# Harmonizing Hit Predictions: Leveraging Audio Features to Forecast Spotify Top Songs.

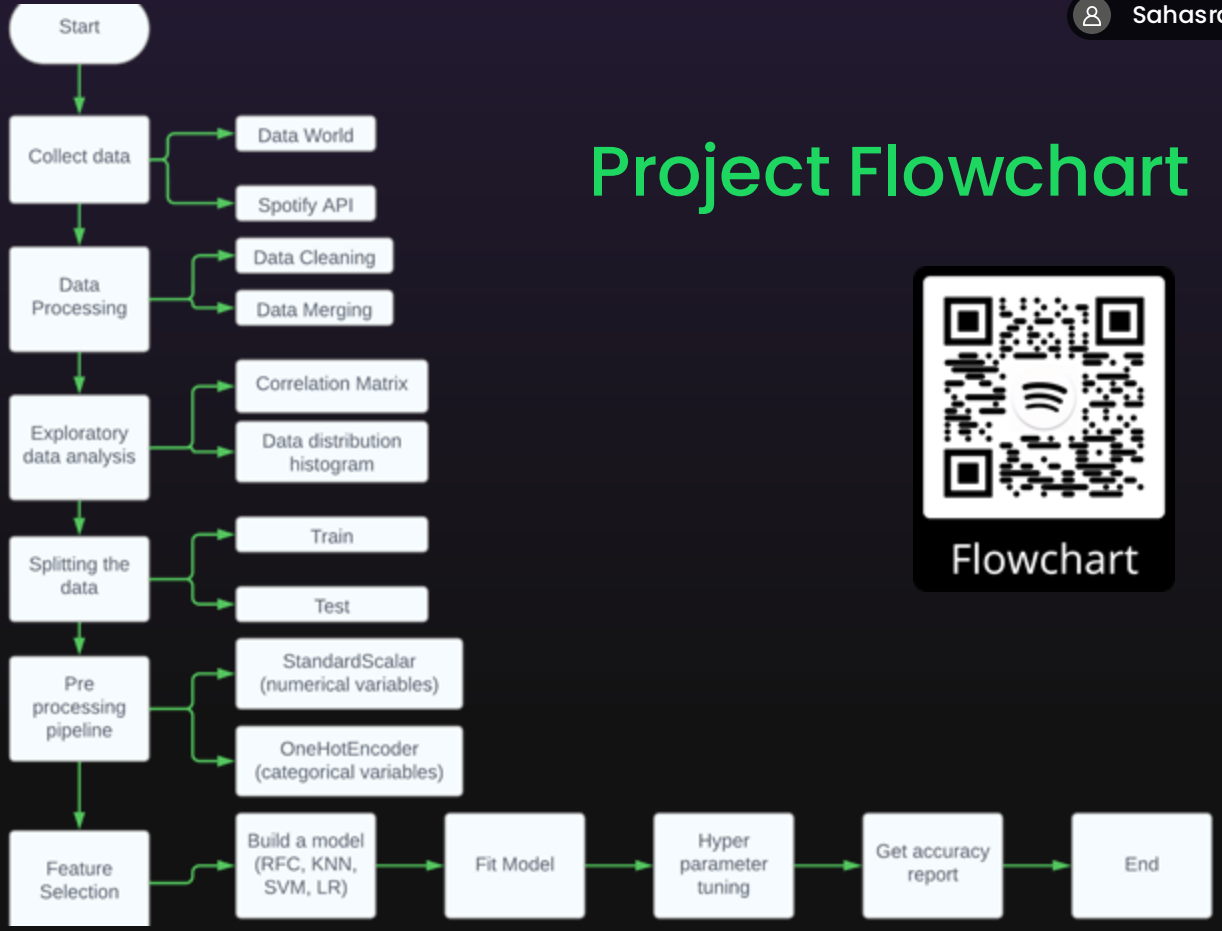
Team A05
Dylan Kakkanad
Priyanka Chaudhari
Saachi Dholakia
Sahasra Konkala

# Importance of This Analysis:

1. Strategic Release Tactics
2. Optimized Investment
3. Global Promotion
4. Enhanced User Recommendations

Play Follow

# Project Flowchart




Flowchart

# Data Description

Shape: 92233 Rows x 20 Columns

## Overview of the Data

- Highly Imbalanced on the target variable
- Low Correlation between target variables and predictor variables.

