Predicting Music Popularity Using Machine Learning Algorithm and Music Metrics Available in Spotify

Journal of Development Economics and Management Research Studies (JDMS) *A Peer Reviewed Open Access International Journal*
ISSN: 2582 5119 (Online)
Crossref Prefix No: 10.53422
09(11), 10-19, January-March, 2022

Predicting Music Popularity Using Machine Learning Algorithm and Music Metrics Available in Spotify

Dr. Prashant Pareek[1], Dr. Poorna Shankar[2], Mr. Pushpak Pathak[3], and Ms. Nidhi Sakariya[4]

Abstract

The exponential growth of online music streaming has given birth to many new platforms among which, the widely used platform is Spotify. The most popular music streaming app's data can be used to predict the capability of a song to be popular before its release with the help of attributes like loudness, energy, acousticness, etc. which is defined when thesong is being made. This study helps to predict the popularity of the song using the song metrics available in Spotify by applying Random Forest classifier, K-Nearest neighbour classifier and Linear Support Vector classifier to compare which of these models can effectively predict the popularity. The results found that Random Forest works the best for predicting popularity with high accuracy, precision, recall and F1-score.

Keywords: Music streaming, Spotify, Exploratory data analysis, K-Nearest Neighbor (KNN), Random Forest, Linear Support Vector Classifier (LSVC)

1. Introduction

In the last few years, multimedia has gained a large amount of user attraction because of the growth of technology and tinternet. Hence, multimedia plays a huge role in enlarging user experience by recommendations, advertisements, searches and retrievals and so, predicting that the content will be able to gain popularity and will be able to reach to the maximum users becomes utmost importance so that the content makers can work on something which can help them to grow instead of falling apart. There are many research efforts to understand and predict content popularity, mostly for videos [1,2].

Music is one of the stress releasing, relaxing, entertaining and career-building platforms in today's world. Music streaming platform is an area that is growing tremendously, such as Spotify, Lastfm, Gaana, Resso, etc. Therefore, it is an efficient and powerful approach to predict the popularity of a song on such platforms, which can help the musicians and listeners to increase

[1]Assistant Professor, Shanti Business School, Ahmedabad, Gujarat.
[2] Professor, Keystone Global, Ahmedabad, Gujarat.
[3] Pursuing MSC in Big Data, FOM University, Germany.
[4] Pursuing MSC in Big Data, FOM University, Germany.

profits by promoting their products and quality of experiences respectively. Today, Spotify is the world's most popular audio streaming subscription service with 356 million users, including 158 million subscribers, across 178 markets. It consists of 70 million tracks along with 2.6 million podcasts created by over 1.2 million artists. Almost 40,000 tracks are loaded daily on Spotify. This study aims to make use of a huge dataset available on Kaggle's website to build a model which can predict the popularity of English songs. The dataset consists of a variety of song features like energy, loudness, dance ability, acousticness with which it could be possible.

After collecting the data, we tried to find outliers, performed some exploratory data analysis to understand the data more in-depth using data visualization through the seaborn library and model the prediction. Classification refers to classifying the data based on a particular constraint. In this study, we classify the data based on the label of popularity, that whether a song will be popular. For that purpose, we used 'Random Forest classifier', 'K- Nearest neighbor' and 'Linear support vector classification algorithm. In this study, the focus will be only on using the song matrix data available, while there are many different factors which may affect the popularity of a song. For example, the social information, number of Twitter posts with hashtags indicating that users are listening to a particular song [3].

