Supervised Learning
Top Hits Spotify

Inteligência Artificial

Grupo 12_1B:
- André Flores - up201907001
- Diogo Faria - up20907014
- Tiago Rodrigues - up201906807

Especificação

A partir de um conjunto de dados sobre músicas do Spotify vamos utilizar algoritmos de aprendizagem supervisionada, especificamente de classificação, para poder prever a sua popularidade.

No conjunto de dados que nos foi dado, existe informação sobre 2000 músicas, dividida entre 18 diferentes parâmetros.

Algoritmos e Ferramentas

Algoritmos: Decision Trees, K Nearest Neighbor, Support-Vector Machine e Neural Networks.

Ferramentas: Python 3, Jupyter Notebook, pandas, Seaborn, Matplotlib e Scikit-learn.

Trabalho Implementado

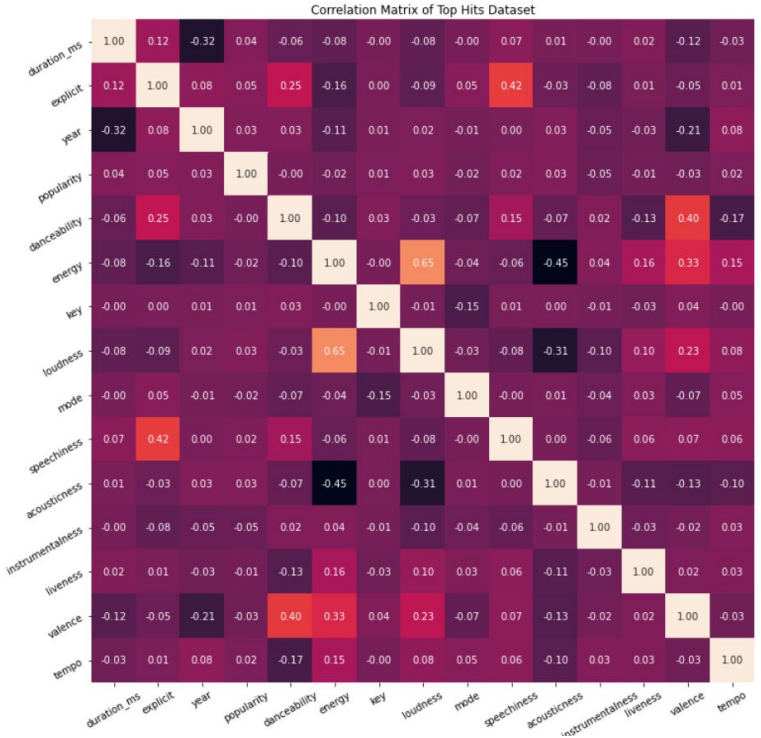
Foi realizado um pré-processamento sobre os dados, sendo que se fez a verificação de dados nulos, os quais não se encontraram nenhum, e também a eliminação de dados de erro no campo de ‘genre’, onde se encontravam valores como: ‘set()’. Também foram removidos outliers dos dados.

Sendo que vamos utilizar algoritmos de classificação e o atributo que pretendemos que os modelos possam prever está dado com valores entre 0 e 100, foi necessário uma transformação sobre eles, sendo que se agregaram em 4 diferentes valores: ‘Low’, ‘Average’, ‘High’ e ‘Very High’, de forma a que cada valor apresentasse um conjunto de dados semelhante.

Trabalho implementado (Cont.)

De forma a reduzir o número de atributos a utilizar, criamos uma matriz de correlação que nos permitisse agregar atributos.

No entanto, sendo que a maior correlação encontrada foi de apenas 0.65/1, decidimos que nenhum dos atributos poderiam ser agregados.



Trabalho implementado (Cont.)

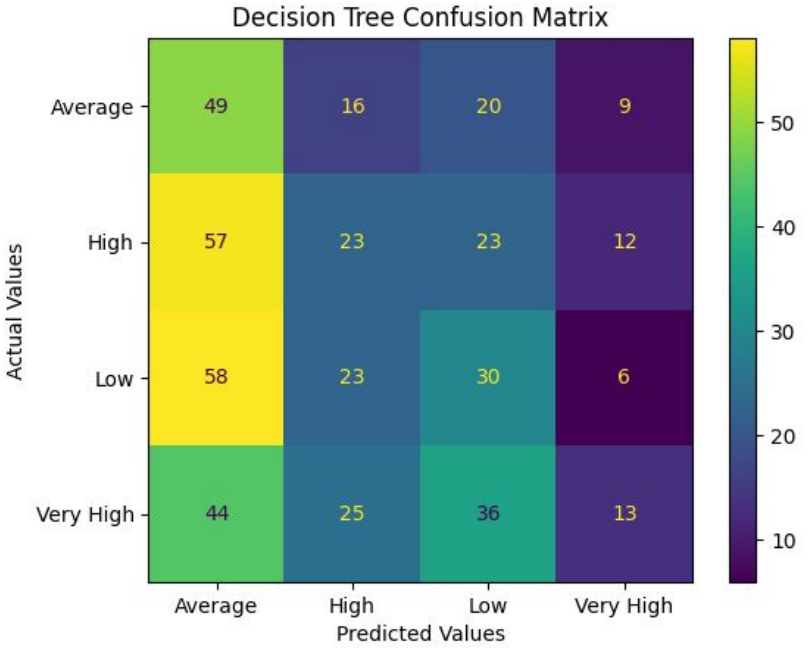
Para melhorarmos os resultados das nossas previsões criámos uma função baseada no GridSearchCV to módulo sklearn que calcula os melhores parâmetros para cada algoritmo.

