International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 10 | Oct 2021 www.irjet.net p-ISSN: 2395-0072

Prediction of an Artist's Success on Spotify

Apoorva Shete[1], Nishant Mohanani[1], Sheetal Gondal[2]

[1]Department of Electronics and Telecommunication Engineering, Thadomal Shahani Engineering College, Mumbai, India
[2]Assist. Prof., Department of Information Technology Engineering, Thadomal Shahani Engineering College, Mumbai, India

---

Abstract - *The popularity and success of artists can depend on multiple features such as age, gender, location of listeners, number of streams, previous playlist success, audio features and more. The dataset used for the research was the WMG database which contains the streaming data of Spotify of almost 3 million entries. This paper focuses on using and adapting the features of the dataset using three machine learning algorithms namely Decision Tree, KNN, Random Forest in order to predict the success of an artist on Spotify. With the help of previously researched and analyzed algorithms the aim is to create a highly accurate model to predict the artist's success.*

***Key Words*: Spotify, Playlist, Decision Tree, KNN, Random Forest**

1. INTRODUCTION

In this project the aim is to develop a model which allows the most accurate prediction of an artist's success on the song streaming app Spotify. Popularity defined as the number or streams or the number of listening events of that particular artist based on time. The research also includes the data of the users including the age, gender, and location of the streamer in an attempt to get a higher accuracy and audience classification of the audience of the artist.

The dataset used the WMG database that consists of Spotify streaming data of a tenure of 11 years from 2010 to the present with over 50 billion rows of information. For streamlining the process and for better accuracy a reduced dataset of 3 million entries was used with 45 attributes. The machine learning algorithms used were Decision Tree, KNN and Random Forest.

1) Decision Tree: The Decision Tree is a supervised learning algorithm that uses simple decision rules taken from preceding data to predict future class or values of the target variable.

2) KNN: k-Nearest Neighbor is one of the most basic Machine Learning algorithms and it is also based on supervised learning algorithm. It checks for likeness between new data and previously used data and categorizes the data based on similarity.

3) Random Forest: Random Forest is a machine learning algorithm that is used to solve regression and classification using the ensemble technique which combines multiple decision trees and other classifiers to give a solution to complex problems.

The paper contains the methods and results of all three models and a comparative study to relate the results achieved through the different types of algorithms and to determine which model generates the highest level of accuracy.

2. LITERATURE REVIEW

This section helps in reviewing and understanding the contents of the preceding papers and research in the field and go through a variety of relevant approaches.

The paper [1] shows the listening history of an individual by collecting user information such as their age, gender, and country etc. The dataset implemented for the project was a subset of the LFM-1b dataset. This dataset allowed the collection of user profile information (including age, gender, and country) and the listening events of the users. For the implementation, the regression algorithm was successful in providing an error rate under the baseline error rate of 33.7%. Paper [2] uses a dataset of approximately 2,000 hit and 2,000 non-hit songs and in order to predict the Billboard success of a song each song is extracted for its audio features from the Spotify Web API. With the help of six ML classifiers namely Expectation Maximization (EM), Logistic Regression (LR), Gaussian Discriminant Analysis (GDA), Support Vector Machines (SVM), Decision Trees (DT), and Neural Networks (NN) majorly focusing on the accuracy, precision, and recall resulting in the prediction of accuracy of the final experiment.

Paper [3] uses Spotify data to explain the popularity of a song by its audio features and provides the relation between the song, using the song audio features and the success of the song quantified by the number of streams on Spotify. The dataset included 1000 songs from 10 different genres, and they were analyzed for audio features such as 'Acousticness', 'Danceability', 'Energy', etc., measured from 0-1 and are (continuous) ratio variables, and a linear

regression algorithm was developed. Paper [4] is a project to predict a song's popularity based only on its musicality features such as its key, modality, loudness, dissonance, and dynamics variation. The dataset of this model was generated from the Free Music Archive (FMA) and sample song clips from Spotify consisting of 3 genres (Hip-Hop, Jazz, Pop) of 20 popular and 20 non-popular songs each. Different classifiers were used for each dataset to evaluate if a song would be popular were decision tree variation of a C4.5 tree, RIPPER ruleset, Naive Bayes, Logistic Regression, and SVM using Polynomial and RBF kernels. In conclusion to the model the popularity of Jazz and Hip-Hop were best predicted by the algorithm based on features such as speechiness, instrumentalness, and valence. The genre that the model was least able to predict the popularity of was Pop.

