*Classification Project Write-up.*

# Song Popularity Prediction


## Abstract

The goal of this project is to build a predictive classification model and identify factors/music features that make a song popular on a streaming platform such as Spotify. I worked with a dataset of music tracks collected from the South by Southwest official Spotify playlist. I performed my analysis using a number of machine learning models such as KNN, Logistic Regression, Random Forest, XGBOOST, and Naïve Bayes.


## Design

The dataset used in this analysis was obtained using Spotify API from Spotify platform. I focused on tracks featured on South by Southwest (SXSW) 2022 official playlist. By using the dataset I was able to analyze tracks of artists who were at a relatively similar career stage when they were selected to perform at SXSW. I used track popularity index to create a target variable for the analysis.

A classification model producing meaningful results could be used by artists and labels to select tracks for promotion of albums. Knowing which track has a higher potential to gain popularity would help them increase marketing efficiency.

## Data

The original dataset contains 1097 music tracks with 14 features. Some of the feature highlights are danceability, energy, tempo, release date, key, mode, and track popularity.

## Algorithms and Tools

Feature Engineering

- Converted categorical features such as key and mode to dummy variables.
- Dropped music tracks released before 2019.
- Used ‘Track Popularity’ to classify tracks into ‘popular’ and ‘unpopular’ classes and created a target variable.

Models and Tools

- KNN, Logistic Regression, Random Forest, XGBoost, Naïve Bayes.
- Grid search to optimize KNN, Logistic Regression, and Random Forest.
- Oversampling and SMOTE to address the issue of class imbalance.
- Shapley values to analyze feature importance
- Focused on ROC AUC and F1 scores to measure model performance.

## Communication

- 5 minute presentation to client.