Universidade de Brasília

Departamento Ciência da Computação

Computação Experimental

Autores:

Antônio Henrique de Moura Rodrigues 15/0118236

Maria Fernanda do Carmo Oliveira 14/0153641

Rafael Dias Silveira 14/0030433

Yan Victor dos Santos 14/0033599

Professora:

Genaína Nunes Rodrigues

Brasília

4 de Julho de 2018

Conteúdo

1 Introdução

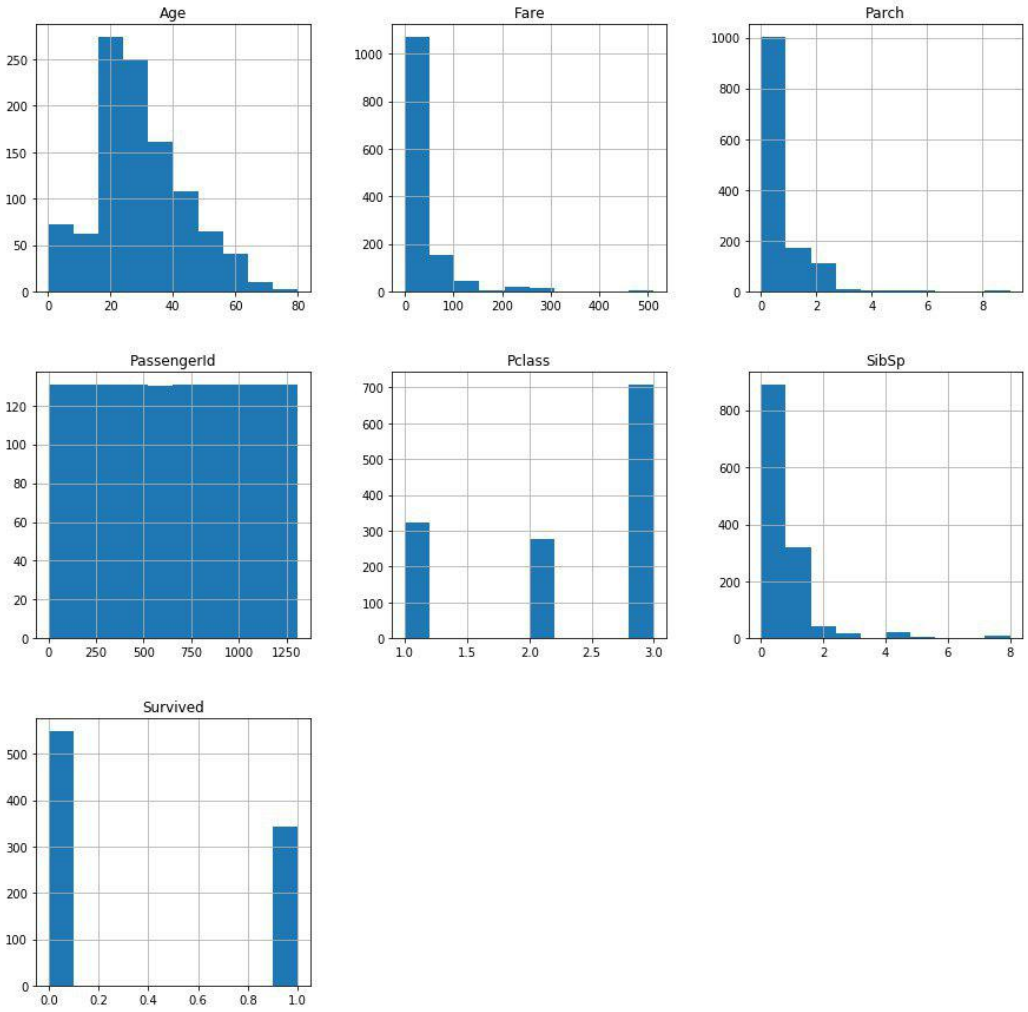
O trágico acidente do Titanic teve repercussão mundial cuja discussão dura até os dias atuais. Diversas famílias sofreram perdas lastimáveis. Após as buscas pelos sobreviventes, foi liberado o número de sobreviventes, esses dados até hoje percorrem a internet. Portanto, estes dados serão usados como objetos principais deste trabalho, com o objetivo de analisá-los para encontrar padrões de fatores que tiveram influência no resultado final do trágico cenário do Titanic: sobreviência ou não sobrevivência dos passageiros. Como exemplo de perguntas, temos: existiam muitos familiares a bordo? Quem pagou mais caro para gozar do luxo de ter uma cabine teve mais chance de sobrevivência? Tudo isso poderá ser visualizado por meio dos gráficos disponibilizados durante as seções a seguir.

