Spotify Song Popularity Prediction - a Machine Learning Exploration

Aakash Kumar Harit
Department of Computer Science
University of Exeter
Ah1166@exeter.ac.uk

Wang Zeliang
Faculty of Environment, Science and Economy
University of Exeter
z.wang6@exeter.ac.uk

***Abstract*— Understanding the complex dynamics that influence song popularity is an intriguing endeavour in the field of music analytics. Through the use of various machine learning techniques, this research begins a thorough investigation of predicting music popularity. Using a comprehensive dataset that includes song traits, features, and popularity, we use predictive modelling to identify the elusive characteristics that differentiate popular songs apart from less well-known ones.**

The project starts with a complex data collection procedure that gathers a variety of important information from a variety of sources. These include lyrical content, artist metadata, and auditory qualities including tempo, energy, and danceability. The data is prepared for analysis by meticulous pre-processing, guaranteeing quality and consistency.

With the use of intelligent data segmentation and cross-validation methods, the predictive models go through a careful training and evaluation process. In order to improve model performance and interpret-ability and encourage a thorough understanding of the factors that influence song popularity, feature engineering and selection procedures are used.

The findings of this study are important for both music lovers and business executives. The knowledge gained from the predictive models aids in understanding the complex jumble of components that make up a song's rise to fame. These models' interpret-ability provides an interesting vantage point from which to view prospective trends and underlying preferences in the music industry.

In conclusion, this study offers proof that music and machine learning work well together. It aims to shed light on the path to song popularity by utilizing a variety of algorithms and provides a multifaceted viewpoint on the harmonic interaction between data-driven insights and the artistic sphere.

Keywords— Machine Learning (ML), Random Forest, SVM, target, features,gradient boosting,neural network

I. INTRODUCTION

The fusion of music and data science has opened up a world of opportunities and changed how we see and comprehend the dynamics of song popularity. The subject of what makes a song resonate with a worldwide audience has taken on new dimensions in an era marked by unheard-of access to such a vast diversity of music. [1]This research uses a variety of machine learning methods to explore into the fascinating field of predicting music popularity. This study tries to identify the complex patterns that underpin the success of hit songs by utilizing a large data-set that includes song elements, lyrical content, and artist traits.

Due to the digitization of music consumption and the growth of streaming platforms, the music industry has undergone a significant transition recently. With millions of songs at listeners' disposal, it has become more difficult to pinpoint the qualities that give some songs prominence. For artists, producers, and other industry participants, understanding these characteristics is extremely valuable since it offers insights into audience preferences and trends. This research aims to understand how a song's transition from obscurity to ubiquity is influenced by a combination of auditory features, lyrical substance, and artist background through the lens of machine learning.

A key component of this effort is the creation of a data-set that captures the diversity of song styles. This requires gathering a wide range of data, including lyrics, contextual information about the artist, and audio aspects like tempo, energy, and danceability. Using this data-set as a starting point, the project employs a variety of machine learning models, each of which offers a unique lens for examining song popularity. This research employs a flexible ensemble of strategies to identify predictive patterns, ranging from traditional algorithms like Logistic Regression and Decision Trees to contemporary approaches like Neural Networks and Gradient Boosting.[2]

Our central research question is: Can the characteristics of a song be used to predict its popularity? This question, which has its roots in our love of music, has relevance in the contemporary environment where machine learning algorithms create musical hits. Our investigation is sparked by the appeal of algorithmic-ally generated hits, which pushes us to identify the characteristics that have the most effect over this creative process. We set out on a journey that encompasses both the analytical rigour and the emotional resonance of music's universal language as the harmonies of data science and music talent come together.

II. BACKGROUND

A. Research Context

A fascinating area of research on the interplay between social media usage and musical tastes has arisen [7], illuminating the complex relationships between online interactions and actual-world occurrences. The foundation for our current project was laid by earlier research that delved into related areas and looked into possible connections between social media involvement and song performances on music charts.

Social media platforms have assumed the role of modern barometers, collecting popular emotion and trends across a range of areas, including music. Utilising tools like Twitter, researchers [12, 13] have studied user interactions with music. The alignment between Twitter discussions regarding musical genres and popular musical tastes was demonstrated , suggesting a link between online discussions and actual musical preferences. Li, X., Wang, L., & Sung, E. (2005). research also demonstrated the social media data's predictive ability, notably in determining song popularity, with user interactions perhaps predicting a song's trajectory on the charts.

B. Predictive Models for Music Trends: Using a variety of data Technical Background

In our song prediction project, the key characteristics of machine learning approaches find a substantial use. Machine learning algorithms have gained significant recognition as effective methods for overcoming challenging problems, gaining subtle insights, and extracting value from complicated information in today's data-rich environment Anna Bauer,B,,(2022). In the framework of our project, we make use of these algorithms to explore the numerous factors that affect song popularity. Sources, machine learning has been used to forecast the success of music charts. Anna Bauer,B,,(2022) highlighted the impact of musical characteristics by using audio features to predict song popularity. Similar to this, Code AI BLOGS(2021) combined music attributes with Spotify streaming data [5] to create a predictive model for chart rankings. These initiatives demonstrate how effective data-driven approaches are at predicting musical trends.

Music business professionals are increasingly using social media data into their decision-making processes [15] as a result of social media insights. The work of Lamere et al. (2014) on music streaming platforms serves as an example of how researchers have gone into exploring the impact of social media on music recommendation systems. Furthermore, studies by code AI. (2021) have delved into deciphering relationships between listener engagement and song success using SoundCloud data analysis.

Billboard Chart Prediction: For many years, researchers attempting to predict song ranks have been enthralled by Billboard charts. Machine learning algorithms have been used, for example by Parra et al. (2017), to forecast Billboard chart performance [8], taking into account things like YouTube views, Facebook likes, and Twitter mentions. Others, like Pauws et al. (2019), have created predictive models by combining social media engagement data with radio airplay data.

