CSE343 Machine Learning - Interim Project Report

Kunal
2022260
kunal22260@iiitd.ac.in

Akshat Gian
2022051
akshat22051@iiitd.ac.in

Dhairya
2022157
dhairya22157@iiitd.ac.in

Ashwin Verma
2022117
ashwin22117@iiitd.ac.in

Arun Singh Rawat
2022106
arun22106@iiitd.ac.in

1. Motivation

We all have a fondness for music, don't we? Most of us have favorite songs, some we truly enjoy, and others we simply don't like. While it might seem that our musical preferences are based on chance, this project seeks to investigate whether there is a data science element influencing why certain songs are more popular than others. Our objective is to analyze musical characteristics to predict a song's popularity among listeners and gauge the extent of that popularity. This research could also assist music composers, singers, and instrumentalists in crafting new music that resonates with audiences. The entire team finds this topic both intriguing and innovative, with ample opportunities for exploration, as music and its traits can be analyzed through various factors. Additionally, as we enhance our skills in machine learning and signal processing, we anticipate numerous possibilities for further exploration in this field, making it a valuable project for ongoing development and iteration in the future.

2. Introduction

This project focuses on analysing if there exists a relationship between music popularity and song characteristics. This project dives deeper into trying to answer the question of whether a particular song's popularity has some underlying data-driven factors leading to it. The main objective of this project is to leverage attributes of a song to predict its popularity and also predict the degree of popularity it may achieve.

3. Literature Review

3.1. Reference 5

In this paper, they are basically building a methodology that can predict whether a song will appear on Spotify's Top 50 Global ranking after a certain amount of time. They approach the problem as a classification task and use the data from the past platform's Top 50 Global ranking collected using Spotify's web API, The model uses information on the songs previously observed in that list, Support Vector Machine classifier with RBF kernel reached the best results in our experiments with an AUC higher than 80% when pre- dicting the popularity of a song two months in advance.

3.2. Reference 6

This project focuses on predicting the popularity of songs on the Million Song Dataset, a crucial aspect for maintaining competitiveness in the expanding music industry. The dataset contains audio features and metadata for around one million songs, the study assesses various classification and regression algorithms to determine their effectiveness in forecasting song popularity. The investigation also identifies the specific types of features that possess the greatest predictive capability in this context.

In this paper, they evaluated different classifications and regression algorithms on their ability to predict popularity and determined the types of features that hold the most predictive power.

3.3. Reference 7

This paper primarily focuses on the Popularity Prediction of Music by related Python tools and various ma- chine learning models like xgboost, XGBRegressor, and Polynomial Regression, the dataset for the study contains 114000 data points and 19 features. The research evalu- ates models using metrics such as mean-squared error and R-squared. Ultimately, XGBoost emerges as the model, demonstrating a strong correlation between music attributes and popularity. The paper also discusses potential areas for improvement, including refining categorical variable encoding and exploring interactions between independent variables.

3.4. Mathematical Concepts

Mean: It is the total sum of values in the dataset divided by the total number of data points, it is a measure of the central tendency of a probability distribution along median and mode.
Standard deviation: It is the measure of the spread or dispersion of a set of data points from their mean, It helps assess the consistency and reliability of data which in turn draws meaningful conclusions.
Covariance: It is a measure of the relationship between two random variables and to what extent, it indicates whether an increase in one variable corresponds to an increase or decrease in another variable. A positive covariance suggests a positive relationship, while a negative covariance indicates an inverse relationship and a covariance of zero implies no linear relationship between the variables.
Correlation: It is a statistical measure that expresses the extent to which two variables are linearly related.