Foi com essa motivação que esse banco de dados foi selecionado para ser trabalhado, assim, foram obtidas as seguintes informações:

Variável	**Definição**	**Valor**
survival	Sobrevivente	0: Não 1: Sim
pclass	Classe do ticket	1: 1st 2: 2nd 3: 3rd
sex	Gênero	
sibsp	nº de irmãos/parceiros a bordo	
parch	nº de pais/filhos a bordo	
ticket	nº do ticket	
fare	Tarifa	
cabin	nº da cabine	
embarked	Porto de embarque	C: Cherbourg Q: Queenstown S: Southampton

As informações podem ser obtidas do site Kaggle, o mesmo site que foi retirado o banco de dados a ser trabalhado durante esse projeto.

Com esse dicionário dos dados e o banco de dados, conseguimos gerar os seguintes histogramas para começar a análise:



Os experimentos realizados, baseando-se no uso de *Machine Learning*, com bibliotecas de *Python*, permitiram a geração de modelos para analisarmos os dados. Portanto, as próximas seções dispõem o uso de métodos de análise para verificarmos a importância de cada fator para o experimento sobre os passageiros.

2 Fundamentação Teórica

2.1 Matriz de Correlação

Em uma matriz de correlação, a relação dos fatores nos retorna uma matriz n x n cuja correlação é calculada entre todas as variáveis entre si, duas a duas. Valores muito próximos de 1 tem uma forte relação entre os fatores.

Valores próximos de zero, tem pouca relação entre si. Valores negativos indica variáveis inversamente proporcionais. Para o nosso trabalho, utilizamos os seguintes fatores: identidade do passageiro, idade, sobrevivência ou não sobrevivência, a classe social, preço da passagem, e duas divisões para família uma considerando irmãos e cônjuges, outra pais e filhos.

2.2 Regressão Linear

Análise de regressão linear tem como objetivo descrever a relação entre duas variáveis. Esta recebe o nome de linear porque tenta aproximar essa relação a uma função de primeiro grau, ou seja, uma reta.

É utilizado regressão linear para estimar valores de uma variável com base nos valores conhecidos de outra, tentando assim, prever os valores futuros da variável em questão.

A variável que se deseja fazer estimativa é chamada de variável dependente, ou y. Ela depende de outra variável para poder ser estimada. A variável que gera a previsão, e consequentimente a correlação, é chamada de variável independente, ou x.

Para gerarmos uma reta de regressão linear, utiliza-se as seguintes funções:

$$y = ax - b$$

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

$$b = \frac{1}{n}(\sum y_i - a\sum x)$$

2.3 Projeto Fatorial - Importância de Fatores

O projeto fatorial auxilia na escolha de fatores, se baseando na importância deles, para decidir quais são considerados bons para aplicarmos em nossos testes e pesquisas. A escolha destes fatores se tornam relevantes ao evitarem alto custo de experimentos. O objetivo do projeto é enriquecer a quantidade de informações que se pode obter a partir de um conjunto de dados.