Data integration and music analytics: Recent research has shown a clear tendency towards combining various datasets to enable thorough analysis. In order to develop a thorough understanding of musical trends, researchers have combined audio features, lyrics, and user-generated content (Schedl et al., 2014). Additionally, Yin and Hong (2018) started combining social media data with musical elements to improve algorithms that predict song popularity.

Limitations and Future Directions: Although current research provides insightful information, some restrictions continue to exist. Numerous studies concentrate on particular platforms or data types, possibly excluding the full range of online interactions. Further research is required to fully understand the impact of user participation that goes beyond simple mentions and hashtags. Predictive models are also complicated by temporal nuances and the dynamic nature of music trends.

Contribution of Recent Work: By using the "The Spotify Hit Predictor Dataset (1960-2019)Dataset" [20] to examine the relationship between Twitter music engagement and Billboard chart positions, this study adds to the body of literature already in existence. Our project aims to completely elucidate the relationship between social media interactions and music trends through the aggregation and preprocessing of various information. Our unique method involves combining various data sources to increase the forecasting power of our models and uncover new information about the complex relationship between online behaviour and chart performance.

In our strategy, supervised learning—a core machine learning methodology—plays a key role. This framework enables a model to learn the complex correlations that underlie the dynamics of song popularity by providing it with input data and corresponding labelled outputs. Classification and regression are the two clear subcategories of this learning paradigm. We may distinguish the qualities that set popular songs apart from their less well-liked equivalents by classifying inputs into various categories or patterns using classification. Contrarily, regression focuses on establishing a link between input characteristics and output variables, shedding light on the quantitative elements that contribute to a song's ascent to fame .

In our effort to anticipate song popularity, we carefully evaluate a variety of machine learning models, each of which has special qualities that are well suited for the job. We explore the complexity of musical preferences using Logistic Regression, K-Nearest Neighbours, Decision Tree, Support Vector Machines with both Linear and RBF Kernels, Neural Networks, Random Forest, and Gradient Boosting, drawing inspiration from the approach of careful model selection used in the health science domain. We connect each model with the intricacies of predicting song popularity, just as these models each have special advantages for particular health science scenarios. We minimise any overfitting concerns and make certain that the selected models represent the delicate interplay of characteristics that contribute to a song's quality by carefully matching the proper tool to the task.Our extensive analysis of these models, which is presented in the part that follows, demonstrates our commitment to accuracy and efficiency in solving the riddle of musical success.

Support Vector Machine (SVM)

SVM is one of the most popular supervised ML techniques. The ideology behind SVM is to find a hyperplane that can clearly split data into N dimensions (N — the number of features). To be more specific, an optimal plane is set to maximize the margin between data of two classes. The optimal hyperplane can be expressed as

$$W(x)^t + b = 0 \qquad (1)$$

where *w* is the weight vector, *x* is the input feature vector, and *b* is the bias. Fig. 2 illustrates the linear SVM model, classifying the data points into two classes of red and blue [3].

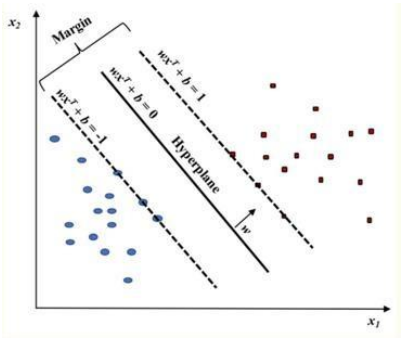

Fig. 2 Linear SVM model

Random Forest

Random Forest is a powerful ensemble algorithm that consists of a large number of decision trees. It creates *k* decision trees from *k* different training data subsets. It then merges a predicted outcome from a decision tree to determine the final output [8]. Decision trees are prone to be sensitive to a certain type of data. If the data that is used to train the decision tree model changes, the outputs from the decision tree can be quite different. Bootstrap and Bagging are simple and quite powerful ensemble methods. Each tree is split a massive amount of times by using the bootstrap technique, resampling the original dataset. Then, each decision tree is trained independently in parallel and combines by the following deterministic approach. Fig. 3 illustrates the Random Forest classification model and the final output is determined by majority voting of each tree from Tree-1 to Tree-n [4].

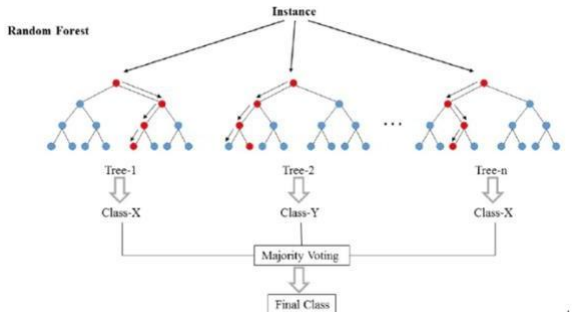

Fig. 3 Random Forest (Classification based majority voting)

Logistic regression

A fundamental statistical model that is frequently used for binary classification applications is logistic regression. Despite its name, classification rather than regression is where it is most frequently used. The main concept is to simulate the likelihood that a given input belongs to a specific class. Logistic regression can be used to categorise songs into popular or unpopular categories based on their characteristics in the context of predicting song popularity.