4. Dataset Description

The primary source of our dataset is the Spotify API which analyses the songs uploaded on its platform on the basis of various sound characteristics. The Spotify API essentially provides us with 2 useful datasets, one which primarily contains the numerical parameters such as:

1. song_duration_ms: This refers to the length of the track in milliseconds.
2. acousticness: This is a confidence measure from 0.0 to 1.0 that assesses whether the track is acoustic. A value of 1.0 indicates high confidence that the track is acoustic.
3. danceability: This describes how suitable a track is for dancing, considering elements like tempo, rhythm stability, beat strength, and overall regularity. Least danceable is 0.0 to 1.0 being most danceable.
4. energy: This is a measure from 0.0 to 1.0 that represents the perceptual intensity and activity of the track. Energetic tracks typically feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale.
5. instrumentalness: Detects the number of different instruments involved in the songs.
6. key: This represents the estimated overall key of the track, with integers mapping to pitches using standard Pitch Class notation.
7. liveness: Detects the presence of an audience in the recording. Value is proportional to probability that music was recorded in a live event
8. loudness: Measured in decibels (dB). It is basically the decibel value for each instance averaged across the entire track and is useful for comparing relative loudness of tracks. Values typically range between -60 and 0 db.
9. audio mode: This indicates the modality (major or minor) of the track, the type of scale from which its melodic content is derived. Major is represented by 1, and minor is represented by 0.
10. speechiness: Measure of the presence of spoken words in a track. When the track is something like an audiobook or podcast where there are mostly clearly said and complete words with a lot of understanding and speech involved this parameter tends to 1 and the other part of the spectrum is the songs which have absolutely no wordings involved in them.
11. tempo: beats per minute (BPM). i.e. the pace of a given song
12. time signature: This provides an estimated overall time signature of the track, specifying how many beats are in each bar (or measure).
13. audio valence: range: [0, 1] is a measure of the musical positiveness conveyed by a track. Higher value means happier the song is while lower the value means that the song is sad, intermediate values convey the balance between positivity and negativity of the song
14. song popularity: Based on rating and number of listens and relistens by the audience.

This dataset is the song data.csv, other than this there is an- other dataset for some additional information such as:

1. artist name: is the name of the singer of the track
2. album names
3. playlist: is the playlist name in which the song is launched in spotify.
4. song name

This dataset is named song info.csv

For our project, we started off by making a new dataset using their primarily available datasets. To form our dataset we have used the concat() function horizontally ie: on axis1 to both the datasets, since the artist name column is common in both of them we dropped it from the second dataset, while since our target variable is the song popularity we added it as the target column towards the end.

Final dataset dimensions: (18835, 19)

This is followed by classifying the numerical score into categories to turn it into a bounded classification problem:

Category	Popularity Score
Potential Masterpiece	78-100
Popular	69-78
Mildly Popular	56-69
Potential Flop	0-56

Table 1. Category wise popularity score

These boundaries have been decided based upon certain percentile calculations which has been discussed more in detail further.

Category	Percentile
Potential Masterpiece	90%
Popular	75-90%
Mildly Popular	50-75%
Potential Flop	0-50%

Table 2. Category wise percentile

We found out what are the percentile scores for each and also plotted distribution plots and quartiles to decide on the numbers

Field	Value
mean	52.991877
std	21.905654
min	0.000000
5%	8.000000
10%	21.000000
25%	40.000000
50%	56.000000
75%	69.000000
80%	72.000000
90%	78.000000
95%	85.000000
max	100.000000

Table 3. Caption

Other than this we have done the Standardization of the dataset to improve the model predictions and accuracy later on however minor tweaking is expected to be needed as we move on further into the project. Encoding was needed for the varchar values in the dataset ie: song_name, artist_name, album name, playlist we have used one-hot encoding for it. Missing values were printed however the dataset has no NaN or missing values hence not needed to be removed/replaced with the average. Distribution curve was plotted to find the percentile and get an idea of the distribution. The resultant plot is a normal distribution as expected.

Feature selection and outlier detection has also been done which has been described in more detail in the methodology.

5. EDA

5.1. Pattern Detection in Data

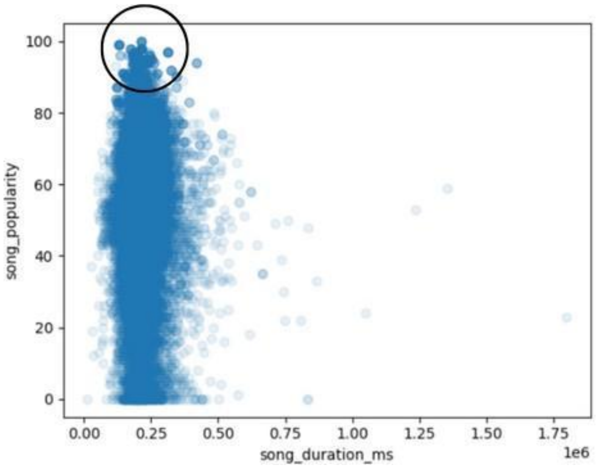


Figure 1. Song Popularity vs. Song Duration

Insights related to figure 1. are:

1. Most of the songs lie in a certain duration band which is less than 400 seconds.
2. Most of the popular songs are of duration around 250 +- 100 seconds.
3. The songs with very high duration are generally less likely to be popular and the same is true with very low duration songs.
4. Since the score required to be a “masterpiece” is more than 85, from the plot it seems that if a song is way too much in duration it is not possible for it to be a masterpiece in terms of popularity.

In case of figure 2 we can observe that most of the songs in our dataset have acousticness on the lower side.

Insights related to figure 3. are

1. More danceable songs are more likely to have higher popularity.
2. Most of the songs in our dataset have danceability in the range of 0.3 to 0.9.

Insights related to figure 4. are

1. The songs in our dataset generally have higher energy.

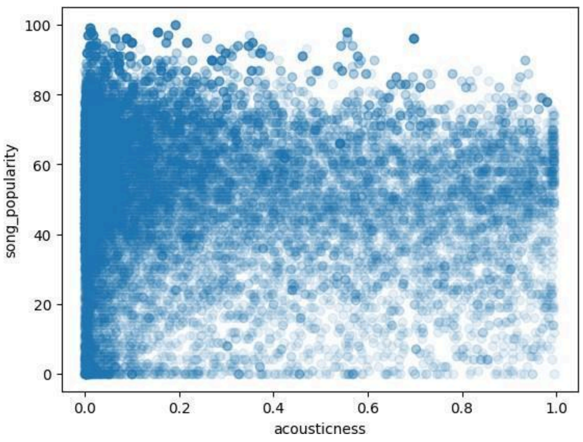


Figure 2. Song Popularity vs Acousticness

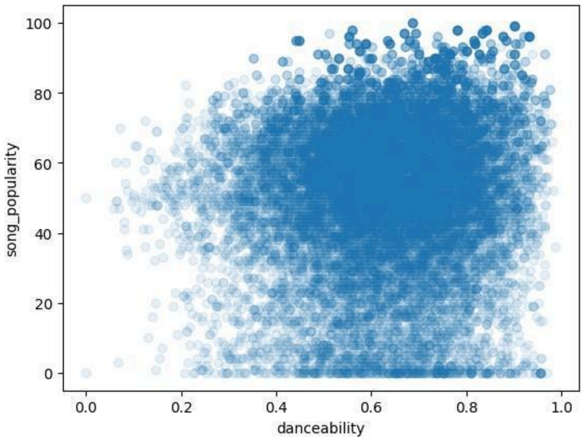


Figure 3. Song Popularity vs Danceability

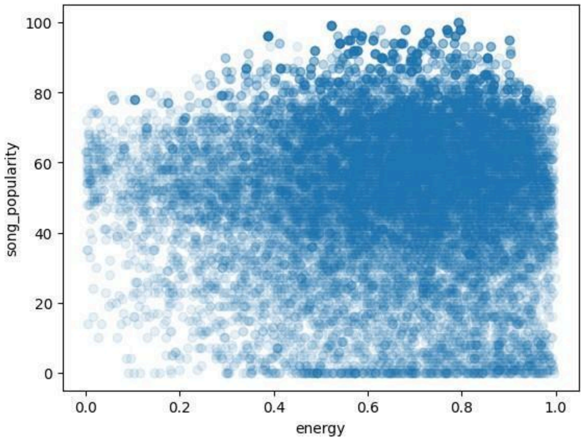


Figure 4. Song Popularity vs Energy

2. None of the overly energetic songs Ie: with an energy score greater than 0.95 seems to be a masterpiece.

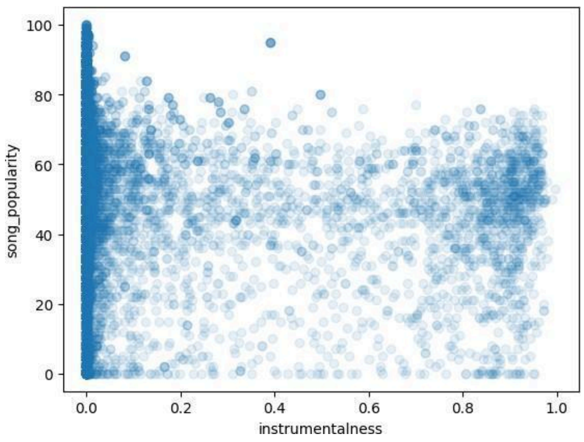


Figure 5. Song Popularity vs Instrumentliness

Insights related to figure 5. are

