Abstract

This project plans to perform a comprehensive exploration and preprocessing of a large Spotify audio dataset (3 GB) from Kaggle, which focuses on audio characteristics and streaming statistics. Our approach will include statistical summaries and correlation analysis across various dimensions such as geography, time, and genre to determine patterns and anomalies in global music preferences. The preprocessing phase will handle the cleaning and normalization of categorical data to prepare for advanced analytics, while the data visualization phase will help with understanding and hypothesis formulation regarding streaming trends. Supervised learning methods like linear regression will be employed to model and predict user engagement and streaming trends, while unsupervised learning methods like k-means clustering will be utilized to perform grouping of songs and user behaviors to uncover hidden patterns in our data. Ultimately, this effort will optimize the large dataset for predictive modeling that can be used to improve content personalization and strategic decision making for the music streaming industry.

Dataset: https://www.kaggle.com/datasets/sunnykakar/spotify-charts-all-audio-data/data