Baseando-se na ideia de fatores mais relevantes, este trabalho analisa quais fatores foram considerados importantes, de forma a encontrar padrões na população, para passageiros que sobreviveram/morreram no acidente do Titanic. A utilização deste método, para fins didáticos, nos permitiu uma melhor visualização dos padrões citados sobre fatores que podem ser considerados influentes para a sobrevivência de um passageiro. Portanto, a quan-

tidade de fatores mostrou-se suficiente para a aplicação de testes e geração dos modelos.

3 Metodologia

3.1 Machine Learning - Titanic

Nesse projeto, vamos utilizar os dados dos passageiros do RMS Titanic para prever, com base nos seus dados, se um passageiro irá sobreviver ou não ao naufrágio. Os dados encontram-se na plataforma Kaggle.

3.2 Matriz de Correlação

Aqui está a matriz n X n. Com os valores de correlação calculados considerando as influências de todas as variáveis em todas as variáveis, agrupando-as de duas em duas.

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

Matriz de correlação:

Out[120]:

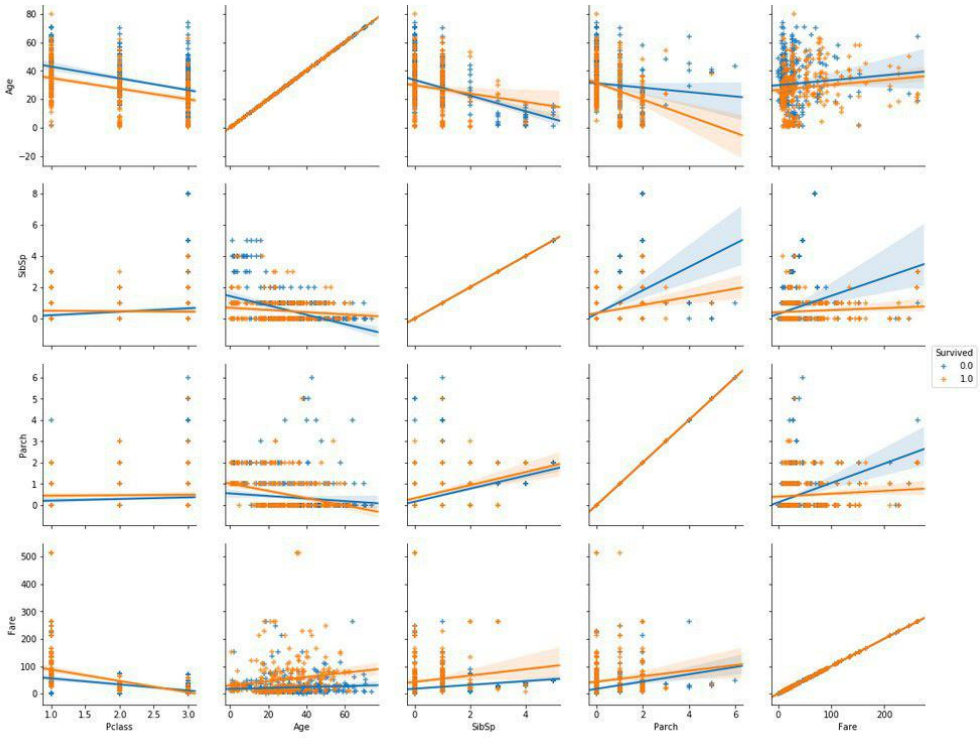
	PassengerId	Survived	Pclass	Age	SibSp	Parch	Fare
PassengerId	1.000000	-0.005007	-0.038354	0.028814	-0.055224	0.008942	0.031428
Survived	-0.005007	1.000000	-0.338481	-0.077221	-0.035322	0.081629	0.257307
Pclass	-0.038354	-0.338481	1.000000	-0.408106	0.060832	0.018322	-0.558629
Age	0.028814	-0.077221	-0.408106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

Como podemos observar na figura acima, as duas variáveis referentas à parentesco tem relação mais forte. Por outro lado, a relação de quase todas as variáveis com a identidade são mais fracas, pois o fator identidade é apenas um número arbitrário de identificação e não faz parte do problema original.