2. State of The Art

In this section, we have provided a comprehensive overview of different types of prediction analytics and algorithms. Predictive analytics has two types of models:

1. Classification: That can predict class membership
2. Regression: That can predict a number

Some widely used predictive models are decision trees, Neural networks, support vector classifiers, linear regression, etc. In this study we have used mainly three classifiers, that is, Random Forest, K-nearest neighbor and Linear Support vector classifier because we need to classify a song whether it falls into a category of a popular song or a flop song. In paper [4], authors have described that they used C4.5, CART and Random Forest to predict the popularity of online news before release and they describe that Random Forest gives the highest accuracy. Similarly, in the paper [5] authors try to predict the future usage of tags in Stock Exchange websites, where Random Forest gives the best result. In paper [6], authors try to predict the song popularity using acoustic features including MFCC features, but their predictions had room for improvement. An end-to-end deep learning architecture named Hit Music Net is presented in the paper [7], which gives better results than other machine learning algorithms. Though it cannot be very helpful for particular music streaming app like Spotify. Paper [8] uses Logistic Regression, Decision Tree, Naïve Bayes and Random Forest to predict the song popularity but they couldn't get good results due to lack of time and data. This paper [9] presents (i) a novel regression approach towards hit song prediction using neural networks which combines wide (high-level) and deep (low-level) acoustic features; (ii) it tells that mood and vocals (the features identified as being crucial when it comes to liking and disliking a song) are also of high relevance for the hit prediction task.

3. Data Description

The dataset has 2, 32,725 records and 18 attributes. The data has all English songs of 26 genres, it has been collected from Kaggle and was extracted from Spotify API. This data needs to be considered for data preprocessing and further prediction operations. However, 14 attributes are such which plays a vital role in the prediction model and 4 attributes namely genre, artist name,

track name and track id do not affect any changes whether they are included or not so these attributes are neglected while modelling. Those 14 attributes which are important and which will define the model predictions are as mentioned below with descriptions:

- POPULARITY: The popularity of a song must range on the scale of 1-100 in numbers or percentages.
- ACOUSTICNESS: A confidence range should be from 0.0 to 1.0 on the scale of whether the track is acoustic or not.
- DANCEABILITY: Dance ability defines how suitable a track is, its scale varies from 0.0 to 1.0, possibly for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity.
- DURATION (MS): Duration of the track must be very less i.e., in milliseconds.
- ENERGY: The energy range must within 0.0 to 1.0 and represents a perceptual measure of intensity and activity.
- Typically, energetic tracks like hip-hop songs or beatboxing music feel fast, loud, and noisy. 6. INSTRUMNETALNESS: In the song's lyrics sounds like "Ooh" and "Aah" sounds are treated as instrumental in thiscontext. Their Value must range from 0.0 to 1.0 on the scale.
- KEY: The songs represent the key of which type key it is. The estimated overall key of the track is thentransformed to the format of a number. (E.g., C, C#, D, D#, G……transformed as 1, 2, 3, 4, 5…, etc.)
- LIVELINESS: This factor detects the presence of an audience in the song recording. The higher liveness values represent an increased probability that the track was performed live.
- LOUDNESS (dB): The overall loudness of a track must be in average decibels. The songs loudness must not be veryhigh in max decibels and not very low in decibels however, can be well hearable to people to understand lyrics.
- MODE: Mode indicates the modality (major or minor) characteristic of a track depending on its key value, this isthe type of scale from which its melodic content is derived or can say how melodic the song is. 11. SPECHINESS: This is one of the important characteristics to be considered. The more exclusively speeches-like the recording (e.g., talkshow, audiobook, poetry, etc.), its attribute value must also be within 0.0 to 1.0, closer to 1 is always a good result.
- TEMPO: As we know, the overall estimated tempo of any track is always in beats per minute (BPM). In musical
- Terminology, the tempo is known as the speed or pace of a given piece and which derives directly from the average beatduration of the song.
- TIME SIGNATURE: An estimated overall time signature of a track.
- VALENCE (float): Any track positiveness is described from its valence measures (i.e., high or low) which alsoranges from 0 to 1 as like other attributes.

4. Proposed Work

Our proposed work helps the music industry in making the predictions that whether the song being created by them will be a hit or a flop. The ability of making a song prediction can also be implemented for customized music suggestions. It can also be helpful in knowing the preferred song for a given population using the Spotify platform. For this study, we collected data from Kaggle, which was extracted from Spotify API. We tried to pre-process the data by transforming the discrete data into continuous. We found some insights from the data by Exploratory Data Analysis with the help ofseaborn and matplotlib libraries available in Python.

