International Journal of Advances in Engineering and Management (IJAEM)
Volume 5, Issue 5 May 2023, pp: 296-304 www.ijaem.net ISSN: 2395-5252

Spotify Data Analysis and Song Popularity Prediction

Sivasai Bhavanasi[1], Sahil Malla[2], V Manichetan [3], CVNJ Dhanush[4], Dr B Prakash[5]

1,2,3,4 Student - Department of Computer Science and Engineering, Gitam University, Visakhapatnam, Andhra Pradesh, India

5 Associate Professor - Department of Computer Science and Engineering, Gitam University, Visakhapatnam, Andhra Pradesh, India

---

Date of Submission: 01-05-2023 Date of Acceptance: 10-05-2023

ABSTRACT— Automated systems that are capable of gathering and analyzing massive quantities of data from the internet are revolutionizing music distribution in the current digital era. The current goal of this project is to use different machine learning and statistical methods to analyze the audio features of songs.

We have tried a number of models that include random effects because we are especially curious about what makes a song popular and to gain more knowledge about these factors. Our research will be valuable for those who want to forecast the success of new products in the music business because it identifies the crucial factors that influence a song's success.

With the help of this research, we aim to shed light on the association between a song's audio features and its popularity and identify the elements that contribute most to a song becoming a success. We think that our study will offer an essential contribution to our comprehension of the music industry and its future course in this fascinating area of research.

Index Terms—Machine Learning, Song Popularity Prediction, Statistics

I. INTRODUCTION

People have always placed a high value on music, and the digital era has made music more accessible than ever. Due to the growth of music streaming services and digital music libraries, music lovers can now amass sizable song compilations that they can listen to whenever they want and from wherever they are. It has gotten harder to keep track of all the music accessible and the connections between different tunes as the amount of music continues to increase.

The use of machine learning and ensembling methods has become essential for navigating the complicated world of music. These methods have the capacity to analyse enormous amounts of musical data, offering insights into the characteristics of a successful song, the variables influencing a song's success, and the potential to forecast the success of new songs.

Our current research focuses on exploring the relationship between song popularity and song audio attributes taken from the Spotify database, such as key and tempo. Specifically, we will be analyzing the amount of streams a song receives on Spotify as a measure of its popularity. By examining the audio attributes of songs and their popularity levels, we aim to identify the features that contribute to a song's success and popularity. Our objective is to develop a model that that can help music professionals and enthusiasts better understand the music industry and make predictions about the success of new songs.

Our research will involve using various machine learning and ensembling techniques to analyze large amounts of musical data. By employing statistical methods and identifying the most significant features that contribute to a song's success, we hope to provide valuable insights into the music industry's future direction. We will also explore questions such as whether the success of previous songs can be used to predict the success of new songs and what factors have the most significant impact on a song's success. Ultimately, our project aims to deepen our understanding of the music industry and how it is evolving in response to technological advancements.

II. PROBLEM IDENTIFICATION AND OBJECTIVES

Music streaming services like Spotify have been changing the way people consume music. However, the vast amount of data that is generated by these services has created an opportunity for data analysts to gain insights into consumer behavior and preferences. The problem is how to predict the popularity of a song based on various features such as artist, genre, danceability, energy, tempo, etc. This information can help music streaming services to understand the factors that contribute to a song's success and tailor their playlists and recommendations to their users' preferences. Moreover, it can also help record labels to identify the characteristics of popular songs and help them in their decision-making process for investing in artists and songs.

The issue is crucial for up-and-coming and independent artists who wish to increase their chances of success by comprehending the essential elements that lead to a song's popularity. Yet, due to the subjective nature of music preferences and the intricacy of the variables that contribute to a song's success, predicting song popularity is a difficult undertaking. Hence, to evaluate the data and discover the crucial factors that influence song popularity, the challenge necessitates the application of advanced machine learning techniques including regression analysis, decision tree classification, random forest, and neural networks. By resolving this issue, music streaming services, record labels, and artists will be better able to make decisions that will significantly affect their ability to succeed as a business.

