Kathmandu University

Department of Computer Science and Engineering

Dhulikhel, Kavre

Report on

"Comparison of Data Mining Algorithms on Spotify Dataset"

[Course Code: COMP 482]

Submitted by:

Sushant Adhikari (05)

Ayush Aryal (07)

Submitted to:

Dr. Rajani Chulyadyo

Department of Computer Science and Engineering

Submission Date: 18th December, 2023

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF TABLES

1. Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) on the "spotify_songs" dataset, consisting of 32,833 rows and 23 columns, was conducted to gain insights into the musical attributes and trends present in the data. The dataset includes a mix of categorical and numeric variables, contributing to a comprehensive understanding of the songs and playlists represented.

1.1. Categorical Data

Column Name	Description
track_id	Unique identifier for each track.
track_name	The name of the track.
track_artist	The artist or group associated with the track.
track_album_id	Unique identifier for each album.
track_album_name	The name of the album.
track_album_release_date	The release date of the album.
playlist_name	The name of the playlist.
playlist_id	Unique identifier for each playlist.

playlist_genre	The genre of the playlist.
playlist_subgenre	The subgenre of the playlist

Table 1: table describing different categorical columns

1.2. Numeric Data

Column Name	Description
track_popularity	Popularity score assigned to the track.
danceability	A measure of how suitable a track is for dancing.
energy	Represents the intensity and activity of the music.
key	The key the track is in.
loudness	The overall loudness of the track in decibels.
mode	Indicates the modality of the track (major or minor).
speechness	Measures the presence of spoken words in the track.
acousticness	Represents the acoustic quality of the track.

instrumentalness	Measures the likelihood of the track being instrumental.
liveness	Indicates the presence of an audience in the recording.
valence	Describes the musical positiveness conveyed by a track.
duration_ms	The duration of the track in milliseconds.

Table 2: table describing different numeric columns

	track_popularity	danceability	energy	key	loudness	mode	speechiness	acousticness
count	32833.000000	32833.000000	32833.000000	32833.000000	32833.000000	32833.000000	32833.000000	32833.000000
mean	42.477081	0.654850	0.698619	5.374471	-6.719499	0.565711	0.107068	0.175334
std	24.984074	0.145085	0.180910	3.611657	2.988436	0.495671	0.101314	0.219633
min	0.000000	0.000000	0.000175	0.000000	-46.448000	0.000000	0.000000	0.000000
25%	24.000000	0.563000	0.581000	2.000000	-8.171000	0.000000	0.041000	0.015100
50%	45.000000	0.672000	0.721000	6.000000	-6.166000	1.000000	0.062500	0.080400
75%	62.000000	0.761000	0.840000	9.000000	-4.645000	1.000000	0.132000	0.255000
max	100.000000	0.983000	1.000000	11.000000	1.275000	1.000000	0.918000	0.994000

Fig 1: Table showing the description of the dataset

Fig 1 shows that the track popularity metric exhibits a range from 0 to 100, with an average popularity score of 42.48. Notably, there is a slight disparity between the mean and median values, suggesting potential variations in the distribution. Approximately 25% of the tracks surpass a popularity threshold of 62. Moving on to danceability, the dataset's songs have an average danceability value of around 0.65. Similarly, the energy level of the songs averages at 0.70, with a quarter of the songs possessing an energy level exceeding 0.84. Regarding the duration of the tracks, half of them have a duration of 216,000 milliseconds, equivalent to 3 minutes and 36 seconds. Additionally, a noteworthy observation is the presence of a track in the dataset with an extensive duration of 517,810 milliseconds, approximately 8.63 minutes.