Next, we did feature extraction where we divided the data into train and test and created the model with Random Forest, K-Nearest Neighbour and Linear Support Vector Classifier. Comparing the results of all these algorithms to find which one of them works the best for predicting the popularity of music. Fig. 1 represents the flow of this project.

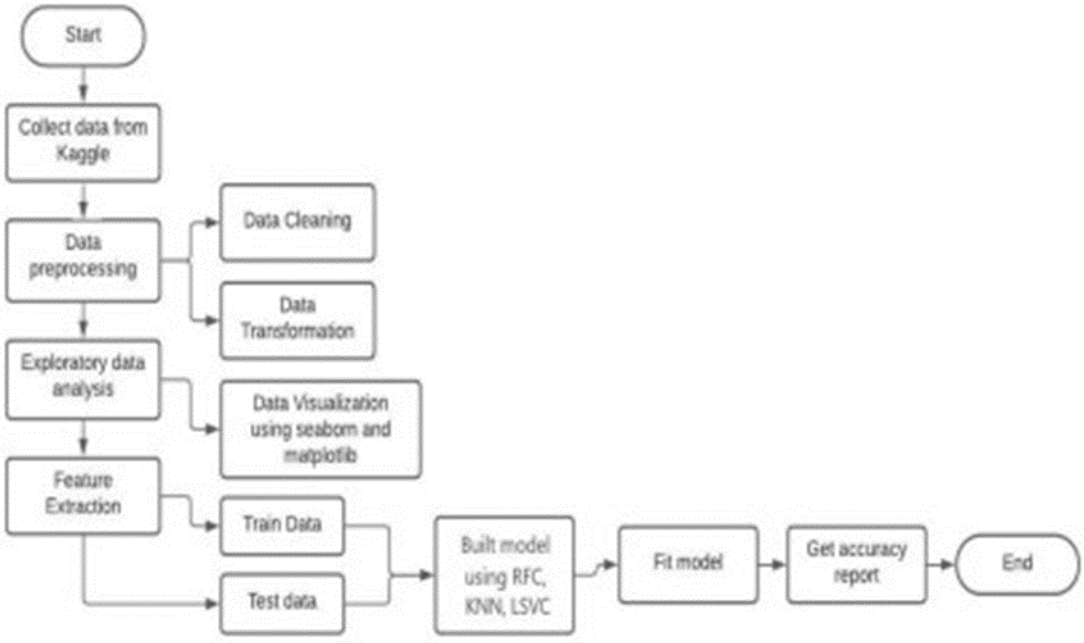

Figure 1 - Implementation flow

4. Implementation

Our implementation includes data preprocessing, exploratory data analysis and classification algorithms. Let's discuss these steps in detail to understand the working of this project

A. *Data pre-processing:* Data pre-processing includes data cleaning and data transformation, the data which we collected from Kaggle has no null values therefore we proceeded to data transformation, our data has three discrete attributes, namely, Time Signature, Key and Mode. These were transformed into continuous as mentioned in table 1,2 and 3.

Table 1 Transformed mode

Original	Transformed
Major	1
Minor	0

Table 2 Transformed time signature

Original	Transformed
0 / 4	0
1 / 4	1
2 / 4	2
3 / 4	3
4 / 4	4
5 / 4	5

Table 3 Transformed key

Original	Transformed
C	1
C#	2
D	3
D#	4
E	5
F	6
F#	7
G	8
G#	9
A	10
A#	11
B	12

B. Exploratory Data Analysis: We did exploratory data analysis to find the correlation of data and to understand how was the data distributed.

Popularity is the target attribute, which is to be predicted. It is shown in Fig. 2 that songs with a popularity of 52 or more are highly popular than others.