This project can have several practical applications. For example, understanding the relationships between variables such as tempo and energy can help music producers and composers make more informed decisions when creating new music.

Similarly, identifying the least useful predictors for a song's popularity can help marketers and music industry professionals focus their efforts on more relevant variables.

Analyzing the association between a song's duration and popularity can provide insights into listener preferences and help guide decisions around song length

Predictive models developed in this project, such as predicting a song's popularity based on its danceability, energy, and loudness, can have commercial applications in music recommendation systems or marketing strategies

The analysis of genre and subgenre frequencies and popularity can provide insights into trends in music preferences over time, and clustering songs by genre using album name, track popularity, and artist name can help music streaming services enhance their recommendation algorithms

Overall, the insights gained from this project can help music industry professionals make informed decisions and develop effective strategies for promoting and marketing new music.

III. SYSTEM METHODOLOGY

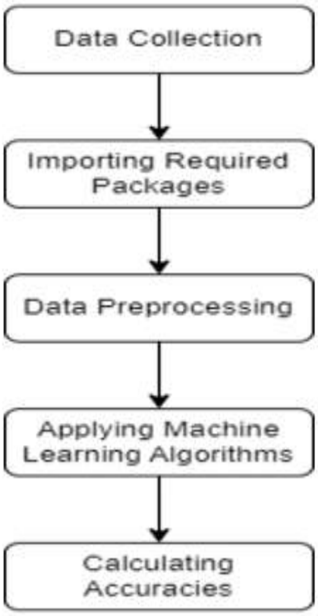

Data Collection
Importing Required Packages
Data Preprocessing
Applying Machine Learning Algorithms
Calculating Accuracies

Fig 1 . Project Flow

IV.IMPLEMENTATION

A. Knowing About the dataset

The dataset used in this study was taken from Spotify and includes a number of characteristics and parameters pertaining to the music. The dataset includes 30000 songs with the attributes listed below.

TABLE I: DATASET COLUMNS

Column	Description
Track ID	Unique ID of each song
Track Name	Name of the song
Track Artist	Name of the song artist
Track Popularity	Popularity of the song
Track Album ID	Unique ID of the album
Track Album Name	Name of the album
Track Album Release Date	Album Release Date
Playlist Name	Name of the playlist
Playlist ID	Unique ID of the playlist
Playlist Genre	Genre of the Playlist
Playlist Sub Genre	Sub-Genre of the playlist
Danceability	Track Suitability for dancing
Energy	Measure for activity, intensity etc.
Key	Key of the track
Loudness	Loudness of the track measured in decibels
Mode	The modality of the track
Speechiness	Presence of spoken words
Acousticness	Confidence measure of the track
Instrumentalness	Value showing the presence of instruments
Liveness	Detects the presence of audience

Valence	Positiveness conveyed by the track
Tempo	Tempo in BPM
Duration	Duration of the song measured in milliseconds

A. Importing the modules

At first, we will load a few libraries, including those required for representing, displaying, and configuring data, as well as those compatible with running models. These modules are all used at various stages of a project. In our project, each module plays a crucial role in providing models and creating the graphs that we need to develop libraries.

1. NumPy is a tool for performing many types of mathematical computations.
2. Pandas: This module is used to read, process, and analyze the data. With pandas, we can quickly manipulate and interpret data as needed.
3. SK-Learn: It is used to import various algorithms and calculate the model's accuracy. With this module, we may fine-tune the parameters to enhance the performance of the model.
It also provides us with various ensembling and advanced algorithms.
4. Matplotlib: It is utilized to create different graphs and visualise the model. It provides us a lot of predefined methods to draw various graphs and metrics.

B. Data Pre-processing

About 150 null values can be found in the dataset. In order to deal with the null values, we implemented the backward and forward fill technique, replacing the null value with the mean of the two forward and two backward values. We made sure to exclude duplicate data from the dataset as well. By doing this, we have made sure that the data is accurate and ready for use.

C. Data Analysis

1. Which characteristics are the least effective for determining how popular a song is?

After adjusting alpha to 0.6 in the Lasso Regression model, we discovered that key, mode, speechiness, valence, and tempo are the least useful criteria for determining a song's popularity. The MSE for training came out to 10.32, while the MSE for testing came out to 11.21 when looking at how the model actually performed.

