



Sobreviventes do Titanic

André Kenji Yai

Índice

- ❖ Contexto do Titanic
- ❖ Dataset
- ❖ Exploração de dados
- ❖ Missing Data
- ❖ Feature Engineering
- ❖ Modelagem
- ❖ Random Forest
- ❖ Evaluation e Validação

Titanic

- ❖ Royal Mail Ship(RMS) Titanic
- ❖ 269.1 metros
- ❖ 825 toneladas
- ❖ Capacidade total:
 - ❖ 3,547 pessoas
 - ❖ 64 barcos salva vidas



Poster of Titanic, 1912
The Granger Collection, NYC-All rights reserved

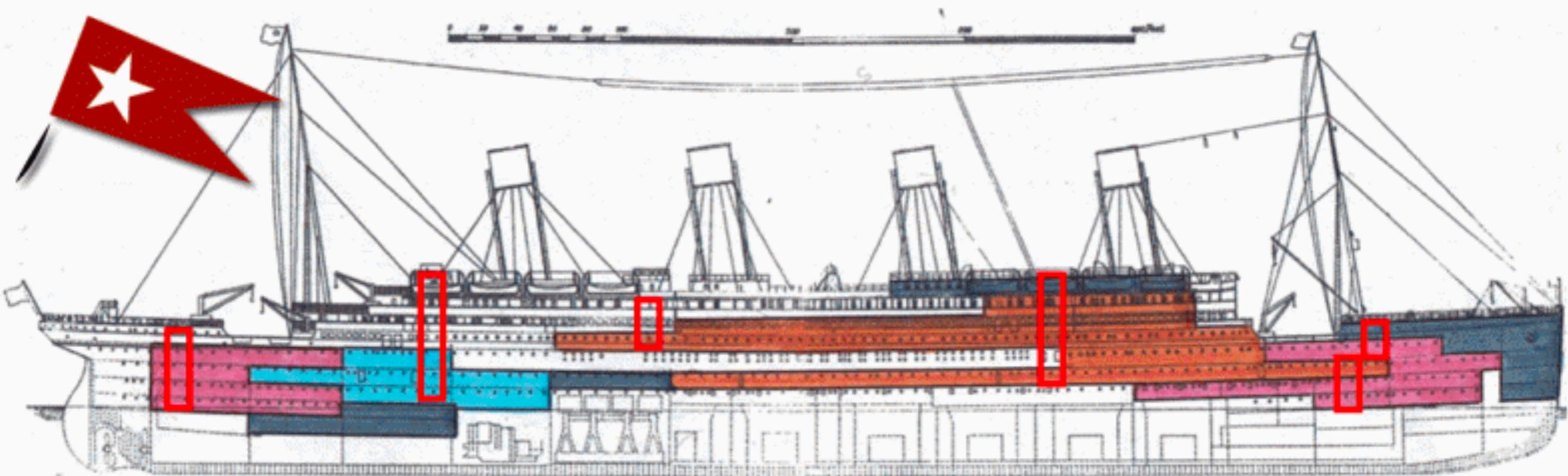


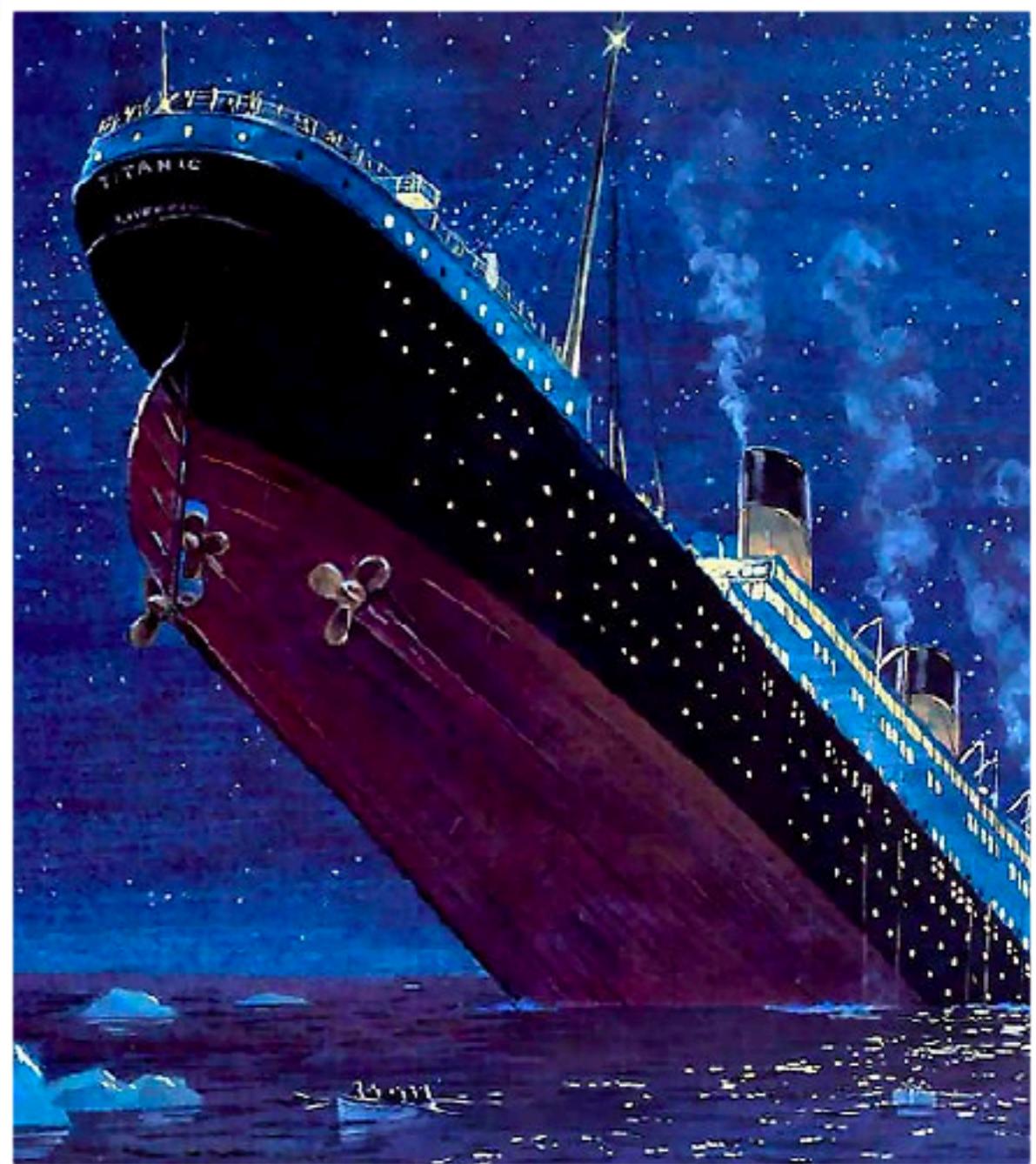
Fig. 1

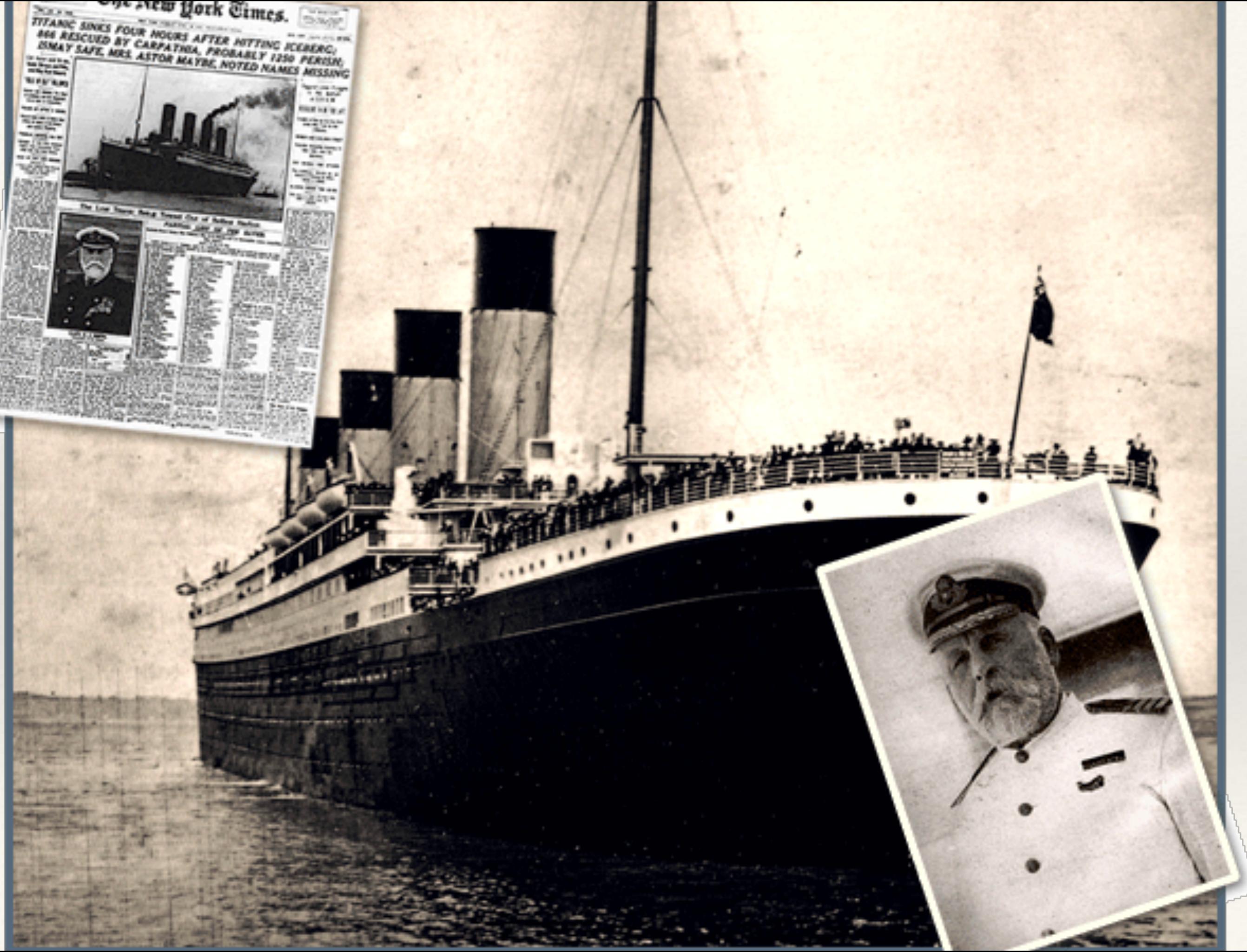
First Class Second Class Third Class Crew Stairs



Acidente do Titanic

- ❖ Em uma das suas viagens entre Cherbourg e Nova York
- ❖ Em 15 abril de 1912 , se colidiu com um iceberg.
- ❖ Na navio tinha:
 - ❖ 2,223 passageiros (62.67% capacidade total)
 - ❖ 20 barcos salva vidas
 - ❖ 702 (31.6 %) passageiros sobreviveram





Dados

- ❖ Questão: Sobreviveríamos no acidente do Titanic?
- ❖ Dados:
 - ❖ Treinamento: 891 passageiros
 - ❖ Teste: 418 passageiros
 - ❖ Sobreviventes: 342 passageiros