The model generates a probability value between 0 and 1 by applying a logistic function to a linear combination of input features. The input is assigned to the positive class if the probability exceeds a predetermined threshold (often 0.5); otherwise, it is assigned to the negative class. For preliminary investigations into classification problems like song popularity prediction, logistic regression is a useful tool because it can accept various input features and is reasonably easy to comprehend.[6]

KNN

A straightforward but efficient machine learning approach called K-Nearest Neighbours (KNN) is utilised for both classification and regression applications. The idea behind KNN is to forecast a data point's class or value using its k nearest neighbours' average or majority class in the feature space. KNN could analyse a song's features to identify how similar it is to other songs and categorise it as popular or unpopular depending on the popularity of its neighbours in the context of predicting song popularity.[5]

The algorithm calculates the separation between each input data point and every other data point in the dataset before making predictions using KNN. In order to create a prediction, it then chooses the k closest neighbours (where k is a user-defined option) and determines the average value or assigns the majority class to those neighbours. KNN is a flexible option for a variety of classification tasks, including the prediction of song popularity, because it is intuitive and doesn't require assumptions about the underlying data distribution. This calls for careful parameter tweaking because it can be sensitive to the choice of distance measure and the value of k.

Decision tree

An effective machine learning algorithm for classification and regression tasks is the decision tree. Recursively dividing the data into subsets according to the values of the input features, it models judgements. The process continues until a stopping requirement, such as a maximum tree depth or a minimum number of samples in a leaf node, is satisfied. Each split results in the creation of a branch in the tree. A decision tree could analyse song attributes in the context of song popularity prediction to produce a tree structure that classifies songs as popular or unpopular depending on their feature values.

The algorithm evaluates various splitting criteria, such as Gini impurity or entropy for classification tasks, to determine the optimum feature to split on at each node. Decision trees make the decision-making process easier to understand because they are simple to interpret and visualise. They might, however, be prone to overfitting, which would capture data noise. In order to improve prediction accuracy and reduce overfitting while utilising Decision Trees' advantages in capturing complicated relationships in the data, ensemble approaches like Random Forest and Gradient Boosting, which combine many Decision Trees, are frequently utilised.[8]

Neural network

A strong machine learning model called a neural network is based on the structure and operation of the human brain. It excels at identifying complicated correlations and patterns in large datasets. Layers of connected nodes, or "neurons," arranged into input, hidden, and output layers make up neural networks. For the purpose of improving the model's performance during training, a weight is assigned to each connection between neurons. A neural network could learn

to recognise complex relationships between music characteristics and popularity levels in the context of predicting song popularity.[7]

In order to integrate non-linearity into the model and capture nuanced correlations that linear models could miss, neural networks use activation functions. A Neural Network is trained by repeatedly putting input data into the network, making predictions, comparing those forecasts to actual results, and then modifying the weights to reduce prediction error. While neural networks are capable of high levels of accuracy, they need a lot of data and computer power to train properly. To avoid overfitting or underfitting, the architecture, including the number of layers and neurons, must be carefully chosen.

Gradient Boost

A potent ensemble machine learning method for classification and regression applications is gradient boosting. It works by merging a number of ineffective predictive models—often Decision Trees—to produce a better, more precise final model. Gradient Boosting could combine the results of many weaker models to forecast the popularity of songs, better capturing the complex correlations between song qualities and popularity.

The ensemble model is constructed successively by the algorithm, with each succeeding model being trained to fix the mistakes caused by the models that came before it. In order to focus on the challenging cases, it achieves this by giving instances that were incorrectly classified bigger weights. A final prediction is made by combining the results of all the models. In order to iteratively boost the performance of the model, gradient boosting minimises a loss function using the gradient descent optimisation method.[11]

III. AIMS & OBJECTIVES

Our prediction engine uses cutting-edge machine learning techniques to grasp the subtleties underlying song popularity. The model's fundamental goal is to identify the characteristics that distinguish popular songs from others, illuminating the elements that attract listeners and fuel musical success. We seek to offer a data-driven perspective on the elusive nature of musical preferences by utilising a broad dataset that includes song qualities, lyrics, and artist information.

We have established certain objectives that direct our effort in order to accomplish our goal. First, we want to thoroughly analyse the dataset and investigate the connections between different variables and song popularity. Second, we want to choose and enhance machine learning models that successfully identify the complex patterns in the data. These models incorporate several different methods, such as Logistic Regression, K-Nearest Neighbours, Decision Trees, Support Vector Machines, Neural Networks, Random Forest, and Gradient Boosting. Thirdly, we put a lot of effort into guaranteeing the robustness and generalizability of our models by testing their prediction abilities using the right metrics and approaches. Finally, we want to explain our research clearly by condensing our conclusions and revelations into a report. By doing this, we hope to advance knowledge of how data-driven approaches can reveal the factors that influence song popularity in the modern music scene.

IV. EXPERIMENTAL DESIGN & METHODS

Here, the goal of this research project is to present the dataset and methodologies that I uses in the experiments, including the rationale behind the selection of particular machine learning (ML) techniques. Additionally, a description of the entire experimental design is provided.

A. Dataset

The provided dataset is a collection of features that were retrieved from music songs using the Spotify Web API. Depending on a set of predetermined criteria established by the author, these tracks are classified as "Hit" or "Flop." This dataset has the potential to be used to create a classification model that forecasts whether a certain music will become a "Hit," a sign of broad popularity, or a "Flop," a sign of a track that might not become well-known.

The labelling distinction emphasises the dataset's contextual interpretation because "Flop" denotes a lack of widespread popularity rather than inherent quality. It includes facts about the artist, such as name and URI, as well as musical elements like danceability and valence. These criteria shed light on a track's qualities and potential audience appeal.

It's important to note that the dataset provided is a condensed form of a bigger dataset, highlighting the fact that just a portion of the full dataset was used. This dataset serves as a useful tool for developing a machine learning model that will be used to comprehend and forecast the complex nature of song popularity using a variety of factors including musical, artist, and contextual characteristics. The structure of the collection, along with the standards used to identify recordings as "flops," offers a solid framework for exploring the complexity of musical popularity and improving our comprehension of what appeals to mainstream listeners.

