Project Title

Supervised Learning for Spotify Music Classification Using Audio Features

Project Overview

This project applies **supervised machine learning** techniques to classify Spotify tracks based on their **audio characteristics**. Using Spotify's audio feature metrics—such as danceability, energy, loudness, and valence—the model learns patterns that distinguish songs into predefined **labels** (e.g., 'sad', 'happy', 'energetic', 'calm')

The main idea of Moodify project is to classify songs not only based on their lyrical and musical features, but also incorporating emotions, in order to provide users with more successful recommendation outputs. Moodify aims to categorize songs into four main emotions and utilize the similarities in musical features within these categories to offer users more effective recommendations, weighting them with emotions.
In this perspective, we build LGBM model to predict emotions of songs and we use this dataset in our both test and recommendation phases. It contains nearly 278.000 songs from Spotify and all songs emotionally labeled.

The project demonstrates the full supervised ML pipeline: **data exploration, preprocessing, feature engineering, model training, evaluation, and interpretation**. The final output can support applications such as music recommendation systems, playlist curation, and music analytics.

Link to Dataset: 278k Emotion Labeled Spotify Songs

Dataset Description

Each observation represents a Spotify track with the following features:

Feature	Description
duration (ms)	Length of the song in milliseconds
danceability	Suitability of a track for dancing (0–1)
energy	Perceived intensity and activity (0–1)
loudness	Overall loudness in decibels (dB)
speechiness	Presence of spoken words (0–1)
acousticness	Probability the track is acoustic (0–1)
instrumentalness	Likelihood of no vocals (0–1)
liveness	Presence of a live audience (0–1)
valence	Musical positivity or mood (0–1)
tempo	Beats per minute (BPM)
spec_rate	Derived spectral/audio rate feature
labels	Target variable 'sad': 0, 'happy': 1, 'energetic': 2, 'calm': 3

Problem Statement

Given a set of **audio features**, predict the **label** of a Spotify track using supervised learning techniques.

Machine Learning Approach

- **1. Exploratory Data Analysis (EDA)**

- Distribution analysis of audio features
- Correlation between features and target labels
- Class balance inspection
- Visualization of feature patterns per label

🔹 **2. Data Preprocessing**

- Handle missing values and outliers
- Feature scaling (StandardScaler / MinMaxScaler)
- Encode categorical labels
- Train-test split (e.g., 80/20)

🔹 **3. Model Training (Supervised Learning)**

The following classification models can be applied:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- Gradient Boosting (XGBoost / LightGBM)

Each model is trained to learn the relationship between audio features and song labels.

🔹 **4. Model Evaluation**

Models are evaluated using:

- Accuracy
- Precision, Recall, and F1-Score
- Confusion Matrix
- ROC-AUC (for binary classification)
- Cross-validation for robustness

Key Insights & Interpretability

- Feature importance analysis to identify which audio attributes most influence classification
- Comparison of model performance
- Insights such as:
    - High **energy + loudness** correlating with upbeat genres
    - High **acousticness + low energy** linked to calm or acoustic categories

Tools & Technologies

- **Python**
- Libraries: pandas, numpy, matplotlib, seaborn
- Machine Learning: scikit-learn, xgboost, lgbm
- Optional: SHAP for model explainability

Expected Outcome

A trained and validated supervised learning model capable of accurately classifying Spotify tracks based solely on their audio features, with actionable insights into the musical properties that define different song categories.

Future Enhancements

- Hyperparameter tuning with GridSearchCV
- Multi-class classification optimization
- Deployment as a REST API or Streamlit app
- Integration with Spotify API for real-time predictions