Como exemplo de análise sobre a tabela, temos muitas correlações negativas, ou seja várias relações com variáveis inversamente proporcionais, como a classe social e o fator de preço pago. É importante destacar que foi considerado a classificação da classe social 1 como as pessoas com renda alta, classe 2 como classe média e a classe 3 como as pessoas de baixa renda. Portanto, pessoas de baixa renda tenderam a pagar por uma passagem mais barata. Portanto, quanto maior a classe social (menor renda), menos o passageiro teve de pagar pela passagem. Todas essas informações são exemplos de análise dos dados da figura acima.

3.3 Regressão Linear

Após aplicar as fórmulas nas variáveis descritas nos dados do Titanic obtidos da matriz de correlação, da tabela gerada na seção anterior, foi gerado uma matriz de regressões, mostrando diversas relações e correlações entre as variáveis quantitativas:



Com a regressão acima, para a análise quantitativa, nota-se:

- Os cruzamentos entre as mesmas váriaveis geram uma reta equivalente a reta de regressão, isso ocorre devido ao fato de uma variável ter um alto índice de relação com ela mesma;
- As variáveis que possuem valores em intervalo contínuo possuem uma dispersão mais aleatória (como é o caso da regressão entre os valores da tarifa e da idade);
- As variáveis que são categoricas seguem um padrão, mesmo fugindo da reta de regressão, isso ocorre devido a associação que é feita entre elas e um número inteiro para análise dos dados;
- É notado um alto grau de dispersão, isso implica que utilizando desse método não prova que as variáveis possuem relação notável.

3.4 Projeto Fatorial

Para verificar quais fatores possuem mais influência no projeto, utilizamos o método de projeto fatorial.

Com base na disciplina, cada variável que descreve os passageiros do Titanic é considerada como sendo fator que influencia em sua sobrevivência. Cada fator pode possuir um ou mais níveis. Dada a complexidade de se trabalhar com muitos níveis, utilizamos o suporte da biblioteca *scikit-learn* para calcular a importância de cada grupo de dados. Os gráficos foram gerados por meio da ferramenta ***seaborn: statistical data visualization***.

Para os gráficos gerados, foi necessário subdividir os fatores em conjuntos. O primeiro conjunto engloba fatores numéricos, tais como *idade*, *parch*, *ticket*, etc. O segundo conjunto de dados engloba fatores categóricos, tais como *sexo*, e *embarked*. Por fim, foi necessário tratar uma coluna separada para a cabine. Portanto, foram gerados 3 tipos de gráficos.

Para cada conjunto de dados, geramos 3 modelos (Árvore de Decisão, Floresta Randômica, Classificação XGB), com o mesmo objetivo, para uma melhor análise dos dados gerados. A variedade de modelos permitiu verificar a acurácia dos testes para cada projeto criado.

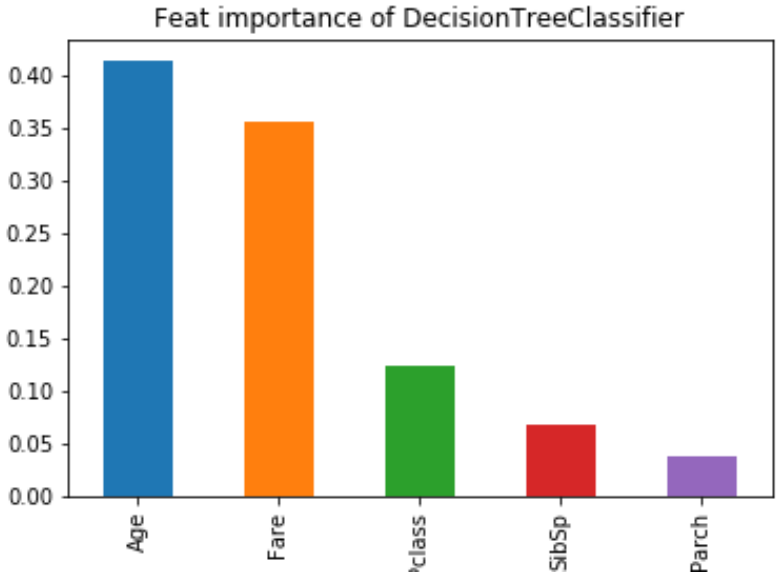


Figura 1: Conjunto de Dados Numéricos 1.1.

