International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

# Identifying Affective Features of Music Tracks to Determine their Popularity using Machine Learning Approach

**Poonam Saini *[1,2], Priyanka Shaktawat [1]**

**Submitted: 10/03/2024 Accepted: 25/04/2024 Accepted: 02/05/2024**

***Abstract:*** Our world of choices for buying products has often been influenced by our friends, peer groups and now the role of technology in building choices cannot be denied. The songs are an integral part of our life and the choice of songs has very often been influenced by our mood, the song's digital presence, its lyrics, singer, the band and many more appropriate attributes. The present work is a study of a very popular online streaming app i.e. Spotify that has a strong base of music content and is popular among all the age groups alike. The songs have a set of attributes like accousticness, danceability, energy, instrumentalness and many more that impinge the listener's mind. The collective set of these features when subjected to the machine learning based techniques bring out the best of the best features out of the songs in the database and that creates a popularity chart. The present work depicts the comparative analysis of the various algorithms for the identification of the popularity. The Random Forest based algorithm shows an accuracy of 80.41%, Logistic Regression with 80.15%, KNN with 77.54% and Decision Tree Classifier shows an accuracy of 68.92%.

***Keywords: Keywords: A****ccousticness, Speechiness, Valence, Loudness, Tempo, Machine Learning, Logistic Regression, Decision Tree, KNN, Random Forest, Spotify App*

## 1. Introduction

Many times, it becomes imperative to make a choice while buying some products through online stores having plenty of items on display, of similar nature, the alternative of the same product that we intend to buy might not be even known to us. The users generally rely on their friends for recommendations of such items and are transpired to them either by word of mouth, recommendation letters, movie and book reviews etc. The recommender systems systematize this process of assistance provided to the users. In some cases, the users give inputs to the system and then the system aggregates it and sends it to the appropriate customer. On the other hand, the system might even match the recommender and the ones in need of the recommendations. With the ever wide spread usage of web entwined with social networks, the e-commerce business has been forthcoming in utilizing this opportunity in connecting with users, approaching them for feedbacks and also tracking their movements on the web, getting feedback indirectly and discretely. Therefore, the feedback turns out to be either implicit or explicit.

There are two important terms in any recommender system, one is the user who gets the advice or recommendation for the products being recommended and the other is the item. The previous interactions of the users with the item often affects the future decisions related to the purchase or selection of choices for the products.

[1] *B. N. University, Udaipur, Rajasthan, India*
[2] *Sir Padampat Singhania University, Udaipur, Rajasthan, India*
*ORCID ID : 0009-0002-6195-9875*
** Email id: Poonam.saini9@gmail.com*

The working of the recommender principle involves dependencies between the user and the item related choices. In general, if talk of movies, if someone is interested in the 1857 revolt of India, it is very likely that the person might also be interested in the related works, but it is very likely that the same person might not be interested in the romantic movies. The degree of correlation between the items may also affect the kind of choices that a person is going to make in the future. Through the research survey, it has been observed that the majority of the recommender system problems are centred around: The matrix-based problems involving m-users and n-items and determining top-k items also called top-k recommender system.

The ultimate purpose of the recommender system is to enhance the sales of the ecommerce companies and accounting to the operational and technical goals. According to [1], the operational and technical goals of a recommender system can be divided into relevance, novelty, serendipity and recommendation diversity.

In the online world, recommender systems have turned out to be a major force to reckon with, especially in the areas of social networking, e-commerce, news recommendation, discovering new hotels etc. The ultimate goal of any recommender system is to recommend the items or render an informative service most relevant to the user's question out of the million choices that need to be made related to the buying of a product from e-commerce website. The sole purpose of any recommender system is to see that the old recommendations are the ones that are least repeated else this may lead to irritated environment and thereby decrease in the overall sales process or even the user might leave the portal for the products elsewhere. There should be an element of surprise for the user, thereby making him filled with awe and surprise. The diversity of items helps in ensuring that the user who needs choices does not get the repetition of the

recommendation of the items and most importantly the recommended items are relevant to the query of the end user.

There is a wide variety of examples of recommender systems which have affected the users to a great extent. The Group Lens is one of the very old systems for recommending articles from Usenet News; the work of Amazon.com involves primarily is to recommend in the area of commercial products. The buyers provide star ratings from 1 to 5 stars. The user behaviour on the web portals is captured through explicit and implicit ratings. Netflix started its journey as a digital video disc (DVD) rental company and then diversified into a streaming business. The users attached with Netflix can recommend on a five-point scale, the user actions of watching the streaming content is recorded for further recommendation. Netflix also provides recommendations for the recommended items, thereby improving the customer loyalty and retention. On the other hand, Google News Personalization system is able to recommend news on the basis of the history of clicks performed by the users. This is an implicit rating system.

The music streaming sites like Spotify, iTunes, Yahoo Music, gaana.com etc have a wide customer base involved in actively listening to the music charts, helping the companies in improving the top 50 or top 100 music lists. Spotify is one of the largest music streaming services in the world. The company started in the year 2006 at Stockholm and has traversed almost every nook and corner of the world. Fig. 1 shows the growth of the userbase of the company from 2015 to 2022[2] and is predicted to be even increasing in the year 2023. There is a continuous upward trend.

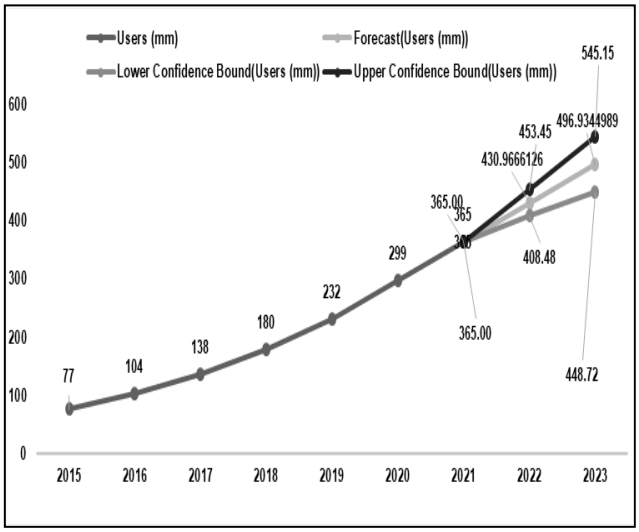


**Fig 1.** The ever-increasing Spotify's User base

The popularity of the Spotify streaming app could be judged by the demographics. The data in the Fig. 2 shows that majority of the users are in the age group of 18 to 34[2], but it has a good presence of 19 % even with the age groups> 50 years, which shows that Spotify has its presence in all the age groups.

The rise in revenue per year as shown through Fig. 3 shows a continuous growth pattern. Spotify had 422 million unique users in Q1 2022, who either use the platform for free with ads or subscribe for ad-free access [2].

The rise in revenue from 2.94 billion Euros in 2016 to 9.66 billion Euros in the year 2021. It has been regressed that the revenue shall be around 10.94 billion Euros in the year 2022 and expected to be 11.7 billion Euros in the year 2023. Spotify increased its annual revenue by 22% in 2021 to €9.66 billion. It has tripled its revenue in the past five years.

Women make up 56% of Spotify's usage in 2021, as shown in the Fig. 4. According to the statistics, there are 422 million people who use Spotify at least once a month, there are 182 million subscribers, 70 million songs and 2.9 million podcasts with Spotify. Its profit earning company.