## Balance the Data

- Using the Spotify API we pulled various playlists of popular songs.
- We then extracted audio features of these songs by a second API call.
- By adding more popular songs into our dataset we hoped to balance the data.

# Data Cleaning

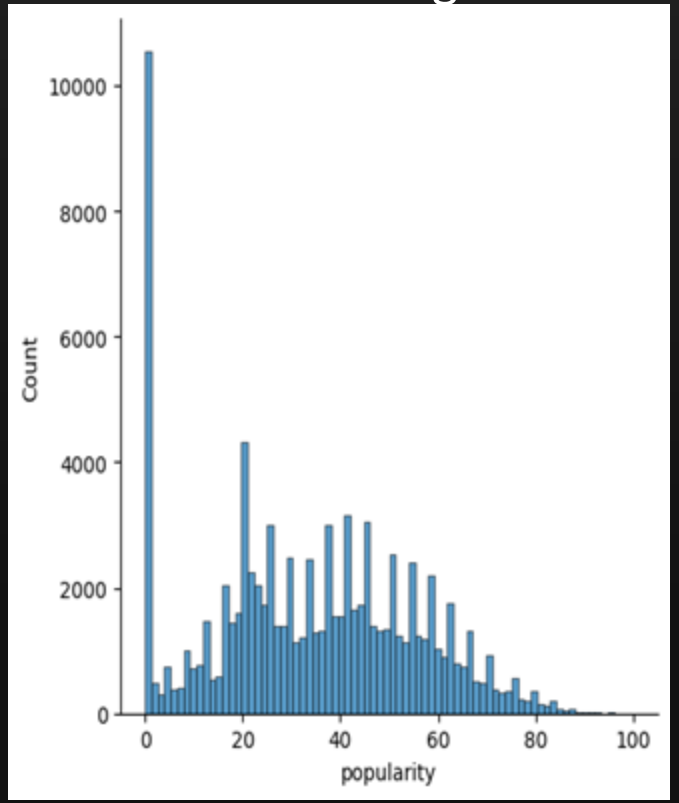


- Popularity was a continuous score ranging from 0-100. We chose a certain cutoff threshold to classify a song as popular or not popular.
- Removed tracks with really low popularity score (0-10).
- From the artists columns we create 2 new columns: Mult-artists and No of artists.
- Drop Duplicate rows based on track_ids.
- Set a cut-off threshold of 55 to classify songs equal to and above as popular and below as not popular

