Árvore de decisão e seus frutos

O que podemos aprender com uma *Decision Tree*

Marcelo Reis
marceloestevesreis@gmail.com
PUC Minas
Belo Horizonte, Minas Gerais, Brasil

Bernardo Lara
be1212lara@gmail.com
PUC Minas
Belo Horizonte, Minas Gerais, Brasil

Raphael Caetano
caetano.missiaggia@gmail.com
PUC Minas
Belo Horizonte, Minas Gerais, Brasil

Bruno Zandona
bruzan57@gmail.com
PUC Minas
Sete Lagoas, Minas Gerais, Brasil

ABSTRACT

Este artigo, intitulado "Árvore de decisão e seus frutos: O que podemos aprender com uma Decision Tree", apresenta um estudo sobre a implementação de algoritmos de árvore de decisão em uma base de dados de drogas.

Foram realizadas etapas de pré-processamento, incluindo o balanceamento da base de dados, utilizando under-sampling para reduzir o número de instâncias das drogas mais representadas.Os atributos numéricos foram mantidos sem alterações, enquanto os atributos categóricos e ordinais foram codificados de acordo com uma transformação específica.

Os métodos utilizados na análise foram descritos, incluindo a implementação da árvore de decisão. Os resultados obtidos mostraram um desempenho sólido na classificação de diferentes tipos de drogas, com destaque para a classe "drugB", que apresentou uma precisão perfeita de 1.00.

KEYWORDS

Árvore de decisão, algoritmos, base de dados, pré-processamento, balanceamento.

ACM Reference Format:
Marcelo Reis, Bernardo Lara, Raphael Caetano, and Bruno Zandona. 2023. Árvore de decisão e seus frutos: O que podemos aprender com uma *Decision Tree.* In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUÇÃO

Na disciplina de Inteligência Artificial, foi proposto um trabalho no qual deveríamos implementar algum algoritmo visto em sala na base de dados escolhida. Para isso, utilizamos a plataforma Kaggle e encontramos a base '*Drugs A, B, C, X, Y for Decision Trees*'. Uma base simples, mas que se encaixou perfeitamente no que estávamos procurando.

Após a escolha da base, passamos para a divisão de tarefas entre o grupo, e, por meio do Google Colab, pudemos começar o trabalho.

2 DESCRIÇÃO DA BASE DE DADOS

A base de dados utilizada neste estudo é intitulada "Drugs A, B, C, X, Y for Decision Trees" e foi obtida por meio da plataforma Kaggle. Essa base de dados é simples, porém adequada para o propósito do trabalho em questão. Ela consiste em informações sobre diferentes drogas, representadas pelas letras A, B, C, X e Y, e foi utilizada para a implementação de um algoritmo de árvore de decisão.

Table 1: Descrição dos atributos do conjunto de dados drug200.csv

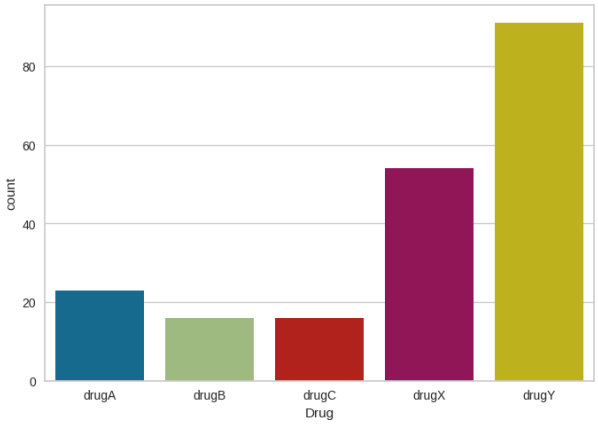
Atributo	Descrição	Tipo de Dado
Age	Idade do paciente	Numérico Max = 15 anos Min = 74 anos
Sex	Gênero do paciente	Categórico F = 0 M = 1
BP	Pressão arterial do paciente	Categórico High = 2 Normal = 1 Low = 0
Cholesterol	Nível de colesterol do paciente	Categórico High = 1 Low = 0
Na_to_K	Sodio e potássio no sangue do paciente	Numérico Max = 6.269 Min = 38.247
Drug	Medicamento prescrito ao paciente	Categórico drugA drubB drugC drugX drugY

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

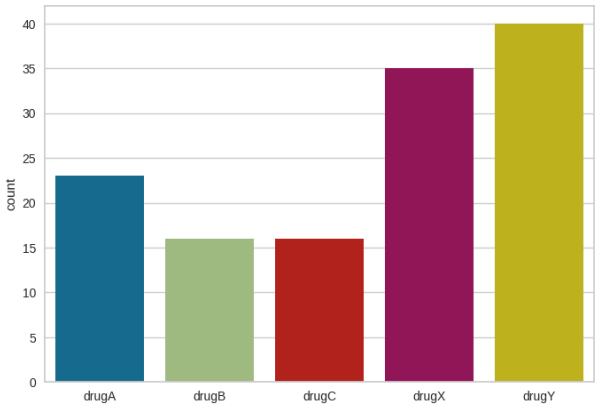
3 ETAPAS DE PRÉ-PROCESSAMENTO

Antes de irmos para a parte avançada, foi preciso averiguar a qualidade da base e suas instâncias. Ao contar a quantidade de instâncias de cada droga chegamos ao seguinte resultado:



Com base nisso, é perceptível um alto desbalanceamento, com a "drugY" tendo 91 instâncias enquanto temos outras drogas com 16 instâncias. Para balancearmos a base, decidimos por usar *undersampling*, reduzindo o número de instâncias de "drugY" para 40 e o de "drugX" para 35.

