DECISION TREE CLASSIFICATION: DATASET TITANIC

Luana Lorena de Matos Tavares[1]

RESUMO

O presente artigo tem por objetivo realizar uma análise, tratamento e classificação do dataset titanic, no qual apresenta informações referentes ao acontecimento do naufrágio do navio, como o nome das pessoas presentes, idade, sexo, cabines, sobreviventes ou não. Diante disso, foi utilizado o algoritmo árvore de decisão para a classificação dos dados com o intuito de estimar a quantidade de pessoas vivas ou falecidas por meio da técnica a ser aplicada.

Palavras-chave: decision tree. dataset. classificação. algoritmo.

ABSTRACT

The present article aims to conduct an analysis, treatment, and classification of the Titanic dataset, which provides information regarding the ship's sinking, such as the names of the individuals present, age, gender, cabins, and survival status. Therefore, the decision tree algorithm was used for data classification in order to estimate the number of survivors and non-survivors through the applied technique.

Keywords: decision tree. dataset. classification. algorithm.

1. INTRODUÇÃO

A análise de dados é uma prática essencial em muitas áreas, fornecendo insights valiosos e auxiliando na tomada de decisões informadas. Em particular, o campo de aprendizado de máquina tem desempenhado um papel significativo na extração de conhecimento a partir de conjuntos de dados complexos. Uma das técnicas mais populares nesse campo é o uso de árvores de decisão, que oferecem uma abordagem intuitiva e eficiente para a classificação e previsão de dados.

Neste artigo, foi explorado o algoritmo das árvores de decisão aplicadas ao conjunto de dados do Titanic. O naufrágio do RMS Titanic em 1912 é um dos eventos mais conhecidos da história e continua a despertar interesse até os dias atuais. Ao longo deste artigo, será

[1] Graduanda em Ciências da Computação pela UFT-Palmas. E-mail: luana.lorena@mail.uft.edu.br

analisado como as árvores de decisão podem ser aplicadas para tratar e analisar o conjunto de dados do Titanic. Inicialmente, será usado o tratamento de dados por meio dos outliers, média aritmética e remoção de colunas com porcentagem significativa de dados vazios e a classificação por meio dos testes, além da apresentação dos resultados.

Os insights obtidos através da aplicação de árvores de decisão no conjunto de dados do Titanic podem oferecer uma compreensão mais profunda dos fatores que influenciaram a sobrevivência dos passageiros como características individuais, como gênero, idade e classe socioeconômica, podem ter influenciado as chances de sobrevivência.

2. METODOLOGIA

Para o tratamento dos dados do dataset, foi importado as bibliotecas utilizadas e posteriormente é utilizado o “pd.read_csv()” para a leitura e início da análise.Além disso, é iniciado o time para calcular o tempo de processamento do algoritmo. Logo após, é realizada a limpeza e preparação dos dados primeiramente removendo as colônias que contenham todos os valores ausentes (NaN) aplicando o método “dropna()”, é importante realizar esse procedimento para remoção de colunas desnecessárias a serem analisadas.

Para o tratamento dos dados foi necessário categorizar os valores por meio do “LabelEncoder()”, seu objetivo é converter as variáveis categóricas em valores numéricos para que seja possível que o algoritmo de árvore de decisão possa processar os dados adequadamente .

Para que sejam tratados os valores atípicos (outliers), é importante remover todas as colunas declaradas na lista que não são relevantes para a classificação e depois cria-se uma cópia do dataset para manter o dataset original.

Após a limpeza inicial, são identificados e tratados os outliers. Isso é feito utilizando o método do intervalo interquartil (IQR) para cada coluna numérica. Valores abaixo do limite inferior (Q1 - 1,5 * IQR) ou acima do limite superior (Q3 + 1,5 * IQR) são considerados outliers e substituídos por valores ausentes (NaN). Depois disso, os valores ausentes restantes são preenchidos com a mediana da respectiva coluna usando o método “fillna()”, removendo as colunas com 90% dos dados faltantes. Dessa forma, os dados estão prontos para serem utilizados na construção da árvore de decisão.