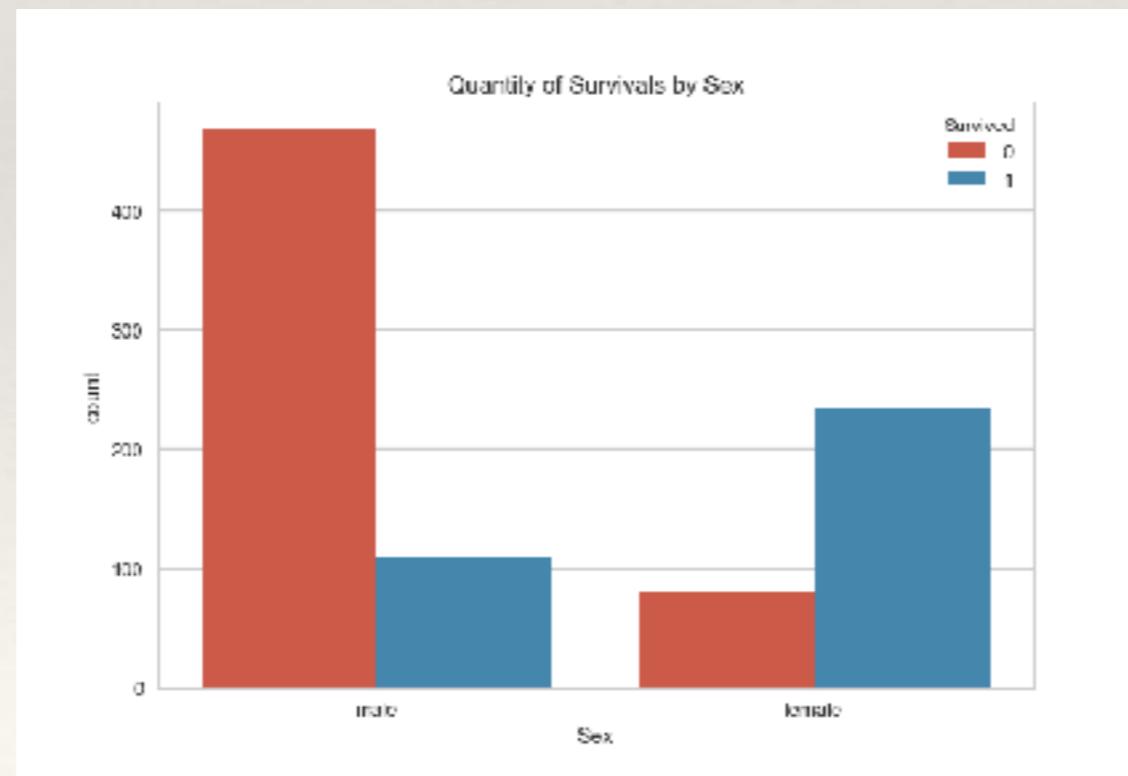


Variáveis

- ❖ Variáveis: Sexo, Idade, Classe Social ,Preço da Passagem, Porto de embarque, Ticket, Cabine, Parch (pais e filhos), SibSp (Irmãos e esposa / o) , Nome e Sobreviveu.
- ❖ Categorias:
- ❖ Numéricos:
 - ❖ Continuos: Idade, Preço da Passagem
 - ❖ Discretos: Parch, SibSp, Sobreviveu
 - ❖ Ordinais: Classe Social
- ❖ Categóricos: Sexo, Porto de Embarque, Ticket, Nome, Cabine