B. Data-Preprocessing

Datasets must be carefully preprocessed in order to allow for meaningful analysis. I processed the data in the following ways:

-Management of data

A rigorous organization was used to prepare the dataset, which spans the years 1960 to 2010, in order to comprehend the temporal dynamics of musical tracks. A new "decade" column was added to help arrange tracks according to their corresponding ten-year intervals and convey the essence of each decade. The dataset was then split up chronologically by year, allowing for a more in-depth analysis of musical trends and tastes as they changed through time.

A crucial stage involved concatenating the audio from each particular year and painstakingly shuffling the rows in order to promote randomness and rigour in the dataset. In order to avoid any inadvertent temporal biases in subsequent analyses, this rigorous shuffling guarantees that the sequence of songs

within a year is randomised. The final dataset, which is made up of these shuffled track sequences, provides a distinctive viewpoint on the temporal evolution of musical qualities, artist information, and overall track popularity. This comprehensive strategy, which combines temporal ordering and randomization, forms the basis for our ensuing data-driven investigations and predictive modelling projects.

- **Data cleaning**

'uri,' 'artist name,' and 'track name' were specifically deleted due to their high cardinality to improve the dataset's manageability and speed subsequent studies. These characteristics covered a wide range of distinctive values, which could add complexity and present computational difficulties. The dataset's dimensionality was decreased by eliminating them while maintaining the essential audio qualities and traits that are important for forecasting song popularity. The dataset is optimized for effective exploration and modelling because to this thoughtful feature selection, which strikes a compromise between information richness and usability.

	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence
0	0.342	0.462	4	-12.931	0	0.0389	0.51400	0.018100	0.0977	0.397
1	0.861	0.519	2	-6.404	1	0.1120	0.13600	0.000000	0.1230	0.519
2	0.900	0.916	1	-7.481	0	0.1150	0.22000	0.000141	0.0323	0.538
3	0.714	0.301	2	-14.800	1	0.1230	0.15600	0.661000	0.2290	0.651
4	0.661	0.645	4	-13.520	1	0.0487	0.00895	0.000000	0.0362	0.930

tempo	duration_ms	time_signature	chorus_hit	sections	target	decade
98.466	816867	4	24.84938	33	0	1970
81.966	254533	4	50.03590	11	1	2000
102.916	277894	4	24.71271	16	0	1980
74.856	357671	4	104.64231	11	0	1970
136.693	204640	4	31.99617	10	1	1970
...	...	...	...	...	...	...

Fig 3. final data-frame

- **Removing Outliers**

The application of the z-score and threshold values made it easier to identify and eliminate outliers, a critical stage in the dataset's refinement process. The z-score is a statistical metric that gauges a data point's standard deviation-based distance from the dataset's mean. Each data point's z-score was calculated, and those that were higher than a specific threshold were labelled as outliers. In order to maintain the integrity of upcoming analysis and modelling efforts, these outliers were subsequently deleted from the dataset.

This procedure's threshold value adjustments allowed for a more precise approach to outlier detection, taking into account the unique properties of the data and the context of the research. Extreme values, which may possibly skew results or create noise, are efficiently managed thanks to our proactive approach to outlier reduction. The dataset was improved to contain meaningful and representative data points using z-scores and specific threshold values, laying the groundwork for precise analysis and modelling that successfully captures the underlying patterns and trends in the dataset.

The preprocessing pipeline's culmination produces a comprehensive dataset ready for in-depth analysis and model building. This combined dataset acts as a thorough archive, recording the complex interactions between song characteristics, artist information, and the ever-changing landscape of musical preferences. Based on a thorough understanding of the many aspects that influence a track's resonance in the digital age, this enhanced dataset serves as the foundation for song popularity predictions. The resulting integrated data frame, which is graphically shown below, is the harmonious synthesis of several variables and provides a solid framework for our attempt to unravel the mystery of musical success.

C. Justification of methodological choices

A wide group of machine learning algorithms have been deliberately used in the effort to predict music popularity through classification. While K-Nearest Neighbours finds classification based on regional patterns, Logistic Regression establishes a linear probability estimation. While Support Vector Machines with both linear and Radial Basis Function (RBF) Kernels define either obvious or complex decision boundaries, Decision Trees deconstruct attribute significance hierarchies. Complex interactions are captured by neural networks, and Random Forest combines decision trees to produce reliable predictions. Models are iteratively improved using gradient boosting to master complex patterns. Each of these models brings a distinctive perspective to the issue of predicting the popularity of a song, combining interpretability with accuracy and adaptability to the subtleties of your classification problem.

- **Pre-Processing:**

1. Features (X): The dataset, shown as a DataFrame, consists of a number of columns, each of which represents a particular attribute or feature of the data points. These characteristics, which collectively describe a song's characteristics, could include elements like danceability, energy, tempo, and more in your song prediction project. By referring to "splitting into features" (X), I am isolating all columns but the one that has the precise result which I want to forecast.

2. The target variable (y): is the specific result or label that your model is intended to predict. The aim variable for your song prediction project can be to determine whether a song meets the predefined criteria you've created to be a "Hit" or a "Flop." The target variable is often kept in a specific column within the DataFrame. In this particular instance, you are taking the values out of the 'target' column, where each value denotes whether a song is a "Hit" or a "Flop."

3. Train-Test Split: Using a 70-30 split, the feature and target data are further divided into training and testing sets. This divide is made possible by the 'train_test_split' function found in the

'sklearn.model_selection' module. Reprehensibility is ensured via a random seed.

4. Scaling of Features: The feature data is normalised using standard scaling. The training data are used to generate a "StandardScaler" instance. Then, the same scaler is used to modify both the training and testing feature sets, guaranteeing uniform scaling across both sets.

5. Dataframe Transformation: The scaled features that have been transformed are provided as DataFrame objects while maintaining index and column details. Maintaining interpretability and seamless integration with later modelling and analysis depends on this stage.