Paper [5] is an attempt to predict if songs make the Billboard hits list by using a dataset of 1.8 million songs from the Spotify API, with a reduced dataset of top 100 hits from 1985 to 2018 of 16,000 songs. The track contained 27 features which include track genre, artist, popularity, explicit, danceability, energy, instrumentalness, key and so on. For the prediction of billboard hits the project featured 4 models namely: Logistic Regression (LR), Neural Network (NN), Random Forest (RF), and Support Vector Machine (SVM). Paper [6] demonstrated the prediction of popularity of music artists based on the number of streams of their tracks in the past and on a daily basis. The experiment used the LFM-1b dataset which contains information of users. Linear regression, support vector machines, and neural networks models were adopted to create, analyze, and optimize predictions of the number of listening events an artist will generate per day. Paper [7] raises the possibility of classifying a song based on its audio features (such as danceability, tempo, key, energy and so on) as a hit or a non-hit. Four machine learning models were used for the experiment namely Logistic Regression, K-Nearest Neighbors, Gaussian Naive Bayes, Support Vector Machine. The classification models built using Naive Bayes and Logistic Regression both had an accuracy of 65%, and the Support Vector Machine had an accuracy of 59%.

This study helped us understand the research done in this field in the past.

3. DATASET DESCRIPTION

Collecting data that is appropriate for the purpose of our research is the first and foremost step in any project. The dataset used for the purpose of this research consists of Spotify streaming data in the WMG database. In this paper, we have used a part of this Spotify data. The selected sample data has 3805499 entries and 45 columns. The Table 1 shown below summarizes the columns that this study is focused on.

Table 1. Summary of the attributes used in this research

ATTRIBUTES	DESCRIPTION
Log Time	Time of every stream
Artist Name	The name of the artists that have created the song
Track Name	Name of the song's track
ISRC	A unique identifier for each version of the song
Customer ID	Customer ID of the listener for Spotify
Location of Customer	Listener's location
Gender of Customer	Listener's gender
Stream Source URI	Where the song was played on Spotify (a unique playlist ID, an album etc)
Birth year	Listener's birth year

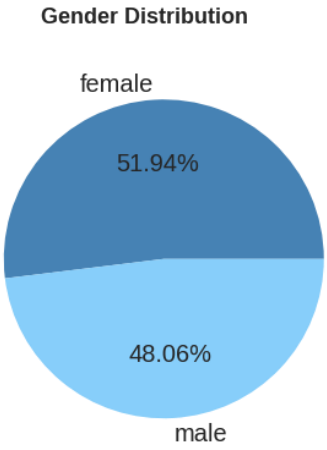

Fig.1 Gender Distribution in the dataset

The plot shown in Fig.2 shows that the listener's birth year is generally between 1980 and 2000, indicating that adults and young adults are the majority of listeners.

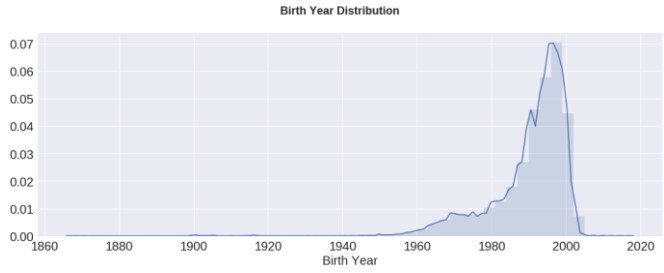

Fig.2 Distribution of Listener's Birth Year

After this, the top 10 most streamed artists were plotted. This is shown below in Fig.3. It can be seen that Charlie Puth is the most streamed artist, followed by Dua Lipa, with over 40,000 and 30,000 streams respectively.

