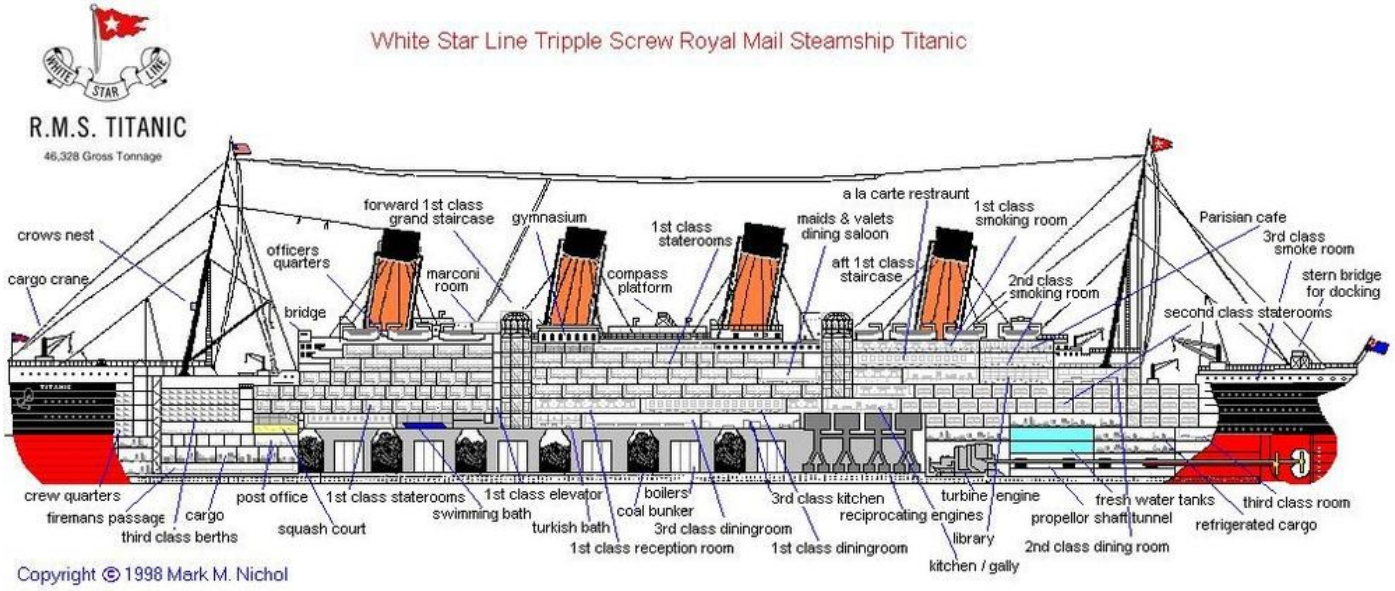
TRABALHO FINAL DA DISCIPLINA DATA SCIENCE

SOBREVIVENTES DO TITANIC



O RMS Titanic foi um navio de passageiros britânico operado pela White Star Line e construído pelos estaleiros da Harland and Wolff, em Belfast.

O Titanic tinha 269 metros de comprimento, 28 metros de largura e 53 metros de altura. Sua arqueação era de aproximadamente 46 mil toneladas, por volta de mil toneladas a mais que o Olympic. Ele operava com uma tripulação de 892 pessoas e podia transportar até 2 435 passageiros dispostos em três classes.

A embarcação partiu em sua viagem inaugural de Southampton para Nova Iorque, em 10 de abril de 1912, no caminho passando em Cherbourg-Octeville, na França, e por Queenstown, na Irlanda.

Colidiu com um iceberg na proa dianteira do lado direito às 23h40 de 14 de abril, naufragando na madrugada do dia seguinte, com mais de 1 500 pessoas a bordo, sendo um dos maiores desastres marítimos em tempos de paz de toda a história.

OBJETIVO

Utilizando técnicas de mineração de dados, você deve encontrar o melhor algoritmo que indicará se um passageiro sobreviveu ao naufrágio ou não.

O que deve ser apresentado:

- Análise exploratória dos dados:
  - Quantidade de passageiros por diferentes classes;
  - Quantidade de passageiros por gênero;
  - Quantidade de passageiros por classe de idades (criar uma nova variável para criar classes de idade, uma vez que essa é uma variável quantitativa);
  - Indicar a taxa de sobrevivência para cada uma das classes acima;
  - Identificar as correlações entre as variáveis (Correlação de Pearson);
  - Criar gráficos para apresentação dos resultados;
- Machine Learning:
  - Treinar algoritmos do tipo classificação (Árvores, Random Forest, dentre outros) para identificar se um passageiro sobreviveu ou não ao naufrágio.
  - Testar a performance do modelo utilizando o cross-validation.
  - Métricas: Accuracy, precision, recall, f1.

- Plus: Utilizar o GridSearchCV para encontrar a melhor performance para os modelos testados utilizando tuning de parâmetros.
- Apresentar a matriz de confusão do modelo treinado.
- Plus: Apresentar a curva ROC
- Apresentar o gráfico de árvore
- Apresentar a importância das variáveis para o modelo.
- Apresentar as predições para o dataset test.csv

Para realização do trabalho você está recebendo dois conjuntos de dados:

train.csv

test.csv

Os dados contêm informações sobre os passageiros do navio Titanic. Conforme apresentado abaixo:

Variáveis	Definição	Valores
PassengerId	ID do passageiro	
Survived	Sobreviveu	0 = Não, 1 = Sim
Pclass	Classe do Bilhete	1 = Primeira Classe, 2 = Segunda Classe, 3 = Terceira Classe
Name	Nome do passageiro	
Sex	Gênero do passageiro	Male = Masculino Female = Feminino
Age	Idade do passageiro em anos	
SibSp	Número de irmãos / cônjuges abordo do Titanic	
Parch	Número de pais / filhos abordo do Titanic	
Ticket	Número do bilhete	
Fare	Tarifa do passageiro	
Cabin	Número da cabine	
Embarked	Porto de embarque	C = Cherbourg, Q = Queenstown, S = Southampton

Pclass: É uma variável do tipo proxy sobre informações socioeconômica dos passageiros;

Age: A idade é uma variável que possui valores fracionários. Se for estimada segue a forma xx.5

Sibsp: O dataset define as relações familiares da seguinte forma:

Sibling = quantidade de irmãos, irmãs, meio-irmão, meia-irmã, marido e esposa;

Parch: quantidade de mãe, pai, filha, filho, enteado e enteada.

* Algumas crianças estavam viajando somente com a babá, logo o valor da variável será 0;