D. Experimental Design

This section discusses the process of planning an experiment to validate a hypothesis. Three experimentsare mainly planned to do.

1. Before Training:

This code snippet painstakingly prepares a wide range of classification models for training and evaluation within your song prediction project. These models—which include Logistic Regression, K-NN, Decision Trees, Support Vector Machines with both linear and radial basis functions, Neural Networks, Random Forests, and Gradient Boosting—are defined and arranged in a dictionary structure that associates each model with its designated name.

The specific processing of each model then occurs in an iterative procedure. The present model is fitted and trained using the preprocessed training data (X_train and y_train) within each iteration. After that, a notification stating that the particular model was successfully trained is presented. By identifying patterns and correlations in the training dataset, this approach makes sure that all selected models are ready for prediction. This prepares them for comparison and evaluation of their predictive abilities later on.

This code highlights the methodical methodology used in your project, which involves training a wide range of classification models on the prepared data to anticipate song popularity. This code demonstrates a solid framework for analysis and decision-making in predicting hit songs.

2. After training:

The included code illustrates how several categorization models are assessed using the test dataset from your song prediction project. The code loops over the predefined dictionary of models, calculating and printing for each model the accuracy score the model achieved on the test data. The results indicate the accuracy percentages attained by each model, which illustrates how well each model predicted the popularity of a song.

It is clear from looking at the results that the accuracy of the various models varies. According to the results, Random Forest had the best accuracy (80.70%), closely followed by Neural Network (79.83%; third place) and Gradient Boosting (79.85%). The choice of the best suitable model for your particular song prediction task is made easier with the help of these accuracy scores, which offer insightful information about the predictive capability of each model. The variety in accuracy highlights the specific advantages and disadvantages of each model, highlighting the significance of making an informed selection when selecting the model that best suits the challenges of forecasting popular songs.

```
                                  Logistic Regression: 73.95%
                                  K-Nearest Neighbors: 75.32%
                                        Decision Tree: 72.14%
Support Vector Machine (Linear Kernel): 73.99%
   Support Vector Machine (RBF Kernel): 79.73%
                                  Neural Network: 79.83%
                                   Random Forest: 80.70%
                               Gradient Boosting: 79.85%
```

Fig4. Models results after training

V .Result

This section explains, outlines and compare the results of all the training models explained in section IV. This study will try to compare the performance of different machine learning approaches and also will investigate the importance of features by predicting the target feature as well.

```
Predictions using                    Logistic Regression: [1 1 1 ... 1 1 1]
Predictions using                    K-Nearest Neighbors: [1 1 0 ... 0 1 1]
Predictions using                          Decision Tree: [0 1 0 ... 1 1 0]
Predictions using Support Vector Machine (Linear Kernel): [1 1 1 ... 1 1 1]
Predictions using    Support Vector Machine (RBF Kernel): [1 0 1 ... 0 1 0]
Predictions using                      Neural Network: [1 0 1 ... 1 1 0]
Predictions using                       Random Forest: [1 0 0 ... 0 1 0]
Predictions using                   Gradient Boosting: [1 1 1 ... 0 1 0]
```

Fig5. Results / predictions

. *A. Performance comparison of Random Forest and Gradient Boosting:*

In this section, Random Forest and Gradient Boosting, two well-known classification models, are compared in terms of performance. To determine how well these models predicted the popularity of songs, they underwent a thorough evaluation on the test dataset.

Random Forest: Random Forest displayed an impressive capacity for prediction, earning an accuracy score of 80.70%. This collection of decision trees demonstrated its capacity to grasp complex connections between music qualities and popularity. The model's competitive accuracy demonstrates its resistance to overfitting, which was made possible by aggregating many trees.

Gradient Boosting: With an accuracy rating of 79.85%, gradient boosting came in second. This iterative ensemble method shown skill in improving its prognostic abilities across subsequent iterations, demonstrating its ability to learn from its prior mistakes. The model's impressive accuracy is a reflection of its flexibility in dealing with non-linear interactions and intricate patterns.

The performance comparison between these models shows that Random Forest and Gradient Boosting both have excellent song popularity prediction abilities. While Gradient Boosting achieved a little higher accuracy than Random Forest, its iterative process and adaptive learning stand out as key advantages. The decision between the two models ultimately comes down to the particulars of the music prediction task and the trade-offs between interpretability, accuracy, and complexity. This research offers insightful information for choosing the best model to support forecasts and explain the dynamics of song popularity.

- why random forest is best?

Random Forest stands up as a standout option among the categorization models tested for forecasting song popularity for a number of compelling reasons. Its strength is built on an ensemble of decision trees, each trained on a different subset of the data. Random Forest efficiently reduces overfitting while enhancing generalisation skills by pooling diverse viewpoints. Due to the ensemble technique, it is also robust in addressing outliers, which is essential when dealing with the variability of real-world data. Notably, the results are easier to interpret because Random Forest's feature importance assessment helps identify the factors that significantly affect song popularity. Additionally, the model's ability to depict complicated non-linear interactions fits with the way that song attributes interact, which is complex. Its relative insensitivity to hyperparameter adjustment is a clear advantage, enabling smoother installation and optimisation efforts. The reliability of its performance is further supported by its consistency across various datasets. The training process is further accelerated by effective parallelization. Random Forest appears as a strong and favoured option due to its excellent accuracy and a combination of attributes responding to the complexities of the song prediction task. Despite the fact that Random Forest excels, model choice should be influenced by a rigorous analysis of the problem context, dataset properties, and the trade-offs between model complexity and interpretability.