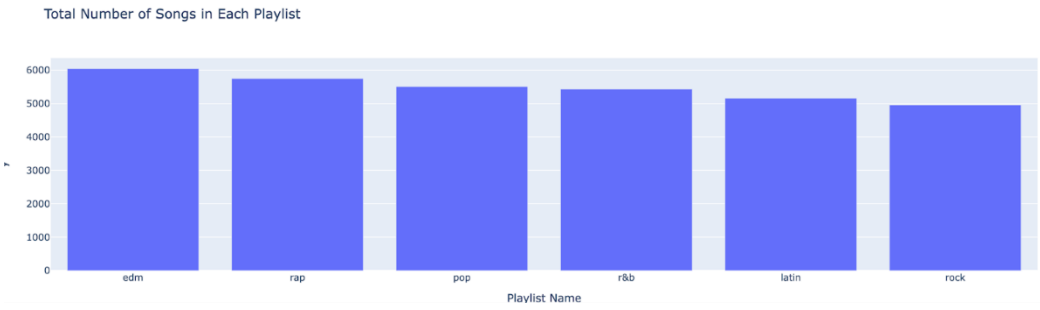

Fig 2: Bar graph showing total number of songs in each playlist

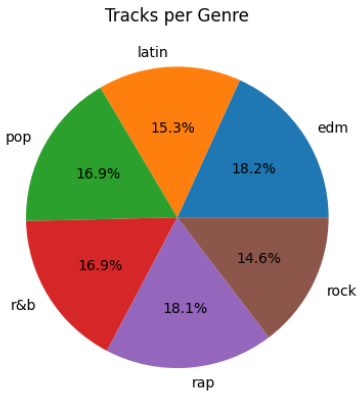

Fig 3: Pi chart showing distribution of songs for different playlist_genre

From the above Fig 2 bar graph and Fig 3 pi chart it can be seen that the playlist edm has the most number of songs (18.2% of total songs) and the playlist rock has the least number of songs(14.6% of total songs).

Fig 4: pi chart showing distribution of songs for different playlist_subgenre

From the above Fig 4 it can be seen that progressive electro house has the maximum number of songs which is 5.51% of the total songs and reggaeton subgenre has least number of songs which is 2.89% of the total songs.

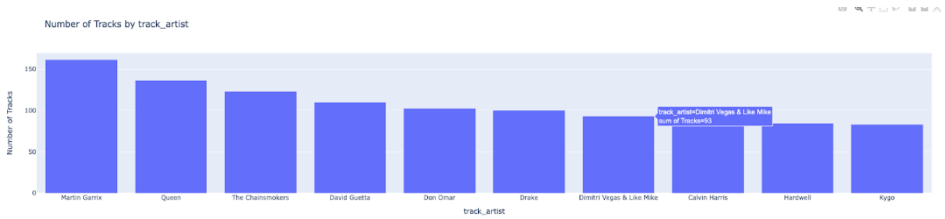

Fig 5: Bar graph showing number of tracks by different track_artist

The plot on Fig 5 shows the total number of track artists. It can be seen that Martin Garrix has the highest number of tracks.

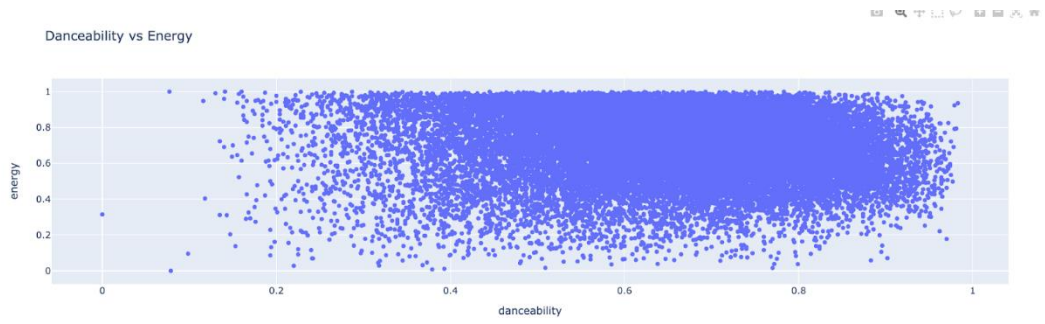

Fig 6: Scatter plot showing relation between danceability and energy

From the above scatter plot we can see that danceability is increased with increase in energy in most of the dataset.

