《Spotify 热门歌曲预测项目 Summary（中文）》

项目概述

本项目基于 *Spotify Tracks 2023* 数据集，旨在构建一个模型，用歌曲的音频属性预测一首歌是否属于热门歌曲（Hit）。由于原始数据中未提供直接的流行度评分，本项目采用 **streams**（播放量）第 **75** 百分位数（**Q3**）作为阈值，将歌曲划分为：

- **Hit（1）**：播放量处于前 25%
- **Not Hit（0）**：播放量较低的歌曲

该项目完整涵盖数据科学流程的三个阶段：
数据处理（**Extract**）→ 模型训练（**Learn**）→ 预测展示（**Predict**）

数据与特征

数据集包含歌曲的多项音频与节奏特征。本项目选取以下特征作为模型输入：

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

目标变量（**Hit**）由 **streams** 高于 **75%** 分位数自动生成。

方法与模型

项目使用了两种主流分类模型进行比较：

1. 决策树（Decision Tree）

- 易解释
- 训练速度快
  作为基线模型，帮助理解特征分裂与决策逻辑。

2. 随机森林（Random Forest）

- 多树集成、鲁棒性更强
- 表现优于决策树
  作为最终模型用于预测演示。

模型评估指标包括：
Accuracy、Precision、Recall、F1-score、混淆矩阵

核心结果

随机森林模型表现最佳：

- 准确率（**Accuracy**）：约 **0.77**
- 热门歌曲（**Hit=1**）的召回率高于决策树
- 混淆矩阵显示模型对热门歌曲（少数类）有一定预测能力，但仍存在漏判情况（**FN**）。

FP/FN 的项目解释：

- **FP**（误判成 **Hit**）：实际不热门的歌被预测为热门 → 会导致推荐系统推送错误歌曲
- **FN**（漏掉 **Hit**）：热门歌曲被预测为普通 → 更严重，因为会错过真正的高价值歌曲

在音乐推荐场景中，**FN** 更需关注。

预测演示（Predict）

本项目最后从测试集中选取 **10** 首歌，使用随机森林模型进行预测，展示：

- 真实标签（**y_true**）

- 预测标签（**y_pred**）

验证模型在实际样本上的可用性。

限制与未来改进方向

当前限制：

- 未使用神经网络等复杂模型
- 特征仅限音频属性，缺少艺术家知名度、发行时间等关键上下文信息
- 未做超参数调优
- **Hit** 的定义仅通过播放量百分位数完成，未考虑更多流行度指标

改进方向：

- 使用 **GridSearchCV** 进行参数调优
- 加入更多外部特征（如社交媒体热度）
- 尝试更高级模型（**XGBoost**、**LightGBM**、深度学习）
- 构建 **Streamlit** 可视化 **web** 应用

总结

本项目成功展示了一个端到端的数据科学流程：
从数据准备 → 特征工程 → 模型训练 → 性能评估 → 最终预测展示。
最终模型（随机森林）可以在一定程度上预测 **Spotify** 热门歌曲，为推荐系统和音乐内容分析提供基础支持。