Laboratory assignment

Component 1

Authors: Leordean Ada Alexandra, Selegean Victor
Group: 242

November 3, 2025

1 Dataset Overview

1.1 Description

The models in this project are trained using a Spotify music dataset [Ame25] created by Solomon Ameh and published on kaggle.com[1]. The dataset provides twelve audio features, eleven descriptive features, and around 5000 rows divided into two categories. The audio features are of a numerical format. Some features, such as 'Instrumentalness' and 'Acousticness' represent a probability and thus take continuous values between 0 and 1, while others like 'Key' take discrete, integer values between 0 and 11.

Below, Figure 1 shows the number of tracks respective to the top genres with most songs associated. The most represented genre is 'electronic', covering over a tenth of the entries, followed in close by 'pop'. Figure 2 shows the distribution of song tempos.

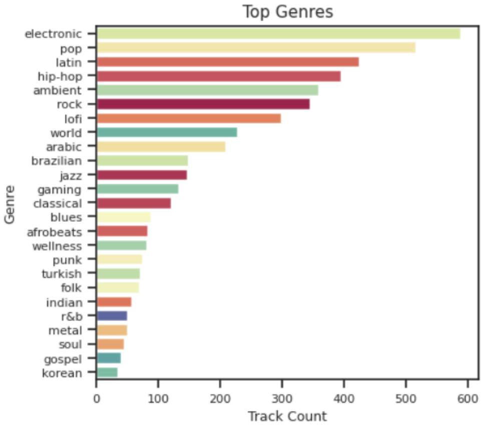

Figure 1: Track count for each genre

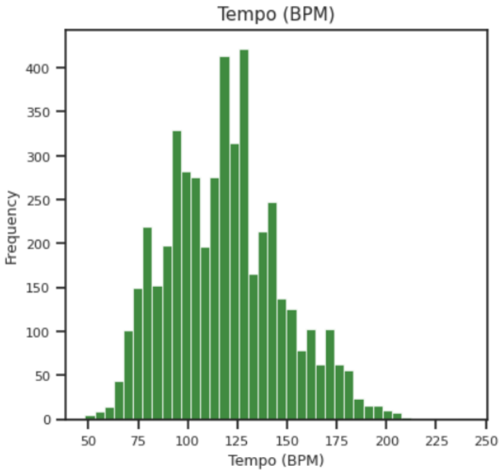

Figure 2: Tempo Distribution

1.2 Quality Considerations

The dataset has a size of 4831 entries. While it does not contain duplicates, it does contain 2 rows with missing values[2]. A decision must be made whether to drop them or insert the missing values.

[1] Full dataset is available at https://www.kaggle.com/datasets/solomonameh/spotify-music-dataset

[2] As of October 2025

2 Definition of the learning task

2.1 Supervised Learning - Linear Regression

2.1.1 Problem definition:

The task is to predict a song's popularity score with the use of regression over the dataset's audio features (e.g.: danceability, energy, tempo). This model will correlate listeners' preferences with different song characteristics, which could be useful for future recommendations.

The objective is a function

$$f_{reg} : R^n \to R$$

that maps to the input of audio features

$$x = [x_1, x_2, \ldots, x_n]$$

to an output value y, which represents the song's popularity.

2.1.2 Problem Specification

The input (x_i) consists of numerical and continuous audio descriptors for each track:

- `danceability`: how suited a song is for dancing
- `energy`: how intense/active a song is
- `loudness`: average decibel level
- `speechiness`: presence of spoken words
- `acousticness`: instrument based music that produces acoustic sounds
- `instrumentalness`: probability the track contains no vocals
- `liveness`: presence of an audience in the recording
- `valence`: emotion of song, how positive it is
- `tempo`: beats per minute (BPM)
- `duration_ms`: track duration in milliseconds

The target variable (y)**:** the track's `popularity`, a numerical score (0–100) assigned by Spotify

Preconditions:

- All missing or invalid data from the dataset must be cleaned;
- All features should be scaled or normalized to ensure comparability;
- The dataset should be divided into training, validation, and testing splits;
- Outliers and extreme values should be identified and handled appropriately.

Output: A continuous value y between 0 and 100 that represents the song's predicted popularity score.

Postconditions:

- Predictions must be a value within the valid range of popularity (0-100);
- Performance should be measured with regression metrics such as MSE, MAE or R^2.

2.1.3 Specification of the Learning Task

Task (T): a Supervised Regression Task where the model learns from given inputs (that represent audio features) to output a continuous value (that represents the song's popularity).

Performance (P): Mean Squared Error (MSE), Mean Absolute Error (MAE) and R^2 score.

Experience (E): the audio features of Spotify songs from the database and their corresponding popularity scores. As such, it is trained on songs with known popularity and learn the connection between the aforementioned and its correlated audio attributes.

2.2 Unsupervised Learning - Clustering

2.2.1 Problem definition:

The task is to group tracks according to their features into genres. This can work similarly to a recommendation system, showing which songs share matching features. The model may then be able to suggest songs similar to one already liked or identify user listening patterns. The task will employ a **partitioning method** for clustering.

Given a number K of clusters that the model will attempt to create, the objective will be a function

$$f_{cluster} : R^n \to \{1, 2, \ldots, K\}$$

that maps to the input of audio features

$$x = [x_1, x_2, \ldots, x_n] \in R^n$$

to an output value $y \in \{1, 2, \ldots, K\}$, which represents the song's designated cluster.

2.2.2 Problem Specification

The hyperparameter K, representing the number of clusters the task will work with, will be fixed to the same number of genres present in the dataset, 35.

The input (x) consists of a subset of each track's attributes: Energy, Tempo, Danceability, Loudness, Liveness, Valence, Speechiness, Instrumentalness, Mode, Key, and Acousticness. All attributes are numerical.

Preconditions:

- Rows with missing data are dropped. The number of rows with missing data represent less than 0.05% of the entire dataset.
- Features are normalized to ensure comparability.

Output: (y): a discrete natural value representing the cluster assignment for each track (from 1 to K), which can be compared to the "playlist_genre" attribute of the original dataset.

Postconditions:

- Predictions must be a value within the valid set of genres ($\{1, 2, \ldots, K\}$).
- Performance is measured using both **external evaluation metrics**, such as Purity, and **internal evaluation metrics**, such as the Davies–Bouldin Index or the Dunn Index.

2.2.3 Specification of the Learning Task

Task (T): an Unsupervised Classification Task where the model learns a (locally) optimal partition of the dataset using the K-Means algorithm.

Performance (P): Davies–Bouldin Index, Dunn Index; Purity score at the end to provide a ”sanity check”.

Experience (E): the aforementioned features of Spotify songs from the dataset. By eliminating the ”playlist_genre” column, the model is forced to find new labels for each entry.

References

[Ame25] Solomon Ameh. Spotify Music Dataset, 2025. Online; Accessed October 2025.