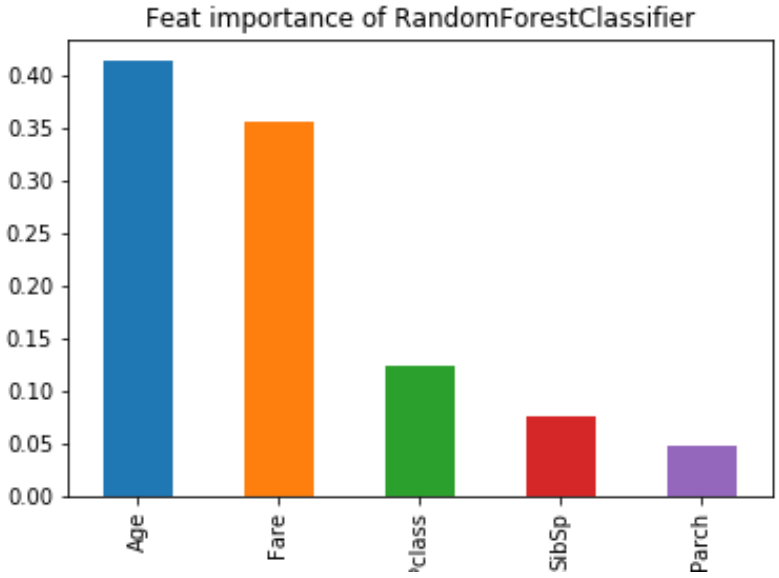


Figura 2: Conjunto de Dados Numéricos 1.2.

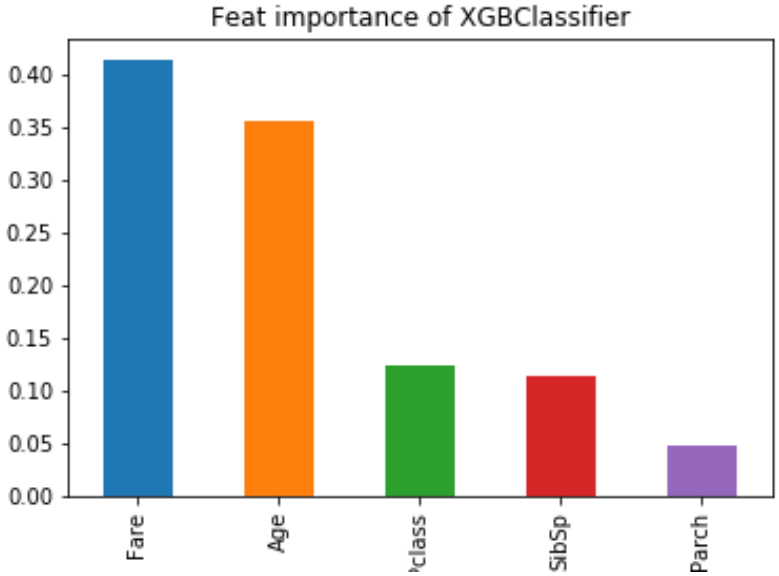


Figura 3: Conjunto de Dados Numéricos 1.3.