Também criámos uma função que calcula e imprime a “precision”, o “recall”, a “accuracy” e o “F-measure”.

Trabalho implementado (Cont.)

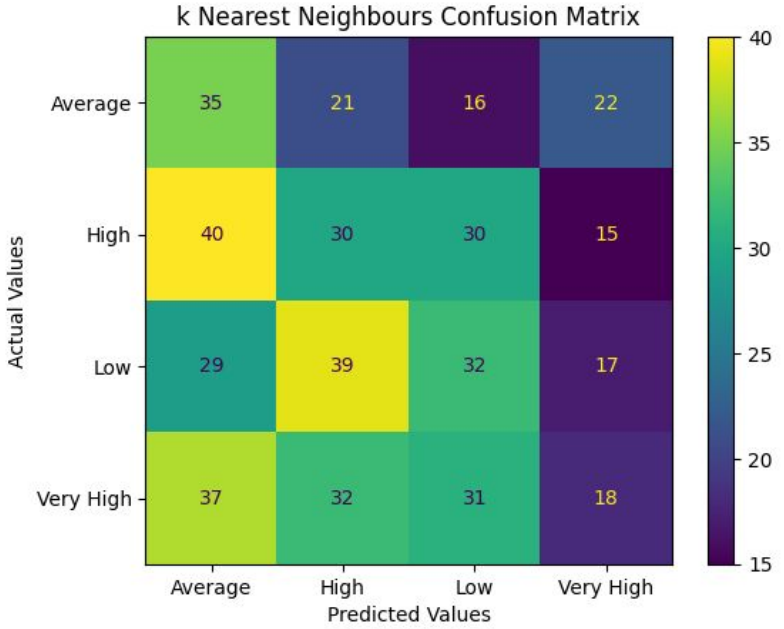
Conseguimos treinar os modelos do sklear com base nos nossos dados e melhore parâmetros, embora os resultados não tenhar sido satisfatórios.

Para Decision Trees obtemos um score d 0.259 com os seguintes resultados.



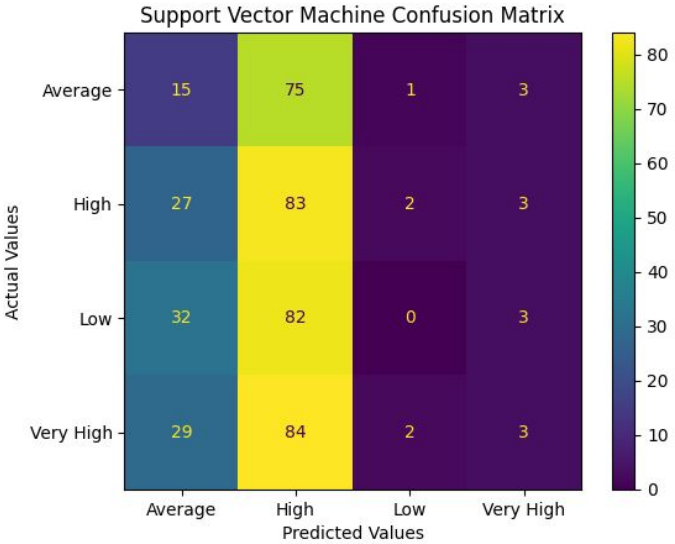
Trabalho implementado (Cont.)

No algoritmo de K-Nearest Neighbour utilizamos o StandardScaler do sklearn dado a este algoritmo necessitar de escalonamento de dados a maioria das vezes. Obtemos um melhor score de 0.282 com os seguintes resultados.



Trabalho implementado (Cont.)

Nas Support Vector Machines, também usando o StardardScaler, escolhemos utilizar a NuSVM do sklearn pois apresenta um parâmetro que controla o numero de vetores de suporte os que nos permitiu obter um resultado melhor com um melhor score de 0.287.



Conclusão

Depois de várias tentativas com diferentes modelos chegámos à conclusão que os algoritmos de classificação exigidos na especificação do trabalho não são adequados a estes dados pois forçam uma divisão do atributo de popularidade em ranks com intervalos muito pequenos e com alta variação de amplitude. Para obter melhores resultados deveriam ter sido utilizados algoritmos de previsão.