Titanic

Machine Learning From Disaster
Universidade de Brasília

Computação Experimental

Professora: Genaina Nunes Rodrigues

Grupo:

Antônio Henrique Moura Rodrigues 15/0118236

Maria Fernanda do Carmo Oliveira 14/0153641

Rafael Dias Silveira 14/0030433

Yan Victor dos Santos 14/0033599

Sumário

- Introdução
- Motivação
- Metodologia
- Análise de Dados
- Resultados e Conclusão

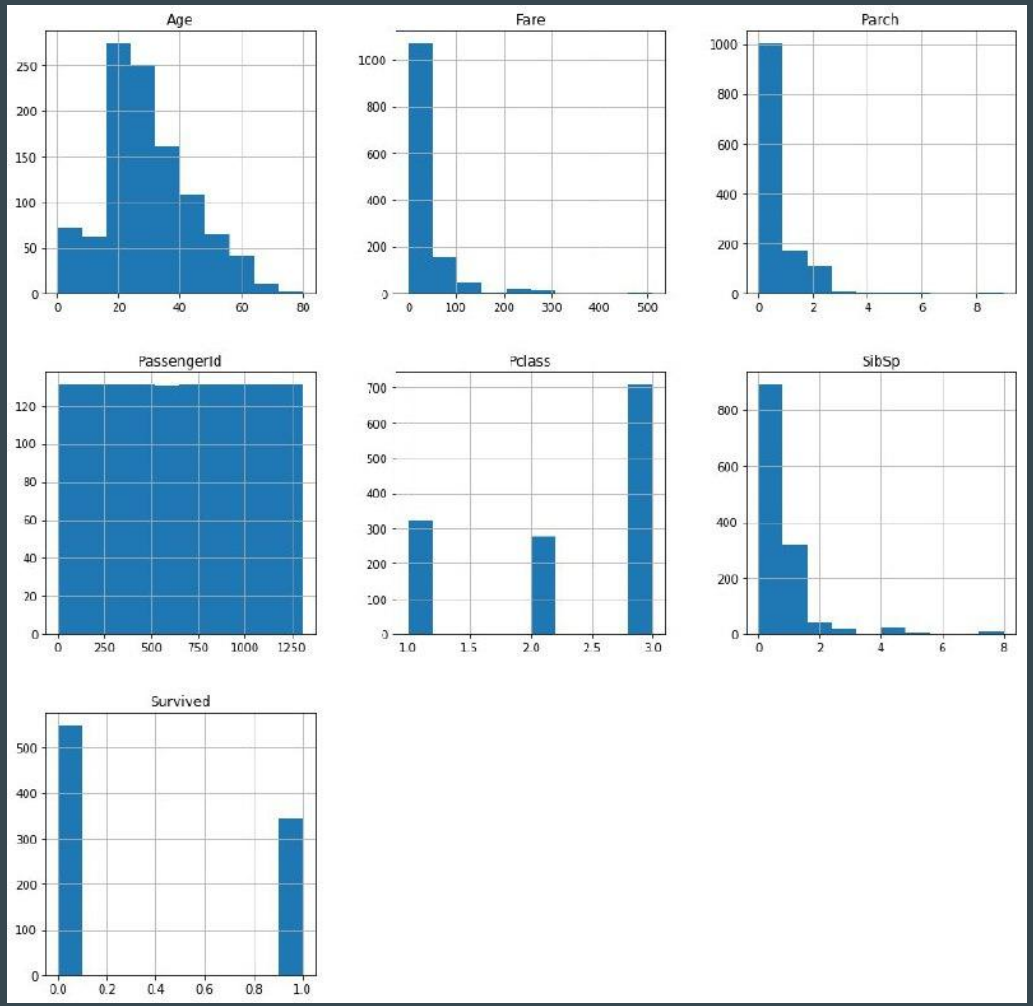
Introdução

- Este trabalho tem por objetivo facilitar o processo de aprendizagem sobre métodos de análise de dados.
- Os dados escolhidos* englobam informações sobre passageiros que estavam a bordo do Titanic.
- O projeto é criar um modelo preditivo, utilizando *Machine Learning*, para prever se um determinado passageiro irá sobreviver ou não, de acordo com as informações dadas.
- Os métodos de análise foram escolhidos para facilitar a visualização de padrões existentes nos dados.

*Fonte: https://www.kaggle.com/c/titanic/data

Histograma de Dados Titanic

Fonte Dados*:
https://www.kaggle.com/c/titanic/data



Metodologia - Matriz de Correlação

- Para fazer a correlação usamos a biblioteca Pandas.
- Para o cálculo foram excluídos valores NA/Null.
- Todas as variáveis são consideradas e no final temos uma matriz N x N.
- Diagonal principal = 1, já que é a correlação de uma variável com ela mesma.
- Quanto mais próxima de 1 for a correlação mais as variáveis são relacionadas.
- Valores próximos de 0 indicam pouca relação entre as variáveis.
- Valores negativos são variáveis inversamente proporcionais.