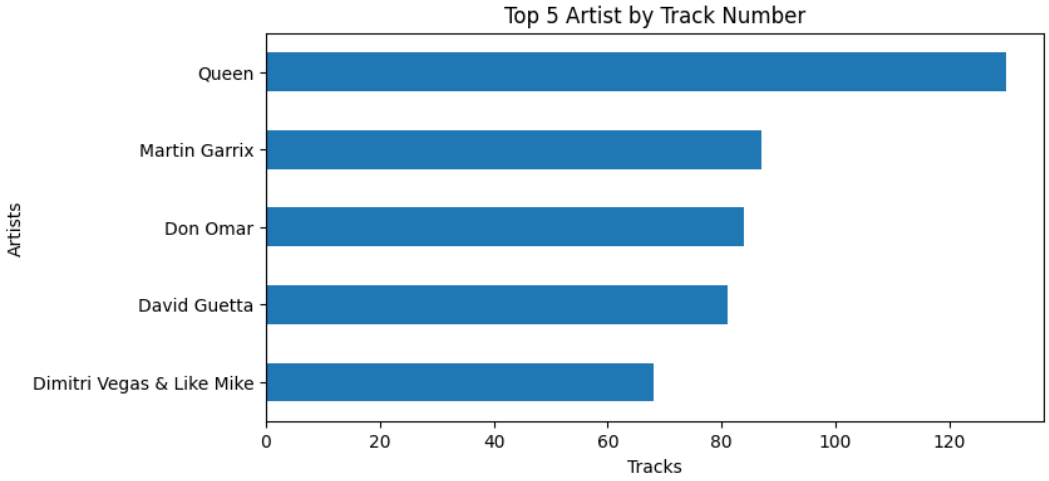

Fig 7: bar graph showing top 5 artist based on number of tracks

From above shown Fig 7 we can see that the artist Queen has the highest number of tracks followed by Martin Garrix. It shows the top 5 artists based on its number of tracks.

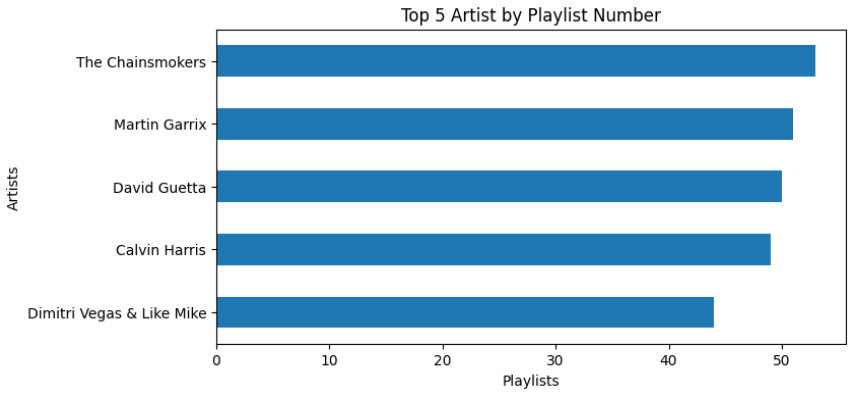

Fig 8: bar graph showing top 5 artists based on number of playlist

We can see from the above fig 8 that the chainsmokers is the top artist in terms of number of playlists which is followed by Martin Garrix.