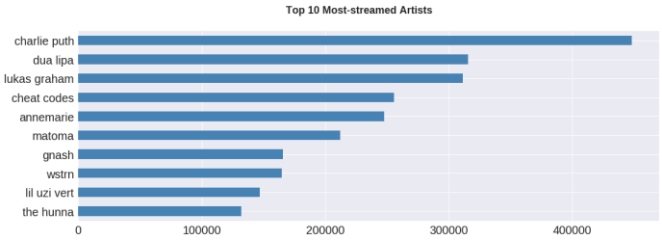

Fig.3 Top 10 Most Streamed Artists

Similarly, the top 10 most streamed playlists were plotted. This is shown below in Fig.4.

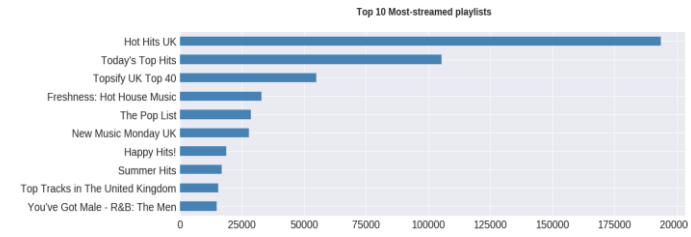

Fig.4 Top 10 Most StreamedPlaylists

Next, the stream count was plotted as per weekdays, and it can be concluded from the Fig.5 shown below that the majority of the streams were on Monday, Saturday, Friday and Wednesday.

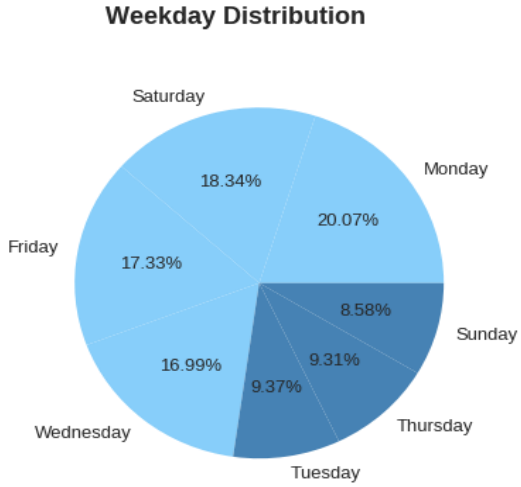

Fig.5 Distribution of Streams as per weekdays

The streaming devices of the users were also plotted, as shown in Fig.6 below. Majority of the users used a mobile for streaming songs.

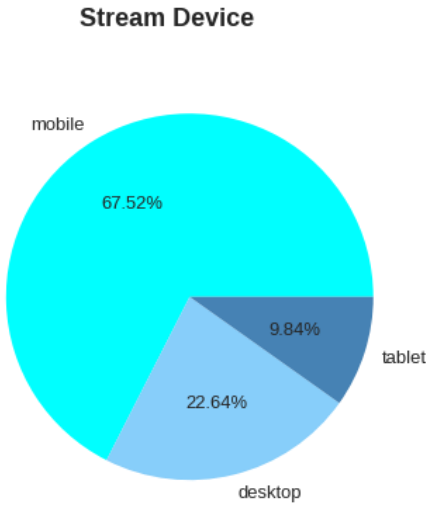

Fig.6 Distribution of Streaming Devices

The most common streaming operating systems were plotted, and it was found that the majority of the users used iPhone OS, followed by Android. This is shown below in Fig.7.