In [120]:
```
print('Matriz de correlação:')
raw_data.corr()
```

```
Matriz de correlação:
```

Out[120]:

	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1.000000	-0.005007	-0.038354	0.028814	-0.055224	0.008942	0.031428
Survived	-0.005007	1.000000	-0.338481	-0.077221	-0.035322	0.081629	0.257307
Pclass	-0.038354	-0.338481	1.000000	-0.408106	0.060832	0.018322	-0.558629
Age	0.028814	-0.077221	-0.408106	1.000000	-0.243699	-0.150917	0.178740
SibSp	-0.055224	-0.035322	0.060832	-0.243699	1.000000	0.373587	0.160238
Parch	0.008942	0.081629	0.018322	-0.150917	0.373587	1.000000	0.221539
Fare	0.031428	0.257307	-0.558629	0.178740	0.160238	0.221539	1.000000

Metodologia - Regressão Linear

Linguagem: Python

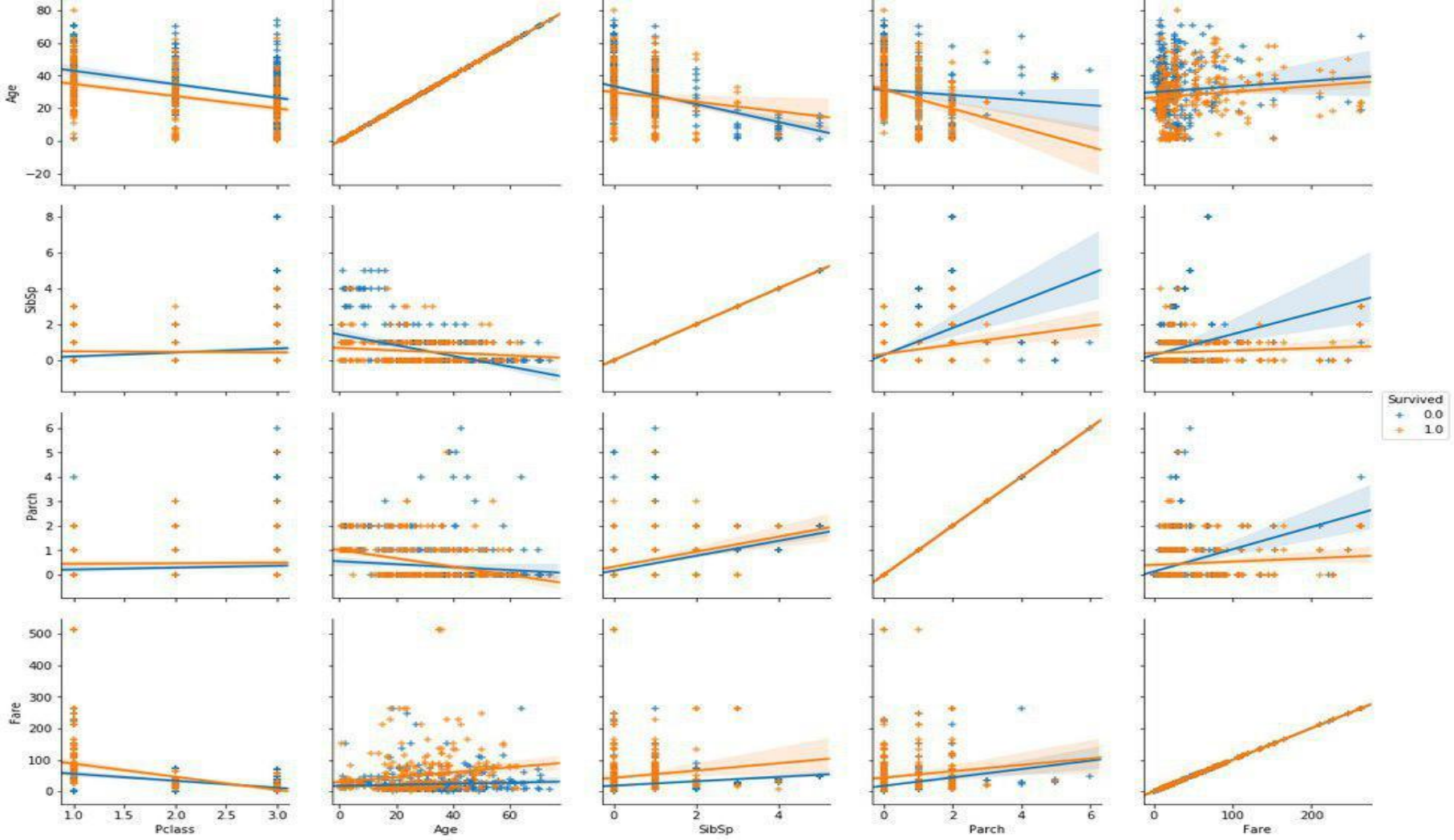
Biblioteca: Seaborn

$$y = ax - b$$

$$a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$b = n^{-1} (\sum y_i - a \sum x_i)$$

*Fonte: R. Jain. The Art Of Computer Systems Performance Analysis. Wiley, 1991. Capítulo 14.


