

**FUNDAÇÃO DE ENSINO “EURÍPIDES SOARES DA ROCHA”**

CENTRO UNIVERSITÁRIO EURÍPIDES DE MARÍLIA – UNIVEM

**Relatório Científico com**

**Análise do Acidente do Titanic**

João Victor Fontes Ribeiro

Luiz Andre Junqueira Mendes

Luiz César Benedito

Marcos Antonio Fernandes Junior

Michel Ikeda Francisco

Paulo Sérgio Bonfim Cristanello

Projeto sobre aprendizado de máquina onde o modelo preveja quais passageiros sobreviveram ao naufrágio do RMS Titanic, Fundação de Ensino “Eurípedes Soares da Rocha, submetido ao professor Fagner Christian Paes.

Marília

2020

Sumário

[1. Introdução 3](#_Toc43278017)

[2. Objetivo 3](#_Toc43278018)

[3. Obtenção dos Dados 4](#_Toc43278019)

[3.1 Definição da Base de Dados. 4](#_Toc43278020)

[3.2 Carregando os arquivos CSV. 4](#_Toc43278021)

[3.3 Importando as bibliotecas necessárias. 4](#_Toc43278022)

[4. Exploração dos Dados 5](#_Toc43278023)

[4.1 Verificando quantidade de variáveis (Colunas) e o tamanho (Linhas) do conjunto treinamento (train.csv). 5](#_Toc43278024)

[4.2 Exibir o tipo de cada feature. 5](#_Toc43278025)

[4.3 Exibir as 5 primeiras entradas do conjunto de dados. 6](#_Toc43278026)

[4.4 Exibir a distribuição estatística dos dados 7](#_Toc43278027)

[4.5 Exibir histograma das features numéricas 8](#_Toc43278028)

[4.5.1 Qual a porcentagem dos passageiros sobreviventes? 8](#_Toc43278029)

[4.5.2 Qual a faixa etária dos passageiros do Titanic? 9](#_Toc43278030)

[4.5.3 As Crianças, sobreviveram mais do que os Adultos?? 10](#_Toc43278031)

[5. Preparação do dados 11](#_Toc43278032)

[5.1 Juntar os conjuntos de dados (Treino e Teste) 11](#_Toc43278033)

[5.2 Selecionar as features 11](#_Toc43278034)

[5.3 Exibir os valores faltantes nos datasets (Treino e Teste) 12](#_Toc43278035)

[5.4 Completar os valores faltantes 12](#_Toc43278036)

[5.5 Preparar features para o modelo de Machine Learning 13](#_Toc43278037)

[5.6 Recuperar os conjuntos de dados (Treino e Teste) 13](#_Toc43278038)

[6. Modelagem & Avaliação 14](#_Toc43278039)

[6.1 Importar biblioteca do modelo de Machine Learning (Regressão Logística) 14](#_Toc43278040)

[6.2 Avaliação do modelo utilizando a Regressão Logística 14](#_Toc43278041)

[6.3 Importar biblioteca do modelo de Machine Learning (Árvore de Decisão) 14](#_Toc43278042)

[6.4 Avaliação do modelo utilizando Árvore de Decisão 15](#_Toc43278043)

[7. Conclusão 15](#_Toc43278044)

# Introdução

O projeto consiste em analisar os dados sobre o acidente do RMS Titanic, construindo um estudo para prever a probabilidade de quais passageiros sobreviveriam ao naufrágio e o que influenciou na morte ou não dos mesmos.

Palavras chave: Regressão Logística, Arvore de decisão, Titanic, Machine Learning.

# Objetivo

O RMS Titanic teve sua construção iniciada em março de 1909, levando 2 anos para ficar pronto, foi lançado ao mar em maio de 1911 e afundou no Oceano Atlântico em 15 de abril de 1912, quando colidiu contra um iceberg, o número de mortos é incerto, pois alguns passageiros desistiram de embarcar de última hora e outros não pagaram. O navio partiu de Southampton com direção a Nova York, com mais de 1500 pessoas a bordo. O Titanic foi feito para ser considerado o navio mais luxuoso e seguro daquela época.

Embora aqueles que escaparam com vida tiveram sua boa dose de sorte, alguns grupos de pessoas eram mais propensos a escaparem da morte do que outros. Por exemplo, mulheres, crianças e passageiros da 1ª Classe. Assim, nota-se que existe algum padrão que pode ser extraído dos dados brutos.