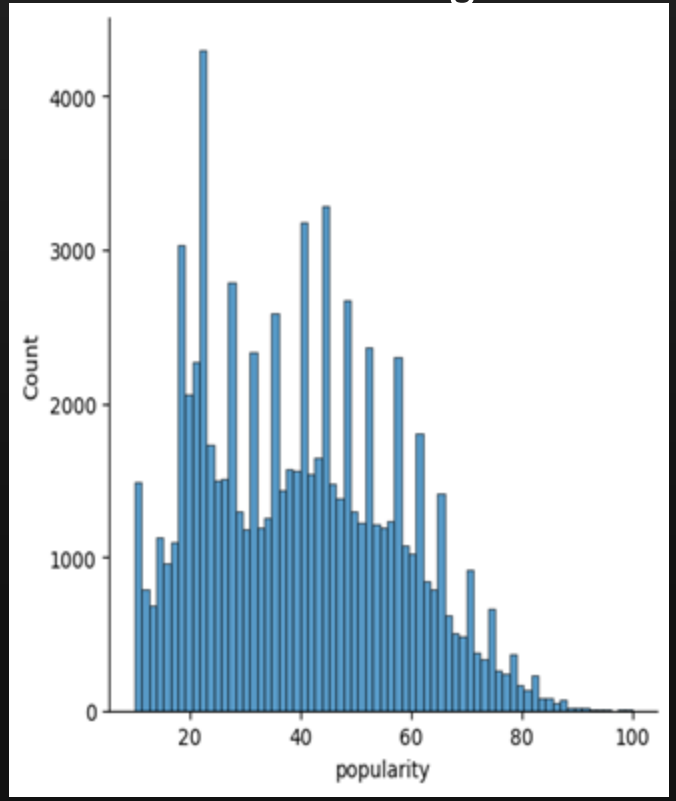
# Distribution of Popularity



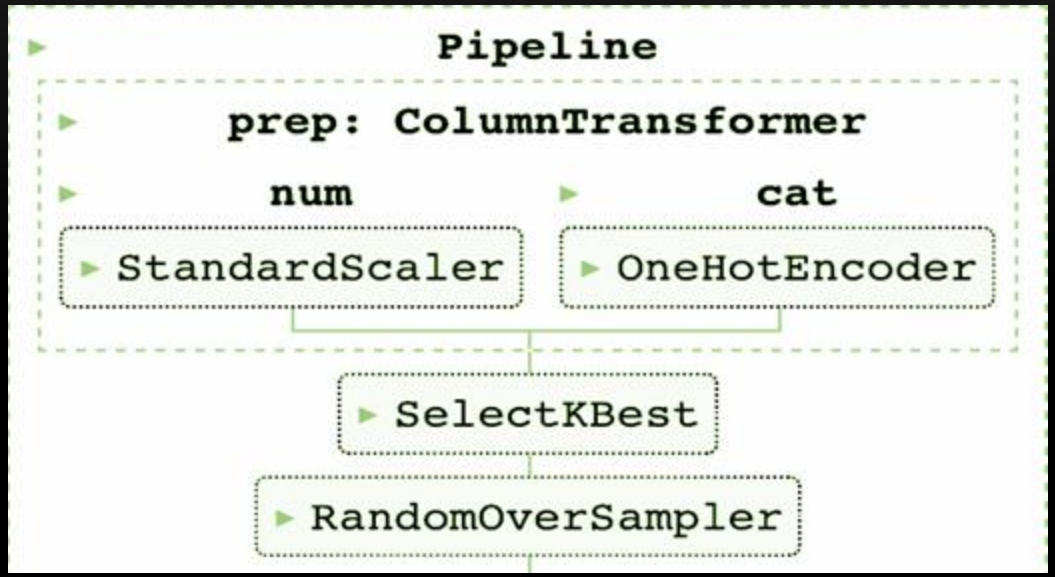
## Before Cleaning



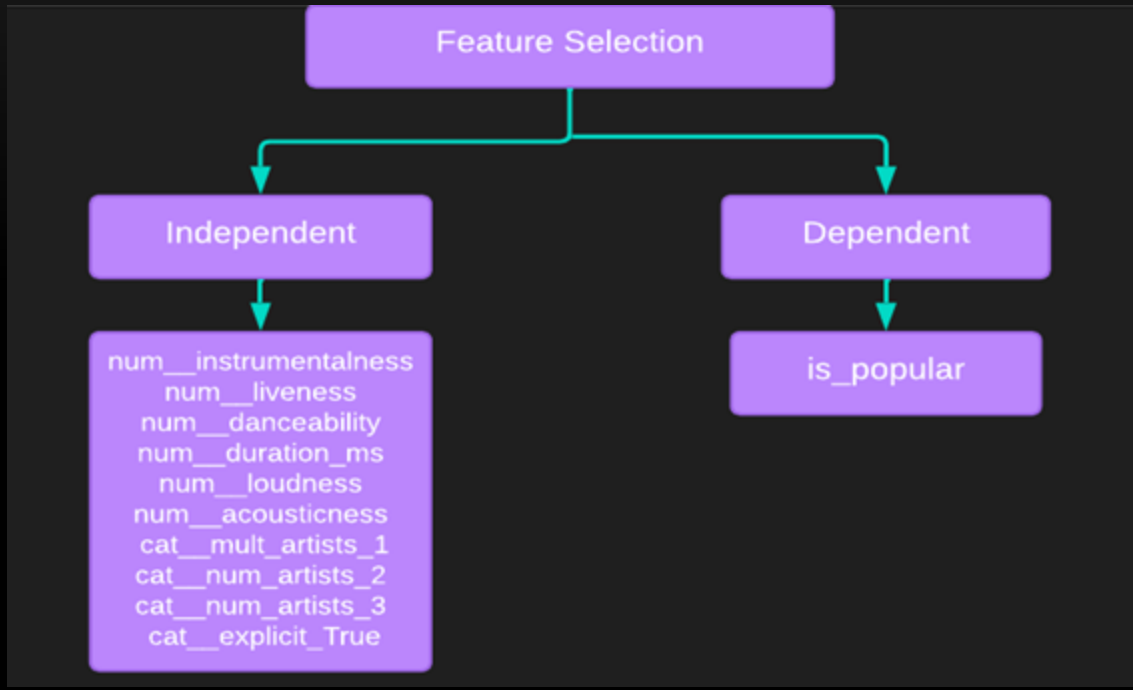

## After Cleaning



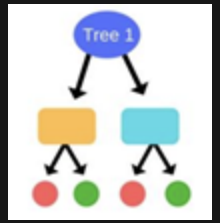
# Pipeline



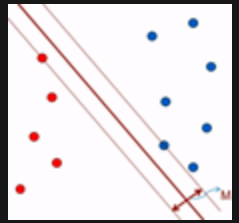
# Feature Selection



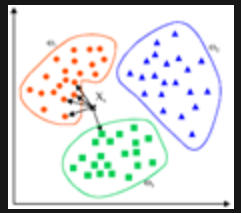
