# PREDICTING BILLBOARD HOT 100 HITS USING SPOTIFY DATA

## INTRODUCTION

The Billboard Hot 100 is the music industry standard record chart in the United States for songs, published weekly by Billboard magazine. The Billboard HOT 100 is still one of the most reliable ways to gauge a song's popularity. This project will be a walkthrough of simple machine learning techniques applied to predict the songs that will become Billboard Hot 100 hits.

## DATASET DESCRIPTION

Our dataset contains data from the following sources:
http:/millionsongdataset.com/
https://www.billboard.com/charts/hot-100

Firstly, ten audio features will be extracted from the Spotify API. Spotify assigns each song a value between 0 and 1 for these features except loudness which is measured in decibels.

AUDIO FEATURES

| | |
|---|---|
| **Danceability** | **Liveness** |
| **Instrumentalness** | Speechiness |
| **Acousticness** | Loudness |
| **Valence** | Tempo |
| **Energy** | Artist Score |

We will also create an artistic score metric, assigning a score of 1 if the artist previously had a Billboard Hit, and 0 otherwise.

## Algorithm Used

Logistic Regression used to predict a data value based on prior observation of a data set.

Supervised Learning: Splitting the data into 75/25 train test ratio.

Decision Tree to create a training model that can be used to predict the class or value of the target variable by learning simple decision rules inferred from the prior data(training data)

GDA for data classification

SVM or Support Vector Machine is a linear model for classification and regression problems.

Neural Networks

## Approach

1. Collect Raw data
2. Exploratory Data Analysis : It is an approach of analysing data sets to summarise their main characteristics, often using statistical graphics and other visualisation methods. We will be performing initial investigation on data to discover patterns, spot anomalies I.e, data cleaning and test hypothesis with the help of summary statistics and graphical representations.
3. Data Modelling: Split data in 75/25 proportion to test it.

OVERVIEW:

In this project, I attempted to forecast whether or not a song will chart on the Billboard Hot 100. I utilised Billboard and MSD Features as my datasets for this project. To begin, Exploratory Data Analysis is applied to the dataset in order to clean it up and convert all of the values to the desired datatype, as well as to gain insights from it. The dataset must be prepared before this. Exploratory Data Findings was performed on the numerical datasets to avoid any conflicts or inconsistencies from compromising my analysis (in csv format). I imported the data and all of the relevant libraries into the EDA. I studied the data, calculated the sum of all missing values. Then, I moved on to changing data into their desired data types because my data collection had no missing values and finally plotted insights from EDA.

I used two algorithms to predict the accuracy:
1) Random Forest
2) SVM

Result :

After applying the algorithms, I got an accuracy of approx. 68% for SVM Model and 72% for Random Forest Model. Hence, Random Forest Model is more precise.