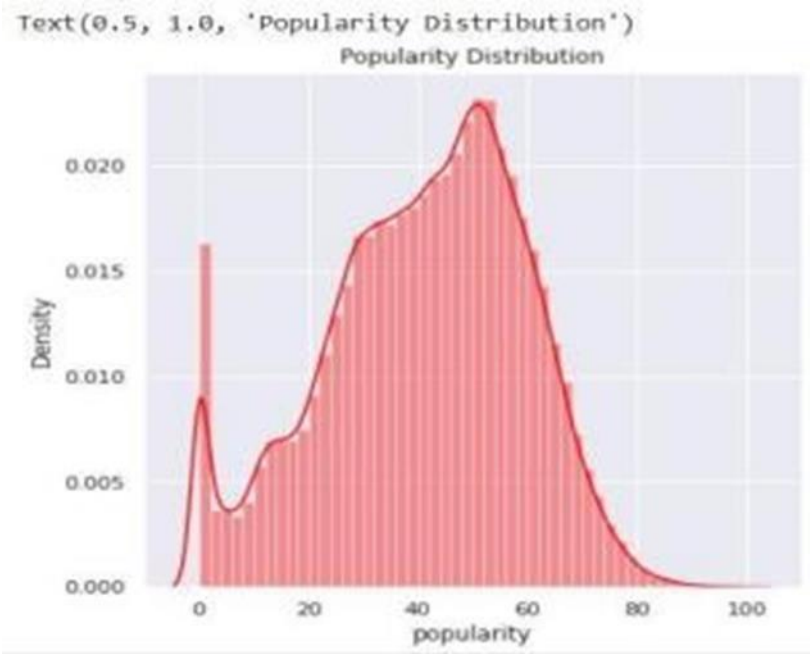

Figure 2 - Popularity Distribution

We found the correlation between the data using the different combinations of attributes

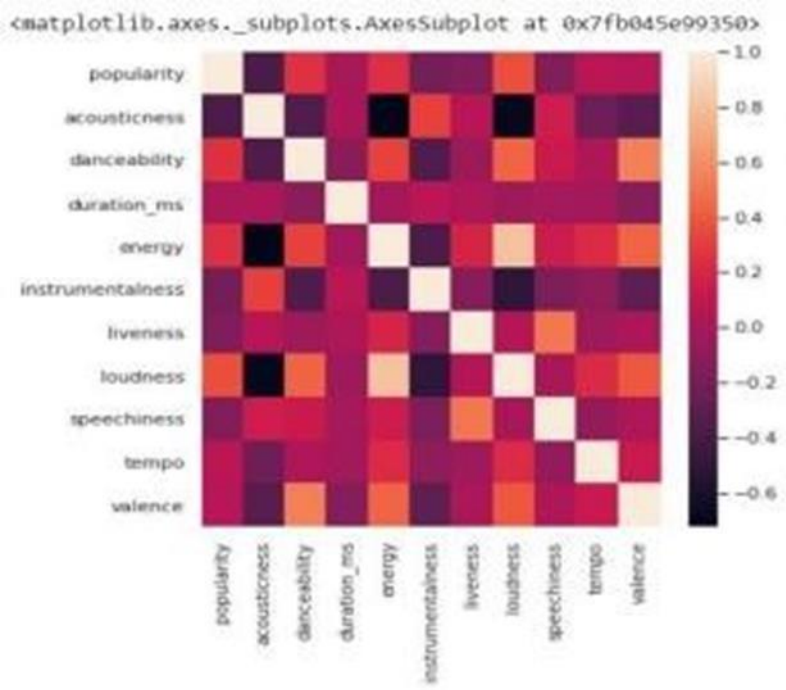

Figure 3 - Correlation of attributes

A. *Classification algorithm:* Classification can be used to group the data into one or more parts. In paper [10] Classification algorithms are used to classify countries based on carbon emission rates. Classification problems were solved using pattern recognition problem, feedforward and cascade forward networks in paper [11].
After understanding data properly, we trained our prediction algorithm one by one and obtained precision,recall, F-score and accuracy.
The random forest classifier showed excellent results with an accuracy of 86%, whereas, the accuracy of K- Nearest Neighbor was 68% and Linear Support Vector Classifier was 64%. Table 4 shows the precision, recall and F-score of each algorithm.