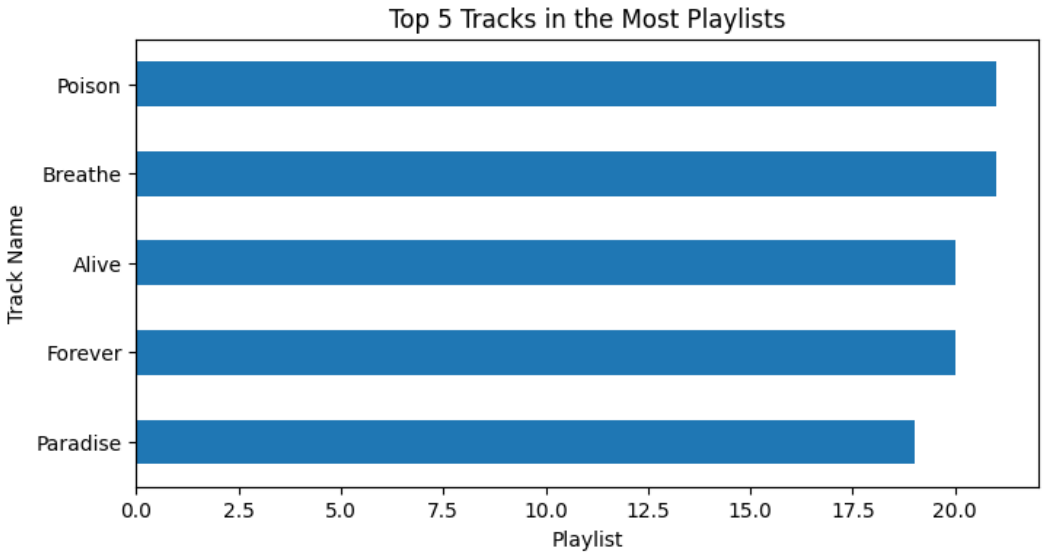

Fig 9: Bar graph showing top 5 tracks in most playlists

From above Fig 9 we can find that Poison is the track which is in most of the playlist. In the second place we have Breathe followed by alive, forever and paradise.

From the song statistics we identified the most popular song as "Dance Monkey" based on its popularity score. Conversely, the least popular song is named "Siren." Additionally, we found that the longest duration song is "47-Remix," while the shortest duration song is titled "Hi, How're You Doin?"

2. Model Training

On the dataset explored above, different data mining algorithms can be applied for clustering, classification,etc. In this report, three different classification algorithms: Decision Tree, Naive Bayes and Random Forest, will be compared for classifying the genre of the songs. For the model training process, following steps were followed:

2.1. Data Cleaning

The dataset contains 23 columns but not all of the columns were valuable for training the classification model. So, the columns 'track_id', 'track_name', 'track_album_id', 'track_album_name', 'track_album_release_date', 'playlist_name', 'playlist_id' were dropped from the dataframe.

Now, on the remaining columns, the null values were checked and found that the column ‘track_artist’ contained 5 null values. So, the rows containing the null values were removed.

2.2. Data Preprocessing

The columns containing numerical values were not on the same scale. Using StandardScaler from the sklearn library, all the columns with numerical values were scaled. As there were 13 columns with numerical values, Principal Component Analysis (PCA) was done to reduce the dimension of feature columns.

Number of Components	Explained Variance Ratio
2	0.29
3	0.38
4	0.47
5	0.55
6	0.63
7	0.70
8	0.77
9	0.83
10	0.89
11	0.94

12	0.98

Table 3: table showing explained variance ratio for different number of components

The explained variance ratio is greater than 0.9 only for the components greater than 10 which means we can only reduce two columns to capture more than 90% of the variance in the actual data. Reducing only two columns to get 94% of the variance captured does not seem to be a valuable trade off so, all the columns with actual data were used.

The three columns containing categorical values: ‘track_artist’, ‘playlist_genre’, ‘playlist_subgenre’ were encoded into integers using LabelEncoder from the sklearn library.

Playlist genre	Encoded label
edm	0
latin	1
pop	2
r&b	3
rap	4
rock	5

Table 4: table showing the actual label and the encoded label of playlist genre

2.3. Training

After all the cleaning and preprocessing of the data, it was split as 80% train data and 20% test data. Then the models DecisionTreeClassifier, GaussianNB and RandomForestClassifier from the sklearn library were used for training the classifier model for classifying the playlist genre using the train data.

3. Model Evaluation

The models trained using different algorithms were evaluated in different criteria: Training time, Classification report, Confusion matrix.

3.1. Training time

Training time is the time taken by the model to fit the training data.

Classification Model	Training time (in seconds)
DecisionTreeClassifier	0.349
GaussianNB	0.032
RandomForestClassifier	10.548

Table 5: Training time for different classification models

From Table 5, it can be concluded that the GaussianNB model was trained the fastest. RandomForestClassifier model takes a lot more time than the other two classifier models.