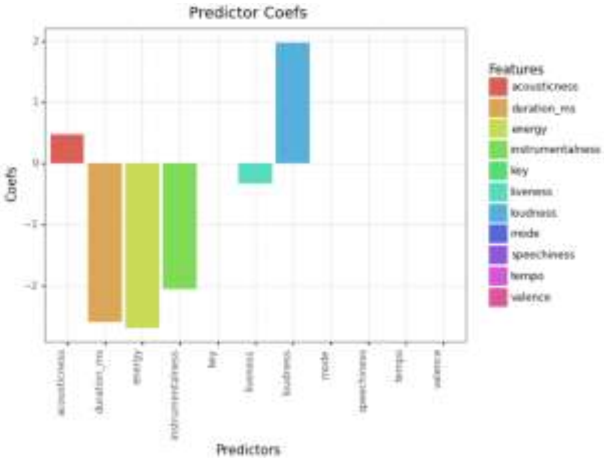


Fig 2 Plot Showing the coefficient of each characteristic

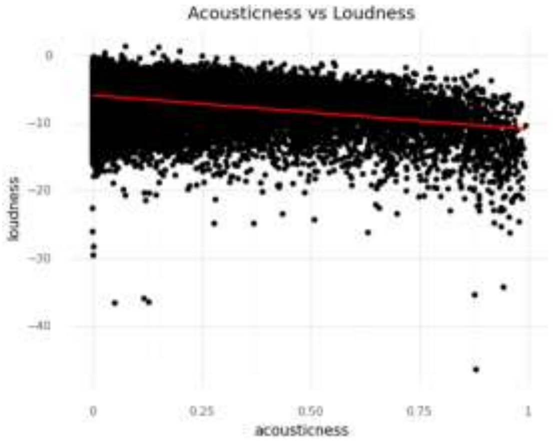


Fig 3 Scatter plot between acousticness and loudness

In Fig 2 To make it apparent which variables were eliminated from the model and which ones remained; we developed this bar graph. It's noteworthy to note that the coefficients for duration_ms, energy, instrumentalness, and liveness were all negative. The only positive coefficients were for acousticness, valence, and loudness. Ultimately, there is no bar for key, mode, speechiness, or pace because Lasso effectively removed these from the model.

In Fig 3 The two most positive coefficients, acousticness and loudness, are plotted against one another in this graph. After analysing the plot, it is apparent that there is a relationship between the two because loudness gradually reduces as acousticness rises. Nonetheless, the data are typically very close to the regression line. While examining acousticness in the range of 0.8 to 1.

2. Is there an association between the duration and popularity of songs?

One of the elements that significantly affects a song's appeal is its duration. we used lasso regression to ascertain whether there was a connection between a song's length and popularity. we wanted to examine the coefficient and determine whether there was any correlation between the two that was favourable. Duration was found to have a coefficient of -3.53. Also, I examined the MSE and R2 values after that. The MSE for this model comes out to 110.32, which indicates that it is operating at an average level. The R2 calculated as 7.9 is likewise typical. Although the data in the residual plot show an upward tendency, the relationship between duration and popularity isn't quite linear