Age
SibSp
Parch
Fare
Pclass
Age
SibSp
Parch
Fare
Survived
0.0
1.0

Metodologia - Projeto Fatorial 2^k

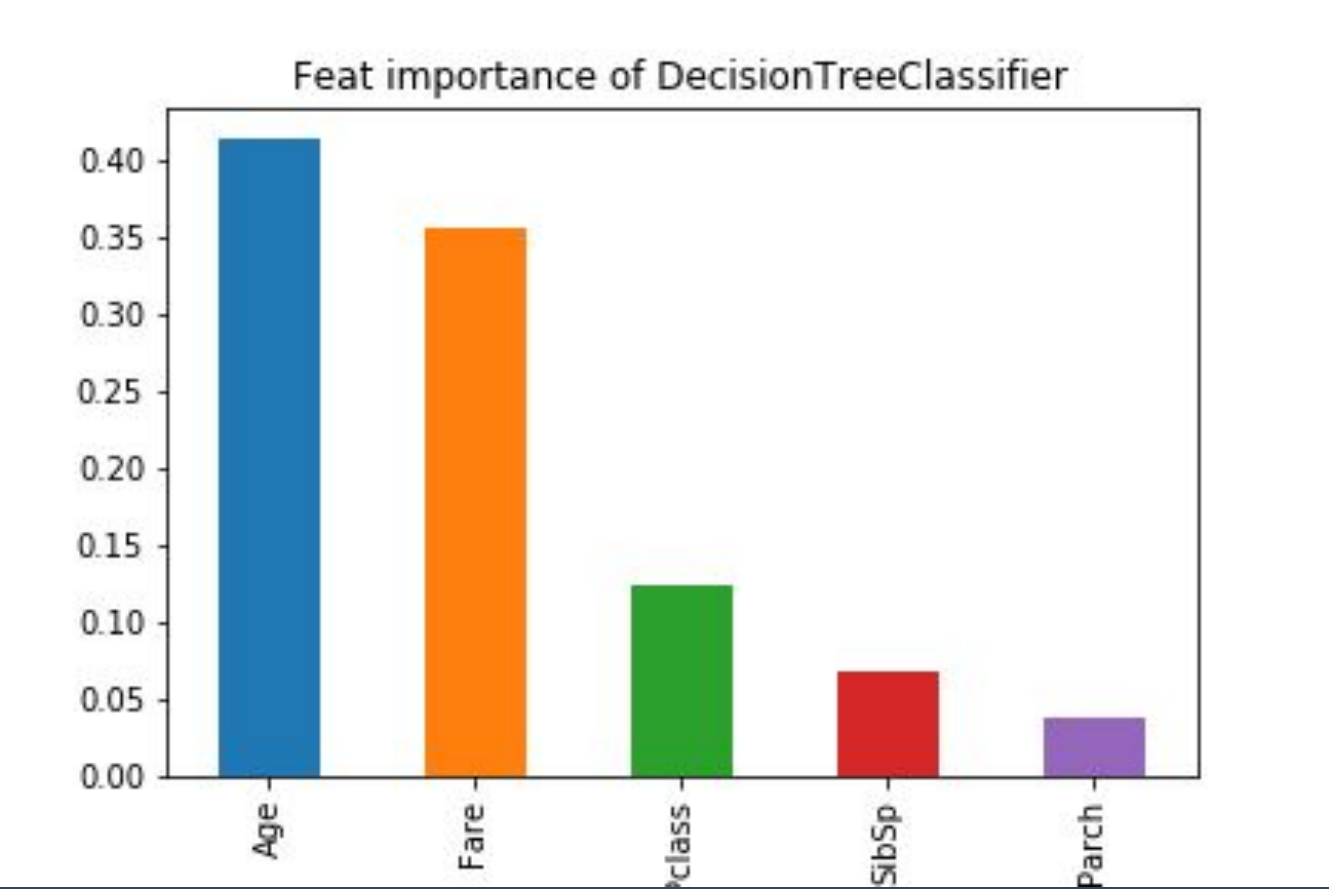
- Método:
  - Após o treinar o modelo, verificamos a importância de cada fator no experimento.
  - Queremos obter a maior quantidade de informação;
  - Foram gerados 3 *projetos*, 1 para cada tipo de dado;
- Tipos de Dados:
  - Numérico: idade, etc;
  - Categórico: sexo e local de embarque;
  - Engenharia de dados (Tratamento): usado para número da cabine ~> nulo = 0, não nulo = 1. Ou seja, se comprou ou não com cabine.

Metodologia - Projeto Fatorial 2^k

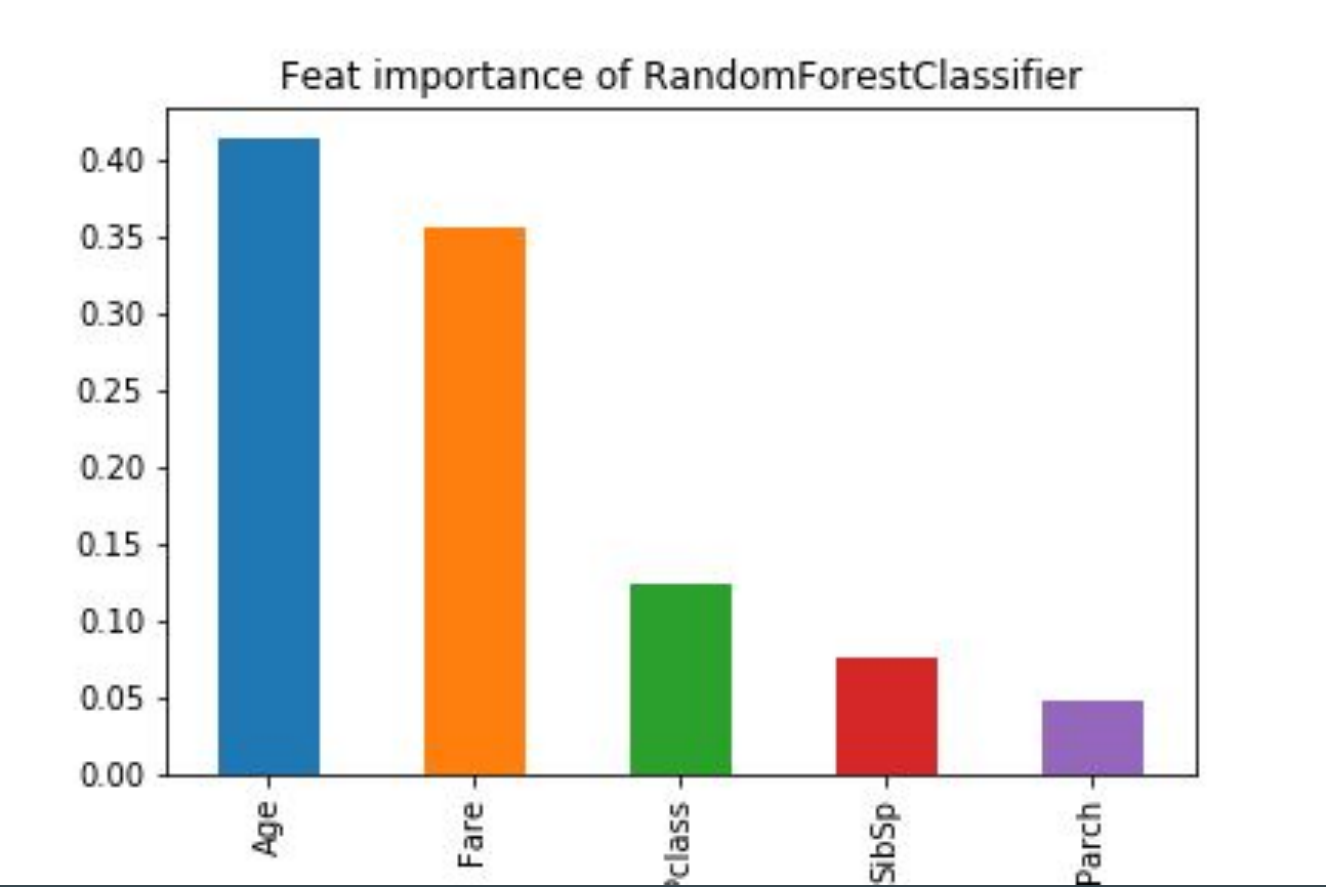
- Fatores: ao total são 9 fatores
- Níveis: sobrevivente (2), tickets (3), outros (n).
- Considerações:
  - Variáveis (fatores) se relacionam, mas não possuem relacionamento forte.
  - A importância é mostrada em gráficos, para cada fator, considerando seu tipo de dados.