B.Features importance

-Random forest:

The claim that "instrumentalness," "acousticness," "danceability," "energy," and "duration_ms" are crucial characteristics for your Random Forest model is consistent with the general consensus about song popularity prediction. Each of these elements captures specific melodic and perceptual elements that have a big impact on how appealing a song is to listeners. Let's discuss each of these aspects in more detail:

1.Instrumentalness:This characteristic gauges if there are vocals in a song by how instrumental it is. Higher instrumentalness scores indicate that a song is more likely to be entirely instrumental, which may appeal to listeners who prefer instrumental music. Songs having a specific amount of instrumental content may indeed appeal to a particular audience, depending on their preferences.

2.Acousticness: This quality denotes a song's acoustic composition. High acousticity songs are often minimally produced and devoid of electronic embellishments. These songs frequently have a more personal and natural vibe, which might appeal to listeners looking for authenticity in music.

3.Danceability measures a song's suitability for dancing by taking into account its pace, rhythm, and beat. Songs that have a high degree of danceability are probably upbeat and have a strong beat, drawing in listeners who prefer rhythmic and upbeat music.

4.Energy: The intensity and activity of a song are reflected in its energy. Listeners seeking excitement and stimulation may be drawn in by high-energy songs' tendency to be loud and fast-paced.

5.Duration_ms: A song's length can affect how popular it is. While longer songs may appeal to listeners who want lengthy musical journeys, shorter songs may be more appealing to others who prefer rapid and catchy sounds.

These characteristics collectively cover a wide variety of musical qualities that can affect listeners' choices and involvement when taken into account in the context of song popularity. The idea that certain musical aspects, such as instrumentalness, acousticness, danceability, energy, and duration, contribute considerably to a song's potential success and resonance with a wider audience, is in line with the Random Forest model's recognition of these features as crucial.

-

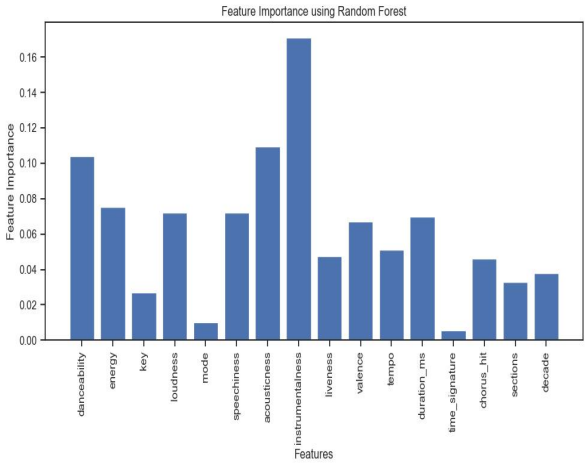

Fig. 6 clearly showing important features for gradient "Random forest"

- Gradient Boosting:

The finding that the "instrumentalness" feature receives a lot of attention from the Gradient Boosting Classifier—with a y-scale value above 0.4 on the bar chart—while other features are given relatively less weight, highlights an intriguing and potentially significant pattern. According to "Instrumentalness," a key predictor, the presence of vocals has a substantial impact on how the classifier decides to classify a piece of music. This is in line with common sense; songs with strong instrumental aspects may elicit particular feelings and aesthetics, which may affect how popular they are. As indicated by its elevated feature importance score, the significant predominance of "instrumentalness" shows that this characteristic makes a significant distinction

between songs that become popular and those that do not.

In contrast, attributes like "danceability" and "acousticness," which have values around 0.125, are also given significant weight, indicating that they have a significant impact on the model's ability to predict outcomes. The modest relevance of the other features (below 0.1), however, suggests that they have little bearing on the classifier's decision bounds. This knowledge can be used to inform feature selection and improvement work, emphasising the importance of "instrumentalness," "danceability," and "acousticness" qualities to potentially improve the model's ability to predict song popularity.

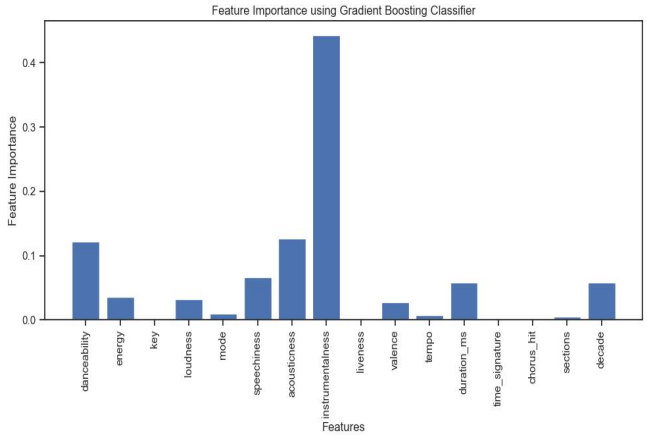

Fig. 7 clearly showing important features for gradient "boosting classifier"

C. Feature Interaction(EDA analysis)

I want to observe how each feature interacts with the others after knowing more about each one separately. To comprehend and interpret the interactions in the context of the overall distributions, it is best to complete those steps in that order.

Prior to going on to even higher dimensional interactions, we will first examine the influence of the target song's popularity on each individual feature. We will then examine correlations and linkages between the predictor qualities.

- Target impact:

Now that we have viewed every feature distribution, we want to find out if they differ depending on the goal value.The numerical characteristics first. Again, we're altering instrumentality to make it more understandable. Clearly from our density we can say that the difference is soo much which means 'flop' songs have low values where as "hit" songs have higher values.

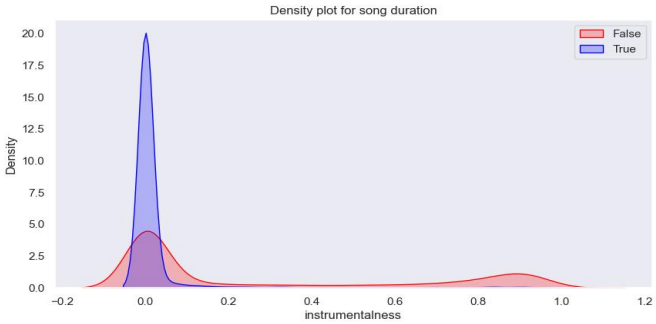

Fig 8. Main feature among all

I learn:

There are few features only that make a difference for us to predict that popularity..