Com isso, o objetivo deste trabalho é aplicar os conhecimentos adquiridos em sala de aula e apresentar um modelo para análise dos dados referente ao acidente. Utilizando métodos e testes estatísticos para prever a probabilidade de o passageiro ter sobrevivido (1) ou não (0), analisando os resultados e escolhendo um bom modelo que seja satisfatório.

# Obtenção dos Dados

## Definição da Base de Dados.

Os dados utilizados foram retirados do site: [www.kaggle.com/c/titanic/data](https://www.kaggle.com/c/titanic/data). Esse site hospeda diversos *datasets* para competições de *Data Science.* O *dataset* contém 12 informações referente os passageiros como idade, nome, sexo e classe de viagem, entre outras.

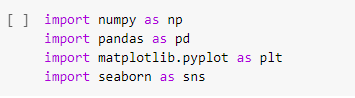
## Carregando os arquivos CSV.

Figura 1 - Carregando Arquivos CSV

Fonte: Desenvolvido pelos autores.

* Conjunto de treinamento (train.csv)
* Conjunto de teste (test.csv)

## Importando as bibliotecas necessárias.

O comando *import* é utilizado para realizar a importação das bibliotecas que contém comandos uteis para o desenvolvimento do projeto.

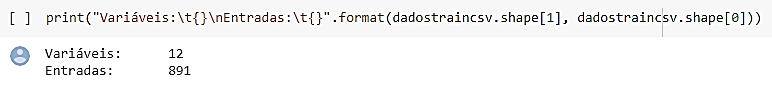
Fonte: Desenvolvido pelos autores.

Figura 2 - Importação das bibliotecas

* O pacote *numpy* possui funções de álgebra linear.
* O pacote *pandas* possui funções de *data* *processing*, CSV file I/O.
* O pacote *matplo*t possui funções para criação de gráficos e visualizações de dados.
* O pacote *seaborn* possui funções para análise de dados estatísticos.

# Exploração dos Dados

## Verificando quantidade de variáveis (Colunas) e o tamanho (Linhas) do conjunto treinamento (train.csv).

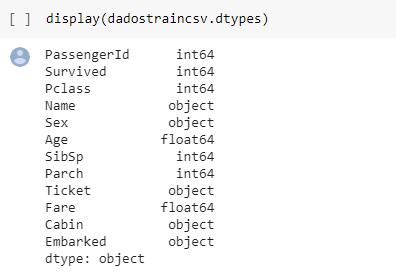


Fonte: Desenvolvido pelos autores.

Figura 3 - Verificando conjunto de treinamento (train.csv)

A Figura 3, mostra que temos 12 variáveis com um total de 891 linhas de informações referente os passageiros no dataset de treinamento.

## Exibir o tipo de cada feature.



Fonte: Desenvolvido pelos autores.

Figura 4 - Exibição do tipo das variáveis

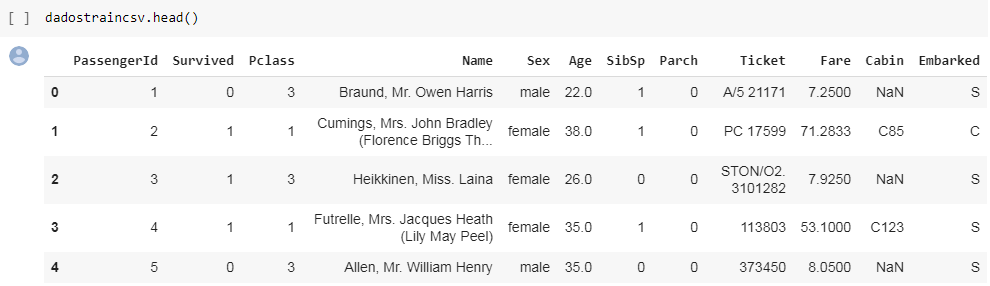
Podemos verificar na Figura 4, que temos 3 tipos diferentes de features no conjunto de dados:

* 5 do tipo int64 - Números inteiros;
* 5 do tipo object - Objetos
* 2 do tipo float64 - Números de ponto flutuante.