Gráficos gerados por utilizando: **seaborn: statistical data visualization**

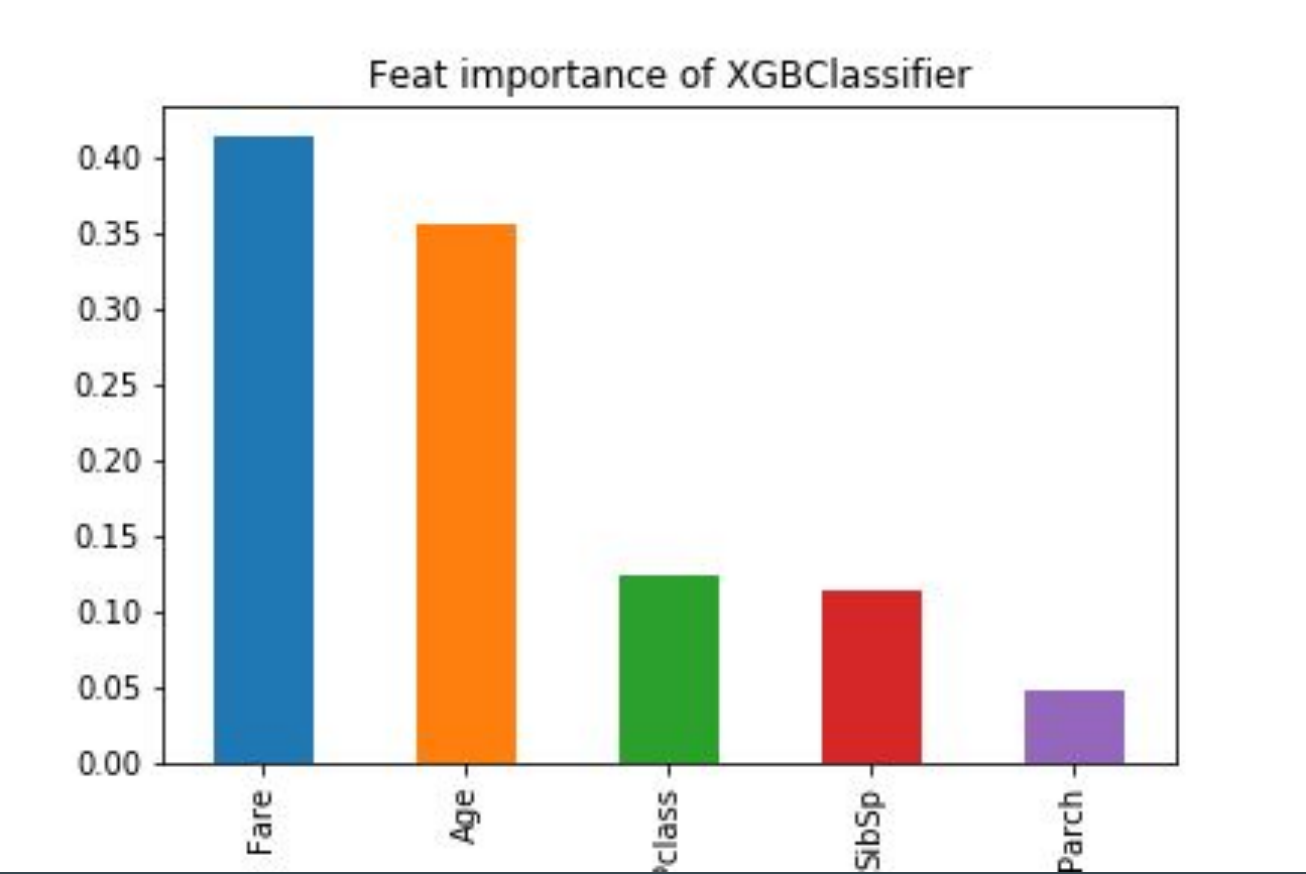
Metodologia - Projeto Fatorial 2^k - Numérico