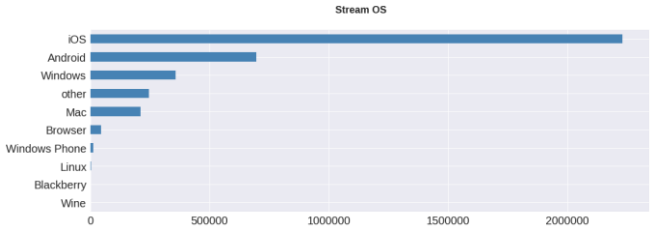

Fig.7 Distribution of Streaming OS

In this research, the criterion of success is decided by whether or not the artist has been listed in either of the four key playlists. Various factors are involved that have an impact on an artist's success. Therefore, to build a predictive model for the success of an artist, we need to change these factors into measurable quantities. These numerically estimated values can be the driver of accurateness of our model. There are several methods by which these features can be generated.

In the dataset used, there is a bias as there are more failure cases than the instances of success. Thus, there may be a strong bias for prediction of 'failure' by our model because of the dataset used. However, to get better results, we have sampled the data and used data with 50-50 instances of success and failure. This is shown below in Fig. 8 and Fig. 9.

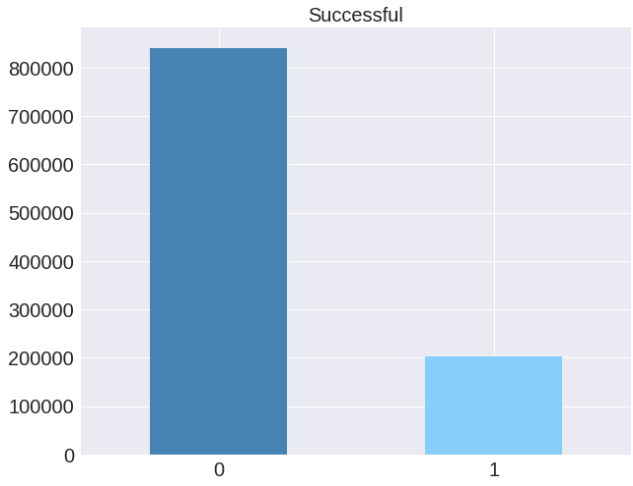

Fig.8 Data before sampling

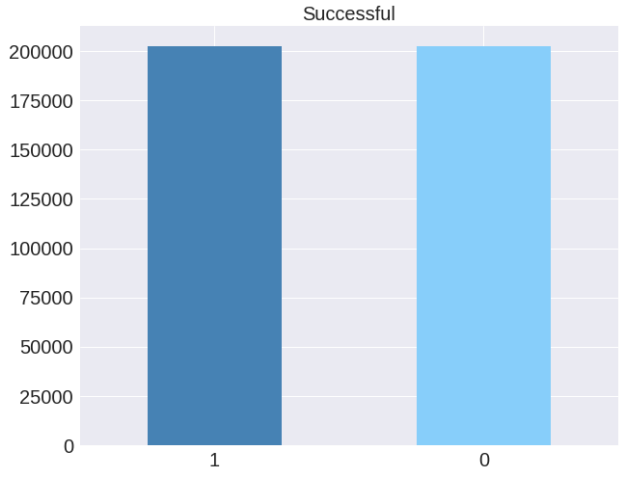

Fig.9 Sampled Data

The next step was the evaluation of data using the different ML algorithms. In this research, we have used Decision Tree, Random Forest and KNN for the evaluation of our model.

The framework for the methodology for the application of this model throughout our research has been shown below inFig.10.

