IE 7275 Data Mining in Engineering

Project Title:

Spotify Data Analysis

Milestone: Project proposal

Group: 8

Student 1: Samiksha Baraskar +1857-390-5787 baraskar.s@northeastern.edu

Student 2: Deepshikha Soni +1857-230-1038 Soni.deeps@ northeastern.edu

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Samiksha Baraskar

Signature of Student 2: Deepshikha Soni

Submission Date: July-10-2023

Problem Setting:

In the age of streaming platforms, music plays a crucial role in people's lives. Understanding users' preferences and analyzing music data can provide valuable insights for personalized recommendations, genre analysis, and other applications. However, exploring and extracting meaningful information from large-scale music datasets pose challenges due to the diverse range of musical genres, audio features, and user preferences.

Problem Definition:

The specific problem addressed in this project is to perform comprehensive data analysis on the Spotify music dataset. Our goal is to utilize data mining principles, machine learning algorithms and processes to gain insights into the relationship between audio features, genres, and user preferences. We aim to explore techniques for genre classification, visualize the distribution of genres based on audio features, and uncover patterns and trends in the dataset. Few questions that we want to address are 1)What are the features that contribute to the popularity of the artists? 2)What music genre posses the highest popularity and how the trend changes affect the popularity of the songs and the artists?

Data Sources:

Few sources we would explore and use the Spotify data are :

I. https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset
II. https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-600k-tracks
III. https://zenodo.org/record/2594557
IV. https://huggingface.co/datasets/osanseviero/top-hits-spotify

Data Description:

The dataset includes 21 columns and approx. 114000 records with following features:

- track_id: The Spotify ID for the track
- artists: The artists' names who performed the track.
- album_name: The album name in which the track appears
- track_name: Name of the track
- popularity: The popularity of a track is a value between 0 and 100, with 100 being the most popular.
- duration_ms: The track length in milliseconds
- explicit: Whether or not the track has explicit lyrics (true = yes it does; false = no it does not OR unknown)
- danceability: Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity.
- energy: Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.
- key: The key the track is in. Integers map to pitches using standard Pitch Class notation
- loudness: The overall loudness of a track in decibels (dB)
- mode: Mode indicates the modality (major or minor) of a track, the type of scale.
- speechiness: Speechiness detects the presence of spoken words in a track.

- acousticness: A confidence measure from 0.0 to 1.0 of whether the track is acoustic.
- instrumentalness: Predicts whether a track contains no vocals.
- liveness: Detects the presence of an audience in the recording
- valence: A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track.
- tempo: The overall estimated tempo of a track in beats per minute (BPM).
- time_signature: An estimated time signature.
- track_genre: The genre in which the track belongs

Project Planning :

Our plan to execute this project is as follows:

1) Data Collection and Pre-processing:
   Retrieve the Spotify music dataset using various available open-source datasets. Pre-process the dataset by handling missing values, normalizing audio features, and addressing any data quality issues.

2) Data Exploration and Visualization:
   Visualize the distribution of music genres based on different audio features using techniques such as bar charts, scatter plots, and histograms. Explore correlations between audio features and track popularity, identifying key features that contribute to popular tracks.

3) Applying Data Mining and Machine Learning techniques :
   Train and evaluate machine learning models (e.g., decision trees, random forests, neural networks) to classify tracks into different genres based on their audio features and to gain other different insights from the data .Each team member will implement different models and techniques.

4) Performance Evaluation :
   Assess the performance of the models using appropriate evaluation metrics. Perform statistical analysis and hypothesis testing to uncover relationships between features. Implementation by each team member will be compared and merged to analyse which model performs best for our dataset.

5) Final Analysis and Reporting:
   Summarize the findings and insights from the data analysis, including visualizations, and results. Present a comprehensive report detailing the project methodology, results, and recommendations for further research.

Few Challenges which may occur :

- Exploring various open-source datasets available may result in a large dataset which may affect the data pre-processing and analysis.
- Dealing with imbalanced genre distributions in the dataset during genre classification.
- Ensuring the interpretability and explainability of the classification models to understand the factors driving genre classification.

Overall, this project offers an exciting opportunity to delve into the world of music data analysis, applying data mining principles and machine learning techniques to gain valuable insights into music genres and user preferences.