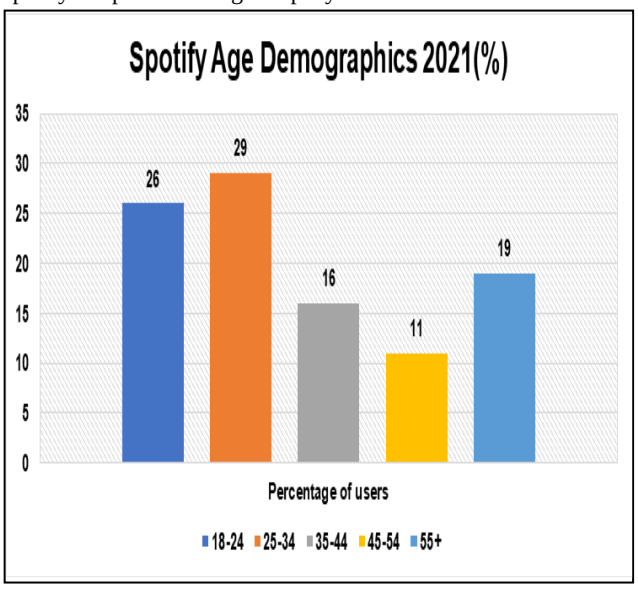


**Fig 2.** Spotify's age demographics

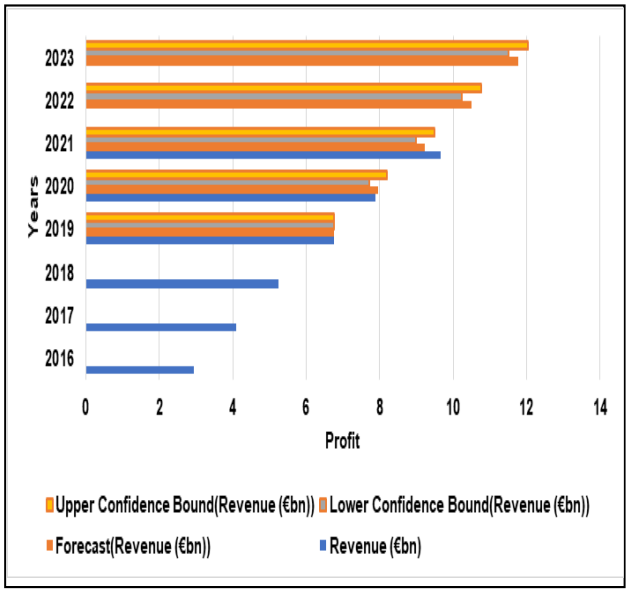


**Fig 3.** Rise in revenue(Q1-2022) for Spotify App

Music has been one of the pastimes for generations together, various artists had a roller coaster ride in the popularity charts, sliding up and down the popularity chart year by year. The songs have nodes and antinodes, their lyrics keep the listeners active and agile during hay times and the sad tunes match with the condition of their hearts. Through this research work, the present work tries to understand the nuances of the songs, their characteristic features, that make them rise and fall on the popularity charts, make or mar the career of the artists, producers directors and financiers. In this research work, Spotify data collected from Kaggle, as a point of study and studied the impact of features on the popularity of the song in the song chart. In section 2, we shall carry out a survey to fathom out the work that has been strictly carried using the Spotify App and the related data. In Section 3, we shall discuss the methodology adopted to

carry out the research work, the Section 4 deals with the result and discussion and finally Section 5 deals with the Conclusion and Future scope.

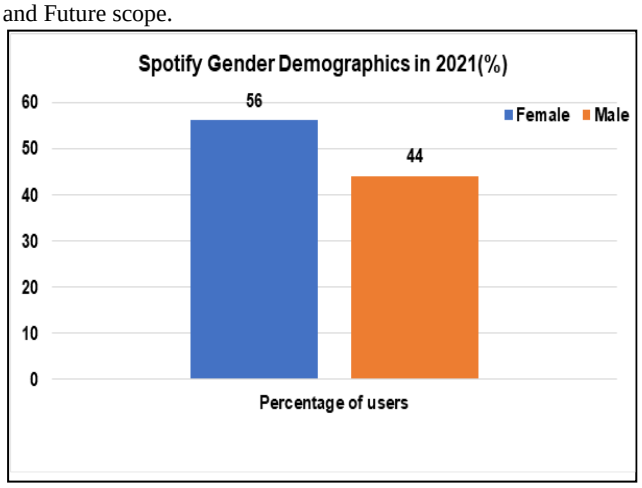


**Fig 4.** Gender demographics users in Spotify App

## 2. Related Work

Spotify app came into existence in the year 2006 in Sweden. It is one of the popular music streaming app around the world, equally popular amongst youngsters and oldies (almost 34%). Fig. 5 shows the number of papers published in reputed journals and surveyed for this research work.

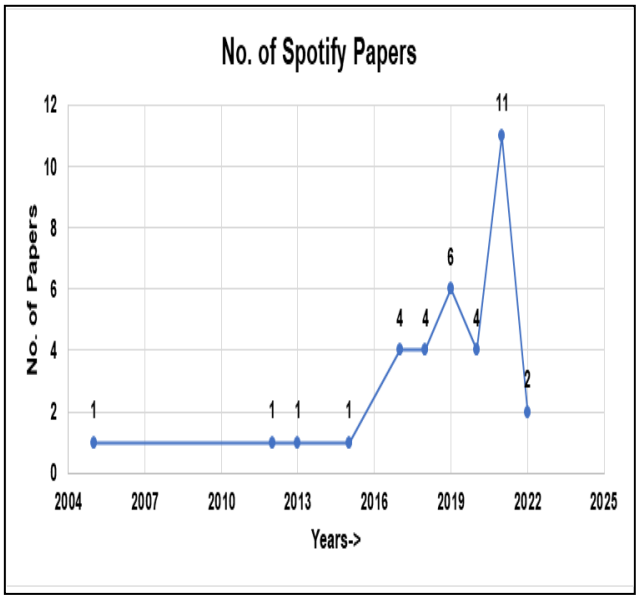


**Fig 5.** Number of Spotify based papers published from 2002 to 2022

Internet and especially the sites like Kaggle are a source of data for the researchers. The Spotify data adopted by [1] has CSV (comma Separated vales) format and the researchers have performed in-depth regression analysis and content visualization to understand the features of data collected along with its characteristic audio features. This work also analyses the most popular and danceable songs that makes it to the top of the music charts in the world.

According to [2], the main purpose of the creation of Spotify is to provide the users with music as per the user requirement on a global scale. Music related features such as name of the artist, genre, album etc. are also made available to the users. The authors have used collaborative filtering technique over MovieLens100k, MovieLens1M and Jester based datasets. By dividing the experimental study into two phases: user-based and item-based techniques, it was observed that the ITR(Improved triangle similarity complemented with user rating preferences) and IPWR(Improved PCC weighted with RPB(rating preference behavior)) are the most suitable similarity measures for a user-based RS while AMI is the best choice for an item-based RS. The main purpose of creation of Spotify is to provide the users with music as per the user requirement on a global scale. The music related features such as name of the artist, genre, album etc. are also made available to the users. The authors[3] provide an overview of the shared features of the Mexico's Top 5o in the year 2019, also the features have been analyzed and results have been compared with global top 50 song chart. The happiness feature has been the most common emotion in Mexico Top50.

The characteristic audio features play an important role in helping to analyze home automations and surveillance systems, speech recognition and music information retrieval etc. "pyAudioAnalysis", is an open-source Python library[4] under Apache License and is available at GitHub. This library helps in various fields including audio event detection, emotion recognition, and content-based movie recommendations. According to the authors, Spotify, a popular streaming app provides its users with millions of tracks enabled through mobile devices, laptops etc. It would be really excruciating for the users to choose the songs from a set of million songs from different singers and genres. Therefore, there is a need for a personalized recommender systems. The Principle Component Analysis (PCA) has been employed as a means of understanding the differences between the different playlists which were utilized to build an SVD based recommender system[5].

The author[6,7] has observed Spotify as the world's most active streaming app, which collects information about users as well and creates a surveillance environment. The app collects information about users and producers and arranges the unstructured information into structured easy to use mood-based playlists. The various music features like danceability, valence, energy, acousticness etc. have been detailed out with their quantitative values.

Some of the authors like[8] have critically examined the Spotify based streaming dataset which it serves to millions of listeners. The authors have deep dived into the connection between the personality traits and music listening behavior. The music has more than 211 characteristic features which could be tweaked to see the desired results.

The authors [9] have expressed that the popularity of the piece of music can be predicted even before it is released on the basis of the features like loudness, energy, acousticness, etc. The work performs popularity analysis using a machine learning approach.

There has been carried out an extensive research on the song lyrics, the kind of words used in the top bracket of the music charts on Spotify. The technique that was followed is semantic analysis based on their meaning. The main point that comes out is the frequent use of Human noun, action verb, mental verb, stative and personal opinion adjective in the popular songs especially in the top five of the music charts[10].

Since the Spotify app has a mass base in the US with listener's age ranging from 18-75 years, with regular activity for 30 days, the work [11] has investigated the likes between human personality and musical listening behaviour. The major outcome of the work is as follows:

- The individuals with neurotic tendency were more

choosy in listening. The emotional standing is negatively correlated with average skip rate [r=-0.71] 95%CI[-0.97,-0.45]

- Openness is correlated positively with all-time track discovery rate[r=0.152

A good number of papers have utilized the machine learning techniques such as RF,KNN and SVM for the identification of popularity of the song using the song metrics. Lyrics have also been one the key parameters of studies taken to understand the streaming media. Music service has formative and cultivating dimensions resulting in the evolution of music Bildung[11].

According to the authors, there are certain observations which could lead to the success of a song or an artist on the popularity chart which is (i) Song's Streaming volume (ii) observation on the top50 playlist (iii) observations on cross country differences in the playlists (iv) usage of instrumentalness variable in explaining the cross country music [12].

The music streaming companies need to have expert reviews to sustain in the market, their reviews have a deep impact over the choices that the people make to listen and this could also impact the popularity charts of the artists ranking.

