ELEMENTOS DE INTELIGÊNCIA ARTIFICIAL E CIÊNCIA DE DADOS

2º TRABALHO PRÁTICO

MARIANA SERRÃO - UP202109927
MAFALDA AIRES - UP202106550

MEDIA DATASET

Foi escolhido a base de dados de media. Trata-se de três datasets, filmes, shows de tv e álbuns, em que o objetivo é juntá-los num único dataset. Este conjunto de dados não é fornecido em formato tabular e, portanto, é necessário passar por pré-processamento.

A este dataset será aplicados dois algoritmos de aprendizagem supervisionada: Árvores de Decisão e K-NN. Os resultados obtidos nos dois serão comparados e, no fim, conclui-se qual o mais adequado.

REFERÊNCIAS

Como referências para a realização do trabalho foram encontrados alguns artigos online, assim como exemplos de código disponibilizados na plataforma GitHub e websites utilizados para obter dados adicionais para o dataset:

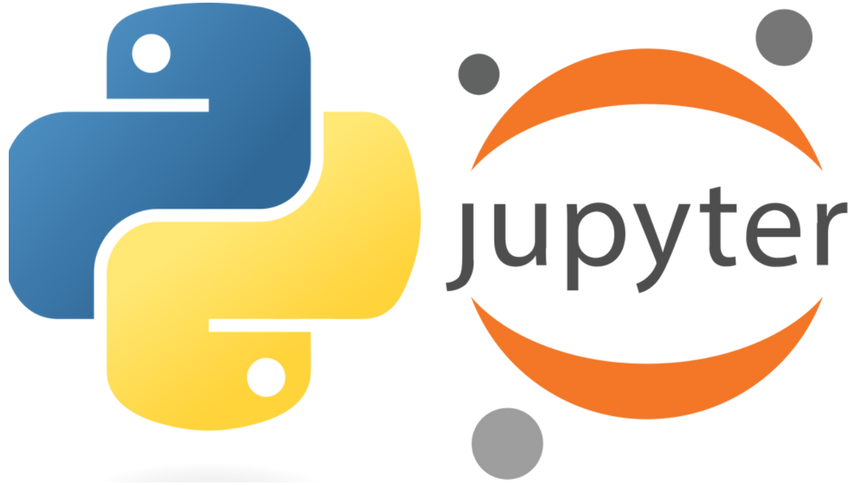
- https://www.kaggle.com/code/ramontanoeiro/cap-tulo-06-machine-learning-arvore-de-decis-o/notebook
- https://github.com/guilhermevaz8/VideoGames-Supervised-Learning/blob/main/Trabalho.ipynb
- https://www.imdb.com/
- https://www.allmovie.com/

FERRAMENTAS E ALGORITMOS UTILIZADOS

Para se organizar os dados, foi utilizado o jupyter notebook, em conjunto com a biblioteca Pandas e scikit-learn.

Foram criados três algoritmos, um para cada dataset, no qual, em primeiro lugar, se converteu os ficheiros txt em ficheiros json (exceto movies, uma vez que já se encontrava neste formato). Em seguida, converteu-se esses ficheiros json em ficheiros csv, acrescentando-se a coluna "Type", para distinguir entre filmes, álbuns e shows de tv.

No fim, criou-se o algoritmo que junta os três datasets num único ficheiro csv (merged_data.csv), no qual as colunas em comum são convergidas.



MERGE DATASET

Type	ReleaseYear	FinalYear	Network	Label/Studio	Genre	Rating	Adaptation	Artist	ReleaseMonth	Director	Runtime	NumOfCriticReviews	CriticScore
Tv-Show	2011	2019	HBO	Warner Bros.	Drama, Fantasy	TV-MA	Based on a book	None	None	None	None	None	Excellent
Tv-Show	2008	2013	AMC	Sony Pictures Television	Drama	TV-MA	Original	None	None	None	None	None	Excellent
Tv-Show	1994	2004	NBC	Warner Bros.	Comedy	TV-PG	Original	None	None	None	None	None	Good
Tv-Show	1999	2007	HBO	Warner Bros.	Drama	TV-MA	Original	None	None	None	None	None	Excellent
Tv-Show	2016	2022	Netflix	None	Horror, Science Fiction	None	Based on a book	None	None	None	None	None	Good
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Movie	2013	None	None	A24	Romance	R	None	None	August	James Ponsoldt	95 min	42.0	Good
Movie	2012	None	None	FilmDistrict	Comedy	R	None	None	June	Colin Trevorrow	86 min	31.0	Good
Movie	2011	None	None	Paramount Vantage	Romance	PG-13	None	None	October	Drake Doremus	90 min	38.0	Mediocre
Movie	2011	None	None	Fox Searchlight Pictures	Romance	PG-13	None	None	June	Gavin Wiesen	84 min	28.0	Bad
Movie	2015	None	None	A24	Drama	R	None	None	July	James Ponsoldt	106 min	35.0	Good