Logo após, é feita a análise de correlação para identificar as relações entre as variáveis e a variável alvo que foi escolhida em y ("Survived"). A correlação é calculada utilizando o

método “corrwith()” e as correlações resultantes são ordenadas em ordem decrescente e crescente, bem como em valores absolutos, para identificar as variáveis mais correlacionadas com a sobrevivência.

Com os dados devidamente preparados, é preparado o treinamento para o modelo. As colunas relevantes para o treinamento selecionadas são atribuídas à variável X (`Pclass, Sex, Age, SibSp, Parch)`, enquanto a variável alvo ("Survived") é atribuída à variável y. Além disso, as variáveis numéricas foram padronizadas utilizando StandardScaler(), para possuírem média zero e desvio padrão unitário. Esse método é realizado para garantir que todas as variáveis tenham a mesma escala e não influenciem indevidamente o processo de aprendizado do modelo.

Após a preparação dos dados, o conjunto de dados foi dividido em conjuntos de treinamento e teste. Nesse caso, 66% dos dados são usados para treinamento e 33% são reservados para a validação do modelo. Essa divisão é realizada utilizando a função “train_test_split()”.

Com os dados de treinamento e teste prontos, criou-se a estrutura do modelo de árvore de decisão utilizando o “DecisionTreeClassifier()”. Em seguida, o modelo é treinado usando o método “fit()” com os dados de treinamento.

Por fim, foi realizada a avaliação do modelo. As previsões foram feitas no conjunto de teste utilizando o método “predict()”. A matriz de confusão então é calculada utilizando o método “confusion_matrix()” para avaliar o desempenho do modelo em termos de verdadeiros negativos, falsos positivos, falsos negativos e verdadeiros positivos. Além disso, também foi calculada a acurácia do modelo usando o “accuracy_score()”. O relatório de classificação é gerado utilizando “classification_report()”, fornecendo métricas como precisão, recall e F1-score para cada classe. Por fim, o tempo de processamento calculado foi imprimido no fim da execução do código para medir a eficiência do modelo.

RESULTADOS

Este artigo analisou o dataset do titanic para classificação por meio do algoritmo de árvore de decisão. Ao final do código é apresentado os resultados:

```
Model Decision Tree:

tn 155, fp 14, fn 49, tp 77

Accuracy: 0.7864406779661017

Classification Report:
              precision    recall  f1-score   support

      morreu       0.76      0.92      0.83       169
  sobreviveu       0.85      0.61      0.71       126

    accuracy                           0.79       295
   macro avg       0.80      0.76      0.77       295
weighted avg       0.80      0.79      0.78       295


O tempo de processamento foi de 1.3633391857147217 segundos.
```

Figura. 1 Resultado da classificação

Ao analisar o relatório de classificação, observamos que o modelo apresentou uma precisão de 76% para a classe "morreu", o que significa que ele classificou corretamente 76% dos casos em que as pessoas não sobreviveram. Além disso, o recall para essa classe foi de 92%, indicando que o modelo identificou corretamente 92% dos casos em que as pessoas não sobreviveram. O F1-score para a classe "morreu" foi de 83%, que é a combinação entre o precision e o recall.

Para a classe "sobreviveu", o modelo alcançou uma precisão de 85%, indicando que 85% das previsões positivas para sobreviventes estavam corretas. No entanto, o recall para essa classe foi de 61%, significando que o modelo identificou corretamente apenas 61% dos casos em que as pessoas sobreviveram. O F1-score para essa classe foi de 71%.

Em resumo, o modelo obteve um desempenho razoável na classificação dos passageiros do Titanic. Ele apresentou uma boa precisão para identificar aqueles que não sobreviveram, mas teve uma performance inferior na identificação dos sobreviventes. O tempo de processamento do modelo foi rápido, levando aproximadamente 1.36 segundos para concluir a análise dos dados.

REFERÊNCIAS

https://www.kaggle.com/datasets/yasserh/titanic-dataset