Spotify Playlist Extension with Pattern Mining and Clustering: Project Proposal

1. **Team Members:**

Team Member 1: *Adarsh Singh [G39508544]*

As this is an individual project, all responsibilities, including data acquisition, processing, analysis, modeling, evaluation, and documentation, will be handled independently. Time will be managed to allocate focused phases for exploratory analysis, experimentation, and result synthesis. Due to the dataset's scale, special care will be taken to prioritize manageable subsets and efficient computation to stay within local processing capabilities.

2. **Business Problem:**

The music streaming industry continually faces the challenge of delivering personalized and contextually relevant recommendations that keep users engaged and satisfied. This project aims to address the problem of automatic playlist continuation, a critical feature that predicts and recommends tracks to extend user-created playlists with thematic and musical coherence. By providing accurate and meaningful playlist continuation, the platform can enhance user experience, increase listening time, and strengthen user loyalty—ultimately driving higher retention and competitive advantage for a streaming service.

3. **Data Sets:**

3.1. Spotify Million Playlist Dataset Challenge: Challenge Set

3.1.1. Contains 10,000 incomplete playlists with partial track information and/or titles.
3.1.2. Focused on in this project due to its manageable size and rich metadata conducive to answering research questions on playlist continuation.

3.2. Spotify Million Playlist Dataset (MPD)

3.2.1. Contains 1,000,000 playlists with large-scale, detailed information.
3.2.2. Used as a supplementary dataset to validate and extend findings from the challenge set without processing the entire dataset at once.

Both datasets are publicly accessible and appropriate for data mining tasks involving large, complex, and structured playlist and track metadata.

Link to dataset: https://www.aicrowd.com/challenges/spotify-million-playlist-dataset-challenge

4. Proposed Solution/Experiments:

4.1. Research Questions

4.1.1. How often do songs co-occur or co-disappear in playlists, and how can this knowledge inform recommendations?
4.1.2. Can playlists and tracks be effectively clustered by genre or other features to improve recommendation relevance?
4.1.3. How does playlist metadata (titles, partial track seeds) influence recommendation quality?

4.2. Methodology

4.2.1. Perform quantitative analysis of song co-occurrence and mutual exclusivity patterns to uncover frequent and rare track associations.
4.2.2. Apply clustering techniques (e.g., k-means, hierarchical clustering) on track metadata and audio features to group songs/playlists by genre or other musical attributes.
4.2.3. Develop recommendation algorithms constrained or informed by clusters and pattern-mined associations.
4.2.4. Use the challenge set for initial analysis and experimentation, then verify promising results on sampled subsets from the full MPD.

4.3. Techniques

4.3.1. Association rule mining and graph analysis for co-occurrence patterns.
4.3.2. Unsupervised clustering based on audio features and playlist metadata.
4.3.3. Collaborative filtering and hybrid recommendation models incorporating cluster constraints.
4.3.4. Use of evaluation metrics such as R-precision, NDCG, and recommended song clicks aligned with the challenge standards.

4.4. Evaluation

4.4.1. System performance evaluation on the challenge test set's withheld tracks using challenge metrics.
4.4.2. Comparisons between baseline recommenders and cluster- or co-occurrence-informed methods.
4.4.3. Analysis of recommendation diversity and thematic coherence.

This project plan leverages structured data mining techniques and scalable experimentation approaches suited for an individual researcher processing large-scale music playlist data. It strategically utilizes the challenge dataset for manageable experimentation and the full dataset for validation, ensuring feasibility without sacrificing depth.