**SPOTIFY SONGS HIT PREDICTION**

**USING MACHINE LEARNING ALGORITHMS**

# **A PROJECT REPORT**

### **Submitted by**

**KAUSHIK B (21ADR020)**

**SUJIT R B (21ADR053)**

**PRASSHANTHINI R (21ADR034)**

**SIVAKARTHIK A T (21ADR047)**

**SANDEEP K (21ADL065)**

**Team Mentor:**

Mrs. Senthilvadivu(AP/AI)

***For***

**20ADT44 APPLIED MACHINE LEARNING**

# **DEPARTMENT OF ARTIFICIAL INTELLIGENCE**



**KONGU ENGINEERING COLLEGE**

# **(Autonomous)**

# **PERUNDURAI ERODE – 6****38 060**

# **MAY 2023**

**PROBLEM STATEMENT:**

* Music is now an international phenomenon which doesn’t have any boundaries.
* It is impossible for producers to predict whether a song that they are releasing is going to be a hit or a flop one.
* If they know the result, they can enhance their standards of their production.
* If they produce flops without properly predicting the outcome of the song that they are producing and releasing, it may end in a huge loss as Music is not just something that involves instruments and voices but a huge market of business which involves more money than any other in the entertainment industry.
* Rather than movies, predictions are tough to make, as movies show some obvious results with the trailers, star value, production house involved. But on the other hand, Music doesn’t hold much other than the star value, thus the success of a song or an album only depends on the music itself.

**LITERATURE SURVEY:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **PAPER NAME** | **OBJECTIVE** | **ALGORITHM USED** | **PERFORMANCE** | **DEMERITS** |
| Predicting Music Popularity Using Machine Learning Algorithm and Music Metrics Available in Spotify | To predict the popularity of the song using the song metrics available in Spotify. | 1.Random Forest classifier  2.K-Nearest neighbor classifier 3. Linear Support Vector classifier. | 1.The maximum accuracy of 89% is achieved by the Random Forest algorithm 2. The lowest performance was observed in the Linear Support Vector Classifier with an accuracy of 64% | The limitation of this model is that the dataset which is used for the prediction of songs is only for English songs available on the Spotify platform. |
| A Study For ML-Based Song Hit Prediction Using Spotify | The main focus of this topic is to predict whether a song will become a hit. Hit songs help labels and music vendors to increase their profits. | 1.Random Forest classifier. 2.K-Nearest neighbor classifier. 3. Support Vector classifier.  4. Logistic Regression. | 1.The maximum accuracy of 86% is achieved by the Random Forest algorithm 2. The lowest performance was observed in the Logistic Regression with an accuracy of 66% | If a user frequently listens to some music, the algorithm may recommend similar songs or that the user may like. |
| Hit predict: Predicting hit songs using Spotify Data | to predict which songs will become Billboard Hot 100 hits | Expectation Maximization (EM), Logistic Regression (LR), Gaussian Discriminant Analysis (GDA), Support Vector Machines (SVM), Decision Trees (DT), and Neural Networks (NN) | The most successful algorithms were Logistic Regression and a Neural Network with one hidden layer. | Objective is narrow as it aims only to predict Billboard hot 100 hits. |
| A Model for  Predicting Music  Popularity on  Spotify. | To build a model for predicting whether or not a song will appear in Spotify’s Top 50 ranking. | 1. Gaussian Naive Bayes 2. K-Nearest neighbor classifier 3. Support Vector Machine 4. Logical Regression. | 1. When tested  with data from June and July 2019, an SVM classifier with  RBF kernel obtained accuracy and AUC above 80%. | It only considers the songs in the Top 50 as popular and the ones  on Viral 50 as unpopular, so it will not be effective for producers to judge. |
| Predicting a Hit Song with Machine Learning | To predict the hit song based on the data source | 1.Logistic Regression  2.Decision Trees  3.Naïve Bayes and  4.Random Forests | The accuracy of Logistic regression is highest 52% whereas that of Decision trees is the lowest (50.5%). | The algorithms are not trained based on standard api's |