Pitchfork is an American music publisher launched in 1995, having extensive coverage of indie-rock-music."P4KxSpotify"[13] is the dataset having album reviews and corresponding Spotify audio features. This dataset has 18403 records in 18 columns. The reviews are scaled on a scale of 0.00 to 100. The dataset can be used for investigation of bias in the reviews and their effects on the design of recommender systems based on skewed data reviews.

When we compare the statistics of Spotify app, a music streaming app, having global presence, with a user base of 100 million with a considerable number of 60 million paid users. Every music streaming platform finds unevenness in streaming behaviour due to internet infrastructure of the geographical location, the rules governing the music content streaming. Spotify database could be used to study sociotechnical and economic process associated with digital media distribution[14].

The Spotify app has two APIs (i) Analytics API (ii) Public API. The ISRC attribute (International Standard for Record Code for Identification) can be used for different listening behaviour of male and female users. The daily listening pattern can also be observed. It has been observed that the mobile users used their own playlists while the people using the desktops used other playlists[15].

Armada is one of the biggest independent dance music labels in the world. In order to automate the process of digital music distribution, there is a need to continuously capture the songs and analyse the web-based data. Spotify has developed web API with several programming languages including R[16].

Hit Science Song$^{TM}$(HSS) is a term that has been used since 2005 to determine the unique set of features, the term HSS was given by Mike McCready. The Hit Song Science aims at predicting whether a song will be a hit before its distribution, by analyzing its audio features through machine learning algorithms17].

The Music Information Retrieval databases and HSS have been able to predict hit songs. The LASSO and XGBOOST have been cited as one of the best approaches for classification models. According to the authors, in order to predict the music chart, the song writer and producers must have some 100 published credits, the individual must be credited with either writer or producer or both, the individual must have at least three number 1 hits on the Billboard 100[18].

The acoustic and lyrical features of the songs have unique place in the process of identification of success of songs on the music charts. In the case of acoustic feature-based techniques, the songs are converted into vectors, we can employ unsupervised techniques. In the lyrical analysis, the transcription of the songs is collected from the internet and then each song is converted in to a vector and then Probabilistic Latent Semantic analysis can be performed[19].

A good number of papers are available which undertake to understand the mechanism involving the individual and social music preference from experimental psychology viewpoint. The studies show that there is deep impact of repeated exposure on liking the songs. The liking is also affected by context, type of music or listening conditions. This "mere exposure effect" is akin to the familiarity principle. The Wundt curve as shown in the Fig. 6 describes the arousal due to mix of repetition and surprise[17].

As we know that there is a large variety of recommender systems for the users, but knowledge about one recommender systems may not be suffice to judge all other in the basket. The black box nature of the recommender system brings a feeling of untrust on the users and provides them with minimum control. The users are given the options of Radar charts and sliders to control the Spotify recommendations. The participants with high musical sophistication interacted significantly more with the radar chart in comparison to the sliders, it has been observed that the users got more choices for new songs, when they are able to manipulate more the musical attributes[18].

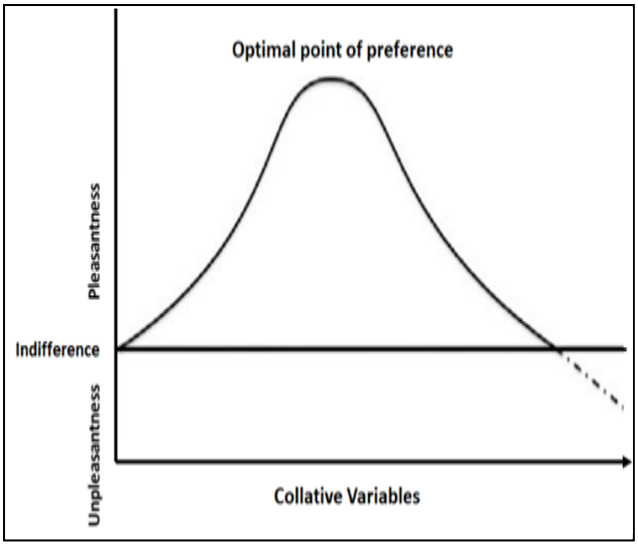


**Fig 6.** Exposure Effect

Spotify has provided the users with tweaking features like related artists, discover, and browse, organize and represent gender for getting proper recommendations for the songs from a vast store of music collections. Spotify has a complex algorithmic base to control its functioning, it has a visual interface of photos, text, clickable links, and graphics to lure the users and remain loyal customers[19].

In a recommender system based on music streaming, it can be predicted that whether the user will listen to the complete song or apply to the skip button depends upon the previous actions during the listening session along with the acoustic features of the songs. The authors have studied the Spotify Sequential Skip prediction data set, which provides in-depth information for performing the analysis, has been used for creation of sequential-based models like GBTs, LSTMs, and Bi-LSTMS[20,21].

Spotify's tagline is" Music for every Mood". Researchers have performed SWOT analysis that shows how the Spotify's brand

culture affects its customer engagement in the US online market. It has been identified that brand culture affects customer engagement. Despite worldwide presence, the Spotify has not been able to make market in the Chinese territory due to different cultures and traditions[22,23].

In order to further promote the presence of Spotify app amongst the users, the company launched Spotify Wrapped in the 2020 campaign, the major effect was that the app downloads increased by 21%. Spotify allows its users to support their favourite singers and hear new songs as well. The music apps generally keep track of the ratio of DAU (Daily Active Users) to MAU (Monthly Active Users). It is an engagement metric that measures the number of days in each month that users performed an activity that qualifies them as active users. This ratio is very important for any music streaming player to judge its userbase[24,25]. The spotify has three business models namely: Network Effects Business Model and Network Orchestrator Business Model, Freemium Business Model and Unlimited Subscription Business Model. Here the term Network effects refer to the positive value and the benefit of each additional user, who in this case may be an artist (content creator), or a subscriber (content consumer). The add value to the company[26]. The freemium business model is typically considered as a revenue model[27,28,29] because, in this case, the model defines how Spotify's revenues are generated from its user base. The third model that is the Unlimited Subscription Business model involves two main factors: (i) subscription to Spotify's premium service and (ii) unlimited access to the service.

The generic competitive strategy followed by Spotify is to have a low-cost position that creates strategic advantage for the music streaming business in terms of making the service's price attractive in the international market. This builds the international online userbase for the company. Spotify's main intensive growth strategies are: Market Development and Market Penetration. These two strategies are simultaneously applied in order to strengthen the company's competitive position as the biggest and leading music streaming business in the global market[30,31,32]

The users create a playlist as a way of cultivating affect. This involves producing, capturing and exploring moods and emotions. There is a relationship between technology, affect and genre. Genres help us in navigating the content on the music app (Spotify in Question) and affect helps in understanding the user experience[33,34]

Through this literature survey, we have covered every minute piece of information from the year 2005 to 2022.

## 3. Proposed Approach

The entire work carried out has the following steps:

- Identification of Data sources and final Data Collection, Preprocessing and Data Refinement process
- Understanding and extracting the essential features of the dataset collected , identifying the correlation between features
- Application of Machine Learning based techniques to find the accuracy of results for popularity.

The main objective of the present work is to understand the features finely ingrained in the songs being streamed by the online apps like Spotify which have a large song base for the listeners and is being modified every now and then. The music streaming data is available from many sources including Kaggle, Nokia DB, and million-song-dataset etc.