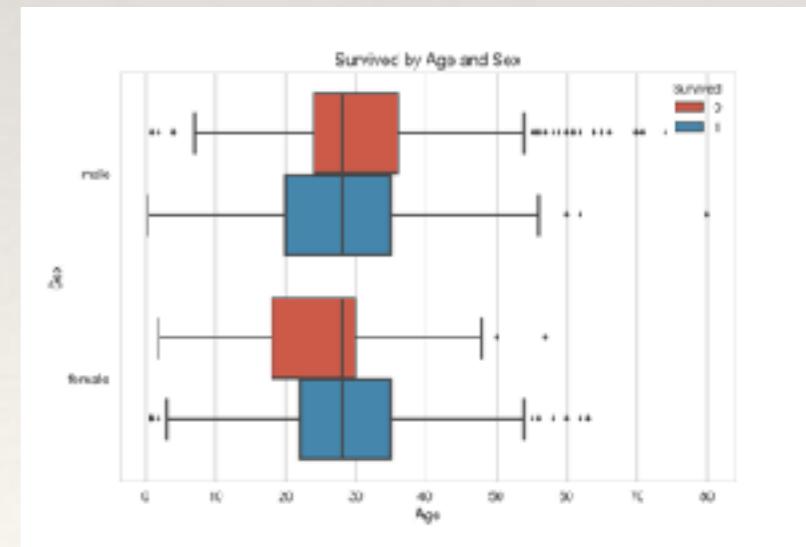
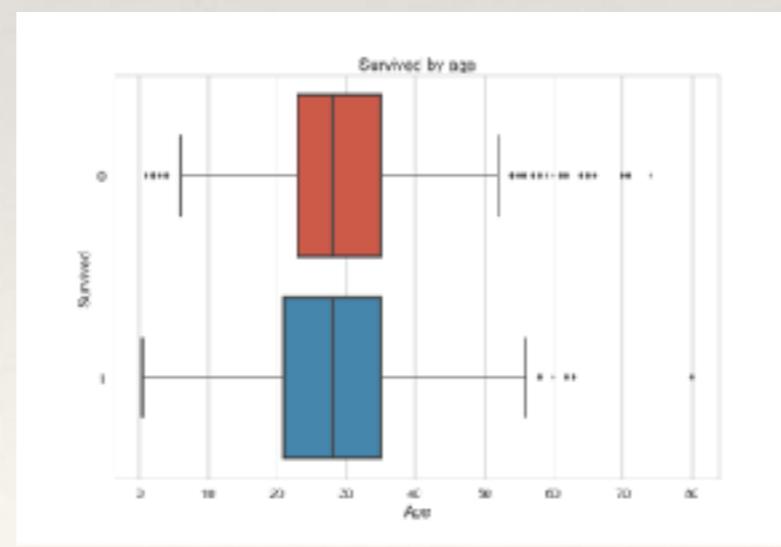
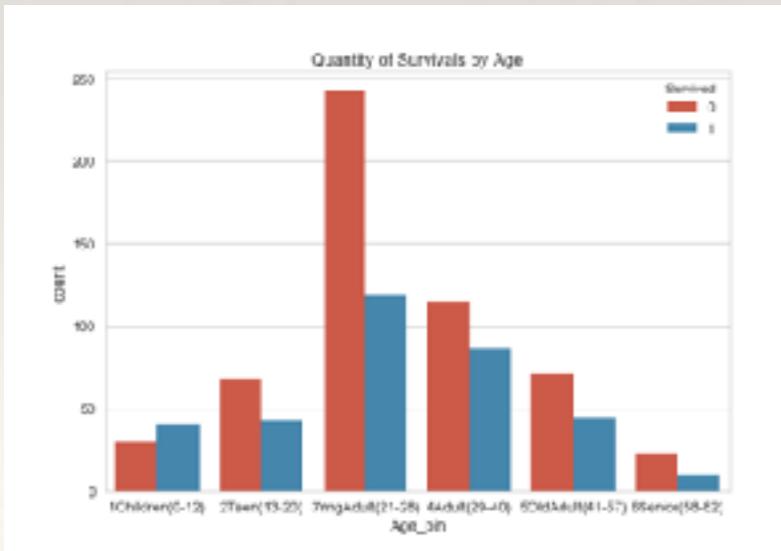
Sexo

- ❖ No nossos dados de treinamento temos:
 - ❖ 577 homens e 314 mulheres
 - ❖ 109 (18.89%) homens sobreviveram e 233 (74.20%) mulheres sobreviveram
 - ❖ 468 (81.11%) homens morreram e 81 (25.80%) mulheres morreram
 - ❖ No teste realizado a acurácia do modelo no qual todas as mulheres sobreviveram obteve uma performance melhor se for comparado com um modelo no qual nenhuma mulher sobreviveu.
 - ❖ Logistic Regression: 79.10% para o modelo onde todos sobreviveram e 58.58% para o todos morreram.



Idade

- ❖ Maioria dos passageiros tinham entre 21-28 anos
- ❖ Maioria das crianças sobreviveu ao acidente.
- ❖ Pessoa mais nova a sobreviver era recém nascido
- ❖ Pessoa mais velha a sobreviver tinha 80 anos