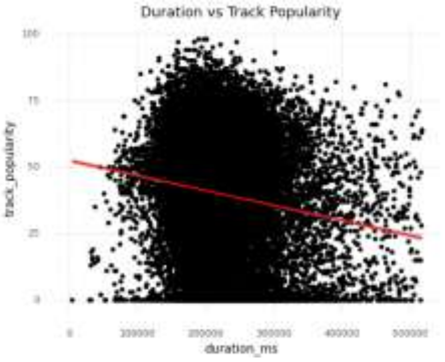


Fig 4 Scatter Plot between Duration and track popularity

3. Which genres and subgenres show up the most in the data? Which are the most popular?

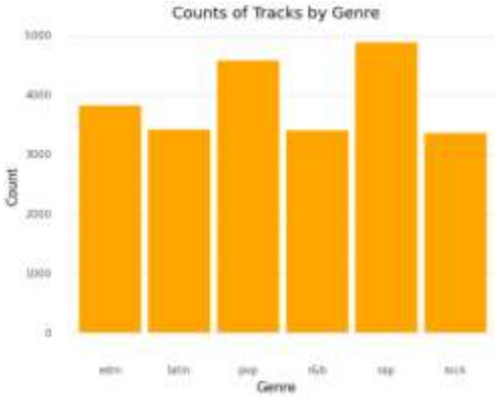


Fig 5 Plot showing count of each track by Genre

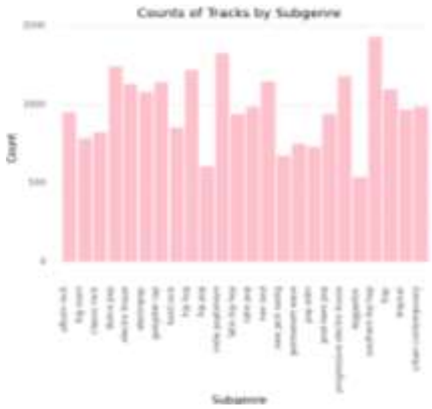


Fig 6 Plot showing count of each track by Sub Genre

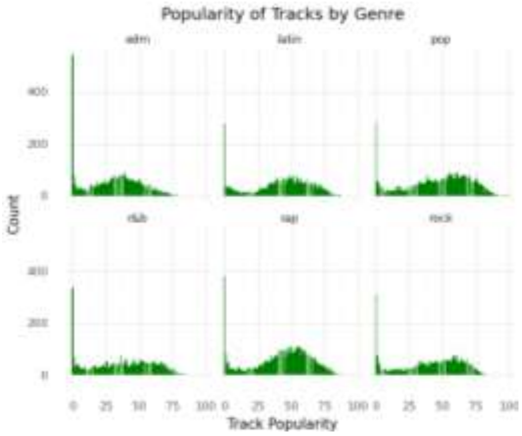


Fig 7 Plot showing popularity of each track by Genre

Popularity of Tracks by Subgenre
album rock
big room
classic rock
dance pop
electro house
electropop
gangster rap
hard rock
hip hop
hip pop
indie poptimism
latin hip hop
latin pop
neo soul
new jack swing
permanent wave
pop edm
post-teen pop
progressive electro house
reggaeton
southern hip hop
trap
tropical
urban contemporary
Count
Track Popularity

Fig 8 Plot showing popularity of each track by Sub Genre

A. Song Popularity Prediction

To predict the song popularity hit prediction we have used the following algorithms

1. Random Forest

A particular kind of machine learning algorithm called Random Forest is employed for classification, regression, and other applications. It is based on the idea of building several decision trees and merging the output to produce forecasts that are more accurate.

Take the situation where you are attempting to forecast whether it's going to rain or not tomorrow. You can accomplish this by making a decision tree by posing a series of queries, like "What is the temperature?" or "Is there any wind?" according to past experience. A single decision tree might not, however, be accurate enough to provide a solid prediction.

The random forest algorithm is used in this situation. It generates many decision trees with various sets of questions and data samples as opposed to only one. Following the combination of these trees, the final forecast is made using the outcomes of all the trees combined.

This approach can produce more precise predictions than a single decision tree and lowers the danger of overfitting, this occurs when the model is very complicated and matches the training data too closely.

2. Gradient Boosting

Gradient boosting is a machine learning algorithm which is used for various classification and regression problems. It works by combining several weak models (called "weak learners") into a single, strong model.

Imagine you are trying to predict the price of a house, and you have a dataset with features such as the number of bedrooms, the square footage, and the location. To use gradient boosting, you would first create a simple model (the "weak learner") that makes a rough prediction based on one or two features.

For example, you might create a weak learner that predicts the price of a house based only on the number of bedrooms. This model would not be very accurate, but it would be better than simply guessing.

Next, you would create another weak learner that tries to correct the errors of the first one. This model might predict the difference between the actual price and the predicted price based on the number of bedrooms.