### 3.1 Data Description and Data Preprocessing Procedures

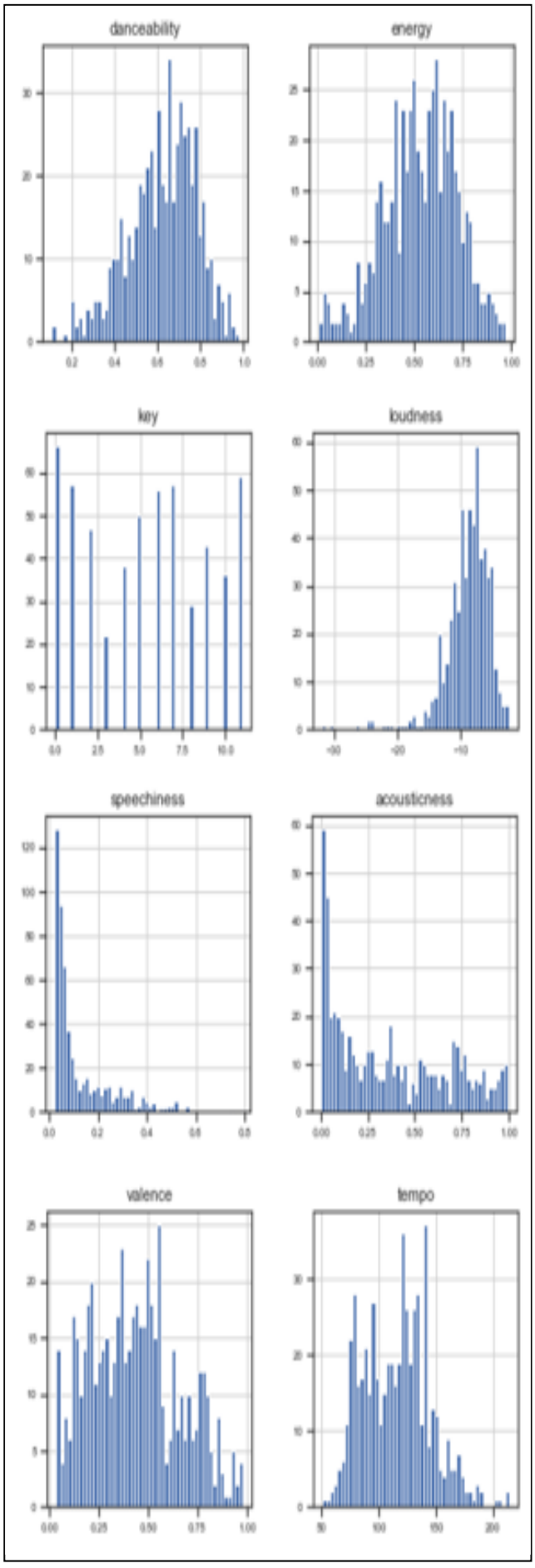


**Fig 7.** Pictorial representation of a few dataset features

The dataset that has been chosen for the present work has 1,70,654 records and 19 attributes. The dataset has been collected from Kaggle and was extracted from Spotify API. This data needs to be considered for data pre-processing which includes splitting of the data set in Training and Validation sets, taking care of missing values and categorical features and Normalization of data set and further prediction operations. This data from the Kaggle consists of several of Spotify's proprietary musical features derived from various attributes of an individual song some of which include:

- Acousticness: The term acousticness refers here to a value representing the probability that a track was created using an acoustic instrument, including voice. The value for this variable is a float value that lies in the range of 0 to 1.
- Danceability: Here danceabilty refers to the song's foot-tapping quality, based on tempo, rhythm stability, beat strength, and isochrony. It has a float range between 0 to 1 and a value of 0.0 means least danceable and a value of 1.0 means the song is most danceable.
- Duration: The term duration here means the duration of a track in seconds as calculated by the Spotify analyzer.
- Energy: Energy is a representative of a perceptual estimation of frenetic activity throughout the duration of the song. The high-energy songs have increased entropy, and tend to feel fast, loud, and noisy (e.g., Death Metal). Energy parameter has a float value in the range of 0 to 1.
- Instrumentalness: It has a Value representing the probability that a track was created using only instrumental sounds, as opposed to speech and/or singing. It has a float in the range of 0 to 1.
- Liveness: It has a value which represents the probability that a track was recorded in the presence of an audience rather than in a studio. It has a float value in the range of 0 to 1.
- Loudness: The average loudness of a track is measured in decibels(dB). Loudness is the psychological correlate of signal amplitude.
- Speechiness: It represents a float value, which represents the presence of spoken words in a track, e.g., talk show, audio book, poetry, rap. Good songs a have a value in the range of 0.33 to 0.66 which means a good mix of music and words.
- Tempo: The estimated tempo of a track in beats per minute (BPM). Tempo has a float value in the range 0 to 250.
- Valence: The term valence means music's positiveness conveyed by the track. The track with a high value of valence is sound positive and expresses a feeling of happiness, cheerfulness and being euphoric. On the other hand, a low value of valence means a feeling of sadness, depression and even angriness to a certain extent. Valence has a float value in the range of 0 to 1.
- Mode : The term mode represents the modality (major or minor) in a track, the type of the scale for which its melodic content is derived. It has a value in the range of 0 to 1. Here 1 is for major and 0 is for minor.
- duration_ms: It generates an integer value which tells us the duration of the track in milliseconds.
- Key: The term key refers to the key the track is in. It maps the pitches using the standard pitch class notation.
- Explict: has a value of 0 or 1 indicating the absence or presence profane words that must not be herd by children of tender age.
- Id - The Spotify ID for the track.
- Type - The object type: "audio_features"
- Popularity: The popularity of the track. The value will be between 0 and 100, with 100 being the most popular. The popularity is calculated by algorithm and is based, in the most part, on the total number of plays the track has had and how recent those plays are. Generally speaking, songs that are being played a lot now will have a higher popularity than songs that were played a lot in the past. Artist and album popularity is derived mathematically from track popularity. Note that the popularity value may lag actual popularity by a few days: the value is not updated in real time.
- time_signature: An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).
- Year: It is the year of release of the track.

The dataset collected with various features and their corresponding values for this research work from []. The features contained in the dataset are represented through the Fig. 7.

Here, Fig. 8(a) shows the actual mean position of the top ten tracks in the dataset collected from Kaggle. The Fig 8(b) shows the popularity distribution and is clear that there is a rising and falling hump of the distribution values.

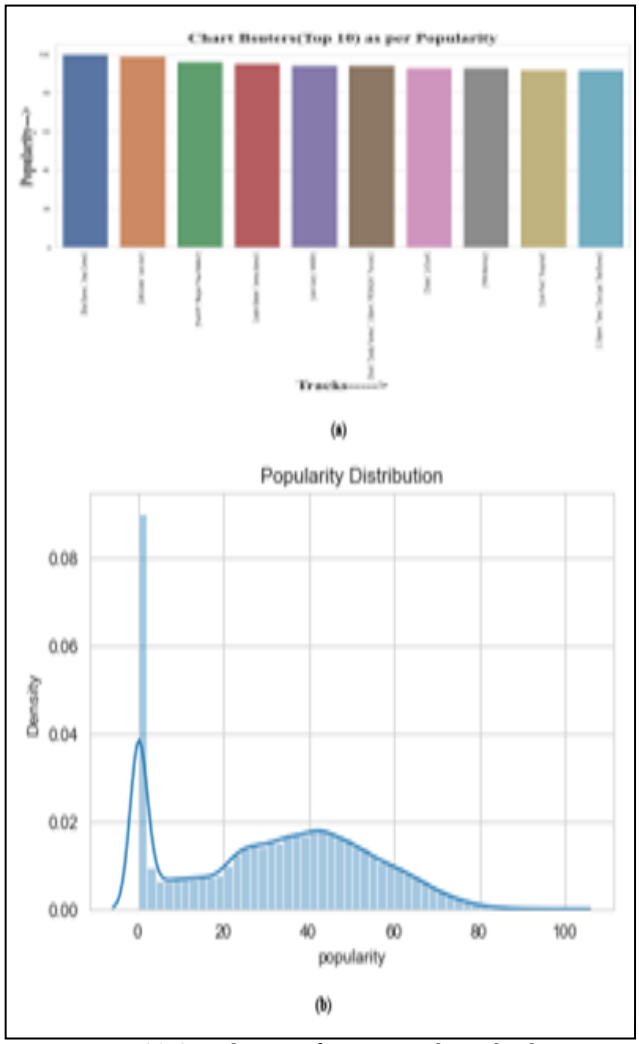


**Fig 8.** (a) Actual mean of top ten tracks in the dataset (b) Popularity distribution of the tracks

Since all the features ingrained in the songs have different scales of measurement and different scales of creation. Loudness is a variable measured in decibels(dB) has a range of -60 to 0 and on the other hand accousticness and valence are measured in the range 0 to 1, this gives rise to the need for normalization of the dataset collected from Kaggle. The normalization process involves computing the difference between the unnormalized data and the minimum value which is then divided by the difference of the maximum value minus the minimum value. In order to compute the importance of each of the features, chi$^2$ test has been used to verify the relevance of the feature and its p-value. The p-value is to establish the interference of the feature with the final result. A low value definitely influences the results and in our present case we have set a limit of 0.05 i.e. 95% confidence level and we eliminate every other feature with p-value higher than 0.05. A Chi-Square computes how expected count **E** and

observed count **O** deviate from each other when c is the degree of freedom and is represented by Eq. (1).

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

Where C=degrees of freedom, O=Observed Values, E=Expected Values.