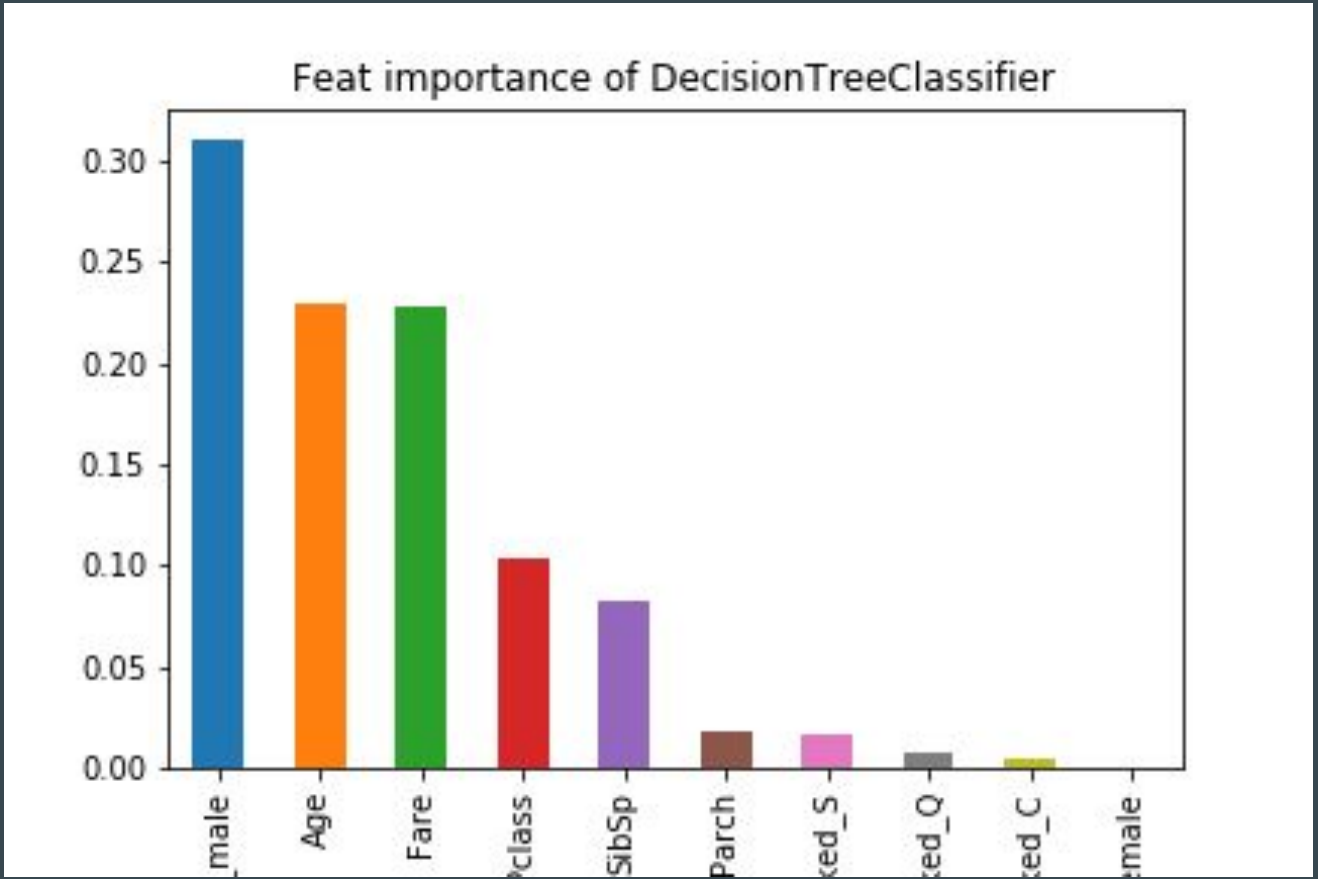
Metodologia - Projeto Fatorial 2^k - Numérico



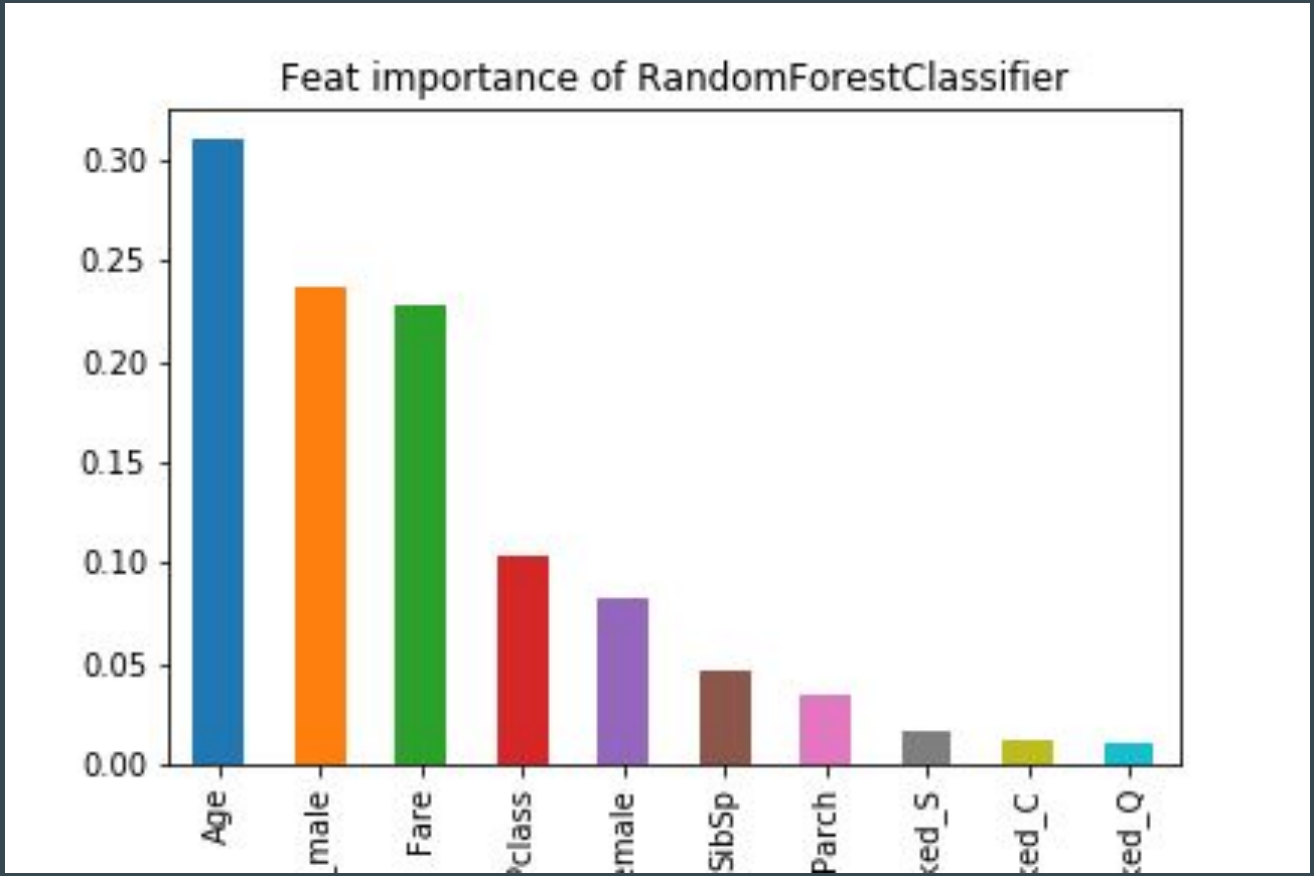
Metodologia - Projeto Fatorial 2^k - Numérico