Table 4 Results of algorithms

	Precision	Recall	F- Score
Random Forest Classifier	0.83	0.70	0.76
K-Nearest Neighbor	0.50	0.45	0.47
Linear Support Vector Classifier	0.46	0.43	0.45

5. Result and Model Evaluation

To evaluate our model, we have used a confusion matrix. It shows the performance of an algorithm on a test dataset for which we have true values and compare it to a training dataset to check its accuracy. Figures 4, 5, 6 represents a 2 x 2 structure of confusion matrix for Random Forest, K-Nearest Neighbor and Linear Support Vector Classifier respectively.

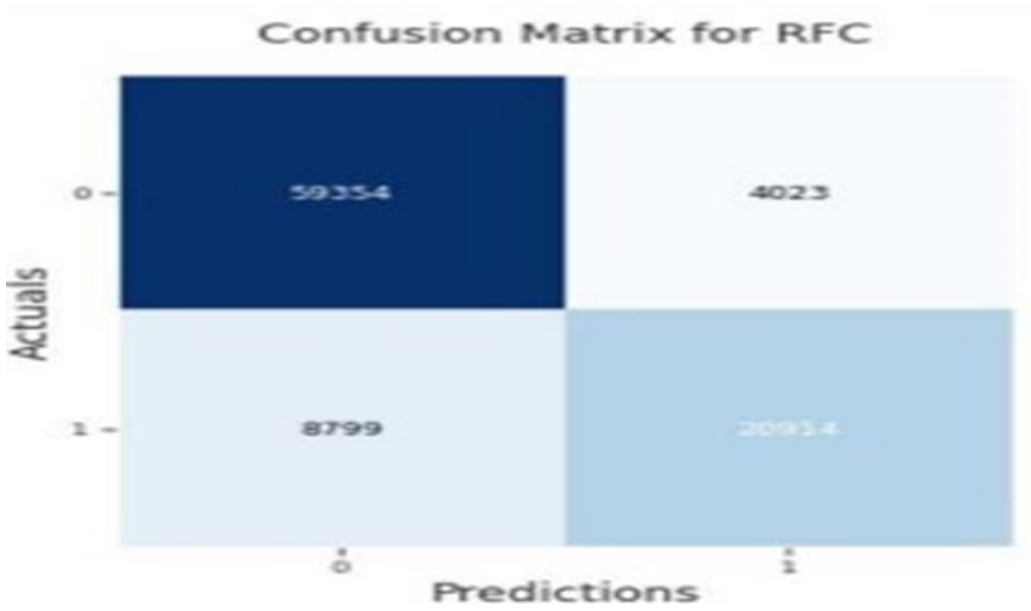

Figure 4 - Confusion matrix for Random Forest

The confusion matrix of Random Forest shows 59354 true positives, 4023 false negatives, 8799 true negatives and20914 false positives.