**Table 1.:** Computation of score and P-values for the feature dataset

| | features | score | pvalues |
|---|---|---|---|
| 0 | popularity | 6.026597 | 1.409188e-02 |
| 1 | key | 8.869689 | 2.899446e-03 |
| 2 | time_signature | 0.665869 | 4.144955e-01 |
| 3 | danceability | 2.654319 | 1.032690e-01 |
| 4 | energy | 32.686717 | 1.082746e-08 |
| 5 | speechiness | 5.301236 | 2.131030e-02 |
| 6 | acousticness | 94.320115 | 2.684088e-22 |
| 7 | instrumentalness | 10.525997 | 1.177069e-03 |
| 8 | liveness | 2.637487 | 1.043685e-01 |
| 9 | valence | 14.588911 | 1.336990e-04 |
| 10 | tempo | 63.301818 | 1.773406e-15 |

The score values calculated using (1) shows that the acousticness tops the score value and time_signature is having a minimum score in the feature set. It is clear from Table 1 that the features like acouticness, energy and tempo shall have a lot to say about the song as compared to the other features.

Figure 9 shows the correlation between the features that make a song more popular than the others. The energy parameter is probably playing a role in influencing a song's popularity with a 0.5 correlation ratio. It may happen that popular songs are energetic but may not necessarily be a good song to dance with. As the dancing correlation is not that high. However, if a song has low energy, that will not mean failure in popularity chart. There is another observation from the above tables that the correlation value for acousticness is the least in comparison to popularity, with a score of -0.59, it helps in understanding that popular songs have a component of electric instruments being played or may be remixes. From the literature survey it is clear that the present chart busters are prepared with latest digital know-how and not from pure orchestra. Utilizing the correlation values, we can also note a few other observations about attributes like Loudness and energy are highly correlated(0.78), acousticness is negatively correlated with energy(-0.74), loudness(-0.56). The correlation plot value between the features energy and acousticness have a highly-correlated inverse relationship, which brings out a fact that the more a song skews towards being acoustic, the less energetic it tends to be and valence and danceability (0.55) are highly correlated.

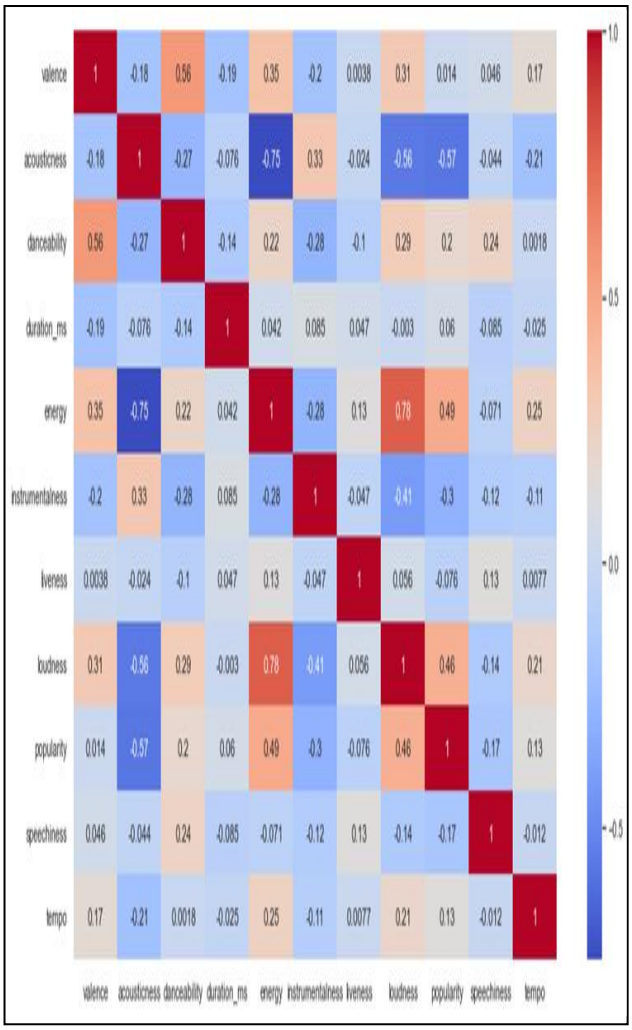


**Fig 9.** Graphical representation of correlation values computed for the selected features

Figure 10(a), shows that energy and popularity are positively correlated and it can be said that the popular songs are full of energy and vigor or it can be said that the songs that have beats and makes people dance are full of energy and they shall be leading the popularity charts. The Figure 10(b) shows the relationship between liveness and popularity, it can be said that the songs which have been sung in the presence of live audience tend to be less popular as compared to the songs created in the closed ambiance. The songs with higher value of liveness may not be able to pull and sway the crowds. To a greater extent. On the other hand the Figure 10(c) shows that the acousticness and popularity bear a positive relation.

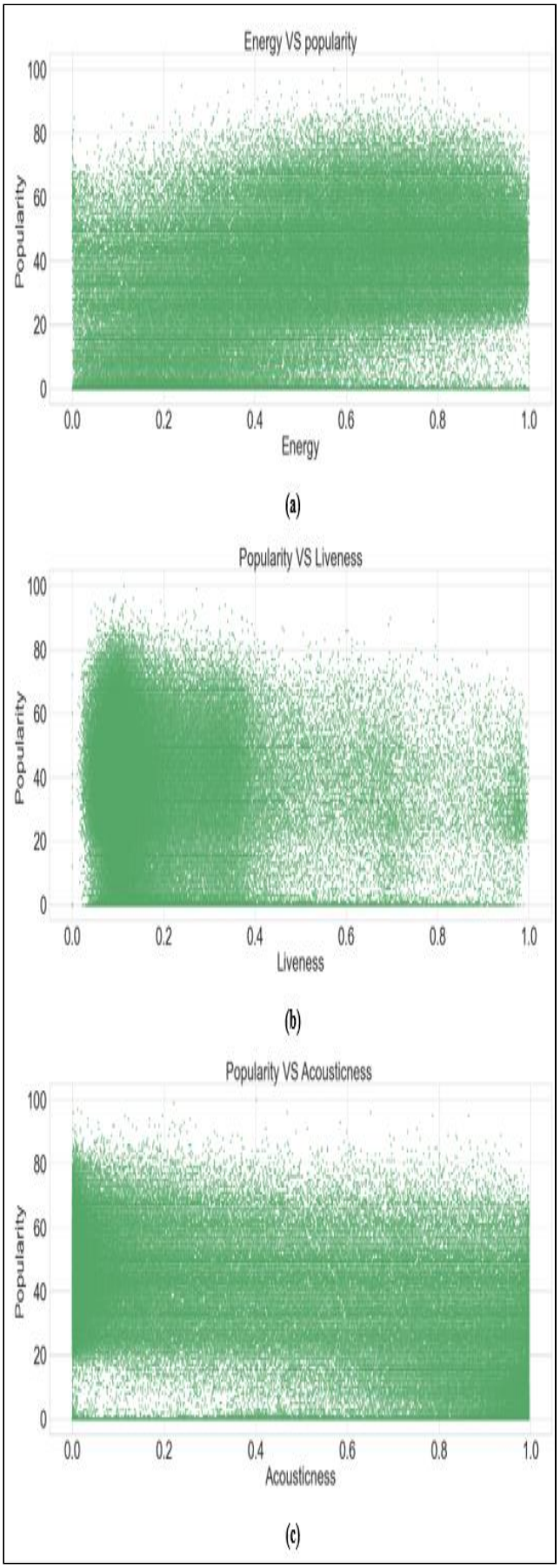


**Fig 10.** Correlation of popularity with other features of the set (a) Energy Vs. Popularity (b) Popularity Vs. Liveness (c) Popularity Vs. Acousticness