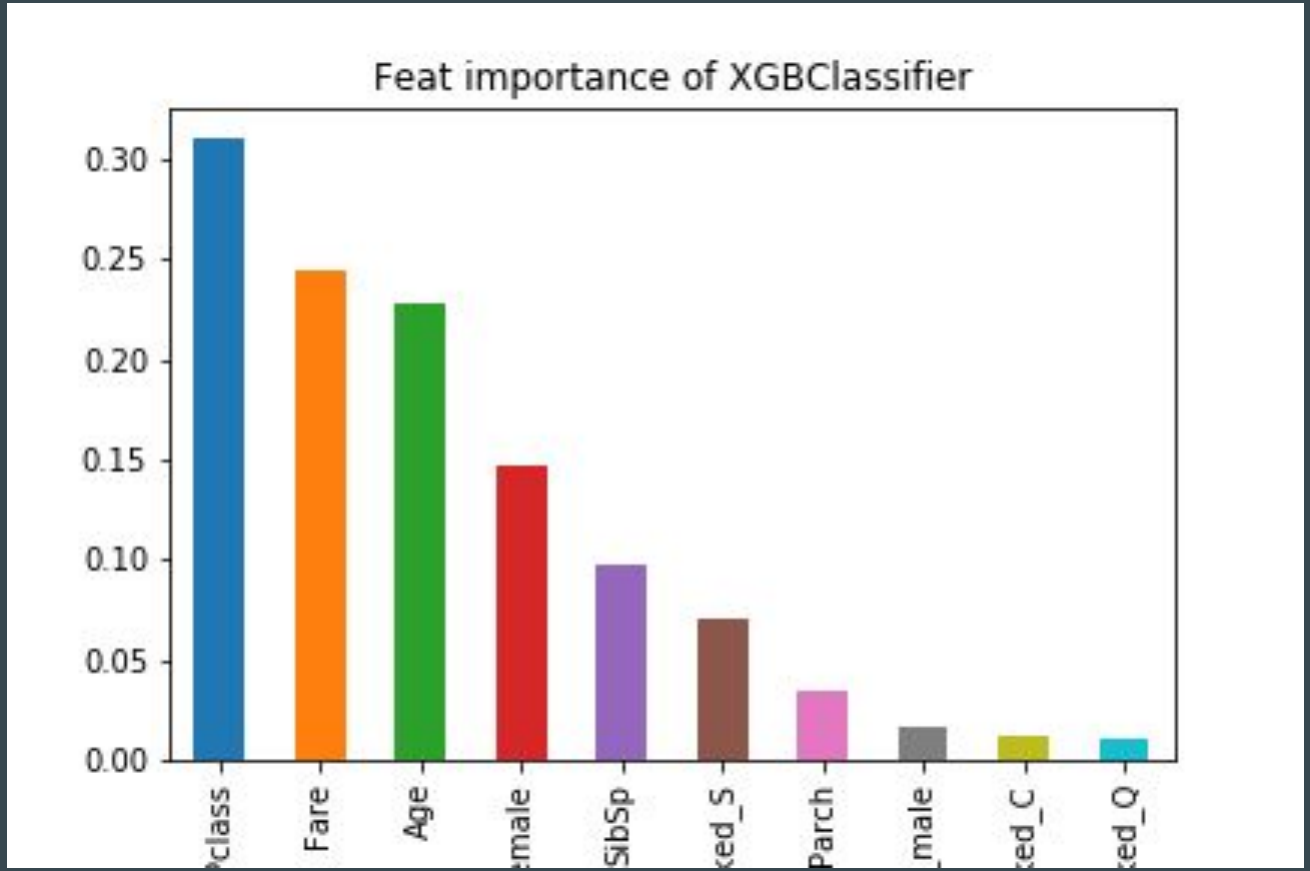
Metodologia - Projeto Fatorial 2^k - Numéricos/Categóricos