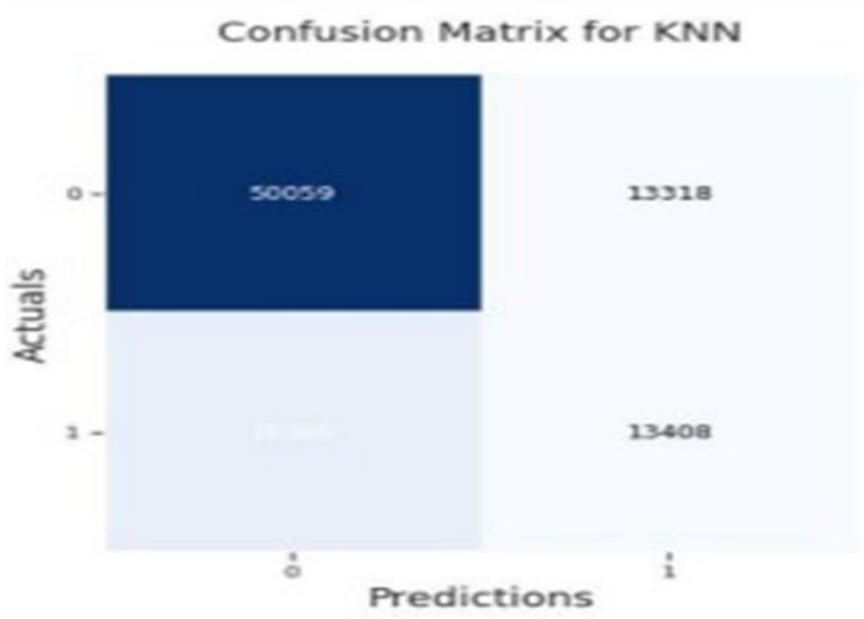

Figure 5 - Confusion matrix for K-Nearest Neighbor

The confusion matrix of K-Nearest Neighbor shows 50059 true positives, 13318 false negatives, 16305 truenegatives and 13408 false positives.

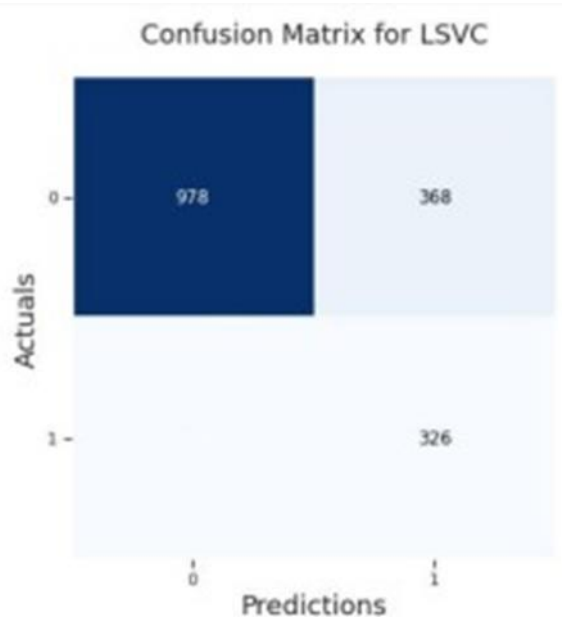

Figure 6 - Confusion matrix for Linear Support Vector

The confusion matrix of Linear Support Vector shows 978 true positives, 368 false negatives, 126 true negatives and 326 false positives.