# Models



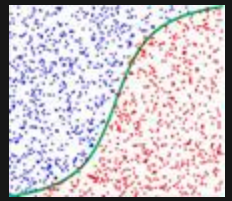

RFC



SVC

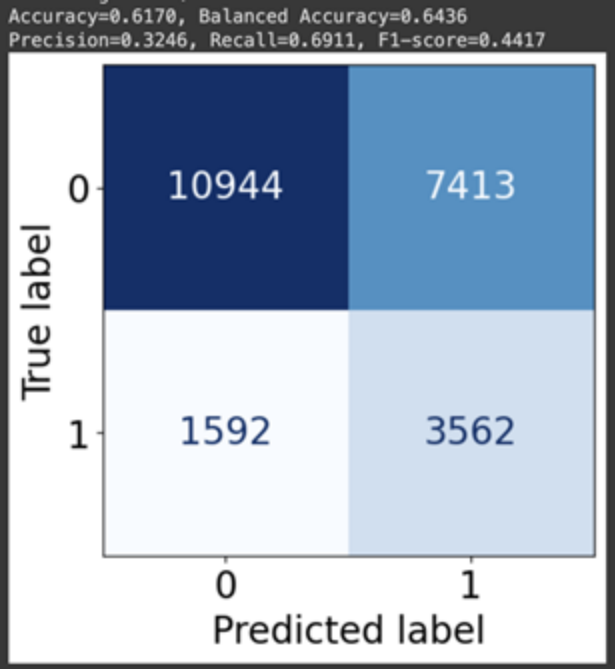


KNN



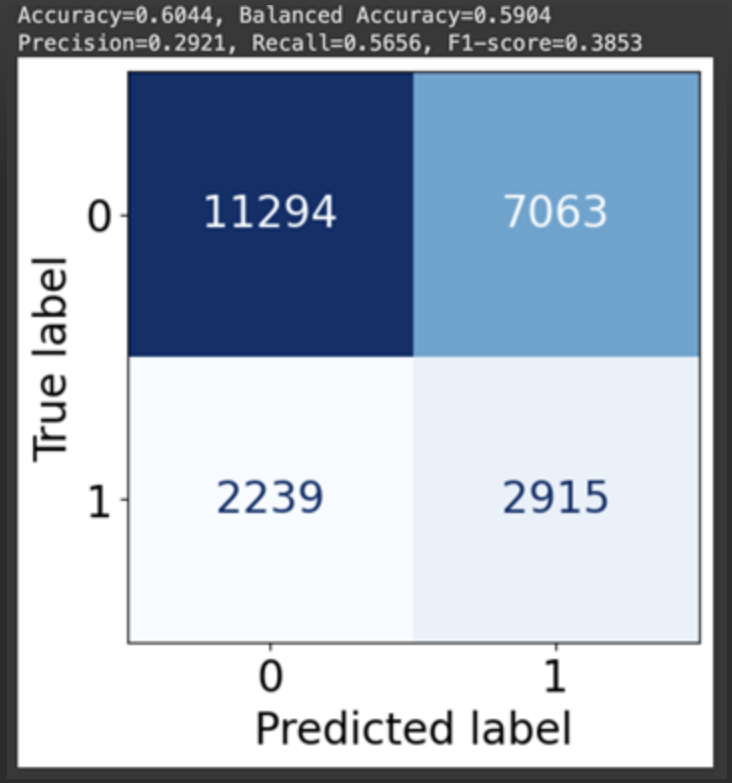
Log Reg

# Random Forest Classifier Model




- Balanced Accuracy after Hyperparameter tuning
- Instances of model correctly identifying songs and misclassification
- Challenges

