***Isac Fernandes Costa***

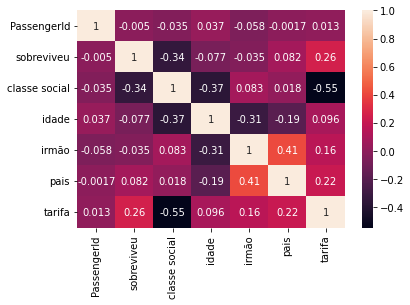
***João Felipe Nunes***

**RELATÓRIO: PROJETO DE MACHINE LEARNING**

O presente relatório tem a intenção de documentar o projeto realizado no início do dia 11 de novembro de 2021 e término no dia 16 de novembro deste ano. Nosso projeto tem por objetivo construir o melhor modelo de machine learning capaz de predizer os dados do naufrágio do navio Titanic, tendo seu foco na predição se a pessoa morreu ou sobreviveu. Estes dados públicos foram retirados de um dataset no aplicativo Dropbox.

Tecnologias utilizadas: Google Colab (análise de dados) e Orange Canvas 3.30.2 (criação do modelo de machine learning).

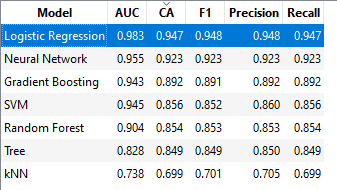
O projeto foi dividido em duas etapas: à análise de dados e a criação do melhor modelo de machine learning para predição de possíveis mortes ou sobrevivências. Nesta primeira etapa o projeto foi desenvolvido na plataforma do Google Colab com a linguagem de programação Python 3.10.0, juntamente com suas bibliotecas facilitadoras: pandas 1.2.2 (para exploração dos dados no dataset e cálculos estatísticos), metplotlib 2.2.4 (para visualização de dados em gráficos) e seaborn 0.11.2 (para visualização de dados em gráficos). O objetivo desta análise é validar a existência de correlação entre a pessoa ter sobrevivido com as outras variáveis (classe social, sexo, idade, nome, porto de embarque, cabine, tarifa, código do bilhete, quantidade de pessoas a bordo e quantidade de pais a bordo), para determinar se essas colunas são essênciais ou não para a predição do modelo de machine learning, visando a maior taxa de acertividade e precisão. Chegamos a conclusão de que a correlação de sobrevivência com a classe social, sexo e custo pago na tarifa, são importantes e tem grande influência na porcentagem de acertividade do modelo. Todas as demais colunas mostram-se não essenciais para o aprendizado. Abaixo, temos o resultado da correlação feita com o Pandas e o Seaborn:



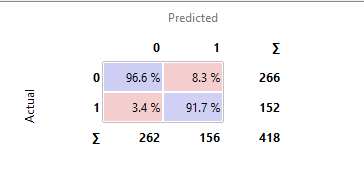
*(Gráfico de correlação gerado pelo seaborn no Google Colab)*

Após ser concluída a análise dos dados no Google Colab com Python, nós criamos dois modelos no Orange Canvas onde o que difere um do outro são as colunas selecionadas para a aprendizagem. No modelo 1 mantemos as colunas de sexo, classe social, custo pago na tarifa, irmãos a bordo, idade e pais a bordo; já no modelo 2 deixamos apenas o sexo, custo pago na tarifa e classe social. Em ambos os casos foram destinados 891 linhas de dados para o treino, e 418 linhas de dados para a validação da assertividade. Em ambos os modelos usamos os seguintes métodos de aprendizagem: KNN, SVM, Gradient Boosting, Neural Networks, Random Forest, Logistic Regression e Tree.

O modelo 1 teve seu melhor percentual de assertividade com o método de aprendizagem Logistic Regression, onde alcançou cerca de 94.7% de acurácia e 94.8% de precisão.

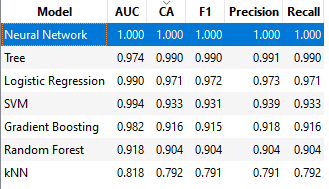


*(Acurácia-”CA” e precisão-”Precision” do modelo 1)*

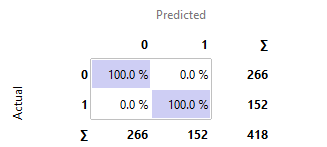
**

*(Matriz de Confusão onde 1-Sobreviveu e 0-Morreu, modelo 1)*

Já o modelo 2 teve seu melhor percentual de assertividade com o método de aprendizagem Neural Network, onde alcançou cerca de 100.0% de acurácia e 100.0% de precisão.



*(Acurácia-”CA” e precisão-”Precision” do modelo 2)*

**

*(Matriz de Confusão onde 1-Sobreviveu e 0-Morreu, modelo 2)*

**Conclusão**

Tendo em vista todos os dados apresentados acima, podemos afirmar que o projeto satisfaz o objetivo de construir o melhor modelo preditivo possível, tendo alcançado um percentual de assertividade e precisão de 100% em suas predições de teste. Nós supomos que os fatores mais determinantes para o alcance desse resultado, foram a remoção de algumas colunas que estavam interferindo na capacidade de aprendizagem do modelo, pois foi comprovado na análise de dados a correlação fraca entre algumas colunas no dataset. Outro fator determinante para a montagem do melhor modelo preditivo, foi a utilização do método de aprendizagem chamado de ‘Neural Network’, que foi o único dentre os sete modelos que atingiu resultados tão expressivos.