E as features (atributos) dos passageiros:

* PassengerId: Número de identificação do passageiro
* Survived: Informa se o passageiro sobreviveu ao desastre
* Pclass: Classe em que estava viajando
* Name: Nome
* Sex: Sexo
* Age: Idade
* SibSp: Quantidade de cônjuges e irmãos a bordo
* Parch: Quantidade de pais e filhos a bordo
* Ticket: Número do bilhete
* Fare: Preço do bilhete
* Cabin:  Número da cabine do passageiro
* Embarked: Porto no qual o passageiro embarcou

## Exibir as 5 primeiras entradas do conjunto de dados.



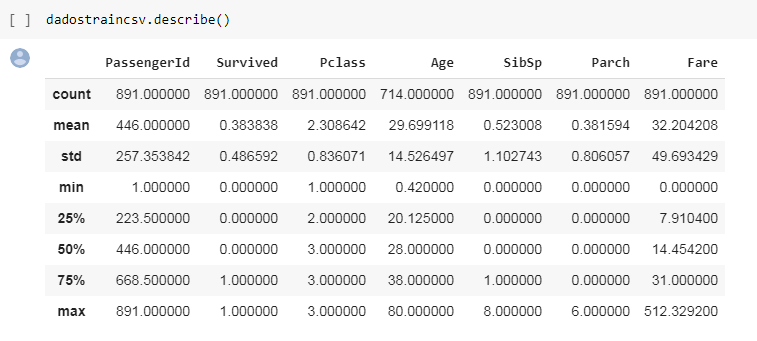
Fonte: Desenvolvido pelos autores.

Figura 5 - Exibindo as 5 primeiras linhas dos dados

Na Figura 5, temos as 5 primeiras linhas do nosso conjuntos de dados. Para exibir uma quantidade maior de entradas, deve-se adicionar o valor como parâmetro da função *head()*. Por exemplo: *dadostraincsv.head(50)*

## Exibir a distribuição estatística dos dados

Utilizamos o método *describ*e() para ter uma visão estatística dos dados e dos valores faltantes.

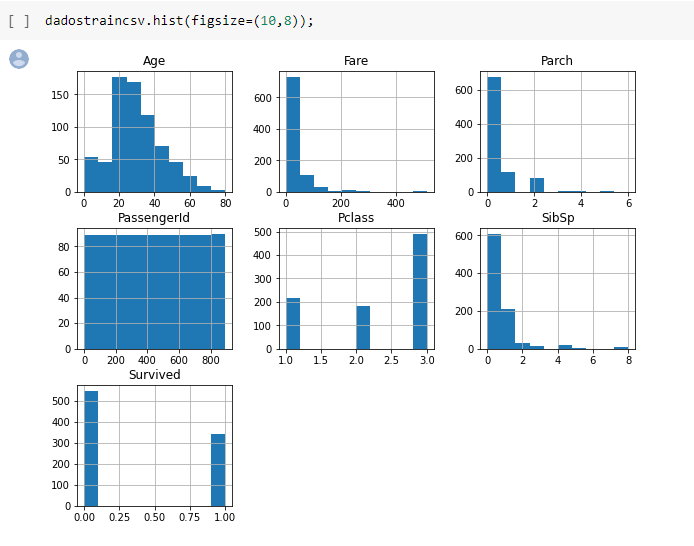


Fonte: Desenvolvido pelos autores.

Figura 6 - Visualizando estatísticas dos dados

Acima na Figura 6 podemos identificar numa análise superficial, que algumas idades estão faltando, que a porcentagem de sobreviventes foi de aproximadamente 38% e que as idades dos passageiros variam de 0,4 a 80.

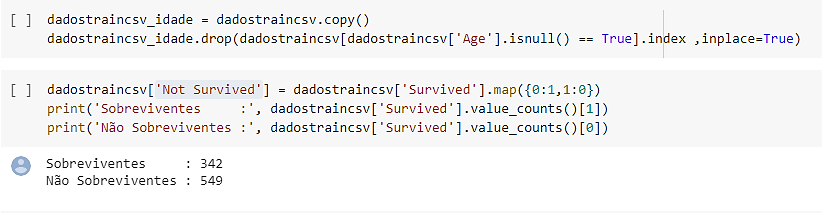
## Exibir histograma das features numéricas

Na Figura 7 logo abaixo, geramos gráficos das features numéricas, para auxiliar na análise de possíveis valores faltantes ou inconsistentes.

Fonte: Desenvolvido pelos autores.

Figura 7 – Histogramas das features numéricas

### Qual a porcentagem dos passageiros sobreviventes?



Fonte: Desenvolvido pelos autores.

Figura 8 - Quantidade de Sobreviventes e Não Sobreviventes



Fonte: Desenvolvido pelos autores.

