Implementação de classificador para determinar probabilidade de sobrevivência no acidente do Titanic através do algoritmo de Bayes ingênuo

Alansidney da S. Júnior
Layse Roberta da Silva Araujo
Marco Aurélio Ferreira Santana

Estudantes de graduação em Engenharia da Computação
do Centro de Informática da Universidade Federal de Pernambuco

22 de Março de 2022

I. Objetivos

Este projeto tem como objetivo o estudo, treinamento, construção, e validação de um modelo classificador utilizando o algoritmo de Bayes ingênuo que seja capaz de receber informações de uma pessoa e concluir se a mesma teria sobrevivido ou não ao desastre do Titanic. Para atingir tal objetivo será utilizado como linguagem de programação Python e como ferramentas para implementação a biblioteca scykit-learn e a plataforma Google Colab. Com a finalidade de criar um programa capaz de fornecer dados que nos ajudem a entender o que aconteceu ao navio e qual o perfil de pessoa provavelmente sairia viva deste.

II. Justificativa

Em 14 de abril de 1912, o navio RMS Titanic naufrágou depois de se chocar com um iceberg. Hoje, quase 110 anos depois, o naufrágio ainda desperta curiosidade do mundo. Apesar disso, de acordo com a BBC, durante esse tempo os acidentes navais tiveram uma diminuição de cerca de oitenta e cinco por cento.

O desastre do Titanic é um marco histórico mundial que rende filmes, artefatos culturais e curiosidade dentro do imaginário social do mundo até os tempos atuais. Várias suposições são feitas sobre o processo de salvamento das pessoas dentro do navio. Será que mulheres e crianças realmente tiveram prioridade no salvamento? Os passageiros tinham prioridades sobre a tripulação? Essas perguntas moveram o time para achar um perfil de quem seria salvo e desenvolver um classificador para saber se uma pessoa provavelmente seria salva ou não.

III. Metodologia

A metodologia utilizada para a implementação do programa consistirá de uma execução sequencial de várias etapas, dentre elas, a análise e tratamento da base de dados, onde serão selecionadas as features mais importantes e a limpeza de dados nulos e inválidos. Em seguida será feita uma visualização e separação de amostras usadas tanto na fase de treinamento quanto na fase de teste do algoritmo, com o auxílio de métodos da biblioteca scykit-learn para manipular os dados previamente tratados e fornecer os retornos de funções necessários para a implementação do modelo, e por fim, ser capaz de predizer com certa acurácia, a partir dos dados de entrada, se uma pessoa seria capaz de sobreviver ao desastre de acordo com suas características de sexo, classe social e idade.

No relatório final, em formato de Jupyter Notebook, será abordado por meio de gráficos e métricas estatísticas como o é o desempenho do nosso algoritmo nos melhores e piores casos, juntamente com os pontos fortes e fracos do mesmo, mostrando

o que se pode ser trabalhado para sua melhor performance.

Dados da base utilizada:

Variável	Descrição	Valores
survival	Se foi um sobrevivente do desastre	0 = Não, 1 = Sim
pclass	Qual classe o passageiro pertencia	1 = 1ª, 2 = 2ª, 3 = 3ª
sex	Sexo do passageiro	male, female
Age	Idade em anos	inteiro
sibsp	Nº de irmãos/cônjuges a bordo do titanic	inteiro
parch	Nº de pais/filhos a bordo do titanic	inteiro
ticket	Nº do Ticket	String
fare	Tarifa do passageiro	Ponto flutuante
cabin	Nº da cabine do passageiro	String
embarked	Indica o porto de embarcação do passageiro	C = Cherbourg, Q = Queenstown, S = Southampton

IV. Cronograma de Atividades

Início	Fim	Atividade
10/03/2022	14/03/2022	Pesquisa e escolha da Base de dados
15/03/2022	17/03/2022	Decisão do tema e título do projeto
18/03/2021	22/03/2022	Planejamento e construção da Proposta de projeto
23/03/2022	30/03/2022	Análise da base de dados e estudo das ferramentas scikit-learn e Google Colab
31/03/2022	07/03/2022	Separação dos dados em grupo para treinamento e grupo para teste
08/04/2022	13/04/2022	Construção do modelo proposto usando o classificador de Bayes ingênuo
14/04/2022	21/04/2022	Fase de testes
22/04/2022	04/05/2022	Análise de Elaboração do relatório de projeto
05/05/2022	10/05/2022	Estudo crítico dos resultados obtidos e gravação da apresentação

V. Referencias

BBC NEWS. Cem anos após desastre do Titanic, acidentes com navios têm redução de 85%
https://www.bbc.com/portuguese/noticias/2012/04/120413_titanic_riscos_atuais_acidentes_jp

KAGGLE. —Titanic - Machine Learning from Disaster
https://www.kaggle.com/c/titanic/data?select=test.csv

Machine Learning in Python. Scikit-learn
https://scikit-learn.org/stable/

Paulo Vasconcellos. O que o Naufrágio do Titanic nos ensina até hoje — Data Science Project
https://paulovasconcellos.com.br/o-que-o-naufrágio-do-titanic-nos-ensina-até-hoje-data-science-project-2fea8ff1c9b5

G1. Titanic foi 'exceção' no perfil dos mortos em naufrágios, diz pesquisa
https://g1.globo.com/ciencia-e-saude/noticia/2012/07/titanic-foi-excecao-no-perfil-dos-mortos-em-naufragios-diz-pesquisa.html

Thiago G Santos. Google Colab: o que é e como usar?
https://www.alura.com.br/artigos/google-colab-o-que-e-e-como-usar

How to Get Started with Kaggle's Titanic Competition — Kaggle
https://www.youtube.com/watch?v=8yZMXCaFshs

On the possibility of short-term traffic prediction during disaster with machine learning approaches: An exploratory analysis
https://www.sciencedirect.com/science/article/abs/pii/S0967070X20304194