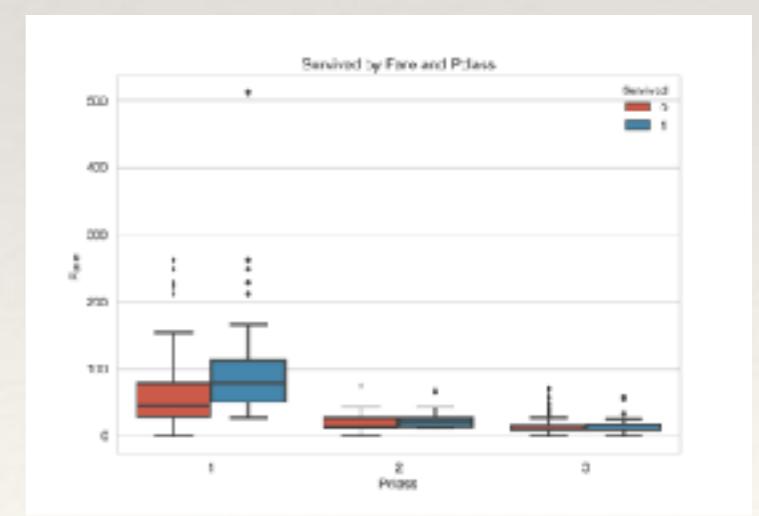
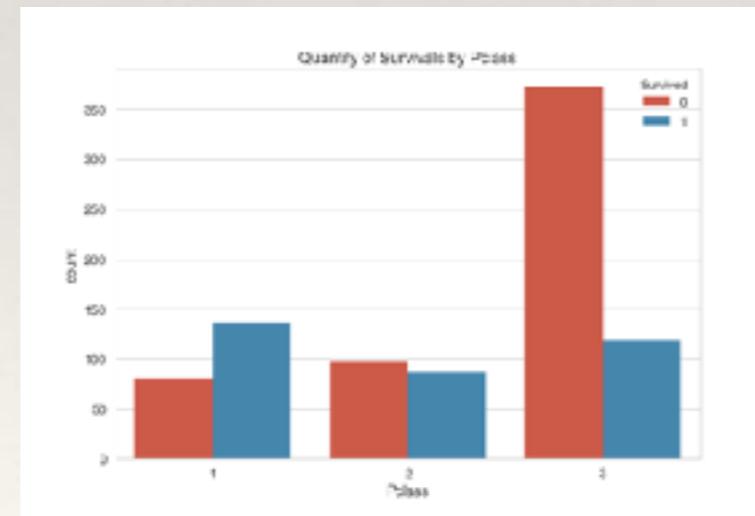
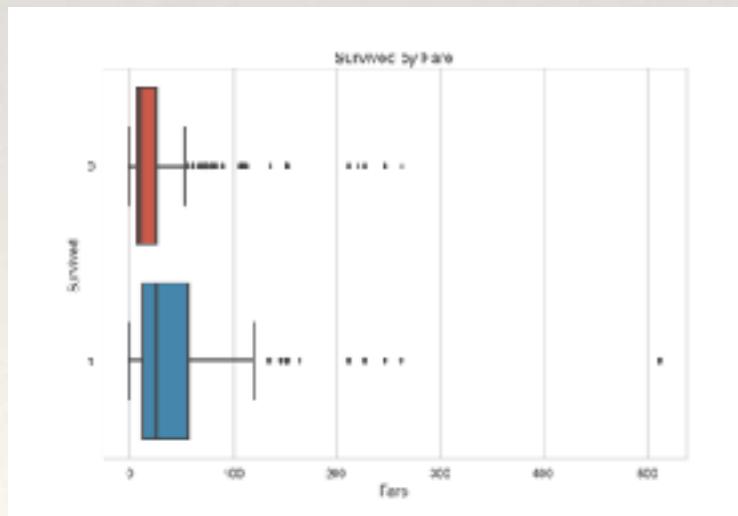


Survival Option is Given To Children and Women



Preço das Passagens

- ❖ Passageiros da 1 classe foi um outro grupo que teve uma maior taxa de sobrevivência
- ❖ Como mostra no gráfico pessoas que adquiriram passagens mais caras tiveram uma maior chance de sobrevivência.



Missing Data

- ❖ Cabin (1014 valores - 77.46%), Age(263 valores - 20%), Fare (1 value), Embarked (2 valores)
- ❖ Cabin - Associei com U (Unknown)
- ❖ Fare e Embarked - Média das linhas com features parecidos
- ❖ Age - Técnica Preditiva - Random Forest Regression