Then, you would combine the predictions of both models, giving more weight to the second model's predictions, and use the combined prediction to update the first model. This process would be repeated multiple times, each time creating a new weak learner that corrects the errors of the previous models.

By combining the predictions of multiple weak models in this way, gradient boosting is able to create a strong model that can make accurate predictions.

3. Bagging Classifier

Bagging Classifier is a type of machine learning algorithm that is used for classification problems. It works by creating multiple subsets of the original dataset (called "bags" or "bootstrap samples"), training a separate classifier on each subset, and then combining their results to make a final prediction.

Imagine you are trying to classify emails as spam or not spam. To use bagging, you would first create multiple random subsets of the original dataset, each containing a different set of emails.

Next, you would train a separate classifier (such as a decision tree or a logistic regression model) on each subset, using features such as the email's subject line, sender, and content to make a prediction.

Then, you would combine the predictions of all of the classifiers, giving equal weight to each classifier's prediction. If most of the classifiers predict that an email is spam, then the bagging classifier will classify it as spam as well.

By creating multiple subsets of the data and training separate classifiers on each subset, bagging helps to reduce the risk of overfitting (when a model is too complex and fits the training data too closely).

4. XG Boost

A machine learning algorithm called XG boost combines the predictions of several insufficient models, such as decision trees, to produce a more robust and precise model.

In XGBoost, weak models are trained iteratively in a way that each new model corrects the errors made by the previous models. This is done by assigning higher weights to the data points that were misclassified in the previous iteration.

Due to the fact that the algorithm minimises a loss function that assesses the discrepancy between predicted and real values, it is known as "Extreme" Gradient Boosting. Additionally, a number of regularisation strategies are included to avoid overfitting

XGBoost has become very popular in machine learning competitions and is widely used in industry because of its high predictive accuracy and speed.

5. Decision Tree Classifier

For many classification and regression issues, the Decision Tree method is used. It operates by breaking the data down into progressively smaller subsets according to a series of rules that are discovered through the data.

Imagine you want to classify whether or not a person likes to go to the beach, based on their age, gender, and income. To create a decision tree, you would start by selecting the feature that is most informative for the task, such as age.

Then, you would split the data into two subsets based on the age threshold that provides the best separation between people who like and don't like to go to the beach. For example, you might split the data into two groups: one for people under 30 and one for people over 30.

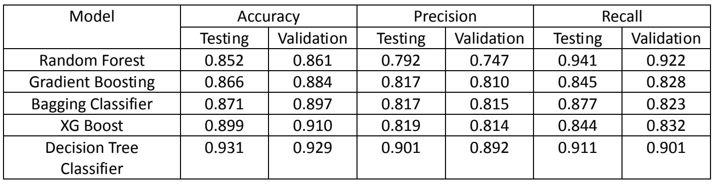
Then, for each subset, you would repeat this procedure by choosing the subsequent most informative feature (such as gender) and dividing the data once more. This procedure would be carried out repeatedly until a stopping criterion was reached, such as the minimal number of data points required for each subset or the maximum depth of the tree.

At the end of this process, you would have a decision tree that can be used to predict whether a new person likes to go to the beach or not, based on their age, gender, and income.

Decision trees are popular because they are easy to interpret and can handle both numerical and categorical data. English-speaking colleague to proofread your paper.

V.RESULTS

Training accounts for 70% of the data, and testing accounts for 30%. When it was given to these models, the following outcomes were attained:



Model	Accuracy		Precision		Recall	
	Testing	Validation	Testing	Validation	Testing	Validation
Random Forest	0.852	0.861	0.792	0.747	0.941	0.922
Gradient Boosting	0.866	0.884	0.817	0.810	0.845	0.828
Bagging Classifier	0.871	0.897	0.817	0.815	0.877	0.823
XG Boost	0.899	0.910	0.819	0.814	0.844	0.832
Decision Tree Classifier	0.931	0.929	0.901	0.892	0.911	0.901