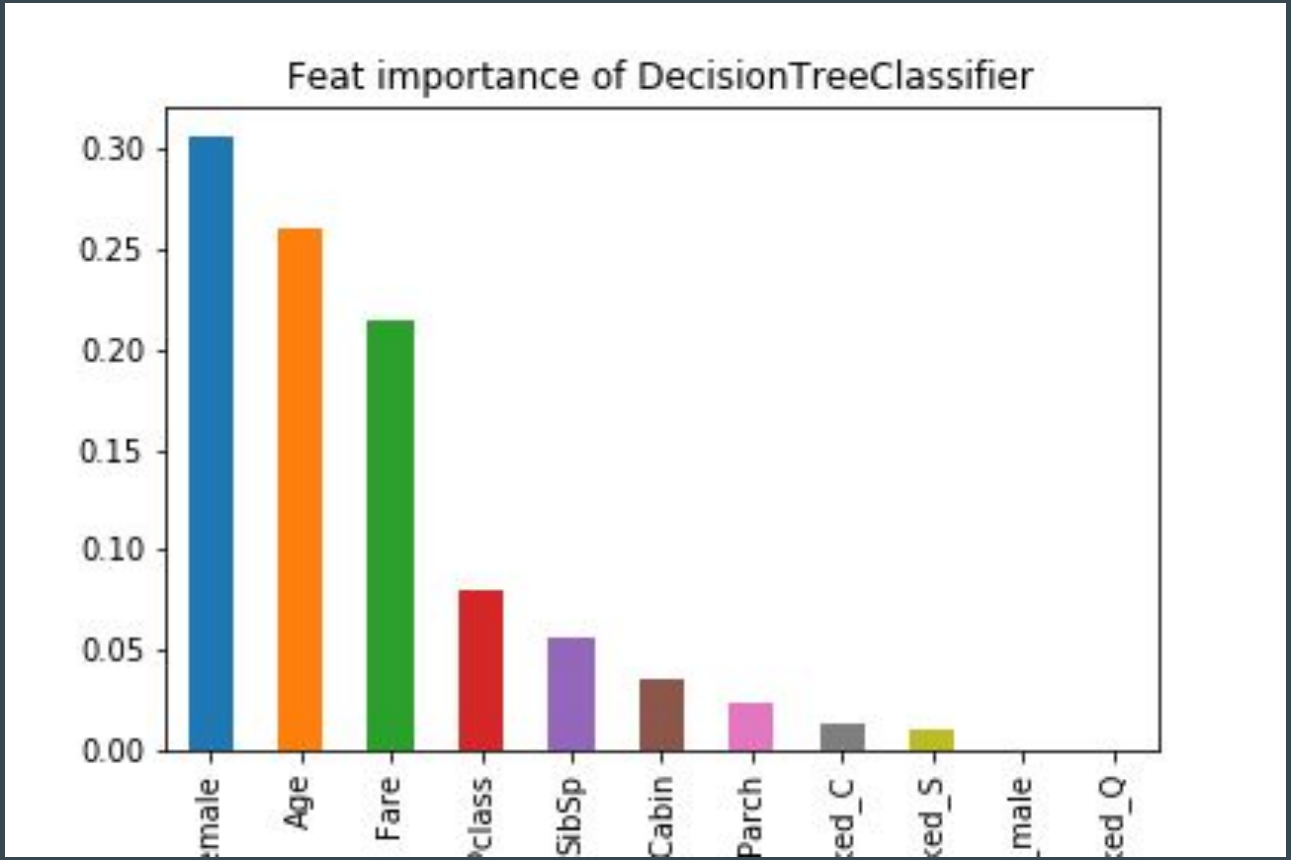
Metodologia - Projeto Fatorial 2^k - Numéricos/Categóricos



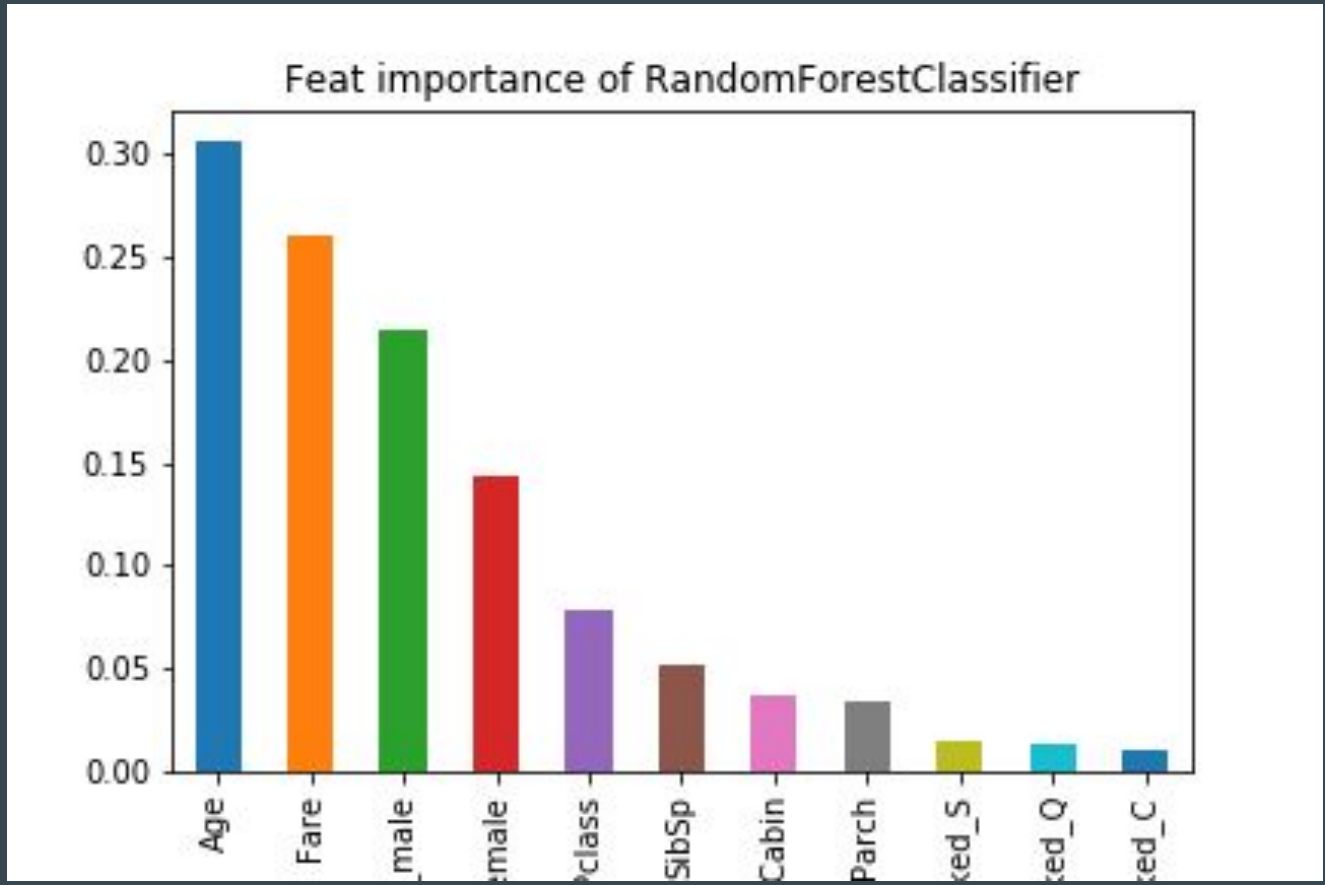
Metodologia - Projeto Fatorial 2^k - Numéricos/Categóricos