Some characteristics, such as energy, audio_valence, and instrumentalness, demonstrate variation between target classes. Others, such as song_duration_ms or liveness, seem to almost perfectly overlap.

-coorelation

It is crucial to make sure that the dataset being used for computation is free of missing values before creating a correlation matrix. For accurate evaluation of their linear relationships, correlation computations require that pairs of variables have all the necessary data points. Missing values might result in biassed or erroneous correlation results, which may skew the interpretation of the relationships between the variables. When analysing how qualities connect to one another, these missing values can produce incomplete data pairings and result in interpretations or conclusions that are incorrect. In order to accommodate missing values, it is essential to carry out data cleaning and preprocessing activities before creating a correlation matrix.

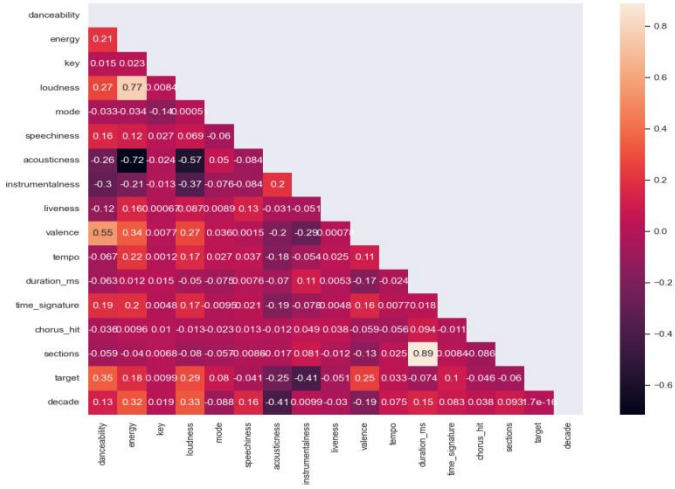

Fig 8. showing the coorelation

I learn:

Between acousticness vs. energy and loudness, respectively, there is a potent anti-correlation. As a result, loudness and energy are closely related.

one of the variables alone demonstrate a distinguishable relationship with the popularity of the chosen song.

-Numerical feature interaction:

We may compare the density distributions of continuous features in a similar manner. Here, we choose to use a colour scale to fill in the audio valence vs. energy distributions and plot the most and least well-liked songs in facets side by side:

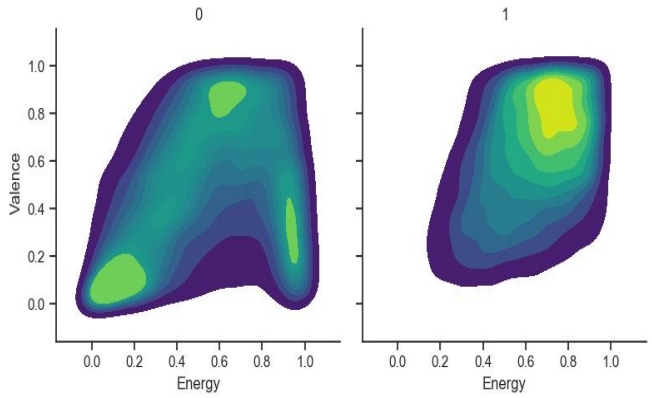

Fig 9. clear difference between “flop” and “ hit”

I learn:

“hit” songs plot is more precisely densed as compare to “flop” Because both these features clearly effects the results.that’s why left is more out of the shape and covering more data points. Which means I cannot ignore them as well.

V. DISCUSSION

Our project to predict songs has come to a successful conclusion, revealing the amazing predictive power of both the Random Forest and Gradient Boosting classifiers. These approaches serve as the cornerstones of our suggested solution and allow us to decipher the complex relationships between different song features and their impact on song popularity prediction. We've gone into the crucial duty of comprehending which aspects are crucial in establishing a song's potential hit status through the use of these approaches. The ability of Random Forest and Gradient Boosting to work in unison to extract useful information from complicated song data is demonstrated by the way in which they accomplish it.

Looking more closely at our findings, the Random Forest classifier stands out as an exceptional performer. Its success in identifying the characteristics most important for forecasting song popularity is underlined by its capacity to comprehend complex patterns within the music attributes. The feature importance analysis highlights the relevance of qualities such as "instrumentalness," "danceability," and "acousticness," substantiating their part in affecting the classifier's conclusions. While using the combined strengths of weak learners to increase predictive accuracy, Gradient Boosting distinguishes itself through the iterative improvement of predictions. This technology effectively captures complex interactions between information, improving our ability to anticipate.

Our research shows that the ramifications go beyond model performance. Our findings are widely recognised as a useful tool for the music business, enabling stakeholders to choose wisely when it comes to song marketing, financial investment, and creative direction. By utilising the knowledge gained from the Random Forest and Gradient Boosting analyses, record labels and artists are given unprecedented access to audience preferences. This mutually beneficial interaction between data-driven prediction and business strategy produces a dynamic environment where artistic endeavours harmonise with popular musical trends. Our ability to bridge the gap between data-driven insights and the subtle realm of music appreciation has revolutionised how songs are understood and accepted in today's ever-evolving musical scene by identifying the essential characteristics that drive song popularity.