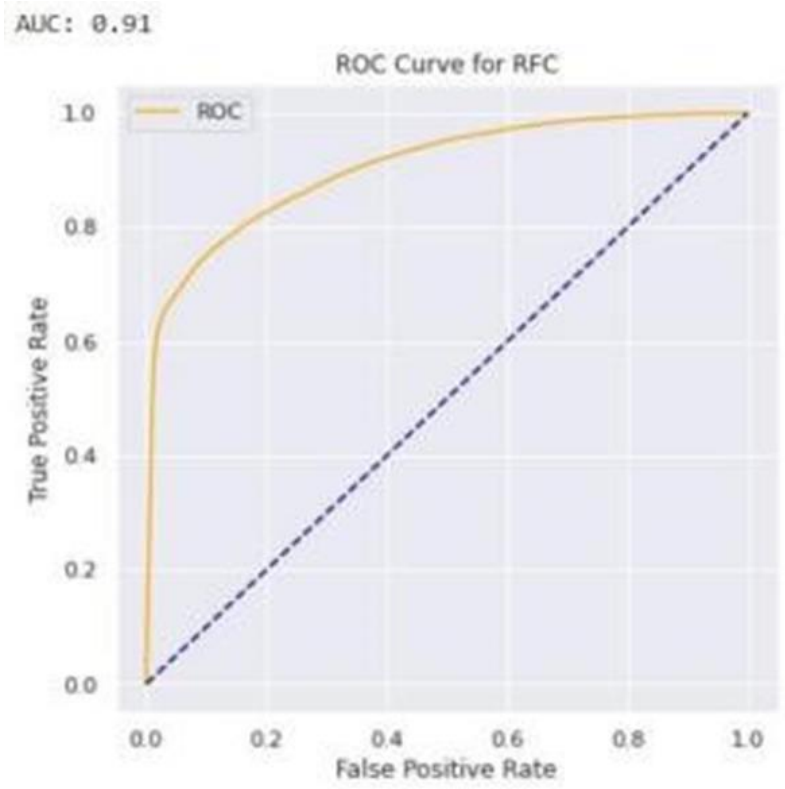

Figure 7 - ROC Curve for Random Forest

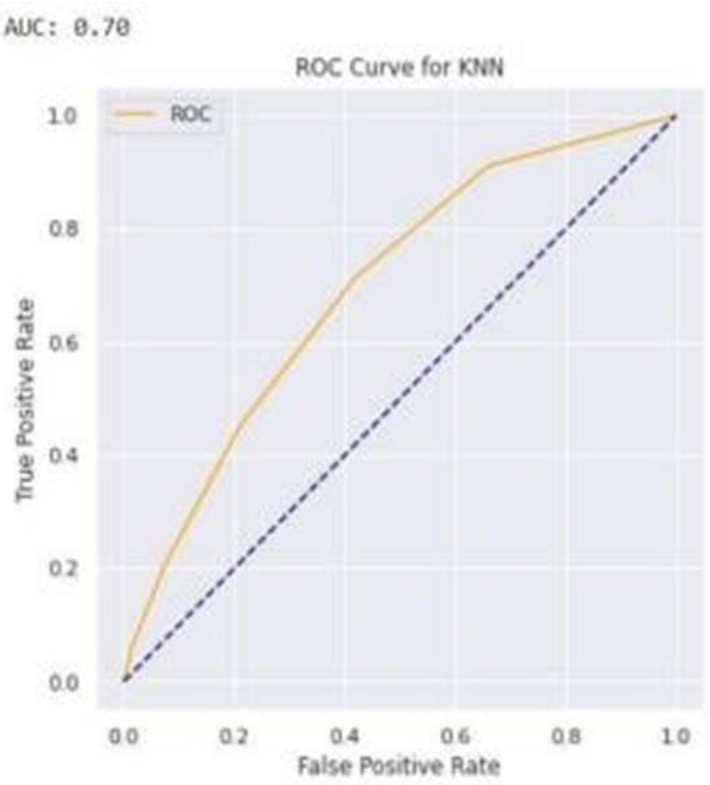

Figure 8 - ROC Curve for KNN

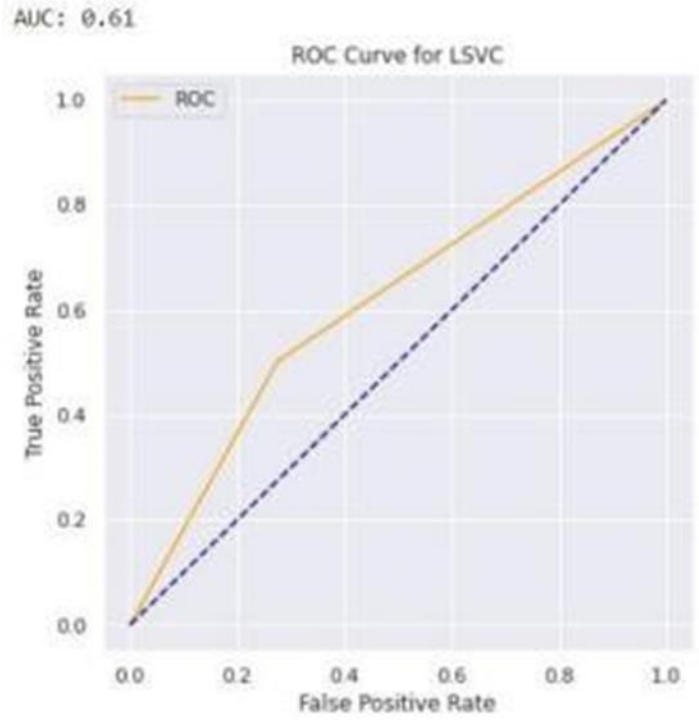

Figure 9 - ROC Curve for LSVC

The maximum accuracy of 89% is achieved by the Random Forest algorithm, it has also given higher precision,recall and F1 score than the other two algorithms. Even the AUC score is 0.91 which is higher. The lowest performance was observed in the Linear Support Vector Classifier which has an accuracy of 64% and relatively low precision, recall, F1 score and AUC

score. It can also be observed in the confusion matrix and ROC Curve given above.

6. Conclusion and Future Work

This paper presents a methodology to predict whether a song will be popular or not using data collected frommusic metrics. The data was collected from Kaggle and trained and compared with the help of three classification algorithms namely, Random Forest Classifier, K-Nearest Neighbour and Linear Support VectorClassifier, that can make predictions of Song before its release whether it will be popular. Among thesethree models, we have found that Random Forest Classifier gives the best results and accuracy which was up to 89%. Hence, we conclude that the Random Forest model is good for future songs popularity predictions. This research includes an array of parameters like loudness, acousticness, energy, key, etc. that could help predict a song being popular. This concludes the popularity predictions of the song using python and song metrics data available from Spotify API through Kaggle.

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. However, for future scope, we can also use this model for different languagesongs and also for different app platforms with help of various datasets.

References