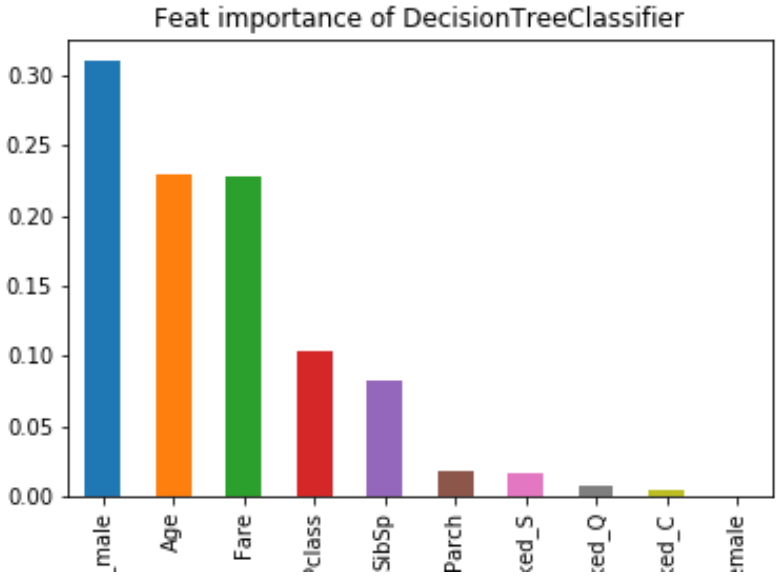


Figura 4: Conjunto de Dados Categóricos 2.1.

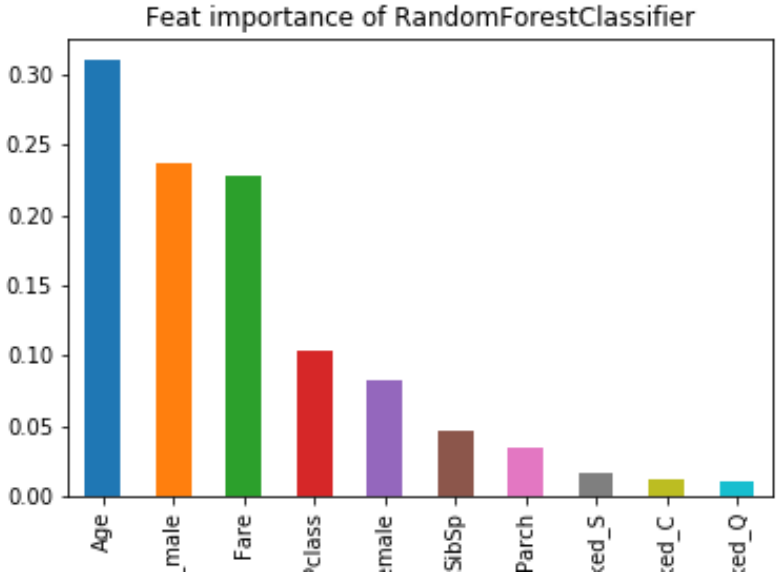


Figura 5: Conjunto de Dados Categóricos 2.2.

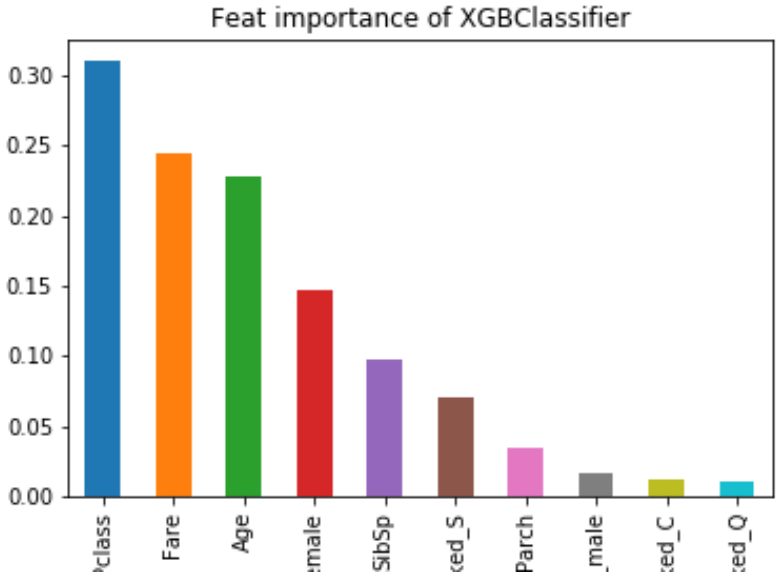


Figura 6: Conjunto de Dados Categóricos 2.3.

