Project Title: Most Streamed Spotify Songs 2023

1. Introduction:

Data on the most-streamed songs provides us about the most popular songs or trending songs among listeners by looking at user preferences, and it also assists artists in writing songs that would appeal to a wider audience. This project tries to use machine learning techniques to predict the best song based on track name, artist name, release date, Spotify playlists and charts, streaming statistics, Apple Music presence, Deezer presence, Shazam charts, and numerous audio attributes. by creating an accurate forecasting model. We may utilize the accurate prediction model for music analysis, platform comparison, artist impact, temporal trends, and cross-platform presence.

2. Problem statement:

Principal Goal:

The major objective of this research is to create a machine learning model that can correctly forecast a song's popularity and success.

Reliability of Prediction:

The model's forecasts' accuracy will serve as a yardstick for the project's degree of success. It would be clear that the model is successful in predicting song popularity if it was very accurate.

Important Feature:

Finding the particular characteristics or qualities of a song that have the most bearing on predicting its popularity is another crucial component of this research. This can offer insightful information for comprehending the elements that lead to a song's popularity.

3. Approach:

Data collection and Preprocessing:

We utilized Kaggle's Most Streamed Spotify Songs 2023, which includes elements like streaming data, Apple Music and Deezer presence, Shazam charts, and numerous audio features.

Exploratory Data Analysis:

Data gathering and preprocessing: we utilized the Kaggle Most Streamed Spotify Songs 2023 tool, which includes features like streaming statistics, Apple Music and Deezer presence, Shazam charts, and numerous audio features.

4. Feature Selection:

1. Adding features like the total number of songs an artist has produced and their average position in the platform chart.
2. Numerical features are scaled and normalized.
3. One hot encoding the categorical variables

5. Model Selection:

We are using machine learning models like

Linear Regression:

1. For the purpose of forecasting a numerical goal variable, such as song popularity, linear regression is a straightforward yet powerful model.
2. You can model the association between your features (such as the total number of songs by an artist or their average position on the chart) and the target variable (song popularity) using linear regression.
3. You may better understand the influence of each feature on song popularity by using the coefficients that linear regression provides for each feature.

Random Forest:

1. A potent ensemble learning technique called random forests combines several decision trees to produce predictions.
2. Because they can handle both numerical and categorical information, random forests are appropriate for your dataset if it has both.
3. They are frequently more reliable and have the capacity to identify intricate links in the data.
4. The random forest model's feature importance scores can be used to determine which features have the greatest bearing on song popularity predictions.

6. Model Training and Hyper parameter tuning:

Splitting the Dataset:

The dataset should first be split into two sections:

1. training dataset
2. Testing dataset.

The chosen models will be trained using the training data, and the testing data will only be utilized to assess how well they perform.

Training the Selected Models:

Then, use the training dataset to run the selected models. The models chosen in this instance are random forest and linear regression.

Fine-Tuning Model Parameters:

To enhance performance, it is essential to change the models' hyperparameters. Random Search and Grid Search Cross-Validation (CV) are two techniques that can be utilized to achieve this. These strategies carefully explore alternative hyperparameter combinations to enhance the model's performance.

7. Model Evaluation:

Root Mean Squared Error (RMSE):

The average size of the differences between anticipated and actual values is measured by the RMSE. Larger errors carry heavier penalties. The performance of the model improves with decreasing RMSE.

Root Absolute Error (RAE):

RAE is comparable to RMSE but employs absolute error values. It offers an alternative viewpoint on the size of errors. Lower RAE denotes higher model performance, similar to RMSE.

R - squared:

The amount of the target variable's variance (song popularity) that can be predicted from the independent variables (features) is measured by R2, or the proportion. It reveals how well the model fits the data by showing how well it matches. A better match is indicated by a greater R2.

8. Deployment and Recommendations:

"By interpreting the model's coefficients, we may determine which characteristics have the most effects on song popularity. This awareness of the essential elements that lead to a song's popularity is quite helpful.

The algorithm can be used to forecast the success of new songs after it performs satisfactorily. With this capacity, we may predict the likely success of upcoming releases based on their qualities and traits.

9. Conclusion:

This project focuses on determining whether a song will become popular.

Artist Insight and Decision-Making:

Artists can learn a lot about the future success of their songs by building a precise forecasting model. They are better equipped to make decisions concerning their artistic endeavors thanks to this information.

Funding and Resource Allocation:

Using this technique, musicians may determine whether a song has a good chance of becoming successful. Their odds of success can be increased by using this information to successfully secure financing and allocate resources.

Radio Channel Programming:

This technique can be used by radio stations to assess the prospective popularity of songs before selecting whether or not to include them in their playlists. This guarantees that the content they are providing for their audience is in line with their preferences and inclinations.