《Spotify Hit Song Prediction Project – Summary》

Project Overview

This project is based on the *Spotify Tracks 2023* dataset and aims to build a model that predicts whether a song is a “hit” based on its audio features. Since the dataset does not include a direct popularity label, this project uses the 75th percentile (Q3) of streams as the threshold to classify songs:

- **Hit (1): Songs within the top 25% in streaming volume**
- **Not Hit (0): Songs below the Q3 threshold**

The project follows a complete end-to-end data science workflow:

Data Processing (Extract) → Model Training (Learn) → Prediction Demonstration (Predict)

Data and Features

The dataset contains a variety of rhythm and audio-related attributes. The following features were selected as model inputs:

- **bpm**
- **danceability_%**
- **energy_%**
- **valence_%**
- **acousticness_%**
- **instrumentalness_%**
- **liveness_%**
- **speechiness_%**

The target variable (Hit) is generated automatically using the 75th percentile of the streaming counts.

Methods and Models

Two major classification models were built and compared:

1. Decision Tree

- **Highly interpretable**
- **Fast training speed**

Used as a baseline model to understand feature splits and decision logic.

2. Random Forest

- **Ensemble of multiple trees with stronger robustness**
- **Better performance than a single decision tree**

This model was selected as the final model for prediction demonstration.

Evaluation metrics include:
Accuracy, Precision, Recall, F1-score, and Confusion Matrix

Key Results

The Random Forest model achieved the best performance:

- **Accuracy: approximately 0.77**
- **Higher recall for Hit (1) compared to the decision tree**
- **The confusion matrix indicates that the model can detect popular songs to some extent, though false negatives still occur**

Interpretation of FP/FN in this project:

- **FP (False Positive): A non-hit song is predicted as a hit**
  → May cause a recommendation system to push irrelevant songs
- **FN (False Negative): A hit song is predicted as non-hit**
  → More critical, as it results in missing genuinely high-value songs

In practical music recommendation scenarios, reducing FN is especially important.、

Prediction Demonstration

In the final stage, 10 samples from the test set were selected and evaluated using the Random Forest model, displaying:

- **True labels (y_true)**
- **Predicted labels (y_pred)**

This validates the model's usefulness on realistic samples.

Limitations and Future Enhancements

Current Limitations

- **Does not use advanced models such as neural networks**
- **Only audio-based features were included; contextual features like artist popularity or release date were not used**
- **No hyperparameter tuning performed**
- **The definition of "Hit" relies solely on streaming percentiles, without incorporating other popularity indicators**

Future Improvements

- **Apply GridSearchCV for parameter tuning**
- **Incorporate external contextual features (e.g., social media engagement)**
- **Experiment with more advanced models (XGBoost, LightGBM, deep learning)**
- **Build an interactive visualization app using Streamlit**

Conclusion

This project successfully demonstrates a complete end-to-end data science process:

Data preparation → Feature engineering → Model training → Performance evaluation → Prediction showcase

The final model (Random Forest) provides meaningful capability in predicting Spotify hit songs and can serve as a baseline for recommendation systems and music content analytics.