**PROPOSED SYSTEM:**

* To predict the song is going to be hit/flop:
* The dataset is subjected to pre-processing methods:
  + Removing categorical values:
  + The categorical values in the dataset are Track name, Artist name and URL which are independent on the success of the song, so they are removed.
  + Standard Scaling:
  + We have data points that are afar from each other i.e. the range is varied and larger. To bring the datapoints closer to one another, Standard Scaling is used.
  + Then the pre-processed dataset is undergone a train-test split where the entire dataset is divided into train, test and validation data.
  + 60% - Training data; 20% - Testing data; 20% - Validation.
* Then the dataset is subjected to three different machine learning algorithms namely:

● LOGISTIC REGRESSION

● SUPPORT VECTOR MACHINE

● RANDOM FOREST

* Then the accuracies of the above algorithms are compared and the best ones is selected

**DATA PREPROCESSING:**

**MODEL DEVELOPMENT**:

**ALGORITHMS USED:**

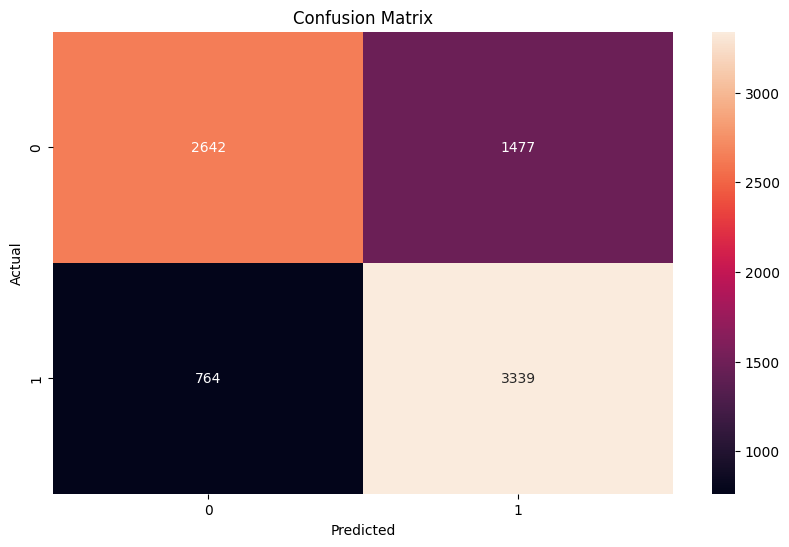
**THE FOLLOWING ALGORITHMS WERE USED:**

* Logistic Regression
* Support Vector Machine
* Random Forest

1. **LOGISTIC REGRESSION:**

* This type of statistical model (also known as a logit model) is commonly used for classification and predictive analysis.
* Logistic regression estimates the probability of an event, such as elected or not elected, occurring based on a given data set of independent variables.
* Since the outcome is a probability, the dependent variable is limited to a value between 0 and 1.
* Logistic regression applies a logit transformation to the odds, i.e., the probability of success divided by the probability of failure.

RESULTS OF LOGISTIC REGRESSION:



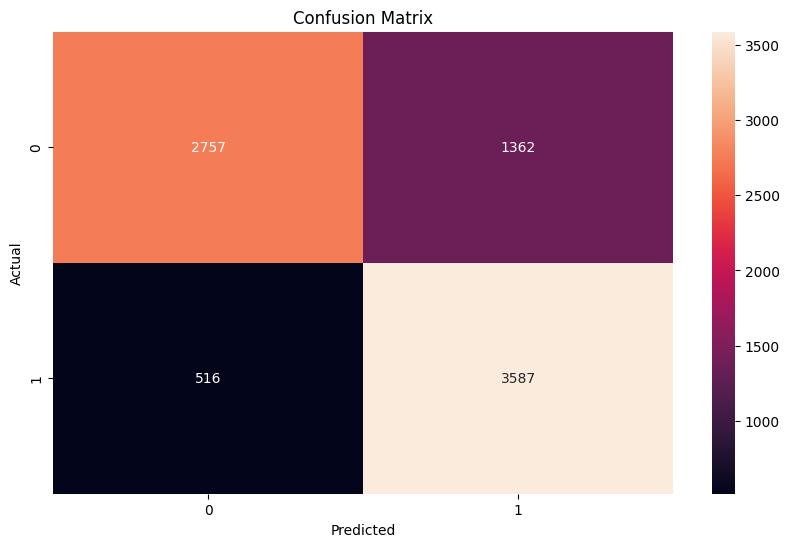
TRAINING ACCURACY: 72.927%

VALIDATION ACCURACY: 72.584%

1. **SUPPORT VECTOR MACHINE:**

* Support Vector Machine or SVM is one of the most popular supervised learning algorithms used for both classification and regression problems.
* However, it is primarily used for classification problems in machine learning.
* The goal of the SVM algorithm is to create the best line or decision boundary that divides the n-dimensional space into classes so that we can easily assign the new data point to the correct category in the future.
* This best decision boundary is called hyperplane.
* SVM selects the extreme points/vectors that help to create the hyperplane. These extreme cases are called support vectors and hence the algorithm is also called Support Vector Machine.

RESULTS FOR SVM:



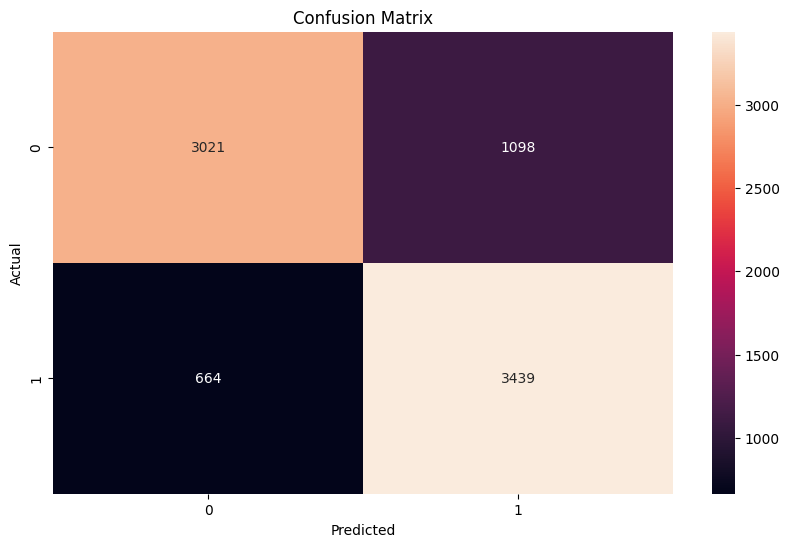
TRAINING ACCURACY: 78.664%

VVALIDATION ACCURACY: 77.158%

1. **RANDOM FOREST:**

* Random Forest is a powerful machine learning algorithm that falls under the category of supervised learning techniques.
* It is widely used for solving both classification and regression problems in the field of machine learning.
* The algorithm is based on the concept of ensemble learning, which involves combining multiple classifiers to solve complex problems and improve model performance.
* As the name implies, Random Forest is a classifier that consists of multiple decision trees, each trained on a subset of the given dataset. The algorithm then takes the average of the predictions from each decision tree to improve the accuracy of the overall model.
* Unlike relying on a single decision tree, Random Forest utilizes the majority vote of predictions to make the final output prediction.