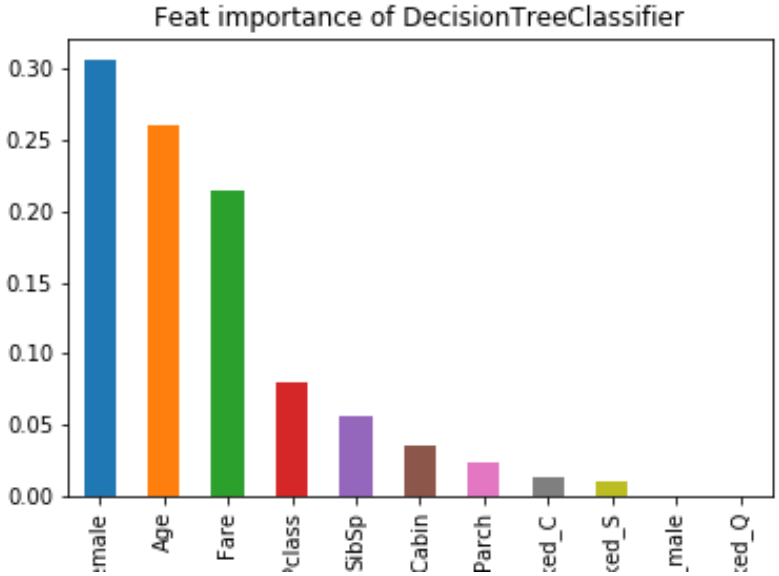


Figura 7: Engenharia de Dados 1.3.

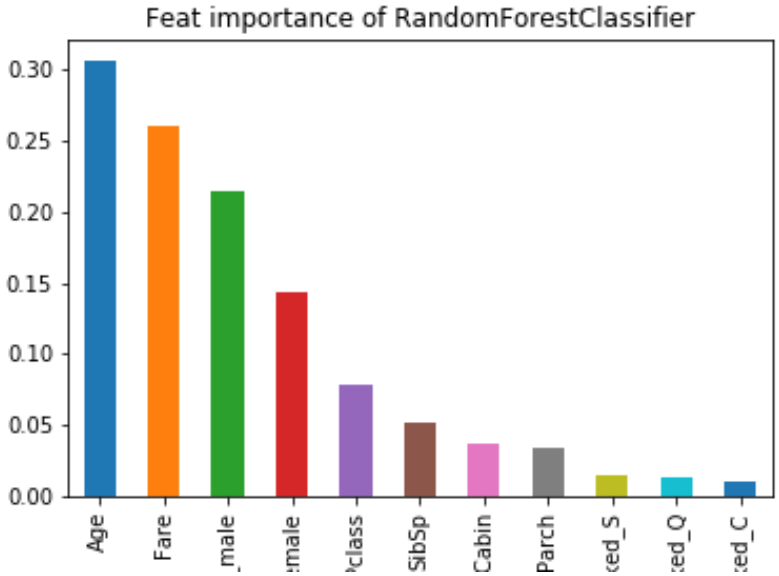


Figura 8: Engenharia de Dados 2.3.