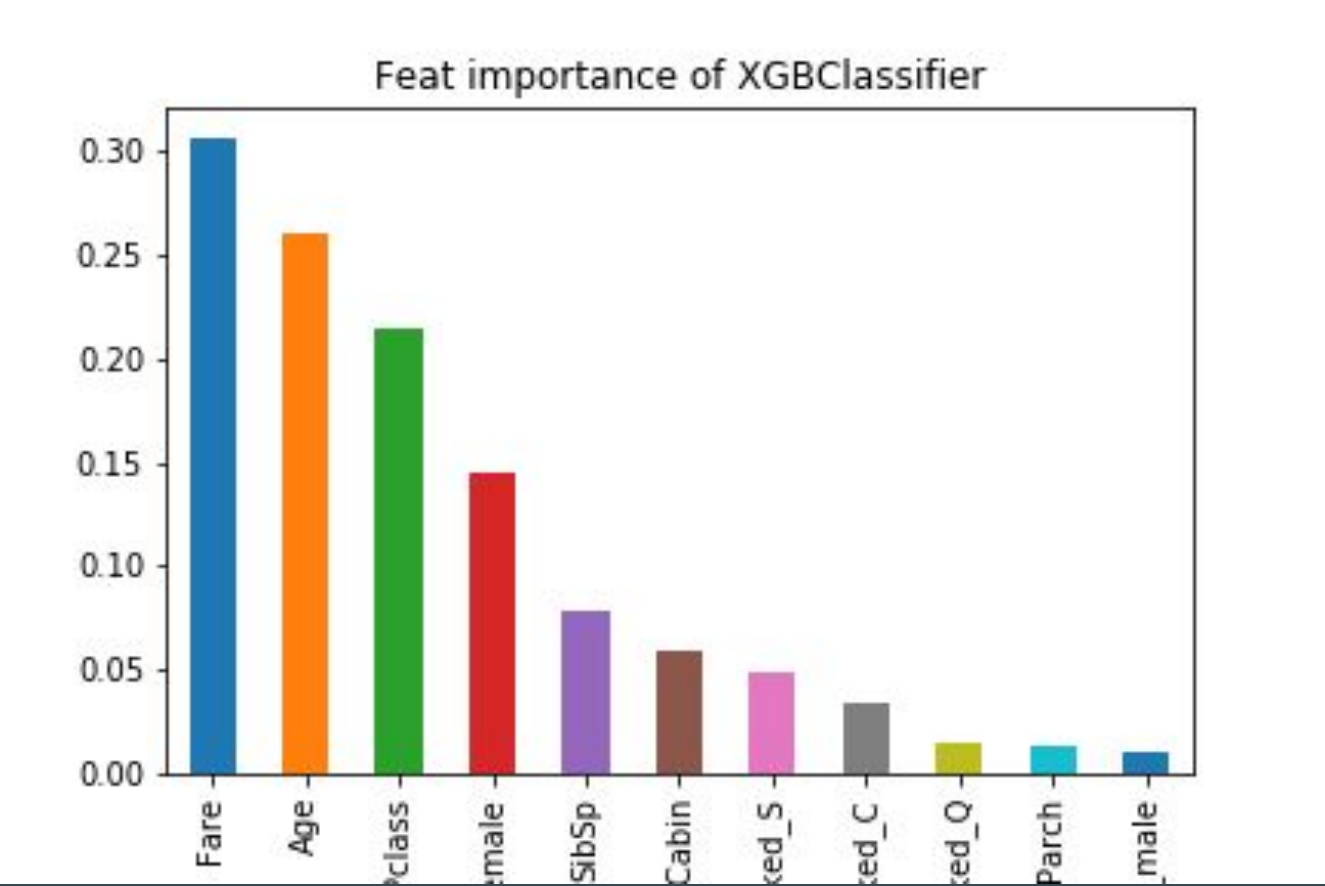
Metodologia - Projeto Fatorial 2^k - Engenharia de Dados



Metodologia - Projeto Fatorial 2^k - Engenharia de Dados



Metodologia - Projeto Fatorial 2^k - Engenharia de Dados



Resultados e Conclusão

Dos três conjuntos de dados:

- Árvore de Decisão;
- Classificação XGB;
- Floresta Aleatória.

A melhor acurácia foi gerada utilizando o terceiro conjunto de dados (flag da cabine) utilizando florestas aleatórias, onde conseguimos 82,83% de acurácia.