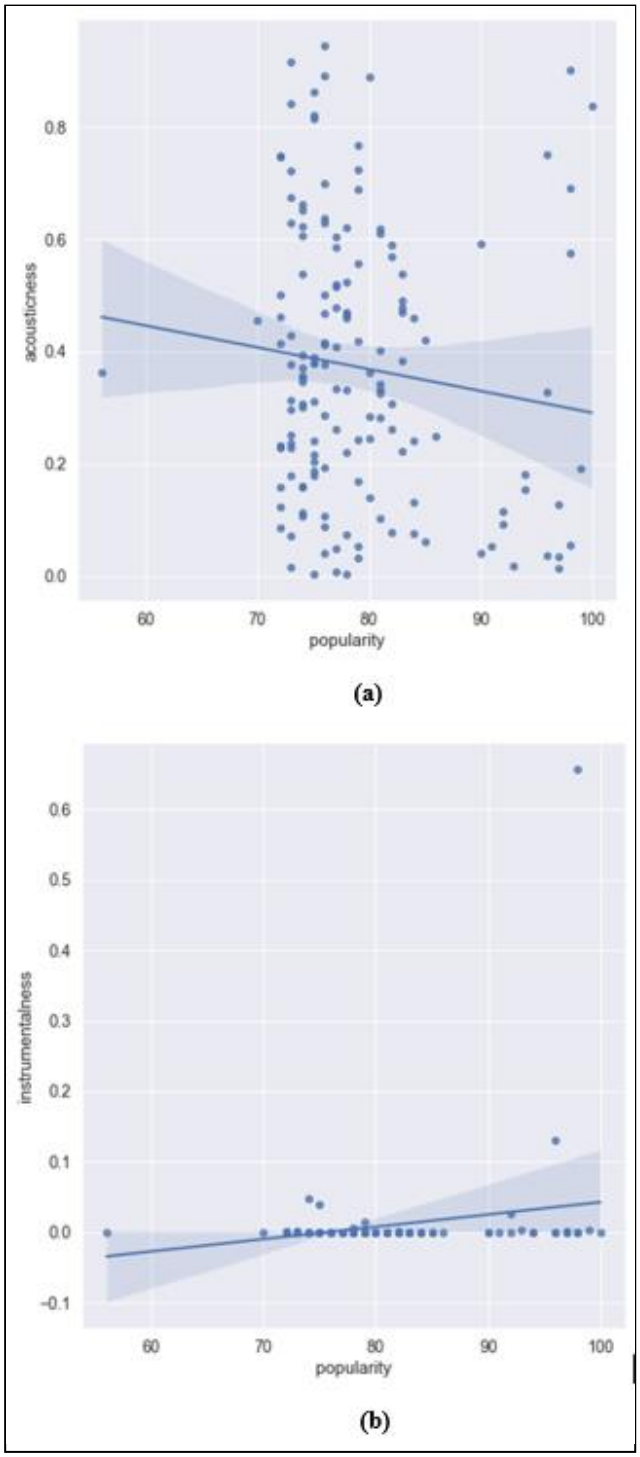


**Fig 11.** Correlation of the popularity feature with acousticness and instrumentalness

Acousticness along with instrumentalness, danceability and duration features influence the popularity of the song. Acousticness decreases as popularity increases.The majority of the songs are at 0 or really close to it as shown in Figure 11(b), when it comes to instrumentalness. However, there is one song that really fishes out and data point, alongside another one that is a little above the 0 line. Acousticness decreases as popularity increases as shown in the Figure 11(a).

From Figure 12(a), it has been observed that the lines drawn for danceability and duration in milliseconds are flat in nature therefore the popular songs with a value greater than 80 are more danceable. The Figure 12(b) shows popular songs have a range of

200 to 300ms of duration.

Figure 13 depicts the analysis of the dataset collected on the basis of the year. Figure 13(a) depicts the counts of the tracks added per year and it is observed that the count is increasing steadily but from 1960 onward the data shows a production of more than 2000 tracks (last observation 2020). In Fig 13(b), it has been observed that there was trend for acoustic based songs till 1970s, but falling continuously, but the songs with danceability, energy and speechiness are consistent in the market. The Fig 13(c) shows that the loudness feature in songs is continuously increasing every year. The Fig 13(d) shows that the influence of tempo-based songs started to rise from 1940s and is ever increasing. The observations on the average overall loudness are that it starts with 0 as the limit at which music starts to clip. As expected, the correlation between loudness and year is close to 1. On an average, the year 2020 has been the loudest year in the past.

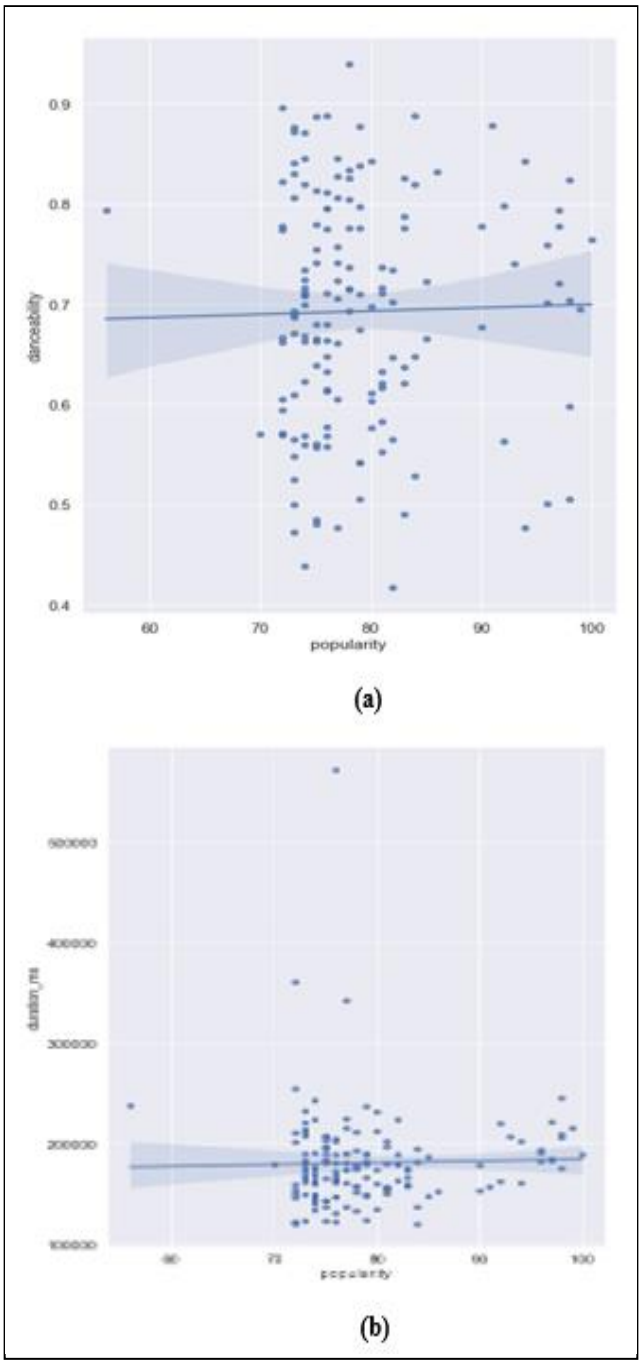


**Fig 12.** Correlation of popularity, danceability and duration_ms features

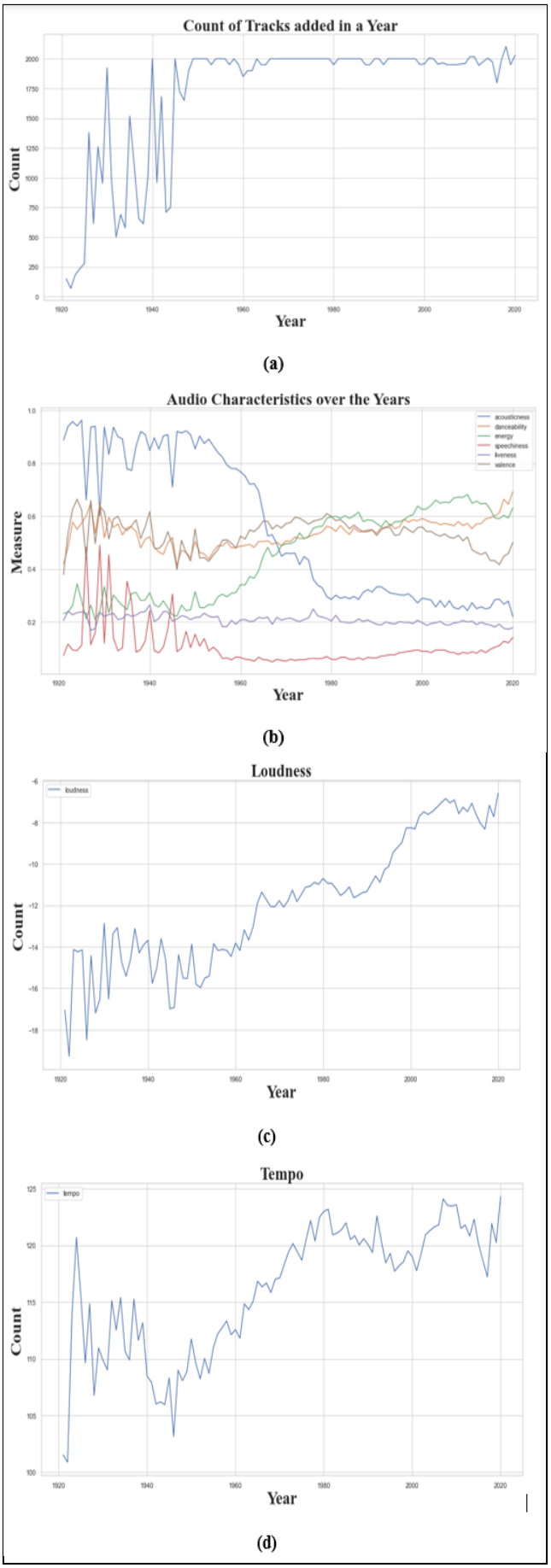


**Fig 13.** (a) No. of tracks added in a year (b) Audio characteristics added over the years (c) Loudness level over the years (d) Tempo level over the years.