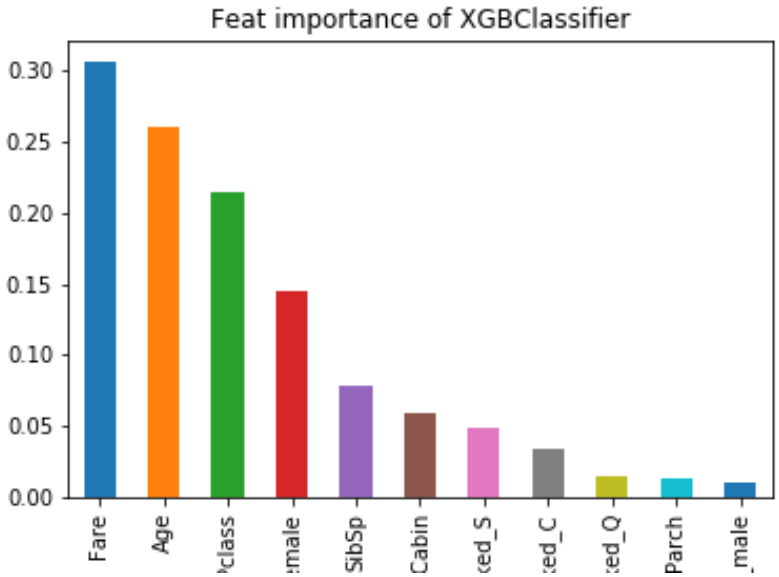


Figura 9: Engenharia de Dados 3.3.

Como podemos observar nas Figuras acima (figuras 1, 2, 3, 4, 5, 6, 7, 8 e 9), os gráficos representam o quanto cada fator tem influência sobre os resultados esperados, no caso, se o passageira irá ou não sobreviver. A idéia

não é testar para novos passageiros hipotéticos, e sim encontrar um padrão tal que o fator x tenha sido frequentemente comum dentre os passageiros sobreviventes.

Na análise das figuras, podemos observar que para o conjunto categórico, ser do sexo masculino forneceu grande impacto para a conclusão de que o passageiro não sobreviveria. As variáveis que tiveram mais impactos, segundo os modelos, foram: sexo masculino, classe, valor pago (*fare*) e idade. Isso revela, para as perguntas feitas no início do trabalho, que o valor pago influenciou bastante na sobrevivência dos passageiros, assim como o sexo e a classe social dos indivíduos.

Vale ressaltar que, para os gráficos gerados, não estão sendo usados todos os fatores em todos os modelos. Apenas destacamos alguns principais para um determinado tipo de modelo. Por exemplo, no conjunto de dados numéricos, gerado pelo XGB, a idade teve grande influência para os resultados.

Sua análise demonstra o quanto se torna mais legível a verificação de influência dos fatores para o projeto final. A utilidade desse modelo sugere eficácia e de grande contribuição para o processo de análise de dados.

4 Resultados e Discussão

Para gerar o melhor modelo preditivo, foram utilizados três modelos de categorização (Árvore de Decisão, Floresta Aleatória e Classificação XGB). Para treinar os modelos, foram utilizados três conjuntos de dados: um somente com os valores númericos, um com valores numéricos e valores categóricos *dummificados* e o último foi gerado uma *flag* para identificar se o passageiro tinha ou não cabine. Após treinar e testar os modelos, verificou-se que a Floresta Aleatória no último conjunto de dados, com uma acurácia de 82,83%.

Apesar de ter conseguido uma acurária relativamente alta, o modelo ainda consegue um desempenho melhor com mais *data engineering* em cima das informações e *tunar* os hiperparametros dos modelos.

5 Conclusão

Por meio deste trabalho, foi possível observar padrões sobre os dados dos passageiros e quais características eram mais frequentes em passageiros sobreviventes. A utilização de métodos sobre o projeto de *Machine Learning* com a utilização de dados de passageiros do Titanic permitiu uma melhor compreensão sobre análise de dados. Por este motivo, cada modelo se mostrou eficaz para ao qual foi proposto.

A utilização de ferramentas gráficas para análise de dados nos permitiu uma aprendizagem mais assertiva a respeito de como devemos analisar uma amostra ou população.

Segue o link do vídeo da apresentação que foi colocada no Youtube: Vídeo.

6 Referência

1. https://www.kaggle.com/c/titanic/data?

2. https://www.youtube.com/watch?v=QXl0EEDCRuY

3. https://www.cse.wustl.edu/ jain/books/perfbook.htm