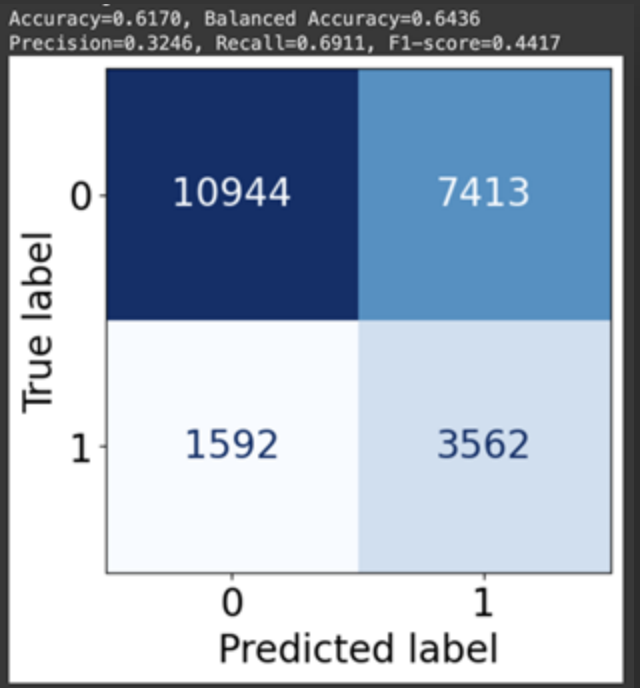
# KNN Model



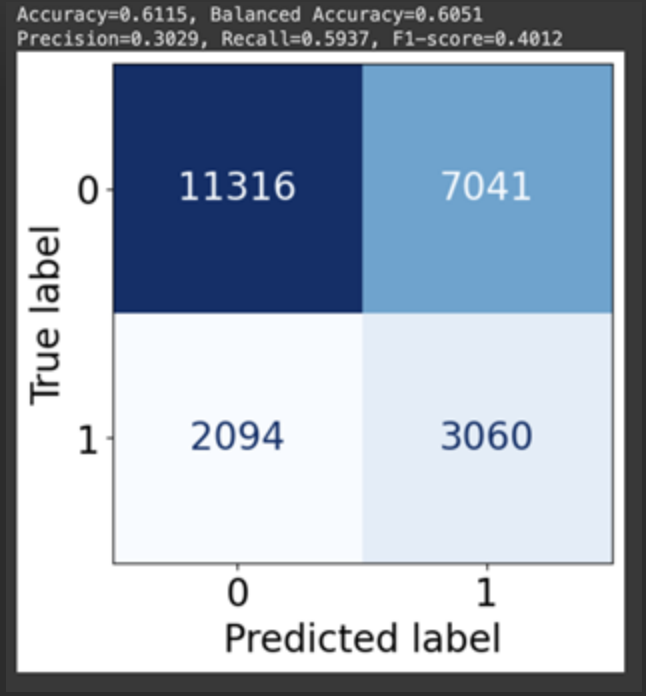

- Principle of Proximity
- Before using Hyper Parameter Tuning
- 3 types of Hyper parameters used:
  I. Grid Search
  Ii. Halving Grid Search
  Iii. Randomized Search

# Hyper Parameter Tuning

## RFC - Halving Grid Search




## KNN - Halving Grid Search




# Model Scores

| # | Models | Accuracy | Balanced Accuracy | Hyper Tuned Balanced Accuracy |
|---|---|---|---|---|
| 1 | Random Forest Classifier | 77.32% | 58.56% | 64.36% |
| 2 | K-Nearest Neighbour | 60.44% | 59.04% | 60.51% |
| 3 | Logistic Regression | 57.22% | 60% | 60% |
| 4 | Support Vector Classifier | 55.89% | 63.45% | 63.45% |

# Limitations

01 Weak correlation of variables.

02 The absence of genre data (Herremans, 2014) limits insights.

03 Exclusion of external factors.

# Thank You

Link to google colab
https://colab.research.google.com/drive/17shSZtfZwrFMzsSofDNQWH9G_cW_xOcv?usp=sharing


Team 05

Harmonizing Hit Predictions

0:23 -3:25