O dataset final (merged_data.csv) trata-se da junção dos 3 datasets. É uma totalidade de 906 linhas por 15 colunas, sendo elas Title, Type, ReleaseYear, FinalYear, Network, Label/Studio, Genre, Rating, Adaptation, Artist, ReleaseMonth, Director, Runtime, NumOfCriticReviews e CriticScore (classe que se quer prever, com os valores: Excellent, Good, Mediocre e Bad)

A coluna CriticScore foi calculada de maneira diferente para os três datasets. Enquanto que, para *movies*, foi utilizado o CriticScore que a base de dados já possuía, para *albums* considerou-se o número de vendas e para *tv-shows* acrescentou-se uma nova coluna com a classificação dada pelos críticos no website IMDB. Para além disso, acrescentaram-se dados na coluna Label/Studio em movies.

ALGORITMOS IMPLEMENTADOS

ÁRVORE DE DECISÃO

Árvores de decisão são algoritmos utilizados para resolver problemas de classificação e regressão. Consistem numa estrutura em forma de árvore, na qual cada nó representa uma decisão baseada num atributo específico. Neste caso, aplicou-se uma árvore de decisão ao dataframe, com a utilização da biblioteca scikit-learn. O algoritmo analisa os padrões nos dados e aprende a tomar decisões com base nos atributos fornecidos. Em seguida, faz previsões nos dados de teste, fornecendo os resultados esperados. Árvores de Decisão são utilizadas pela sua capacidade de capturar relacionamentos complexos e pela facilidade de interpretação.

K-NN

O algoritmo KNN (K-Nearest Neighbors) é uma técnica de aprendizagem computacional utilizada em problemas de classificação. Baseia-se na proximidade entre as instâncias de dados para fazer previsões. No código realizado, o KNN foi aplicado ao conjunto de dados através da biblioteca scikit-learn. Este foi treinado com os dados de treinamento, onde foram considerados os 5 vizinhos mais próximos para tomar decisões de classificação. Em seguida, foi usado para fazer previsões nos dados de teste. Este algoritmo é útil em situações em que a relação entre os atributos e os rótulos é complexa e não linear.

RESULTADOS

ÁRVORE DE DECISÃO

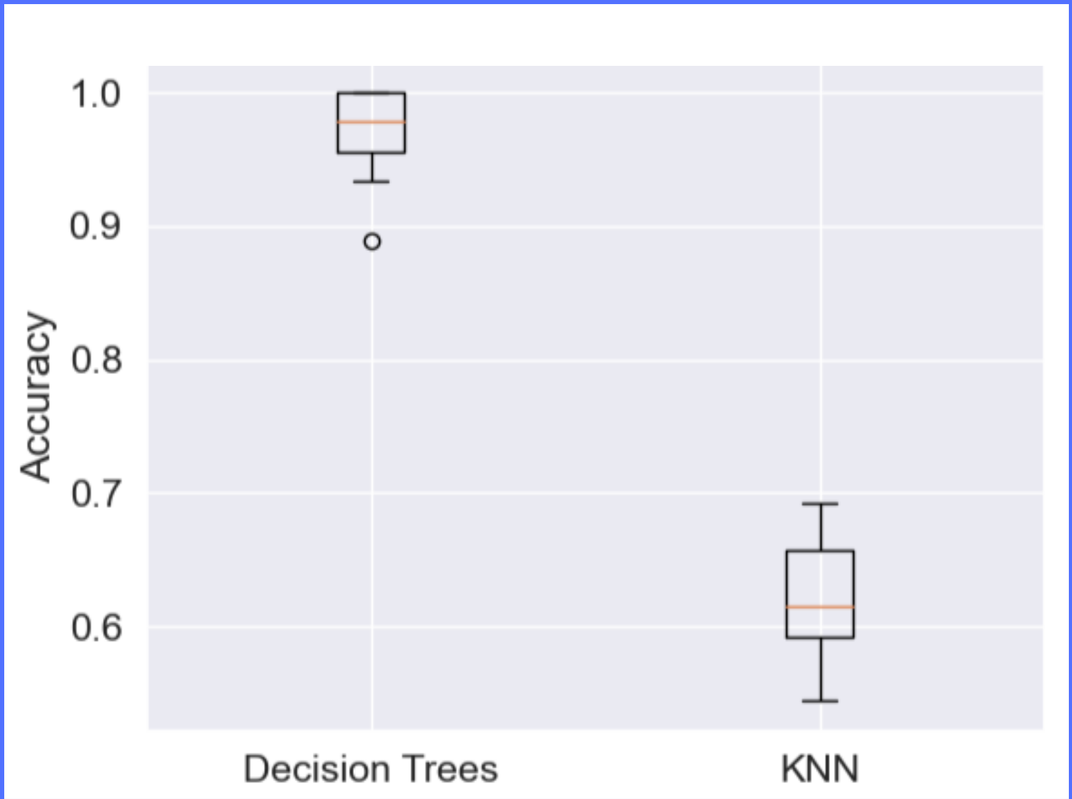
Accuracy	**0.8603**
Precision	**0.8569**
Recall	**0.8579**
F1-Score	**0.8574**