Fig 9 Outcomes of various models

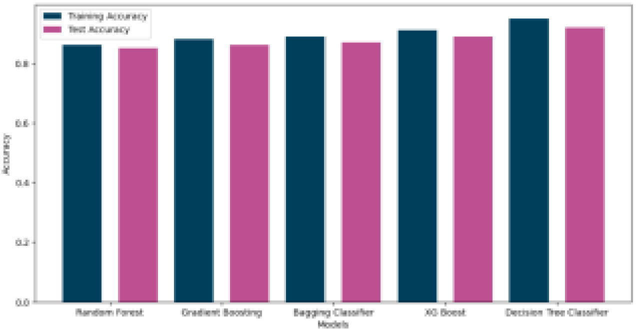


Fig 10 Plot Showing Accuary of each model

CONCLUSION AND FUTURE SCOPE

With the emergence of deep learning models and machine learning algorithms in recent years, the field of music analysis and prediction has undergone a major transformation. Particularly, studies have demonstrated that a song's success can be significantly influenced by a number of its defining characteristics. We have found that the popularity of a song is favourably influenced by factors like acousticness, valence, and loudness through the analysis of large musical datasets.

We used a decision tree classifier, which has been shown to be the most accurate technique for predicting the popularity of a song, to make these predictions. The algorithm has a 93% accuracy rate, which means that in the vast majority of instances, it can accurately determine the popularity of a song.

One potential application of this technology is in the realm of music streaming services. By using machine learning algorithms to predict the popularity of songs, these services can more effectively curate playlists and recommend new music to listeners. This could lead to a more satisfying user experience and help music streaming services stand out in a crowded and competitive marketplace.

Overall, the use of machine learning algorithms and deep learning models is revolutionizing the field of music analysis and prediction. By leveraging these tools, we can better understand the factors that contribute to a song's popularity and create more effective systems for recommending music

REFERENCES

[1] A MODEL-BASED APPROACH TO SPOTIFY DATA ANALYSIS: A BETA GLMM

Mariangela Sciandra and Irene Carola Spera essay manages the auditory aspects of songs from a statistical perspective. Also, it recommends statistical tools for the study of this data. It focuses in particular on the data capturing techniques made possible by Spotify Web API. In order to provide a preliminary response to queries like, "What factors determine popularity?" special consideration is given to song popularity. A Beta model with random effects is provided. Finding a model that may explain this link and identifying the traits that are thought to be most crucial in popularizing a song are both extremely fascinating topics for those attempting to forecast the success of new items.

[2] SONG HIT PREDICTION: PREDICTING BILLBOARD HITS USING SPOTIFY DATA

The Hit Song Science challenge, which seeks to predict which songs will top the charts, is being worked on by Kai Middlebrook and Kian Sheik. They created a dataset of about 1.8 million hit and non-hit songs, and then used the Spotify Web API to extract their audio properties. On their dataset, they examined four models. Their top model, random forest, has an accuracy of 88% in predicting Billboard song success.

[3] HITPREDICT: PREDICTING HIT SONGS USING SPOTIFY DATA

To forecast which songs would top the Billboard Hot 100, Elena Georgieva, Marcella Suta, and Nicholas Burton took on the Hit Song Science challenge. They compiled a dataset of around 4,000 hit and non-hit songs and retrieved each song's audio characteristics using the Spotify Web API. Using five machine learning algorithms, they were able to predict a song's Billboard success with about 75% accuracy on the validation set. Logistic Regression and a neural network with a single hidden layer were the most effective methods.

[4] AUTOMATIC PREDICTION OF HIT SONGS RUTH DHANARAJ1, BETH LOGAN HP LABORATORIES CAMBRIDGE HPL-2005-149 AUGUST 17, 2005

Ruth Dhanaraj, Beth Logan HP investigated the use of artificial music analysis to predict potential popular tunes. They take the acoustic and lyric data from each song and, using common classifiers like

Support Vector Machines and boosting classifiers, sort the hits from the duds. Its characteristics are based on either global subjects or global sounds that have been unsupervised learned from lyric databases or acoustic data. A corpus of 1700 songs used in experiments shows performance that is significantly better than random. When properly recognizing successful songs, the lyric-based elements are somewhat more helpful than the acoustic features. The two characteristics cannot be combined, and no noticeable benefits result. Examination of the lyric-based characteristics reveals that a song is more likely to be a hit when specific semantic information is missing