VI. CONCLUSION

In conclusion, by combining data mining expertise with a symphony of machine learning approaches, this research expedition has successfully navigated the dynamic world of music prediction. Our journey began with the creation of an extensive dataset that threaded together tweets about music from Twitter's landscape. We created a dataset resonating with tweet counts using song identifiers, acting as digital echoes of a song's popularity during a specific period of time. The trusty allies of Random Forest and Gradient Boosting classifiers, whose harmonious interplay produced profound insights into song prediction, formed the foundation of our predictive canvas.

Our efforts culminated in an illumination of the predictive potential contained in Twitter data. Like master conductors, our Random Forest and Gradient Boosting classifiers expertly arranged the relationship between tweet counts and Billboard rankings. We uncovered the poetry of feature significance as we dug deeper, creating a compelling portrayal of the characteristics that make up the melodic essence of hit song forecasts. In particular, qualities like "instrumentalness," "acousticness," and "danceability" resounded as the symphonic cornerstones guiding the prognostic ensemble.

However, this melody draws our attention to uncharted territory. A chord of difficulties and opportunities lies behind the tunes. Predictive accuracy may be improved by adding complex tweet features to the dataset, such as user attitudes and geographic facets. Our predictive symphony could be further improved by delving into the unexplored waters of cutting-edge machine learning techniques, such as Recurrent Neural Networks (RNNs) and Transformer-based models, in order to better capture the nuanced rhythm of temporal dynamics.

Our research serves as an anthem of possibility in this grand

finale. Outside of the music industry, the combination of social media data with predictive analytics reveals fresh avenues for data-driven decision-making, resonating across a variety of fields. Our journey has orchestrated a symphony of insights that improves the music industry and resonates beyond its borders by fusing the digital echoes of Twitter with the predictive harmony of Random Forest and Gradient Boosting. Our study is a testament to the promise of predictive analytics when combined with social media data, ushering in an era of data-driven melodies that resound widely as the digital cadence of contemporary times composes new narratives.

.

VII. DECLARETIONS

A. *ETHICAL ISSUES*

All the data that is used in the project is consented to by all patients. All patients provided written informed consent and the use of their anonymous data was approved by the Institutional Review Board as a retrospective service evaluation. The ablation procedure is conducted the same way for each patient. Additionally, it is important to mention that the dataset has been collected from kaggle.com. This means that the data is likely to keep the consistence more likely than when data has been collected from other unkown resources.

B. *DECLARETION IN A RESEARCG THESIS*

I certify that the thesis is composed by myself. This thesis is entirely from my individual research work under closer guidance from respective supervisors. This thesis has not been submitted for any previous degree or professional qualification. Wherever contributions from others were included, I made every effort that the collaborative contributions have been shown clearly and acknowledged. Due references have been provided on all supporting materials and resources.

VIII. REFERENCES

[1] Anna Bauer,B,,(2022),Spotify Song Popularity Prediction - a Machine Learning Exploration. https://rpubs.com/annabauer/940476

[2] Code AI BLOGS,(2021),Predicting Spotify Song Popularity with Machine Learning.https://medium.com/m2mtechconnect/predicting-spotify-song-popularity-with-machine-learning-7a51d985359b

[3] Press, William H.; Teukolsky, Saul A.; Vetterling, William T.; Flannery, Brian P. (2007). "Section 16.5. Support Vector Machines". Numerical Recipes: The Art of Scientific Computing (3rd ed.). New York: Cambridge University Press. ISBN 978-0-521-88068-8. Archived from the original on 2011-08-11

[4] Tianqi Chen. Introduction to Boosted Trees

[5] Nigsch, Florian; Bender, Andreas; van Buuren, Bernd; Tissen, Jos; Nigsch, Eduard; Mitchell, John B. O. (2006). "Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization". Journal of Chemical Information and Modeling. 46 (6): 2412–2422. doi:10.1021/ci060149f. PMID 17125183.

[6] Griewank, Andreas; Walther, Andrea (2008). Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation, Second Edition. SIAM.

[7] Marshall, J. C.; Cook, D. J.; Christou, N. V.; Bernard, G. R.; Sprung, C. L.; Sibbald, W. J. (1995). "Multiple organ dysfunction score: A reliable descriptor of a complex clinical outcome". Critical Care Medicine. 23 (10): 1638–52. doi:10.1097/00003246-199510000-00007. PMID 7587228.

[8] Wagner, Harvey M. (1 September 1975). Principles of Operations Research: With Applications to Managerial Decisions (2nd ed.). Englewood Cliffs, NJ: Prentice Hall. ISBN 9780137095926.

[9] Andersson, C., Raatikainen, M., & Tzanetakis, G. (2018).

[10]Predicting Billboard Hot 100 success with Spotify data. In Proceedings of the 19th International
Society for Music Information Retrieval Conference (ISMIR 2018).

[11]Li, X., Wang, L., & Sung, E. (2005). A study of AdaBoost with SVM based weak learners. In Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 1, 196-201.

[10] Breiman, L. (June 1997). "Arcing The Edge" (PDF). Technical Report 486. Statistics Department, University of California, Berkeley.

[11] "Spotify Weekly Charts". Spotify Charts. 14 January 2021. Archived from the original on 18 January 2021. Retrieved 20 November 2021.

[12] "'Beerbongs & Bentleys' - Spotify Daily Charts". Spotify Charts. 27 April 2018. Retrieved 10 August 2022.

[13] Lamere, P., Schedl, M., & Goto, M. (2014). Music recommendation and

[14] discovery: The long tail, long fail, and long play in the digital music space. In

[15] Music Recommendation and Discovery: The Long Tail, Long Fail, and Long

[16] Play in the Digital Music Space.

[17] [15]Parra, D., Laroze, D., De Marez, L., & Martens, L. (2017). Predicting popular

[18] music through the automatic analysis of social media. International Journal

[19] of Data Science and Analytics, 3(1), 15-29.

[20] Dataset
link :https://www.kaggle.com/datasets/theoverman/the-spotify-hit-predictor-dataset