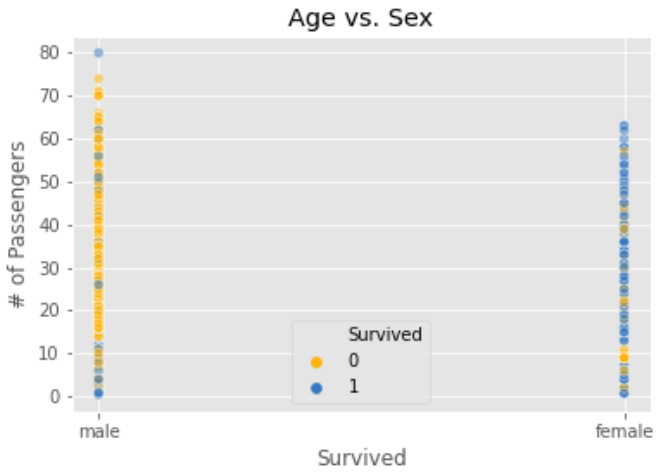

No gráfico com atributo “Sex”, nota-se que há uma grande discrepância nos números de sobreviventes entre os gêneros feminino e masculino, tendo, entre o gênero feminino, aproximadamente, 74% de sobrevivência, e, entre o gênero masculino, apenas 19%, aproximadamente. Portanto, esperasse que, para uma análise preditiva, o atributo “Sex” deva ser uma característica de importância decisiva para estimar as chances de sobrevivência de um passageiro aleatório.

O gráfico em seguida, do atributo “Pclass”, que designa qual é a classe de viagem do passageiro, sendo 1 a primeira classe, 2 a classe executiva e 3 a classe econômica. Como é possível observar, a mortalidade dentro da classe 3 é significantemente mais alta que nas outras classes, o que pode estar relacionado com diversos fatores, como o fato da quantidade de homens dentro da classe 3 (que é igual a 347) ser quase 3 vezes maior que na classe 1 (que é de 122), e como acabamos de ver que a mortalidade entre o gênero masculino é muito maior que o feminino, é um fator importante a se considerar.

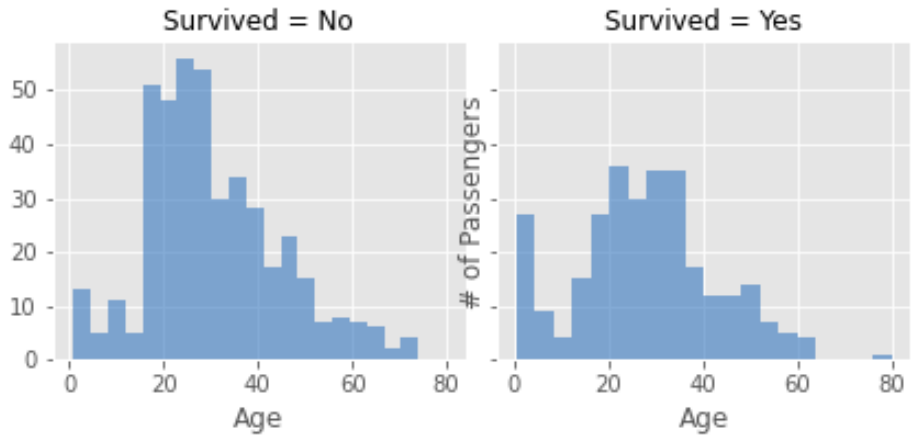
Para o atributo “Parch”, que se refere ao número de pais/filhos que estão abordo de um determinado passageiro, podemos observar em seu respectivo gráfico que as pessoas que não tem pais ou filhos abordo tem uma chance de sobrevivência consideravelmente menor do que as que tem 1 ou mais pais/filhos abordo. Combinando essa informação com o gráfico do atributo “SibSp”, que informa sobre a quantidade de irmãos e cônjuges de um passageiro, o qual também tem uma mortalidade maior para um número nulo de irmãos/cônjuges, podemos inferir que as chances de sobrevivência de passageiros os quais não tem qualquer pai/filho ou irmão/cônjuge abordo é bem baixa. Ainda, agrupando e

relacionando esses dois atributos “Parch” e “SibSp” com os atributos “Pclass” e “Sex”, pode-se notar que passageiros do gênero masculino, na classe econômica e sem qualquer parente abordo, tem uma chance extremamente baixa de sobreviver. Curiosamente, a quantidade de passageiros que tem todas essas características é de 264, o que representa 29,6% da quantidade total de passageiros abordo do Titanic. Uma combinação de atributos com grandes chances de fatalidade e não coincidentemente representada nos dados.
Por fim, neste conjunto gráfico, temos o atributo “Embarked”, o qual informa em qual das cidades um determinado passageiro embarcou, sendo “S” para “Southampton”, “C” para “Cherbourg” e “Q” para “Queenstown”. É possível observar que os passageiros que embarcaram na cidade de “Southampton” têm uma mortalidade maior do que as demais cidades. Novamente, não coincidentemente, isso deve ao fato de que 644 pessoas embarcaram na mesma, o que representa, aproximadamente, 72% do total, e, além disso, das 644 pessoas, 441 eram do gênero masculino, o que é mais um indicativo que contribui para que a quantidade de passageiros que embarcaram em “Southampton” e não sobreviveu seja tão maior do que a quantidade de sobreviventes.

Agora, relacionando o atributo “Age”, que informa a idade de um determinado passageiro, com os atributos “Sex” e “Survived”, temos o seguinte gráfico de dispersão:




Observa-se que no gráfico “Age vs. Sex” acima que há uma dispersão levemente maior na idade do gênero masculino, e que há um padrão certo na mortalidade, a qual é mais predominante entre os 20 e 50 anos de idade, no mesmo gênero. Já no gênero feminino, a distribuição da taxa de sobrevivência é mais uniforme, com uma pequena faixa de quantidade de sobreviventes entre os 45 e 65 anos. Ainda no atributo “Age”, se pôde gerar o seguinte gráfico de histograma, relacionado com o atributo “Survived”:


Survived = No
Survived = Yes
# of Passengers
0
10
20
30
40
50
0
20
40
60
80
Age
Age

4. Análise Preditiva e Machine Learning

5. Resultados

6. Conclusão