[5] MUSIC POPULARITY: METRICS, CHARACTERISTICS, AND AUDIO-BASED PREDICTION

To account for many facets of popularity, Junghyuk Lee and Jong-Seok Lee first construct eight measures for popularity. The features of music popularity in the actual world are then thoroughly understood by examination of each popularity metre using long-term real-world chart data. Moreover, they create categorization algorithms based on auditory data to forecast popularity indicators. They specifically concentrated on analyzing MPEG-7 and Mel-frequency cepstral coefficient (MFCC) characteristics with other traditional acoustic data, such as measures representing music complexity. The results demonstrate that, even if there is still space for improvement, it is possible to forecast a song's popularity metrics based on its audio signal far more accurately than by chance, especially when employing both the intricacy and MFCC characteristics.

[6] SPOTHITPY: A STUDY FOR ML-BASED SONG HIT PREDICTION USING SPOTIFY

The goal of this study was to forecast which songs will become Billboard hits by taking a hit song prediction approach developed by Ioannis Dimolitsas, Spyridon Kantarelis, and Afroditi Fouka. Through using Spotify Web API, we compiled a dataset of about 18500 hit and non-hit songs and extracted their audio attributes. On our dataset, we evaluate four machine learning models. Our accuracy rate in predicting a song's Billboard success was about 86%. Support Vector Machine and Random Forest were the most effective algorithms.

[7] MUSICAL TRENDS AND PREDICTABILITY OF SUCCESS IN CONTEMPORARY SONGS IN AND OUT OF THE TOP CHARTS

In order to comprehend the dynamics of success, associate success with acoustic qualities, and investigate the predictability of success, Myra Interiano, Kamyar Kazemi, Lijia Wang, Jienian Yang, Zhaoxia Yu, and Natalia L. Komarova examined upwards of 500 000 songs released with in UK during 1985 and 2015. There have been shown to be several multi-decadal trends. For instance, "happy" and "brightness" are clearly trending downward, while "sadness" is trending somewhat upward. Also, music is becoming less "masculine." It's interesting how popular songs have their own unique dynamics. They are typically "feminine" and more "happy" than other people, as well as more "party-like" and less "relaxed." It's not easy to tell what makes a song popular vs ordinary. Successful songs tend to anticipate the dynamics of all songs in some contexts, while in others they tend to mirror the historical records. They employed random forests to forecast song popularity, initially using their acoustic characteristics and subsequently including the "superstar" variable (telling us whether the song's artist had recently debuted in the top charts). This made it possible to quantify how much the songs' purely musical qualities contributed to their success and offered a timeline for how popular music trends change over time.

[8] A COMPARATIVE ANALYSIS OF K-NEAREST NEIGHBOR, GENETIC, SUPPORT VECTOR MACHINE, DECISION TREE, AND LONG SHORT TERM MEMORY ALGORITHMS IN MACHINE LEARNING

This report by Malti Bansal, Apoorva Goyal, Apoorva Choudhary mainly discusses in-depth all aspects of five algorithms for machine learning: K-Nearest Neighbor (KNN), Genetic Algorithm (GA), Support Vector Machine (SVM), Decision Tree (DT), and Long Short-Term Memory (LSTM) network. These five algorithms are necessary to enter the field of machine learning (ML) and have been covered in this paper in great detail. By investigation and analysis of recently published papers that conducted quantitative and qualitative research on real-time challenges, mostly predictive analytics in diverse disciplines, this work sheds light on a variety of novel results and conclusions linked to these algorithms. The circumstantial origin of these algorithms is another topic covered in this paper, even though it hasn't been discussed much in prior works but is still a hot topic among ML enthusiasts and novices alike. The algorithms were thoroughly examined and explored in all aspects, both qualitatively and statistically, to explain and comprehend the correctness, robustness, and dependability of the algorithms. The LSTM network and SVM algorithm have projected a superior behavior over the others.