3.2. Classification Report

Classification report is the summary of the performance of a classifier model. It includes 4 metrics: precision, recall, f1-score and support for each class, in this case each playlist genre.

```
Decision Tree
Running time: 0.3493776321411133 seconds
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1236
           1       1.00      1.00      1.00      1062
           2       1.00      1.00      1.00      1102
           3       1.00      1.00      1.00      1112
           4       1.00      1.00      1.00      1112
           5       1.00      1.00      1.00       942

    accuracy                           1.00      6566
   macro avg       1.00      1.00      1.00      6566
weighted avg       1.00      1.00      1.00      6566
```

Fig 10: Training time and classification report of DecisionTreeClassifier

In Fig 10, it is shown that the accuracy, precision, recall, f1-score all have value 1.00 which means the model is predicting all the 6566 test data correctly without any mistake.

```
Naive Bayes Classifier
Running time: 0.03245425224304199 seconds
              precision    recall  f1-score   support

           0       0.58      0.66      0.62      1236
           1       0.41      0.50      0.45      1062
           2       0.43      0.41      0.42      1102
           3       0.44      0.44      0.44      1112
           4       0.57      0.42      0.48      1112
           5       0.65      0.61      0.63       942

    accuracy                           0.51      6566
   macro avg       0.51      0.51      0.51      6566
weighted avg       0.51      0.51      0.51      6566
```

Fig 11: Training time and classification report of GaussianNB

Fig 11 shows that the GaussianNB model has accuracy 0.51 which means it is predicting correctly more than it is incorrect. However, the accuracy of 0.51 is quite low and needs improvement to make use of this model. The metrics:

precision, recall and f1-score on the class 5 (rock) is higher compared to other classes while the metrics are lower on classes 1 (latin), 2 (pop) and 3 (r&b). This means the model is having difficulty differentiating the songs of classes 1, 2 and 3 while it is more confident on songs of class 5.

Random Forest
Running time: 10.547629117965698 seconds

	precision	recall	f1-score	support
0	0.91	0.89	0.90	1236
1	0.93	0.95	0.94	1062
2	0.88	0.87	0.88	1102
3	0.92	0.94	0.93	1112
4	0.96	0.96	0.96	1112
5	0.95	0.97	0.96	942
accuracy			0.93	6566
macro avg	0.93	0.93	0.93	6566
weighted avg	0.93	0.93	0.93	6566

Fig 12: Training time and classification report of RandomForestClassifier

Fig 12 shows that the accuracy of the RandomForestClassifier model is 0.93 which means the model is correct about 93% of the time. The precision, recall, and F1-score are all above 88% for each class, which indicates that the model is doing well on classifying instances correctly, finding all positive instances and balancing the trade-off between precision and recall.

3.3. Confusion Matrix

Confusion matrix was prepared for each model which compares the actual values with the predicted values for each target class.

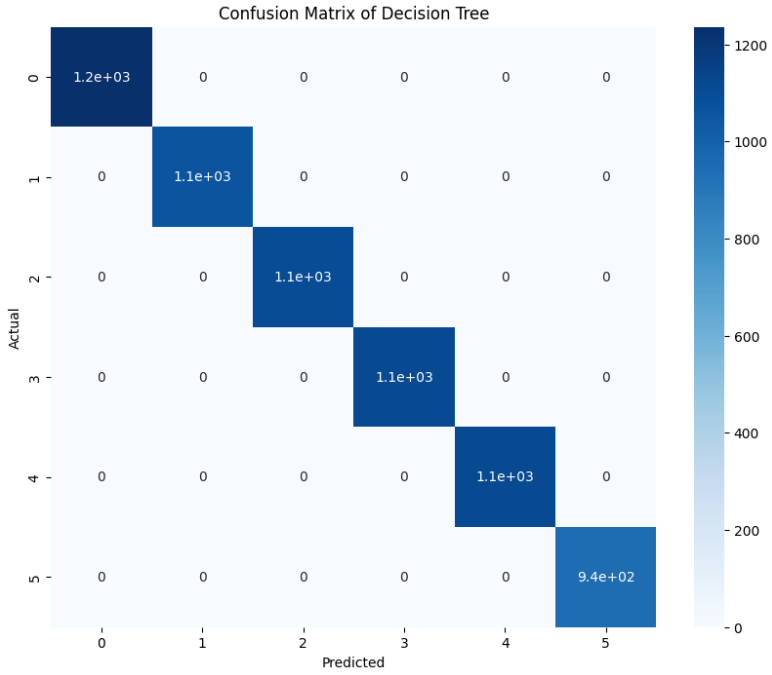

Fig 13: Confusion Matrix of DecisionTreeClassifier

The confusion matrix shown on Fig 13 has value 0 on all the cells except the diagonals. This means that the DecisionTreeClassifier model is predicting every target class correctly without any error.

