Construção de um Modelo Preditivo usando Python

Um modelo preditivo desenvolvido para o curso Linguagens de Programação

Gustavo de Oliveira Freitas
123608640
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brasil
frtgusta.20231@poli.ufrj.br

Henrique Pousa da Rocha Fragoso
123239687
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brasil
henriqueprfragoso.20231@poli.ufrj.br

João Ricardo Monteiro Scofield Lauar
123125642
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brasil
joaoscofield.20231@poli.ufrj.br

João Vítor Pinto Vizeu
123086131
Universidade Federal do Rio de Janeiro
Rio de Janeiro, Brasil
jvpvizeuufrj.20231@poli.ufrj.br

***Abstract*—Este artigo delineia um projeto de análise de dados com ênfase na construção de um modelo preditivo utilizando Python. Baseado nos dados dos sobreviventes do desastre do Titanic, o propósito do modelo consiste em predizer a sobrevivência de um indivíduo.**

***Index Terms*—análise de dados, Python, notebook, Titanic**

I. Introdução

A análise de dados e a construção de modelos preditivos são fundamentais na extração de informações valiosas de conjuntos de dados complexos. Este projeto concentra-se na aplicação de técnicas de aprendizado de máquina para modelar o perfil dos sobreviventes do Titanic, utilizando variáveis como idade, classe e gênero como *features*.

O naufrágio do Titanic em 1912 fornece um cenário único para explorar as relações entre diversas variáveis e a probabilidade de sobrevivência dos passageiros. A análise dessas relações não apenas amplia o entendimento sobre os fatores que influenciam a sobrevivência em um contexto específico, mas também destaca a eficácia das ferramentas analíticas modernas.

O objetivo principal deste projeto é desenvolver um modelo preditivo capaz de discernir padrões nos dados e prever se um passageiro teria sobrevivido ao desastre. Algoritmos de aprendizado de máquina serão empregados para treinar o modelo com base em uma base de dados que inclui informações relevantes sobre os passageiros.

Ao adotar uma abordagem metodológica que envolve desde a preparação e limpeza dos dados até a avaliação do desempenho do modelo, esperamos identificar variáveis significantes e construir um modelo robusto.

II. Visão Geral do Projeto

A. Objetivo

O propósito principal deste projeto é analisar e modelar o perfil dos sobreviventes do naufrágio do Titanic por meio de técnicas de aprendizado de máquina. Especificamente, busca-se construir um modelo preditivo utilizando dados relacionados a características como classe, idade, gênero e outras variáveis da base de dados disponível.

Este trabalho não foi financiado por nenhuma agência específica.

B. Metodologia

A metodologia utilizada no projeto é baseada no método universal para resolução de problemas com análise de dados, o esquemático feito por Silva [1] elucida as etapas:

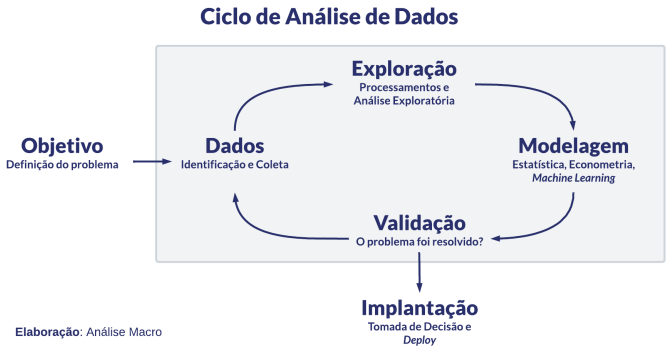

Fig. 1. Metodologia em Análise de Dados.

1) **Dados:** Coletar dados relevantes para a realização do objetivo definido, utilizando ferramentas como nesse caso, Python.
2) **Exploração:** Organizar e analisar dados para identificar padrões e anomalias. Para isso, são utilizadas ferramentas estatísticas, de programação e visualização, como pandas, seaborn e matplotlib.
3) **Modelagem:** Experimentar soluções baseadas em dados, como análises estatísticas e modelos preditivos de machine learning.
4) **Validação:** Garantir eficácia da solução proposta, analisando métricas de acurácia e resultados estatísticos.

Utiliza-se conhecimento em amostragem de dados, interpretação estatística e programação.

5) **Implantação:** Etapa direcionada ao mercado de trabalho, cujo foco diverge do proposto pelo trabalho.

A fim de concluir essas etapas, as bibliotecas escolhidas para a realização foram o pandas, para manipulação de dados, e matplotlib e seaborn para visualização de dados.

III. Desenvolvimento do Projeto

A. Coleta de Dados

Os dados necessários para este projeto são extraídos do conjunto de dados do Titanic, que inclui informações detalhadas sobre passageiros, como classe socioeconômica, idade, gênero, entre outras. Esse modelo foi fornecido pelo site educacional Kaggle, em uma competição de ciência de dados. Esse tipo de conjunto de dados é fornecido em arquivos .db, específicos para bases de dados como a do Titanic. Usualmente, utiliza-se a biblioteca pandas em Python para a manipulação e leitura desses arquivos.