### 3.2 Applications of Machine Learning Techniques

The entire dataset has been modelled around four major algorithms which includes KNN(K-nearest neighbours), RF(Random Forest), DT(Decision Tree Classifier) and LR(Logistic Regression). A brief discussion about them is as follows:

**Logistic regression:** Logistic Regression has more often been employed for classification purpose and is based upon the standard logistic regression function, for predicting the outcome of an observation given a predictor variable (x), is an s-shaped curve defined in (2)

$$p = \frac{\exp(y)}{[1 + \exp(y)]} \tag{2}$$

This can be also simply written as (3)

$$p = \frac{1}{[1 + \exp(-y)]} \tag{3}$$

where:

**y = b0 + b1*x, exp()** is the exponential and **p** is the probability of event to occur (1) given **x**. Mathematically, this is written as **p(event=1|x)** and abbreviated as **p(x)** as in (4)

The output of logistic regression function is always a value lying between 0 and 1. If the output value obtained from the logistic regression model is higher than a threshold value, in that case, the model predicts the positive class, and vice versa for below the threshold.

**KNN:** K-Nearest Neighbour is one of the most basic machine learning algorithms and it is also based on supervised learning techniques. It checks for likeness between new data and previously used data and categorizes the data based on similarity. It can be used for both classification and regression based tasks. The datapoints are plotted in a high-dimensional space, and the distance between all of these points is calculated. Then, a value of K is determined, usually through iterating through several potential values of K within the training data to uncover the optimal value, and the K nearest neighbours are looked at to determine a final prediction.

**RF:** Random Forest: Random Forest is an ensemble-based machine learning algorithm that is used to solve regression and classification which combines multiple decision trees and other classifiers to give a solution to the complex problems. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems. Random Forest is based on ensemble-based technique which means combining many models together and is built around two major techniques names as bagging and boosting. Bagging creates a different training subset from sample training dataset with replacement and the final output is based on majority voting whereas in the case of boosting combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy

**DT:** 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. A decision tree involves a method of classification that relies on a series of descending choices. It is based on the decisions made down each of the branches, a conclusion is reached regarding the question of interest. The decision trees use the CART algorithm (Classification and Regression Trees). In both cases, decisions are based on conditions on any of the features. The internal nodes represent the conditions and the leaf nodes represent the decision based on the conditions.

## 4. Results and Analysis

As discussed through different graphs in section 3, it is clear that the features like accousticness and energy are negatively correlated, year and popularity are positively correlated. The machine learning analysis that was performed on the dataset using four different algorithms are able to achieve the accuracy results as high as 80%, which speaks well of the model trained and tested. The dataset collected shows increase in production of songs having considerable higher level of tempo and loudness content. The songs with danceability, energy and speechiness have always been in the midlevel of regular production system. The future scope of the work involves the use of deep neural network-based techniques to further analyse the characteristics of the lyrics along with the other parameters to predict the performance of the artists and the place in the yearly popularity chart. In the present work involves the comparative analysis of the various algorithms for the identification of the popularity. The Random Forest based algorithm shows an accuracy of 80.41%, Logistic Regression with 80.15%, KNN with 77.54% and Decision Tree Classifier shows an accuracy of 68.92%. We conclude that the Random Forest based algorithms perform the best on the given dataset. The results have been shown graphically through Fig 14, which is a bar-chart showing the individual performance of the machine learning based algorithms.

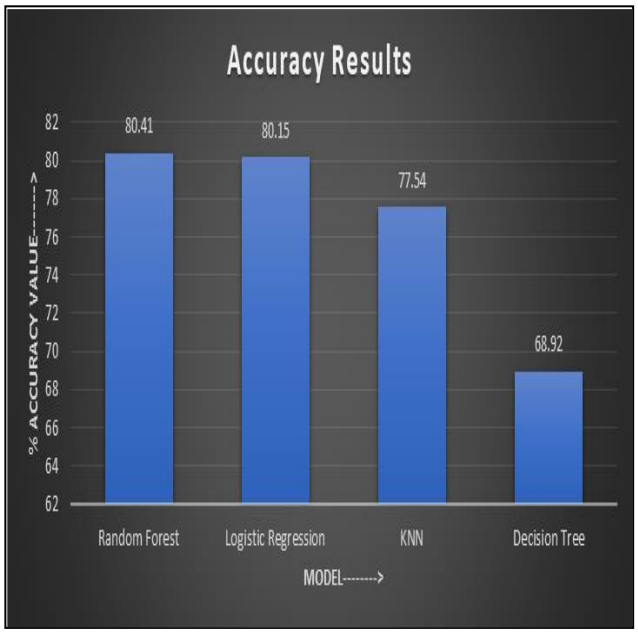


**Fig 14.** Models and their accuracy level for the dataset

The metrics AUC(Area under the Curve) has also been used here to check the performance of the various machine learning models used for computational purpose. Here AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. In the present context the AUC values for the Random Forest algorithm is 73.71%, for Logistic Regression, the AUC value is 73.31%, for KNN the AUC value is 71.37% and for Decision Tree is 64.77%. All the values mentioned here speak well about the performance of the algorithm concerned. Here in our case Random Forest Algorithm has performed the best as shown in the Fig 15.

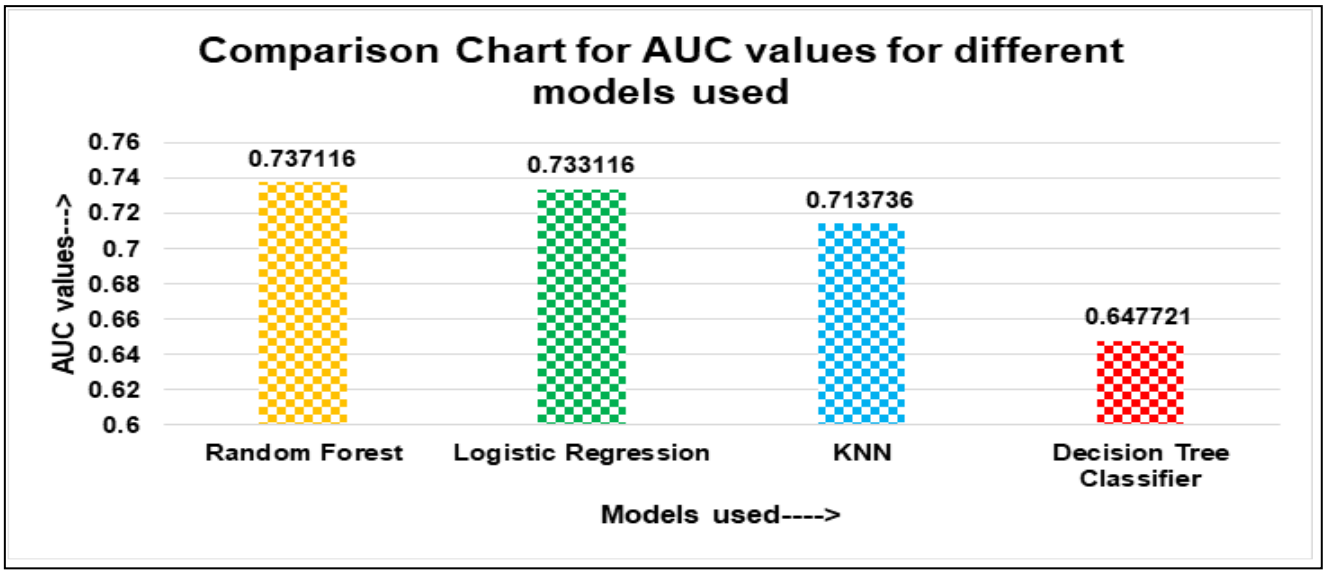


**Fig 15.** Models and their AUC level achieved

**Fig 16.** Comparison of the present research work with [1,2]

Further the Fig 16 shows a comparison between two reference works which have used the same Kaggle based dataset and a similar type of machine learning models used. The Fig 16 shows that Random Forest based algorithm fares better of all.

## 5. Conclusion and Future Scope

The research conducted within this paper demonstrates that underlying features of individual songs can be collected and input into machine learning based statistical models that can predict with a high degree of accuracy whether that song will become a commercial success or not. This scientific approach can be applied to the inherently artistic music industry in order to allow for greater efficiency and scale for individual artists, producers, and record companies. It is finally concluded that songs play a major role in one's life for mood management that shall enhance the particular mood of the individual. The songs also help in self-identity development and express our image to others in the form of type of music being listened and forms a sort of bonding between the groups for enjoyment and fun. The songs have their own genere, the singer has his/her own style, the music companies and the percussionists also have a role in changing the position of a track in the chart. From our machine learning based results it is clear that the features do play a role in affecting the popularity of the songs in the music market. The results have clearly portrayed a choice of 80% and above if the companies really think of employing the machine learning based skills and choice of features influencing the music market. The music industry is very dynamic and ever-growing, the lyrics can be one of the important parameters for improving the position of a track in the yearly chart. The deep learning based techniques shall definitely help in achieving our future goal.