1. Trzcinski T, Rokita P (2017). Predicting Popularity of Online Videos Using Support Vector Regression. IEEETransactions on Multimedia. 2017;19(11):2561-2570.
2. Wu J, Zhou Y, Chiu D, Zhu Z (2016). Modeling Dynamics of Online Video Popularity. IEEE Transactions onMultimedia. 2016;18(9):1882-1895.
3. Ma Z, Sun A, Cong G (2013). On predicting the popularity of newly emerging hashtags in Twitter. Journal of the American Society for Information Science and Technology. 2013;64(7):1399-1410.
4. Zhang Y, Lin K (2021). Predicting and Evaluating the Online News Popularity based on Random Forest. Journal ofPhysics: Conference Series. 2021;1994(1):012040.
5. Chenbo F, Yongli Z, Shidi L, Qi X, Zhongyuan R (2017). Predicting the popularity of tags in Stack Exchange QAcommunities. 2017; 90-95.
6. Lee J, Lee J (2018). Music Popularity: Metrics, Characteristics, and Audio-Based Prediction. IEEE Transactions onMultimedia. 2018;20(11):3173-3182.
7. Martin-Gutierrez D, Hernandez Penaloza G, Belmonte-Hernandez A, Alvarez Garcia F (2020). A Multimodal End-to-End Deep Learning Architecture for Music Popularity Prediction. IEEE Access. 2020; 8:39361- 39374.
8. Agha R, Krsihnadas N (2020). Predicting a Hit Song with Machine Learning: Is there an apriori secret formula? International Conference of Data Science, Artificial Intelligence and Business Analytics. 2020; 111-116
9. EvaZ, Ramona H, Michael V (2019). Hit song prediction: Leveraging Low- and High-Level Audio Features. 2019: 4-8
10. Poorna S, Vrushsen P (2020). Classification of Global Carbon Emission using Artificial Intelligence. International Journal of Computer Application. 29(3):31-38.
11. Poorna S (2013). Performance Analysis of Different Feed Forward Networks in Non-Linear Classification. International Journal of Soft Computing and Engineering. International Journal of Soft Computing and Engineering. 2013; 3:2231-2307.

***