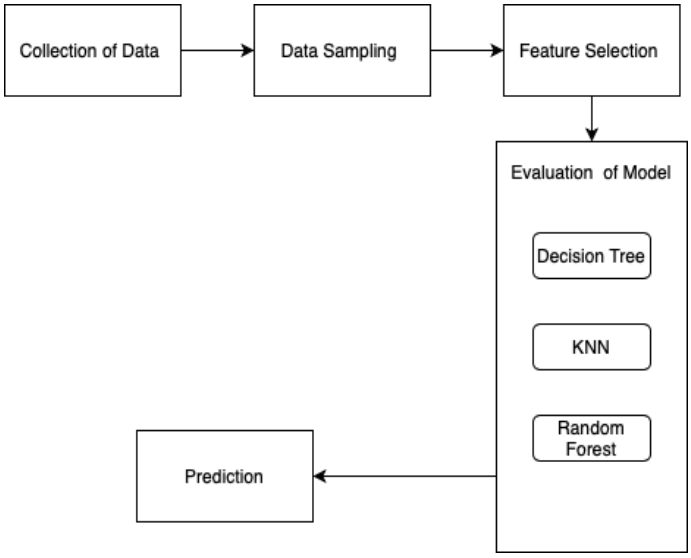

Fig.10 Framework for the implementation of our model

4. RESULTS AND DISCUSSIONS

The results achieved by each algorithm are reviewed in this section.

The accuracy obtained by each of the algorithms used is summarized below in Table.I.

Table.I Accuracy achieved with each algorithm on the test data

ALGORITHM USED	ACCURACY
Decision Tree	71.11%
KNN	67.6%
Random Forest	70%

A comparative graph of the accuracies achieved from each algorithm on the test data has been shown below in Fig. 11.

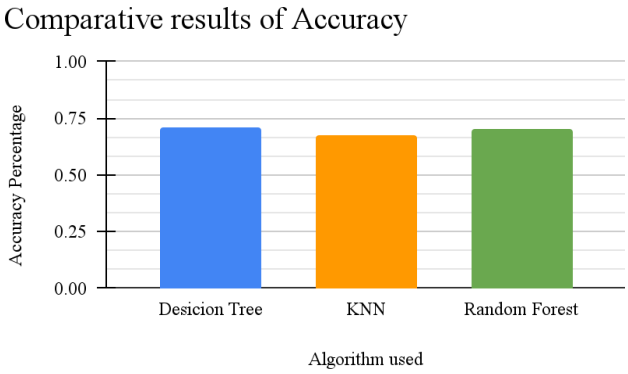

Fig. 11. Comparative results of accuracy obtained by the chosen classifiers

Since the best accuracy was obtained using the DT classifier, the Confusion matrix and ROC curve for the same were plotted.

A. *Confusion Matrix*

This is a method of summarizing the performance of a classification model. It tells us what our classification model is getting right, and the kind of errors made by this model.

The confusion matrix has been shown below in Fig.12. The predicted data results can be understood in the following manner from the confusion matrix:

- Negative classes identified as negative (True Negatives): 2180
- Negative classes identified as positive (False Positives): 18736
- Positive classes identified as negative (False Negatives): 5815
- Positive classes identified as positive (True Positives): 34741

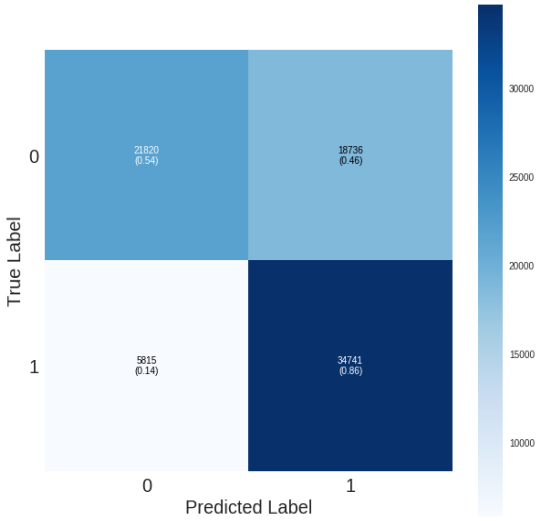

Fig.12 Confusion Matrix for DT classifier

B. *ROC AUC*

The Receiver Operating Curves or ROC curves give the capability of the model to identify and classify the

parameters correctly. It plots the True Positive Rate (TPR) v/s the False Positive Rate (FPR). The AUC explains the results of ROC.