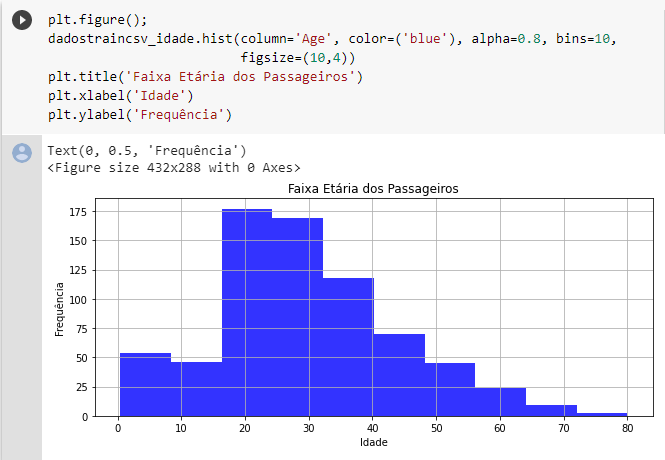
Figura 9 - Gráfico com porcentagem de sobreviventes e não sobreviventes

Do total de 891 passageiros, 342 sobreviveram e 549 não sobreviveram como mostrado na Figura 8.

Para melhor entender, na Figura 9 temos um gráfico com a porcentagem de sobreviventes (38.4%) e não sobreviventes (61.6%)

### Qual a faixa etária dos passageiros do Titanic?

Como citado anteriormente, podemos confirmar pelo gráfico da Figura 10, que a idade dos passageiros está entre 0.4 anos a 80, com uma quantidade maior de pessoas na faixa dos 20 – 30 anos de idade.



Fonte: Desenvolvido pelos autores.

Figura 10 - Gráfico com faixa etária dos passageiros

### As Crianças, sobreviveram mais do que os Adultos??

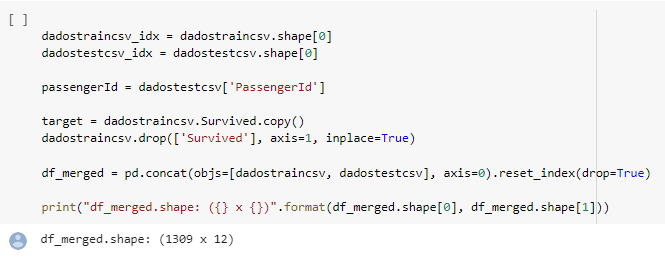
Fonte: Desenvolvido pelos autores.

Figura 11 – Gráfico comparativo entre sobrevivência de adultos e crianças

Para analisar se as crianças sobreviveram mais que os adultos, definimos “crianças” todos os passageiro com menos de 18 anos. Como mostra o gráfico, as crianças tem uma taxa de sobrevivência maior que a dos adultos.

# Preparação do dados

## Juntar os conjuntos de dados (Treino e Teste)



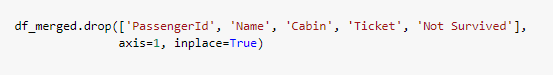
Fonte: Desenvolvido pelos autores.

Figura 12 – Código para Concatenar treino e teste em um único DataFrame

## Selecionar as features

Como qualquer conjunto de dados, pode existir a necessidade de retirar algumas features por falta de informações ou por não serem úteis.

Optamos por retirar as features: *PassengerId, Name, Cabin, Ticket e Not Survived*, por não conter informações que irão influenciar no projeto.

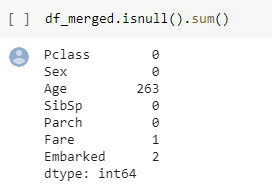


Fonte: Desenvolvido pelos autores.

Figura 13 – Linha de comando onde são retirados algumas features

.

## Exibir os valores faltantes nos datasets (Treino e Teste)

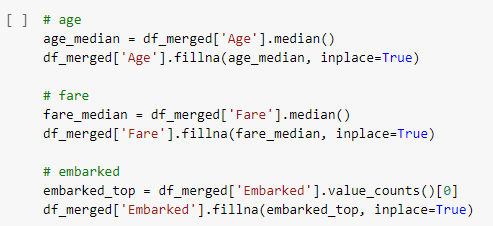


Fonte: Desenvolvido pelos autores.

Figura 14 – Verificação de valores faltantes

## Completar os valores faltantes

Depois de verificado os dados faltantes, iremos imputar essas informações substituindo os valores nulos.