Feature Engineering

❖ Categóricas

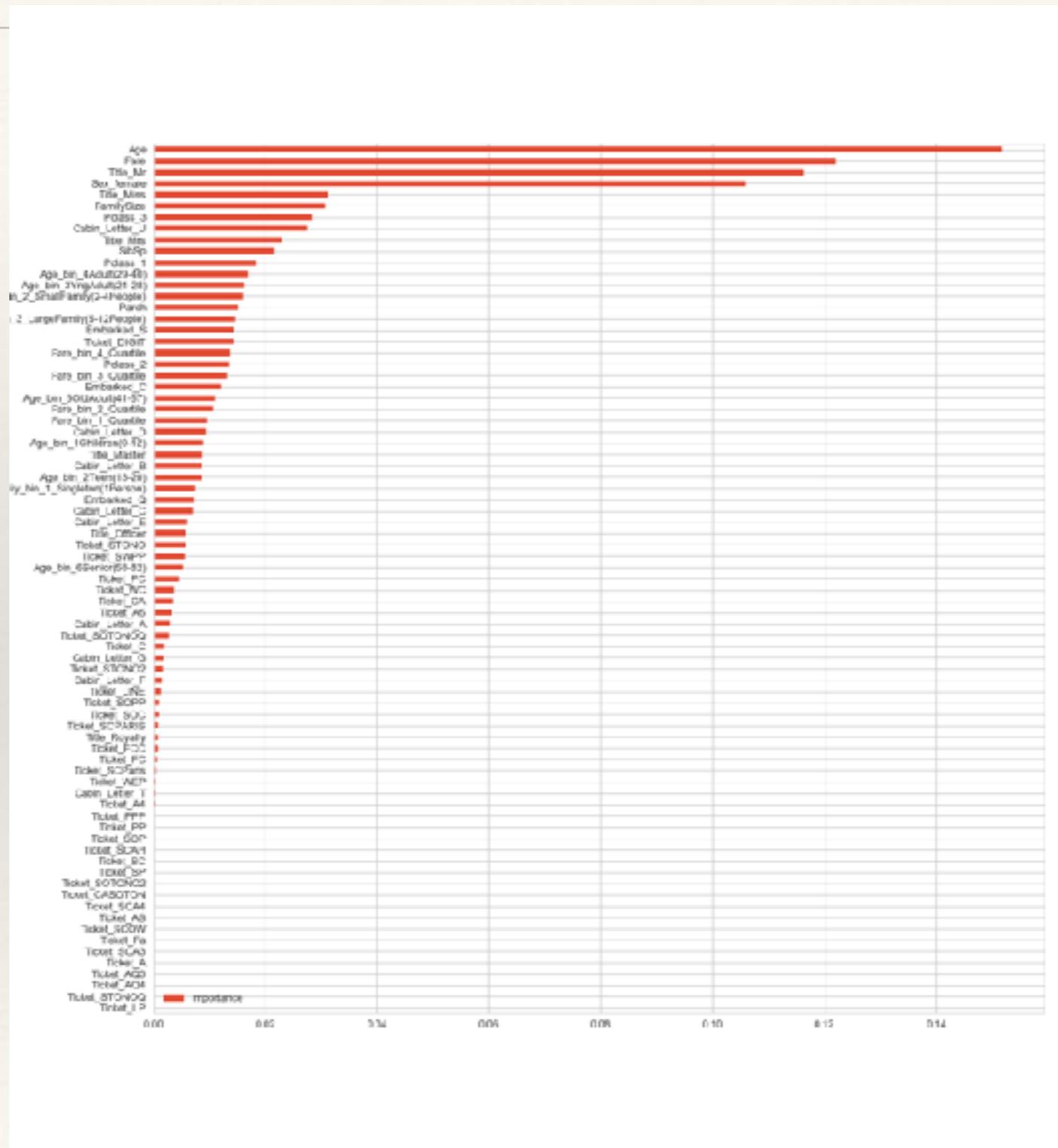
- ❖ Nomes: Título. (Mr, Mrs, Miss) e encode
- ❖ Cabines e Tickets: Prefixos e encode
- ❖ Sexo: Map (Female:1 e Male: 0)
- ❖ Embarked: Encode

❖ Discreto:

- ❖ Parch e SibSp: FamilySize e Family_bins

❖ Continuos:

- ❖ Age e Fare: Dividido em categorias e encode



Random Forest

- ❖ É um algoritmo baseado em arvores de decisão.
- ❖ Pode ser usado tanto para classificação quanto regressão.
- ❖ Nele diversas árvores de decisão são formadas de forma aleatória.
- ❖ A predição é dada através da média dos resultados das arvores geradas.
- ❖ Resultado obtido com RF é cerca de 80% de accuracia.

Referencias

- ❖ <https://www.britannica.com/topic/Titanic>
- ❖ <http://www.titanicfacts.net/>
- ❖ <https://www.kaggle.com/c/titanic>
- ❖ <https://www.kaggle.com/helgejo/an-interactive-data-science-tutorial>
- ❖ <http://www.ultravioletanalytics.com/2014/11/03/kaggle-titanic-competition-part-ii-missing-values/>

Sobreviremos no Titanic?