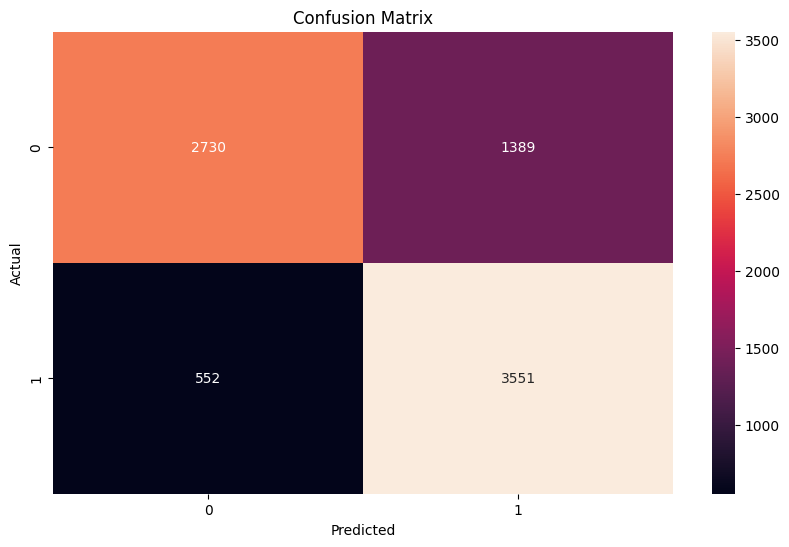
RESULTS:



TRAINING ACCURACY: 99.95%

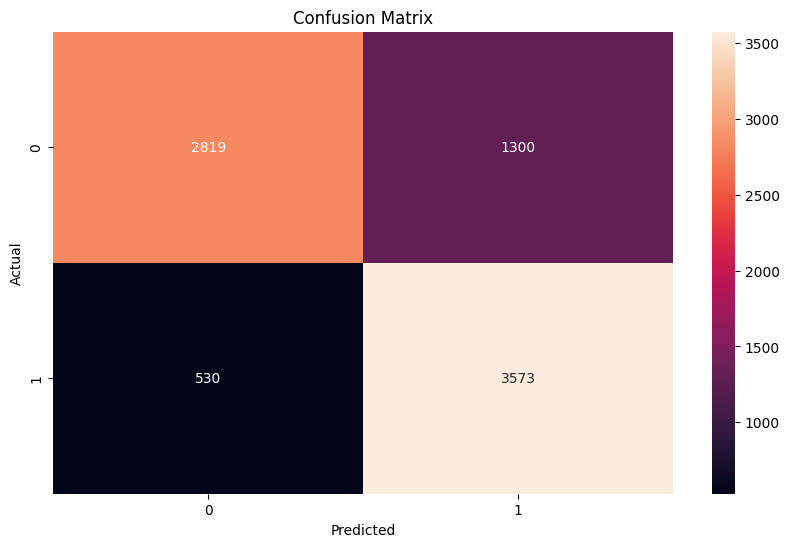
VALIDATION ACCURACY: 78.59%

* Very High training accuracy implies that the model is overfitting of the data.
* In order to correct it we train the random forest with hyperparameters tuning and setting the number of estimators to 500 and depth to 5.



TRAINING ACCURACY: 77.03%.

VALIDATION ACCURACY: 75.82%.



TRAINING ACCURACY: 82.03%

VALIDATION ACCURACY: 77.48%

Feature Importance Ranking:

Feature 8: 0.2949

Feature 7: 0.1469

Feature 1: 0.1315

Feature 2: 0.0845

Feature 4: 0.0647

Feature 6: 0.0633

Feature 12: 0.0587

Feature 10: 0.0528

Feature 15: 0.0284

Feature 9: 0.0194

Feature 11: 0.0180

Feature 14: 0.0151

Feature 5: 0.0088

Feature 3: 0.0078

Feature 13: 0.0052

* The random forest model with hyperparameter tuning and feature analysis has a training accuracy of 0.8334 and a validation accuracy of 0.7880, which is an improvement over the previous model. The feature importance ranking suggests that feature 8 is the most important, followed by features 7 and 1. Features 4, 2, and 12 also have relatively high importance scores. Features 5, 3, and 13 have the lowest importance scores.

**RESULTS:**

**Model created with Random Forest with hyperparameter tuning and estimators set to 100 and depth set to 5 with feature analysis gives out more training accuracy and validation accuracy.**