Dessa forma, conseguimos balancear a base sem uma grande perda de informações. Resultando nesses dados:



Para a codificação dos atributos, aqueles que eram numéricos, como "*Age*" e "*NaToK*" se mantiveram sem alterações. Já os valores de Sex, BP e Cholesterol, por serem valores categóricos e alguns ordinais(BP e Cholesterol), acabaram sendo transformados da seguinte maneira:

Sex - F:0 - M:1
BP - High:2 - Normal:1 - Low:0
Cholesterol - High:1 - Low:0

Essas transformações foram aplicadas para facilitar o uso desses atributos no algoritmo escolhido. Algo notável na base escolhida é ausência de um atributo categórico e nominal, o que acabou por diminuir o trabalho braçal em questão de código.

Por fim, foram usadas as seguintes quantidades para teste:

Table 2: Quantidade para Testes

Drogas	Quantidade
DrugA	7
DrugB	5
DrugC	5
DrugX	10
DrugY	12

4 DESCRIÇÃO DOS MÉTODOS UTILIZADOS

O algoritmo utilizado foi o CART (Classification and Regression Trees). O CART é um algoritmo de aprendizado de máquina que constrói árvores de decisão binárias, onde cada nó interno representa um teste em um atributo e cada ramo representa o resultado desse teste.

No caso específico da implementação em questão, foi usado o critério "Gini", conforme especificado no código:

$(criterion =' gini', max_depth = 3, random_state = 0)$

Também optamos pela geração da árvore de profundida = 3, com o objetivo de evitar resultados falsamente positivos graças a várias regras de decisões. Sendo a nossa principal ferramente de trabalho o Google Colab e a linguagem Python.

5 RESULTADOS E DISCUSSÕES

Aqui estão os resultados obtidos na árvore de decisão para as drogas:

Table 3: Resultados da Árvore de Decisão para as Drogas

Drug	Precision	Recall	F1-Score	Support
drugA	0.88	1.00	0.93	7
drugB	1.00	0.80	0.89	5
drugC	0.62	1.00	0.77	5
drugX	1.00	0.70	0.82	10
drugY	1.00	1.00	1.00	12
Accuracy			0.90	39
Macro Avg	0.90	0.90	0.88	39
Weighted Avg	0.93	0.90	0.90	39

Como é possível ver na Tabela 1, os resultados da árvore de decisão mostram um desempenho sólido na classificação de diferentes tipos de drogas. Analisando os principais indicadores, como a precisão, é perceptível que a classe "drugB" tem uma precisão perfeita de 1.00, o que significa que todos os casos classificados como "drugB" são realmente "drugB". As classes "drugA," "drugC," "drugX," e "drugY" também têm precisões bastante altas, variando de 0.62 a 1.00.

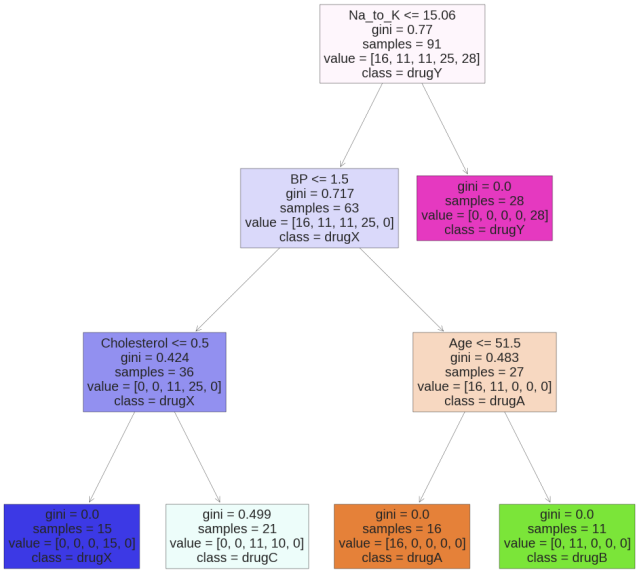
Já se tratando sobre recall, a classe "drugA" possui um desempenho perfeito de 1.00, o que indica que todos os casos reais de "drugA" foram corretamente identificados.

No entanto, "drugX" tem um recall um pouco menor de 0.70, o que sugere que alguns casos de "drugX" podem ter sido classificados incorretamente.

O F1-Score combina precisão e recall para fornecer uma métrica única de desempenho. Em geral, as classes têm F1-Scores bastante equilibrados, variando de 0.77 a 1.00. Isso indica um bom equilíbrio entre identificar corretamente as classes e evitar falsos positivos.

A acurácia geral do modelo é de 0.90, o que representa uma medida global de quão bem o modelo está classificando todas as classes. Ter uma acurácia de 0.90 sugere que o modelo está desempenhando bem na classificação geral.

É importante notar que todos esses resultados são obtidos a partir de uma árvore de decisão simples, com poucas regras de classificação, porém com resultados justos e confiáveis. A árvore está representada na imagem a seguir:



Em resumo, os resultados são bastante promissores, com altas precisões e recalls para a maioria das classes. No entanto, é importante observar que a classe "drugX" teve um recall um pouco menor, o que pode ser um ponto de melhoria.

6 CÓDIGO DESENVOLVIDO

https://colab.research.google.com/drive/14F5cAx7es8M9nzYH4illG9g58vH_9SEm?usp=sharing

7 REFERENCES

Kaggle, 2021. Drugs A, B, C, X, Y for Decision Trees. https://www.kaggle.com/datasets/pablomgomez21/drugs-a-b-c-x-y-for-decision-trees