Song/Music Genre Classification

OBJECTIVE:

To analyse the audio and lyrical features that correlate with a song's popularity (if any) and classify the song's genre among various genres based on these features.

INTRODUCTION:

A music genre classifier is a software program that predicts the genre of music in audio format. Accurately categorising the vast spectrum of musical styles presents a unique challenge. Combining the worlds of music, data science, and machine learning, the goal is to develop a robust system capable of discerning the nuanced characteristics that define genres like rock, pop, jazz, and hip-hop.

STEPS IN THIS PROJECT:

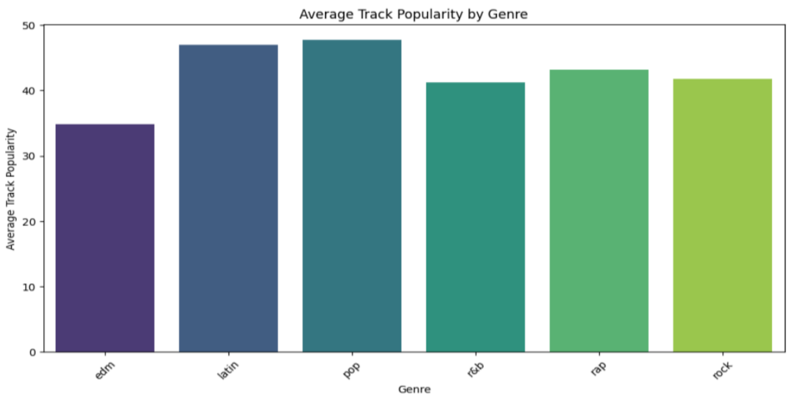
1) Importing the libraries

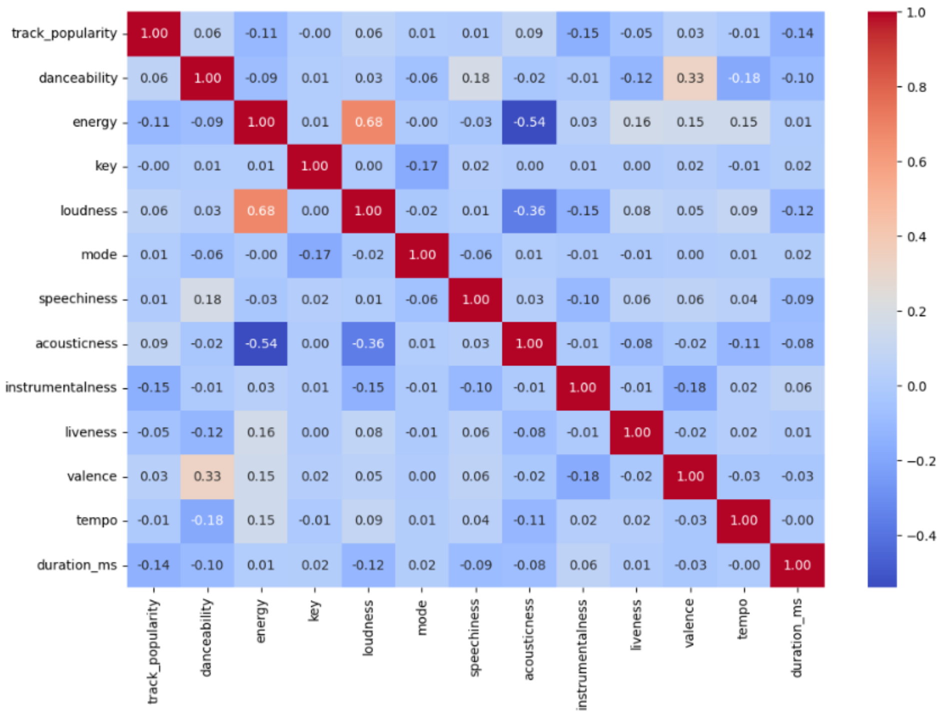
Initially, python libraries such as Pandas, NumPy, Matplotlib, and Seaborn were imported and used. Later, scikit-learn was imported for data processing, model training, testing, and evaluation.

2) **Loading the dataset:** The Spotify songs Comma-separated file (CSV) file containing around 32000 is loaded for analysis.

3) Exploratory Data Analysis:

- A total of 23 attributes are present in the dataset with data types: float64 (9 attributes), int64 (4 attributes), and object(10 attributes).
- The average song duration in the dataset is 225799.811622 ms or 3 min and 45 sec.
- The pie chart visualizes the percentage of different genres in playlists. Through the chart, we can see that EDM has 18.4% of the data, Rock 15.1%, Latin 15.7%, R&B 16.5%, Pop 16.8%, and Rap 17.5%.
- The distribution of data points for energy and genre is visualized using a violin plot.
- Generally, pop and latin tracks tend to be more popular.
- EDM tends to be the less popular genre.