Para o preenchimento da idade (age) e preço do bilhete (fare), optamos por calcular o valor da mediana para ter uma precisão melhor, já no local de embarque (embarked) adicionamos com o valor de maior frequência.

Fonte: Desenvolvido pelos autores.

Figura 15 – Preenchimento de valores faltantes

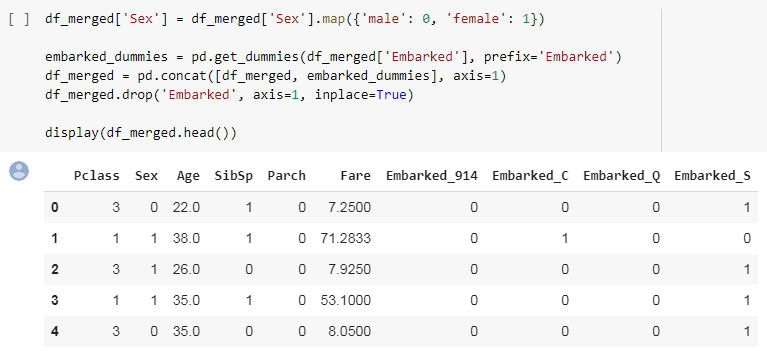
## Preparar features para o modelo de Machine Learning

Convertemos a feature embarked para *dummies*, depois utilizamos o comando *get\_dummies que realiza a* analise dos dados contidos e criar o número apropriado de features, como resultado obtivemos: *Embarked\_C, Embarked\_Q e Embarked\_S.*

A feature *sex* será utilizada de forma binária, sendo (1) mulher e (0) homem.

Fonte: Desenvolvido pelos autores.

Figura 16 – Preparação das features



## Recuperar os conjuntos de dados (Treino e Teste)

Fonte: Desenvolvido pelos autores.

Figura 17 – Linha de comando para recuperação dos datasets



# Modelagem & Avaliação

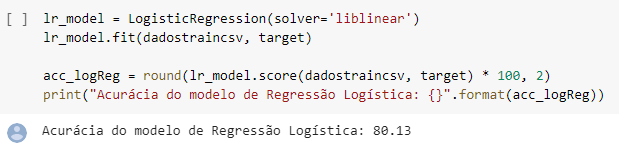
## Importar biblioteca do modelo de Machine Learning (Regressão Logística)

Figura 18 – Importação da biblioteca de Regressão Logística

Fonte: Desenvolvido pelos autores.

## Avaliação do modelo utilizando a Regressão Logística

Utilizando a regressão logística, obtivemos uma acurácia de 80.13%



Fonte: Desenvolvido pelos autores.

Figura 19 - Avaliação do modelo

## Importar biblioteca do modelo de Machine Learning (Árvore de Decisão)

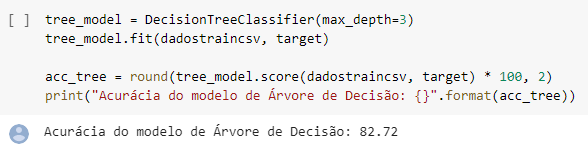


Figura 20 – Importação da biblioteca de Árvore de Decisão

Fonte: Desenvolvido pelos autores.

## Avaliação do modelo utilizando Árvore de Decisão

Utilizando o modelo de árvore de decisão, obtivemos uma acurácia de 82.72%



Fonte: Desenvolvido pelos autores.

Figura 21 - Avaliação do modelo

# Conclusão

Podemos concluir que o projeto alcançou o seu objetivo.

Todo o projeto ocorreu como o esperado, esse Desafio do dataset Titanic nos deu uma excelente oportunidade de testar na pratica, os conhecimentos adquiridos em sala de aula. Suas amostras são confirmados pela própria história do ocorrido, e sua acurácia se mostrou na média entre diversos projetos avaliados na competição da Kaggle. Evidentemente, para que se obtenha um modelo com uma acurácia maior, aconselha-se analisar melhor os dados e realizar mais testes com outras bibliotecas de Machine Learning. Para este projeto, o algoritmo de Árvore de Decisão foi o mais eficaz, retornando um bom resultado. Para projetos posteriores, será essencial o estudo mais aprofundado da exploração e preparação dos dados, assim como de outras bibliotecas de Machine Learning, a fim conseguir melhores resultados do que os obtidos neste projeto.