The ROC curve for this model is shown below in Fig.13.

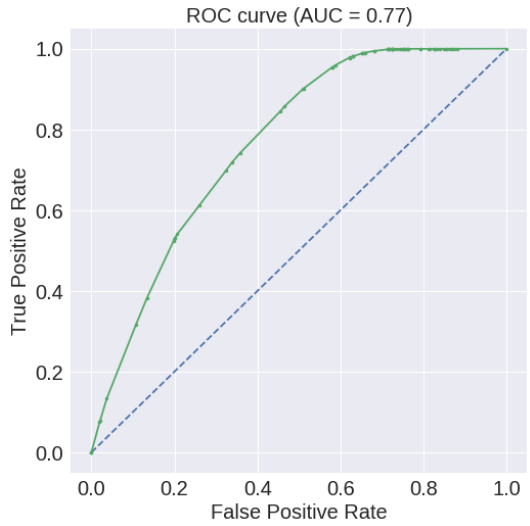

Fig.13 AUROC plot of the implemented model

5. CONCLUSION AND FUTURE SCOPE

Spotify is used by millions of users across the globe. Thus, prediction of the success of an artist's work on Spotify can help song producers and investors a great deal. In this paper, we have successfully built a model that does the same with a good level of accuracy. We have also compared three different ML algorithms and the Decision Tree classifier has given us the best accuracy and results.

In future, we can use a larger dataset to gain much more accurate results about an artist's work on Spotify. Along with this, the comments and information about an artist's engagement on various social media platforms can also be included.

Thus, there is major scope of improvement for research in this field in the future.

ACKNOWLEDGEMENT

We would like to thank Assist. Prof. Sheetal Gondal for her support and guidance throughout this project. We would also like to thank our Head of Department, Dr. Ashwini Kunte, and our Principal, Dr. G.T. Thampi.

REFERENCES

[1] Krismayer, T., Schedl, M., Knees, P. *et al.* Predicting user demographics from music listening information. *Multimed Tools Appl* 78, 2897–2920 (2019). https://doi.org/10.1007/s11042-018-5980-y

[2] E. Georgieva, M. Suta, N. Burton, Hitpredict: Predicting Hit Songs Using Spotify Data Stanford Computer Science 229: Machine Learning. http://cs229.stanford.edu/proj2018/report/16.pdf

[3] Nijkamp, R.. "Prediction of product success: explaining song popularity by audio features from Spotify data." (2018). https://www.semanticscholar.org/paper/Prediction-of-product-success%3A-explaining-song-by-Nijkamp/ba06fe3e0b65e799fcc7b1434dac387f986da1ed#citing-papers

[4] Y. Chen, A. Dixit, S. Sanyal, et al. Show Me What You Got: Song Popularity Prediction Using FMA Dataset. https://www.ischool.berkeley.edu/sites/default/files/sproject_attachments/info_251_-_final_project_report.pdf

[5] K. Middlebrook, K. Shaik. SONG HIT PREDICTION: PREDICTING BILLBOARD HITS USING SPOTIFY DATA. (2019) https://arxiv.org/pdf/1908.08609.pdf

[6] F. Jetzinger, F. heumer, M. Schedl. Towards Predicting the Popularity of Music Artists(2017). http://www.cp.jku.at/people/schedl/Research/Publications/pdf/jetzinger_mml_2017.pdf

[7] M. Reiman, P. Ornell. Predicting Hit Songs with Machine Learning.

https://kth.diva-portal.org/smash/get/diva2:1214146/FULLTEXT01.pdf

[8] "ConfusionMatrix", https://machinelearningmastery.com/confusion- matrix-machine-learning/

[9] "AUCROC", https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-