Group 8 Project Proposal

Team Members → Zhuo Chen, Rutvik Deo, Ishita Dutta, Boquan Fang (Team Leader)

Dataset Link → https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube

Dataset In Use → User Inputed Playlist

Dataset Description → This dataset contains statistics on many songs in Spotify and Youtube. For the purposes of this project, we will only stick to Spotify half of the data, not including the values that are for the corresponding Youtube representation of the song.
Statistics measured:

- Track
- Artist
- Url_spotify
- Album
- Album_type
- Uri
- Danceability
- Energy
- Key
- Loudness
- Speechiness
- Acousticness
- Instrumentalness
- Liveness
- Valence
- Tempo
- Duration_ms
- Stream

We will add one more column, our popularity as a direct measure from the number of streams of the music video (categorize stream values). Data is as recent as Feb 7 2023, can be more recent at time of download.

Problem Statement → Given a playlist and a song that is not included in the playlist, can we determine if the song would fit well and provide a good listening experience if it were added to that playlist? Formerly, can we accurately predict the popularity of a new song? Given an old song, how popular will it be from now on?

Goal → The goal of this project is to use multiple statistical machine learning techniques on spotify data in order to, based on the variables describing the data such as length of the song, and listening characteristics. Our previous goal was to predict how popular a song will be based on those attributes. The Project goal is to determine if the new song fit in the energy with the given playlist.

Potential Data Methods → Classification methods like tree model, KNN, logistic regression, etc. Use K-fold cross-validation in order to get a rounded accuracy score for any and all methods used before deciding on a final model.

Timeline →

- Project Proposal → Apr 16th
- Pre-exploratory analysis → 18th
- Data Cleaning and preparation → 21st
- Literature Review → 22 rd
- Exploratory Analysis → 26th
- Running Models for accuracy → May 9th
- Final model selection → May 13th

- Creating interface → May 19th
- Making the presentation and demo → May 25th, Final Presentation Due May 30