**HitPredict: Predicting Billboard Hits Using Spotify Data**

**Data Wrangling:**

First we will explore the datasets to find out the description of the columns as loading datasets using Pandas library like type of the column etc. Then we will check for missing values in each column and fill those missing values with the statistical measures like mean, median etc for numerical features. For categorical columns we will impute the missing values with Mode etc. If a column is having more percentage of missing values we will drop those columns.

**Exploratory Data Analysis:**

Next we will plot some visualizations for the columns using libraries like seaborn, matplotlib so that we can extract some insights from the data.

**Feature Engineering:**

Then with the features we can extract some additional features which will contribute more towards our target variable. For categorical columns we will apply some encoding techniques like one-hot encoding, label encoding to convert them to numerical values as some of the algorithms cannot accept categorical data. We need to apply clustering techniques like K-Means Clustering, PCA to cluster the data which belongs to the same kind.

**Modelling:**

Finally for modelling we will split the data as training data, test data in the proportion of 70:30. With training data we will train the model and testing data will define how the model is performing on the unseen data. We will apply different classification ML algorithms on training data. It is not a good approach to test a model's performance on testing data. To overcome this problem we will split the training data into training set and holdout set so that we can test model’s performance on this holdout set using Cross-Validation techniques. Every model contains some parameters. To achieve best metrics for the model we need to do HyperParameterTuning. With the best parameters we will train the model. The model with the best metrics will be our final model.