Song Predictive Analysis

Team Members: Adarsh Agrawal (A20517609)
Abhishek Sharma (A20506258)
Dheeraj Goud (A20500290)

Project Description:

Music preferences have evolved over time, and with increased access to new music, music platforms are utilizing machine learning to identify patterns and characteristics that can give them a competitive edge in attracting and retaining users. Our project aims to explore how music platforms leverage the data they collect to enhance user experience through personalized music recommendations and micro-genre categorization. An example of this can be seen in Spotify's mobile app, which has organized its music library into numerous sub-genres using a combination of machine learning and expert curation.

Our project aims to address specific questions, and while we may uncover additional insights from the data, we plan to focus on answering these questions.

- Predict the year of the song based on different characteristics like album cover etc.
- Given a song, what are some other recommendations that would suit the same user's taste?
- Can we predict the popularity of a song based on given features?
- Which genre got famous/changed according to year and why?

Methodology:

- Data Cleaning and Exploration.
    - Working around missing values.
    - Clearing outliers.
- Feature Engineering.
    - Finding Correlation between columns.
- Model Training.

Our project involves training a model to predict various traits, and to assess its effectiveness, we intend to utilize k-fold cross-validation and test the dataset. We will apply regression, classification and clustering techniques to answer our research questions. In assessing classification problems, we plan to use the F1 score and possibly the confusion matrix to evaluate the model's performance. For clustering, we will consider the random index and silhouette score as suitable metrics.

Datasets:

- Dataset 1: Kaggle (https://www.kaggle.com/datasets/lehaknarnauli/spotify-datasets?resource=download&s

+elect=tracks.csv&select=tracks.csv). This dataset has two subsets of data, artists & tracks. Here are the details for each dataset:
    - Tracks: 20 columns and 586,672 rows
    - Artists: 5 columns and 1 million rows
- Dataset 2: Kaggle (https://www.kaggle.com/datasets/mrmorj/dataset-of-songs-in-spotify?resource=download+d). This dataset contains 2 subsets as well.
    - Genres_v2: 22 columns and 42 thousand rows.
    - Playlists: 2 columns and 40 rows.

Models:

Classification:

- Logitboost
- Random Forest

Clustering:

- KMeans

References:

- “Genre Classification of Spotify Songs using Lyrics, Audio Previews, and Album Artwork”, by Tyler Dammann and Kevin Haugh, Stanford University. (2017)
- “Neural Network Music Genre Classification Classification”, by Nikki Pelchat and Craig M. Gelowitz. IEEE, (2020)
- “A Survey of Music Recommendation Systems and Future Perspectives”, by Yading Song, Simon Dixon, and Marcus Pearce, Queen Mary University of London. (2012)