https://github.com/ColitoBandito/cs4412-project-Collin

Mapping Mood Profiles and Feature Associations of Spotify Tracks

February 6, 2026

Student ID: 001042948
Data Mining W01
Collin Knowles

DATASET DESCRIPTION

The dataset chosen for this project is the Spotify Audio Features dataset, found on Kaggle. This dataset is metadata scraped from the official Spotify Web API. It has around 130,000 tracks and includes attributes for each such as danceability, energy, loudness, speechiness, acousticness, instrumentalness, and valence. There are twelve total rows of attributes that describe aspects of each track, but not all will be used for discovery.

Attributes

The analysis is centered on numerical features of tracks that express the "vibe" or structure of each of these tracks. Key attributes include:

- Acousticness: A confidence metric of whether or not the track is acoustic in nature.
- Energy: A measure of musical intensity and activity within the track.
- Valence: A metric that describes musical positiveness (happiness number) conveyed by the track.
- Speechiness: Detects the presence/amount of spoken word within the track
- Instrumentalness: Predicts the amount of instrumentation within the track as contrasted to vocals.

Data Quality Considerations

Early inspection into the data indicates that the data is properly structured but some amount of preliminary work will be required. The attributes loudness and tempo, measured in dB and BPM respectively, exist on significantly larger scales than the standard 0.0 to 1.0 range of valence or energy. First steps in normalizing the data will include standardizing these features and allowing for proper distance based mining algorithms. The aim is to avoid bias by feature magnitude.

Discovery Questions

In alignment with the stated goal of pattern discovery, the aim of this project is to investigate the following:

- Mood Groupings: Are tracks with a high relationship between energy, valence and acousticness able to mathematically be clustered into specific muscial ""moods"? Do these mood profiles transcend traditional genre description?
- Feature Co-occurrence: I will be finding a number of track feature co-occurances but an immediate example would be: Do combinations of high amounts of instrumentalness coincide with specific tempo ranges in low-speechiness tracks? Does this show patterns in ambient music as opposed to vocal music?
- Anomalous Track Profiles: What tracks do not align with established and evidenced mood profiles? A hypothetical example would be a track with high energy and low loudness.

Planned Techniques

I've identified three particular techniques that encompass the data mining categories relevant to the aforementioned questions

Clustering (K-Means and DBSCAN)

K-Means clustering will be used to partition the dataset into k groups in an effort to answer the mood profiling question. Both the Elbow Method and Silhouette Analysis will be used to determine what that number of clusters will look like. For finding outlier tracks or more non-spherical data shapes, DBSCAN will be used.

Dimensionality Reduction (PCA)

As it stands there are over 10 numerical features for each track, which may lead to difficulties in data visualization. I aim to use PCA to reduce this feature space for tracks into two or three primary components. This should give better cluster separation when doing a visual assessment as well as help with identifying global scale patterns.

Association Rule Mining (Apriori)

To enable finding multiple occurances of track features, I will assign discrete categorization to these continous variables. An example would be mapping Energy > 0.8 to "High Energy". Afterwards, the Apriori algorithim will be used to find rules that have strong support and confidence levels, ultimately showing how the presence of some audio attributes imply the presence of other relevant ones.

DATA DICTIONARY TABLE

Table 1: Full Attribute Inventory for Spotify Dataset

Attribute	**Description / Significance**
acousticness	Confidence measure of acoustic nature (0.0 - 1.0).
danceability	Suitability for dancing based on tempo and rhythm.
duration_ms	The length of the track in milliseconds.
energy	Intensity and activity level of the track.
instrumentalness	Predicts the absence of vocal content.
key	The estimated overall key of the track.
liveness	Detects the presence of an audience in the recording.
loudness	Overall loudness in decibels (dB).
mode	Modality (Major = 1, Minor = 0).
speechiness	Detects the presence of spoken words.
tempo	The overall estimated beats per minute (BPM).
time_signature	Estimated overall time signature (3/4, 4/4, etc.).
valence	Measure of musical positiveness (0.0 - 1.0).

TIMELINE

The project will be done within three phases in accordance to the semester milestones and using the KDD process:

- **Milestone 2: Preprocessing and EDA (Weeks 4-7):** Focus on data cleaning, handling the skewed distributions of features like instrumentalness, and using PCA to reduce dimensionality for future data visualizations.
- **Milestone 3: Mining Implementation (Weeks 8-11):** Execution of the K-Means clustering algorithm to identify mood profiles and the Apriori algorithm for generating feature association rules.
- **Milestone 4: Evaluation and Final Report (Weeks 12-15):** Statistical validation of clusters using Silhouette scores and interpretation of the discovered patterns.

Anticipated Challenges

The primary challenge will be creating discrete categories based on the datasets given continuous numbers. I will have to experiment with different "bins" for these categories to ensure results yielded from them are genuinely useful and not obvious musical patterns that could be discerned with common sense.