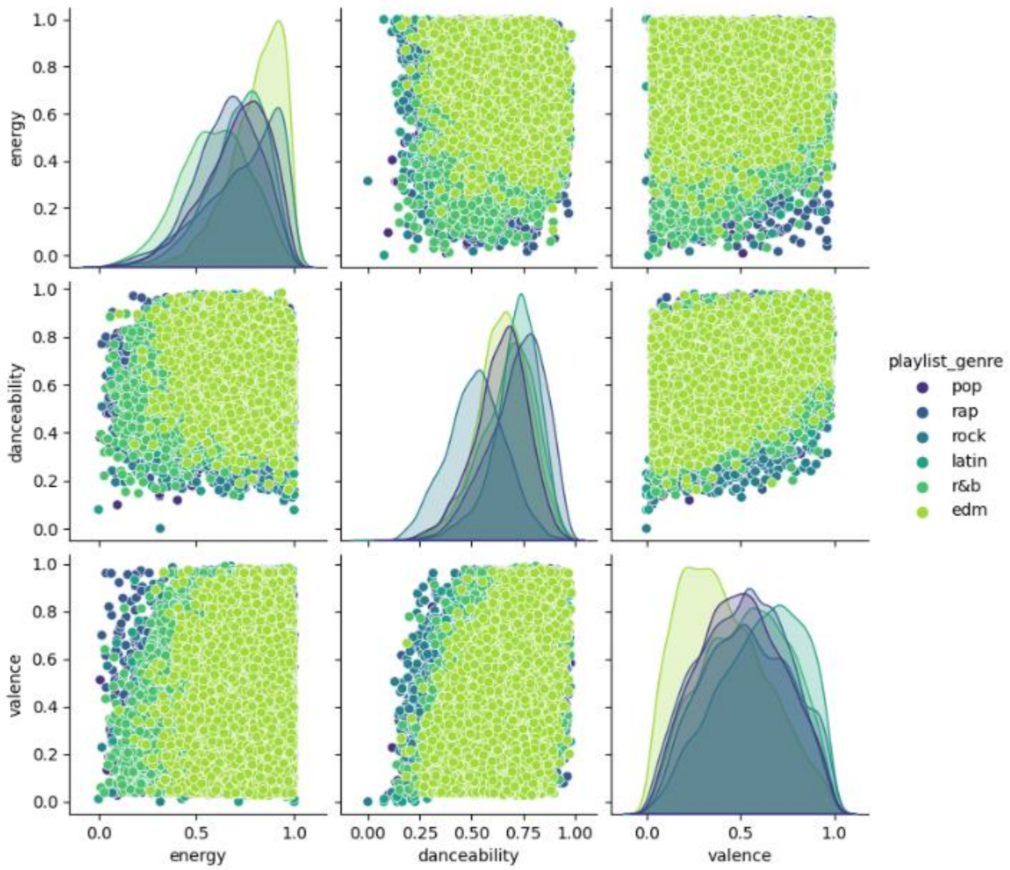




- From the correlation heatmap, we obtain the following inferences:
  1. As the loudness of a song increases, the energy of the song is also likely to increase. Conversely, if the loudness decreases, the energy is more likely to decrease.

2. With a correlation of -0.14 between track_populariy and duration_ms, there is a weak tendency for longer songs to be associated with slightly lower track popularity on average. However, the relationship is not strong, and other factors may contribute to the overall picture.
3. Based on the correlation coefficient of -0.15, there is a weak tendency for songs with higher instrumentalness to be associated with slightly lower track popularity on average. However, the relationship is not strong, and other factors may play a role in determining track popularity.
4. Features: energy and acousticness are inversely proportional with a correlation of -0.54. This means that singing **acoustic songs requires less energy**.
5. Features: loudness and acousticness have a negative correlation of -0.36.
6. Features: danceability and valence have a positive correlation of 0.36.
7. The weak correlations suggest that predicting track popularity is likely influenced by a combination of factors beyond the scope of the current dataset.

- Rock and edm songs tend to be more energetic.
- r&b songs are usually the least energetic.
- A pair plot is used to get the relationship between energy, danceability and valence variables for data points belonging to different genres.



- This analysis serves as a starting point for music industry professionals, researchers, and enthusiasts interested in exploring trends and patterns in the ever-evolving landscape of music.

4) Feature engineering :

4.1) Encoding:

Label encoding is performed. Label encoding is a technique used in machine learning and data analysis to convert categorical variables into numerical values. In the Spotify dataset, 10 attributes of object type are encoded using the label encoder.

4.2) Feature scaling :

- Feature scaling is the process of normalizing the range of independent variables or features in a dataset.
- It's a data preprocessing technique that helps with data analysis and modelling.
- Standard scaler which removes the mean and scales each feature/variable to unit variance is used to scale all the values in the dataset.

5) Train - Test split: The dataset is split into train and test sets with test size = 0.3 i.e., 30%. The remaining 70% will be the size of the train set.

6) Machine learning models:

1) Decision Tree Classifier

- A decision tree classifier is a class that can perform multi-class classification on a dataset.
- This algorithm is used to classify genres based on 22 independent attributes.
- **Building the model:** The decision tree classifier is built with random state 42 for random seed reproducibility.
- **Training & Testing the model:** The model is fitted on the train set and is tested with a test set. The predictions were obtained.
- **Evaluating the model:** The following evaluation metrics were used and scores were obtained:
    1. Accuracy: 0.9984771573604061
    2. Precision: 0.9984791791265867
    3. Recall: 0.9984771573604061
    4. F1-score: 0.9984763933613913
    5. Confusion Matrix: [[1827 0 7 0 0 5]

       [0 1543 0 0 0 0]

       [2 0 1639 0 0 0]

       [0 0 0 1614 0 0]

       [0 0 0 0 1725 0]

       [1 0 0 0 0 1487]]

2) Random Forest Classifier

- Random forest classifier (RF) builds multiple decision trees on several randomly selected subsets of the training dataset.
- Combines the output of a multiple decision tree to reach a single result.
- **Building the model:** In this random forest classifier the number of estimators taken was 100.
- **Training & Testing the model:** The model is fitted on the train set and is tested with a test set. The predictions were obtained.
- **Evaluating the model:** The following evaluation metrics were used and scores were obtained:
    1. Accuracy: 0.9855837563451777
    2. Precision: 0.9984791791265867
    3. Recall: 0.9984771573604061
    4. F1-score: 0.9984763933613913

CONCLUSION:

The decision tree classifier and random forest classifier performed well and gained the best accuracy of about 99.8% and 98.5%. This is very helpful not only in classifying songs based on the genre but also in recommending songs to users based on their preferred genre.