Machine Learning for Spotify Genre Classification

Topic: A comparative analysis of six machine learning models for predicting the genre of songs on Spotify based on their audio features and metadata.

Group 8:

Robera Abajobir
Sanghyun An
Austin Bell
Carter Prince (representative)
Tyler Varma
Anvita Yerramsetty

Dataset

Spotify Tracks Dataset:
https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-db
Contains detailed metadata and audio features for >200k songs on Spotify

Problem type: Classification
Target variable: genre (26 categories)

Plan

Models:

- Logistic Classifier
- Random Forest Classifier: Anvita Yerramsetty
- K-Nearest Neighbors Classifier: Austin Bell
- Gaussian Naive Bayes Classifier
- Gradient Boosting Classifier
- Multilayer Perceptron (Neural Net) Classifier

(these may be subject to change depending on computational and practical constraints)

Project Logistics:

- One model per person, models could be first come first served or assigned randomly.
- Each person is responsible for implementing, training, tuning, and testing their model.
- Separate “data prep” notebook performs all data preprocessing: feature engineering, cleaning, scaling, splitting. To make the model comparisons fair, each person uses the final preprocessed dataset as their training and test data
- Each person works individually in their own notebook, outputting their results in a consistent format (JSON could work). This format would contain all information

necessary for producing the final report, including the confusion matrix and evaluation metrics.
- Another separate notebook reads and synthesizes everyone's outputs and performs our final analysis, comaprison, visualization, etc.

Software

Python:

- pandas
- xgboost (for GBM)
- pytorch (for NN)
- scikit-learn (for other models)

plain Python scripts + Google Colab Notebooks
GitHub as version control