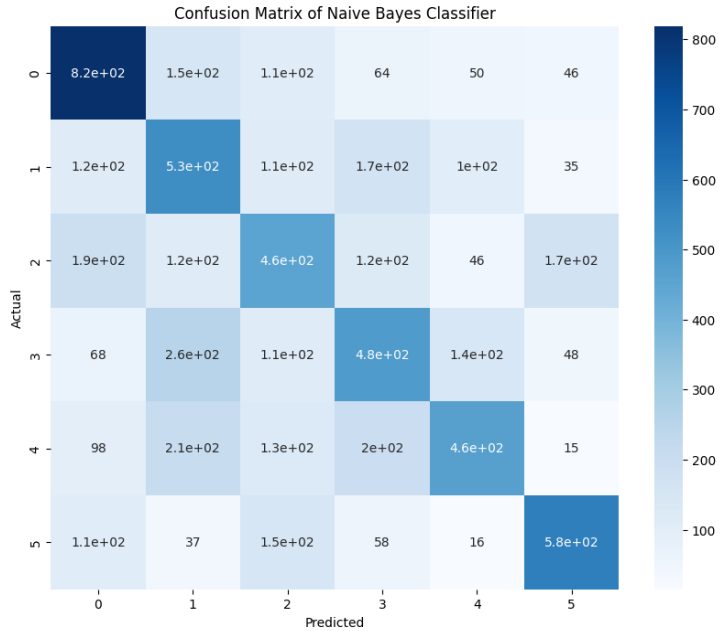

Fig 14: Confusion Matrix of GaussianNB

From the confusion matrix of GaussianNB in Fig 14, it can be seen that all the cells outside of the diagonals also have values greater than 0. This means a lot of misclassification is being done by the model. It can also be seen that the row and column of class 5 has less values on the misclassified cells while the row and column of classes 1 and 2 have higher values on cells other than the diagonal cells. This means that the model has a higher chance of correctly classifying the song as class 5 (rock) and lower chance of correctly classifying the song as class 1 (latin) and 2 (pop).

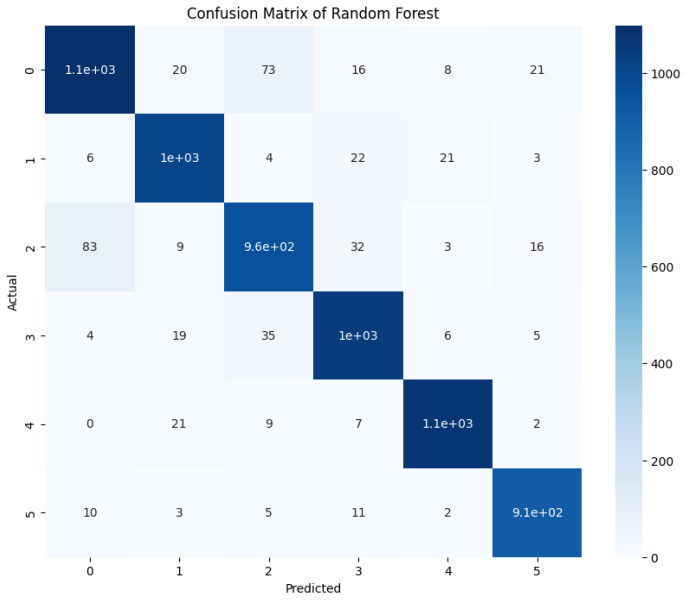

Fig 15: Confusion Matrix of RandomForestClassifier

In the confusion matrix of RandomForestClassifier shown in Fig 15, it can be seen that the values on the cells other than the diagonals are very less. This means a very small amount of data is being misclassified by this RandomForestClassifier model. The most that this class has misclassified is that it has predicted 83 instances of class 2 songs as class 0 songs.

Overall, after following the training process explained on the Model Training section of the report, the DecisionTreeClassifier model seems to be the most accurate among the three models with 100% accuracy and being the model to train the fastest, GaussianNB model is the least accurate with 51% accuracy. The DecisionTreeClassifier model has less training time and highest accuracy, so it seems to be the optimal model for performing prediction.