B. Análise Exploratória

A Análise Exploratória será utilizada para compreender a natureza dos dados coletados através de técnicas estatísticas e de visualização, e dessa forma detectar padrões na estrutura dos dados. A princípio, com métodos básicos de visualização do pandas como *.head*, podemos obter a seguinte estrutura (As primeiras 5 colunas dos primeiros 5 passageiros):

TABLE I
TABELA 1.1: ESTRUTURA BÁSICA DA BASE DE DADOS.

PassengerId	Survived	Pclass	Sex	Age	SibSp
1	0	3	male	22.000000	1
2	1	1	female	38.000000	1
3	1	3	female	26.000000	0
4	1	1	female	35.000000	1
5	0	3	male	35.000000	0

A Tabela 1 é um recorte do resultado que de fato foi obtido, a respeito da visualização eficiente. A partir da Análise Exploratória, é possível obter/deduzir as seguintes informações correspondentes a cada coluna:

TABLE II
TABELA 1.2: INFORMAÇÕES DA BASE DE DADOS

Survival	Variável binária, 1 = sobreviveu
Pclass	Qual classe financeira do navio o passageiro estava
Sex	Sexo biológico
Age	Idade em anos
SibSp	Número de irmãos e cônjuges a bordo
Parch	Número de pais/filhos a bordo
Ticket	Número do ticket
Cabin	Número da cabine
Fare	Tarifa paga
Embarked	Porto em que embarcaram

A partir das duas primeiras análises, já é possível notar que a variável Cabin tinha uma recorrência alta de entradas nulas. Em uma análise mais aprofundada, descobriu-se que 684 linhas de 891 são nulas. Considerada demasiada incompleta, essa coluna foi descartada.

Outra análise útil é a criação de um mapa de calor representando a correlação de variáveis:

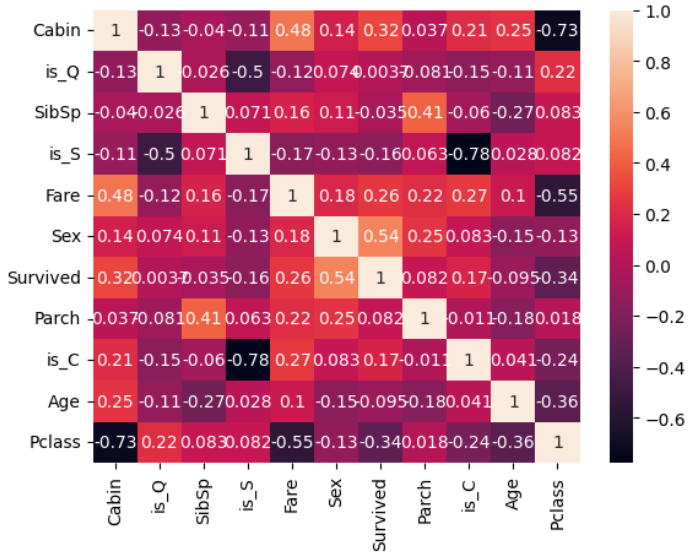

Fig. 2. Correlação de variáveis.

C. Tratamento de Dados

Na etapa de tratamento de dados, procura-se principalmente adaptar variáveis categóricas para números inteiros, já que os modelos de aprendizado de máquina se constituem melhor dessa forma.

Um exemplo de variável categórica se encontra na coluna ”Sex”, como visto na Tabela 1.1. Transforma-se ”Male” e ”Female” em 0 e 1.

Também foram identificadas 177 linhas sem valor na coluna ”Age”, portanto, é interessante encontrar uma forma de preencher essas lacunas antes de criar o modelo, ou caso não seja possível, descartar a coluna. Nessa situação, é possível deduzir uma média de idade pelo pronome de tratamento, por exemplo, ”Mrs.” denota alguém do sexo feminino e casado, provavelmente alguém com idade mais elevada.

Outro ponto relevante é a variável Fare, que possui irregularidades na base de dados, sendo assim melhor transformar os valores nulos na média ou excluir a coluna. Foi escolhido transformar os valores na média.

D. Criação do Modelo

O modelo escolhido foi o *RandomForestClassifier*, que testa diferentes modelos de aprendizado de máquina e se escolhe o resultado mais recorrente. A precisão do modelo na base de dados de treino foi de $82,12\%$, e para a base de dados de teste foi de $78,23\%$.

IV. Conclusão

Na competição do Kaggle análoga a este trabalho, houveram modelos que obtiveram 100% de precisão, mas a maioria dos competidores chega a uma precisão entre 75% e 80%, portanto o desempenho obtido pelo modelo do trabalho pode ser considerado satisfatório.

References

[1] Fernando da Silva. *O Ciclo de Análise de Dados: Um Roteiro para Resolver Problemas*. 2023. URL: https://analisemacro.com.br/econometria-e-machine-learning/o-ciclo-de-analise-de-dados-um-roteiro-para-resolver-problemas/.