K-NN

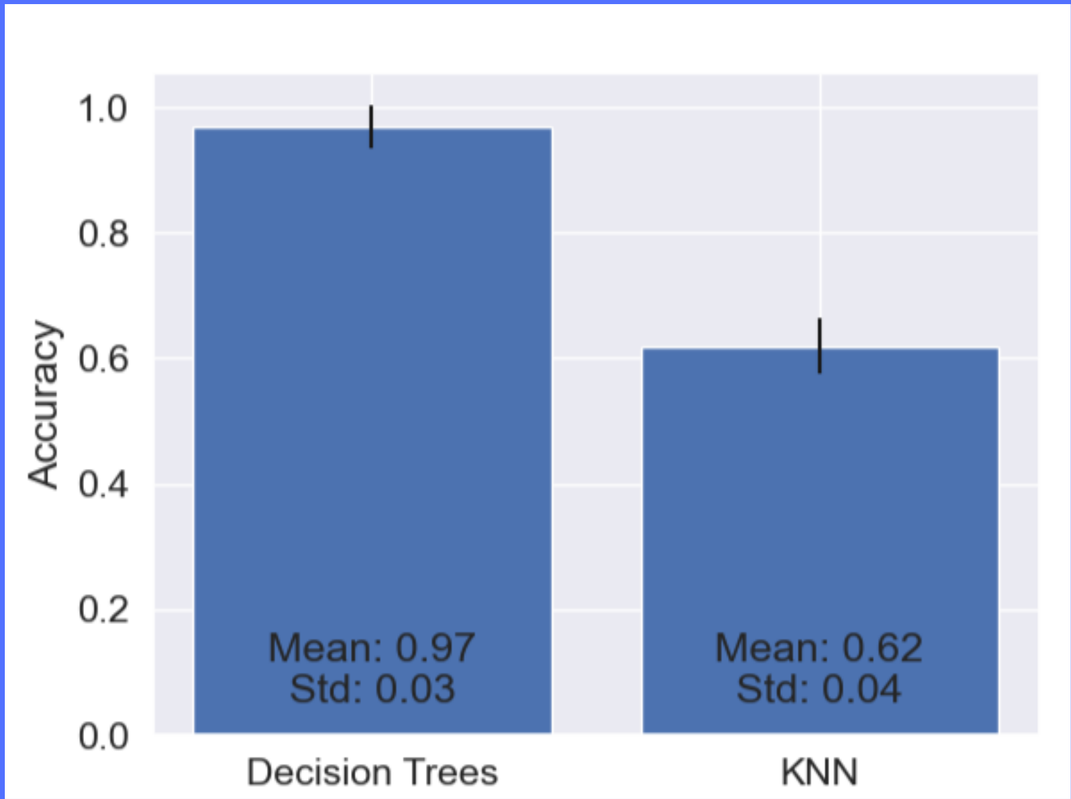
Accuracy	**0.5331**
Precision	**0.3986**
Recall	**0.3939**
F1-Score	**0.3959**

RESULTADOS

RESULTADOS DA ACURÁCIA



ACURÁCIA E DESVIO PADRÃO



CONCLUSÕES

Com base nos resultados obtidos, conclui-se que a árvore de decisão é uma abordagem superior ao algoritmo k-Nearest Neighbors para o conjunto de dados em questão. Ao se comparar o desempenho dos dois algoritmos, a árvore de decisão demonstrou melhores resultados em termos de acurácia, precisão, revocação e F1-Score.

Assim, a árvore de decisão é capaz de capturar melhor a estrutura e os padrões presentes no conjunto de dados, resultando numa classificação mais precisa e confiável.