12/31/2025

Music Recommendation System

Name: Muhammad Hashir, Zohaib Hassan, Umair Yousaf, Muhammad Talha
Sap ID: 70137237, 70138058, 70137484, 70138652

SUBMITTED TO: SIR ASIF AHSAN

Table of Contents

Music Recommendation System Using Spotify Dataset

Detailed Project Documentation

1. Introduction

With the rapid growth of digital music platforms, users are exposed to millions of songs. Manually searching for music that matches a user's mood, taste, or preference is time-consuming and inefficient. Music recommendation systems solve this problem by analyzing song features and user behavior to suggest suitable tracks.

This project focuses on analyzing a large Spotify music dataset and predicting song popularity using machine learning techniques. Additionally, songs are classified based on **mood (valence)** and **acoustic nature**, which helps in understanding music characteristics and building a foundation for recommendation systems.

2. Problem Statement

Music platforms store massive datasets containing audio features, but extracting meaningful insights from this data is challenging. Traditional systems fail to:

- Accurately predict song popularity
- Classify songs based on emotional and acoustic features
- Analyze relationships between music attributes

This project addresses these issues by applying data analysis, feature engineering, and regression models**.**

3. Project Objectives

The main objectives of this project are:

1. To analyze Spotify music data using data science techniques
2. To perform exploratory data analysis (EDA) for feature understanding
3. To classify songs based on:
    - Mood (Valence)
    - Acousticness
4. To predict song popularity using machine learning models
5. To compare performance of different regression algorithms

4. Scope of the Project

- Dataset-based music analysis
- Popularity prediction (not real-time recommendation)
- Machine learning regression models
- Data visualization and insights

Out of Scope:

- Real-time user interaction
- Mobile or web application
- Personalized user login system

5. Dataset Description

5.1 Dataset Source

- Spotify Music Dataset

5.2 Dataset Size

- **Rows:** 155,628
- **Columns:** 22

5.3 Dataset Attributes

Attribute	Description
valence	Musical positivity
year	Year of release
acousticness	Acoustic confidence
danceability	Suitability for dancing
duration_ms	Duration in milliseconds
energy	Intensity level
explicit	Explicit lyrics (0/1)
instrumentalness	Instrumental presence
key	Musical key

Attribute	Description
liveness	Live audience presence
loudness	Overall loudness
mode	Major/Minor
popularity	Popularity score
tempo	Beats per minute
speechiness	Spoken words
artists	Artist names
name	Song title

6. Tools and Technologies

6.1 Software Tools

- Google Colab
- Python

6.2 Libraries Used

- Pandas & NumPy – Data processing
- Matplotlib & Seaborn – Visualization
- Scikit-learn – ML models
- XGBoost – Advanced regression

7. System Architecture

7.1 Workflow

1. Data Collection
2. Data Cleaning
3. Exploratory Data Analysis
4. Feature Engineering
5. Model Training
6. Model Evaluation
7. Result Analysis

8. Data Preprocessing

8.1 Missing Values

- Checked null values
- Very few missing values were present
- Removed or handled appropriately

8.2 Duplicate Records

- No duplicate records found

8.3 Feature Transformation

- Duration converted from milliseconds to minutes
- Categorical labels encoded numerically
- New features created for song classification

9. Exploratory Data Analysis (EDA)

EDA was performed to understand the dataset distribution and relationships.

9.1 Univariate Analysis

- Histograms for numeric features
- KDE plots for density estimation

9.2 Bivariate Analysis

- Scatter plots (Energy vs Loudness)
- Year vs Popularity

9.3 Correlation Analysis

- Heatmap used to identify strong correlations
- Energy and loudness showed strong positive correlation

10. Feature Engineering

10.1 Song Classification by Valence

Valence represents emotional positivity.

Valence Range	Classification
< 0.3	Sad Song
0.3 – 0.6	Balanced Mood Song
> 0.6	Happy Song

10.2 Song Classification by Acousticness

Acousticness	Type
< 0.3	Highly Electronic
0.3 – 0.6	Mixed
> 0.6	Mostly Acoustic

11. Machine Learning Model Design

11.1 Problem Type

- **Regression Problem**

11.2 Input Variables

- All numeric and encoded features except popularity

11.3 Output Variable

- Popularity

12. Models Used

12.1 Gradient Boosting Regressor

- Ensemble learning technique
- Combines weak learners
- Handles non-linear relationships well

12.2 XGBoost Regressor

- Optimized gradient boosting
- High performance and accuracy
- Handles large datasets efficiently

12.3 Decision Tree Regressor

- Tree-based model
- Easy to interpret
- Prone to overfitting

13. Model Training and Testing

- Data split:
    - 80% Training
    - 20% Testing
- Random state fixed for consistency
- Same dataset used for fair comparison

14. Evaluation Metrics

Metric	Description
R^2 Score	Measures prediction accuracy
RMSE	Root Mean Square Error
MAE	Mean Absolute Error

15. Results and Analysis

- XGBoost achieved highest accuracy
- Gradient Boosting showed balanced performance
- Decision Tree overfitted the data
- Popular songs often had balanced valence values

16. Conclusion

This project successfully analyzed a large Spotify dataset and predicted song popularity using machine learning. The study proved that:

- Mood and acoustic features strongly influence popularity
- Ensemble models outperform basic regression models

The system provides a strong foundation for building real-world music recommendation engines.

17. Future Work

- Implement collaborative filtering
- Use deep learning (ANN, LSTM)
- Add user-based preferences
- Deploy as a web application

18. References

1. Spotify Dataset
2. Scikit-learn Documentation
3. XGBoost Documentation
4. Python Data Science Handbook