## References

[1] Ochi V et.al.,"Spotify Danceability and Popularity Analysis using SAP", arXiv:2108.02370 [cs.DC], DOP: August 05, 2021.

[2] Fethi. F.," Similarity measures for Collaborative Filtering-based Recommender Systems: Review and experimental

comparison” , Journal of King Saud University –Computer and Information Sciences-2021.
[3] Pérez-Verdejo et al. “The rhythm of Mexico: an exploratory data analysis of Spotify’s top 50”. Journal of Computer Society Sc 4, pp. 147–161 (2021). Available: https://doi.org/10.1007/s42001-020-00070-z.
[4] Giannakopoulos T.,” pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis”, PLOS ONE, Public Library of Science, vol. 10 no. 12, pp. 1-17, 2015. Available: https://doi.org/10.1371/journal.pone.0144610
[5] Pichl M. et al.,” Understanding User-Curated Playlists on Spotify: A Machine Learning Approach”, International Journal of Multimedia Data Engineering & Management Volume 8, Issue 4, pp 44–59, October 2017 Available: https://doi.org/10.4018/IJMDEM.2017100103.
[6] Diaz F.,” Spotify: Music Access At Scale”, SIGIR '17: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval August 2017 Pages 1349, Available: https://doi.org/10.1145/3077136.3096471
[7] Braun, T. A., " Dance like nobody's paying": Spotify and Surveillance as the Soundtrack of Our Lives (Doctoral dissertation, The University of Western Ontario (Canada)), 2020.
[8] Anderson I.et al., “Just the Way You Are”: Linking Music Listening on Spotify and Personality”, Social Psychological and Personality Science · July 2020, SAGE publishers.
[9] Pareek P. et al.,” Predicting Music Popularity Using Machine Learning Algorithm and Music Metrics Available in Spotify”, Journal of Development Economics and Management Research Studies (JDMS), A Peer Reviewed Open Access International Journal, ISSN 2582 5119 (Online), vol. 09 no. 11, pp. 10 -19, January-March, 2022.
[10] Isnuhoni S., et al.,” A Semantic Analysis on Spotify top Songs for Teens”, Journal of Research on Applied Linguistics Language and Language Teaching, vol. 2, no.. 2, pp. 123 – 131, November 2019.
[11] Almqvist, C. F., Leijonhufvud, S., & Ekberg, N., “Spotify as a case of musical bildung. Nordic Research in Music Education, vol. 2 no. 1, pp. 89-113. 2021. DOI: 10.23865/nrme.v2.3023.
[12] Luis Aguiar, Joel Waldfogel,” Platforms, Promotion, and Product Discovery: Evidence from Spotify Playlists”, National Bureau of Economic Research, 2018. Available: http://www.nber.org/papers/w24713
[13] Brubaker J.R.,” P4KxSpotify: A Dataset of Pitchfork Music Reviews and Spotify Musical Features”, Proceedings of the Fourteenth International AAAI Conference on Web and Social Media (ICWSM 2020), vol. 14, pp. 895-902, 2020.
[14] Fleischer, Rasmus & Pelle Snickars,” Discovering Spotify – A Thematic Introduction”, Culture Unbound, vol. 9, no. 2, pp. 130–145. 2017. Published by Linköping University Electronic Press: http://www.cultureunbound.ep.liu.se
[15] Wang R., “Analysing Spotify listening data from Armada Music”, Radboud University Press, 2018.
[16] Mariangela Sciandra , Irene Carola Spera,” A model based approach to Spotify data analysis: a Beta GLMM”, Journal of Applied Statistics, vol. 49 no. 1, pp. 214-229, 2022.
[17] Vasseur G. et al., “Spotify: You have a Hit!”, SMU Data Science Review: vol. 5, no. 3, Article 9, 2021. Available: https://scholar.smu.edu/datasciencereview/vol5/iss3/9
[18] Dhanaraj, R., & Logan, B.,” Automatic Prediction of Hit Songs.”, International Society for Music Information Retrieval, pp. 488-491, 2005. Available: https://ismir2005.ismir.net/proceedings/2024.pdf
[19] Francois Pachet,” Hit Song Science”, Li, Tao, Mitsunori Ogihara, and George Tzanetakis, eds. Music data mining. vol. 20. Boca Raton: CRC Press, 2012.
[20] Jin Y. et al., “Controlling Spotify Recommendations: Effects of Personal Characteristics on Music Recommender User Interfaces”, UMAP '18: 26th Conference on User Modelling, Adaptation and Personalization Singapore July 8 - 11, 2018, pp. 101-109. ISBN:978-1-4503-5589-6.
[21] Werner A., “Organizing music, organizing gender: Algorithmic culture and Spotify recommendations”, Popular Communication: The International Journal of Media and Culture, vol. 18 no. 1, pp. 78-90, DOI:10.1080/15405702.2020.1715980. Available: https://doi.org/10.1080/15405702.2020.1715980
[22] Hurtado A. ,Wagner M., Mundada,S.,” Thank you, Next: Using NLP Techniques to Predict Song Skips on Spotify based on Sequential User and Acoustic Data”, 2019. Available:https://cs229.stanford.edu/proj2019aut/data/assignment_308875_raw/26510716.pdf
[23] Shukun Yin S., Fu L.,“ The Effectiveness of Brand Culture on Customer Engagement A Case Study of Spotify in the U.S.”, Advances in Economics, Business and Management Research, vol. 203, Proceedings of the 2021, 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021). pp. 2653-2659. Atlantis Press, 2021.
[24] Jain P., “How Spotify Wrapped 2020 Marketing Campaign Boosted Mobile App Downloads And Engagement”,[Online]. Available: https://www.moengage.com/blog/spotify-wrapped-2020-app downloads-engagement/.
[25] Brown C., “Spotify’s Business Model, Generic Strategy & Growth Strategies”, 2019,[Online] Available: https://www.rancord.org/spotify-business-model-generic-competitive-strategy-intensive-growth-strategies
[26] Garcia P., “Spotify SWOT Analysis: Internal & External Strategic Factors”, 2019. [Online] Available: https://www.rancord.org/spotify-swot-analysis-internal-external-strategic-factors
[27] QuesadaS.R.,” Genres as Social Affect: Cultivating Moods and Emotions through Playlists on Spotify”, Social Media + Society, vol. 1 no. 11, pp. 1–11, April-June 2019: Available: https://doi.org/10.1177/20563051198475
[28] Iordanis S.P.,” Emotion-Aware Music Recommendation Systems Mitigating the Consequences of Emotional Data Sparsity”, Doctoral Dissertation, Aristotle University of Thessaloniki, 2021.
[29] LeCun Y. et al.,” Feature Learning and Deep Architectures: New Directions for Music Informatics”, Journal of Intelligent Information Systems, vol. 41, pp. 461–481, 2013.
[30] Pichl M., Zangerle E.,Specht G.,“ Understanding User-Curated Playlists on Spotify: A Machine Learning Approach”, International Journal of Multimedia Data Engineering and Management, vol. 8 no. 4, October-December 2017.
[31] Shete A., Mohanani N., Gondal S., Prediction of an Artist's Success on Spotify”, International Research Journal of Engineering and Technology (IRJET), vol. 08 no. 10, Oct. 2021.
[32] Kalustian K., Nicolas Ruth N.,” Evacuate the Dancefloor”: Exploring and Classifying Spotify Music Listening Before and During the COVID-19 Pandemic in DACH Countries”, Jahrbuch Musikpsychologie, vol. 30, 2021.Available: https://doi.org/10.5964/jbdgm.95.
[33] Malheiro R.,” How Does The Spotify API Compare to the music emotion recognition state of the art?”, Proceedings of the 18th Sound and Music Computing Conference, June 29th – July 1st 2021, pp. 238-245. Axea sas/SMC Network, 2021.
[34] Chmiel A.,Schubert E.,” Unusualness as a predictor of music preference”, Musicae Scientiae, vol. 23 no. 4, pp. 426–441, 2019. DOI: 10.1177/1029864917752545.
[35] Bello P., Garcia D.,” Cultural Divergence in popular music: the increasing diversity of music consumption on Spotify across countries”, Humanities and Social Sciences

Communications, vol. 8 no. 1, pp. 1-8, 2021. Available: https://doi.org/10.1057/s41599-021-00855-1