1. Most of the songs have either very few instruments or too many instruments.
2. Almost all of the masterpieces use a very small number of different instruments.

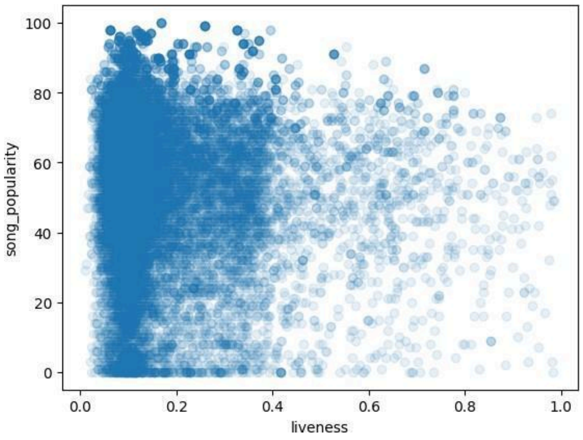


Figure 6. Song Popularity vs Liveliness

Insights related to figure 6. are

1. Most of the tracks have less liveliness.
2. Songs with less liveliness are likely to be very popular.

From figure 7 we can observe that most of the songs generally are on the louder side.

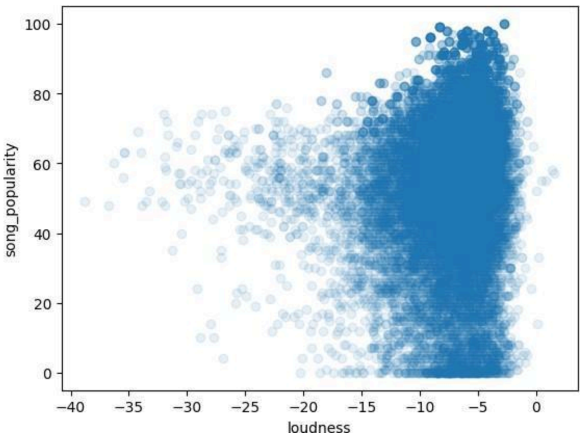


Figure 7. Song Popularity vs Loudness

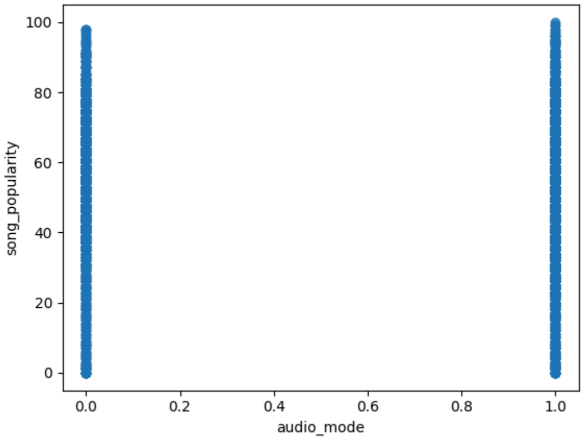


Figure 8. Song Popularity vs Audio Mode

From Figure 8, it is evident that audio mode does not appear to have a significant impact on song popularity, as both audio modes display nearly identical popularity distributions.

From Figure 9, it is apparent that songs usually exhibit low levels of speechiness.

As for Figure 10, the key insights are:

- The tempo of songs tends to be concentrated within a specific range of values.
- There are no songs with extremely high or low tempos.

From figure 11 we can observe that most of the songs have time signature values of 3, 4 or 5.

From Figure 11, we can see that the majority of songs have time signature values of 3, 4, or 5.

Figure 12 reveals that each type of song has a substantial number of listeners, indicating that

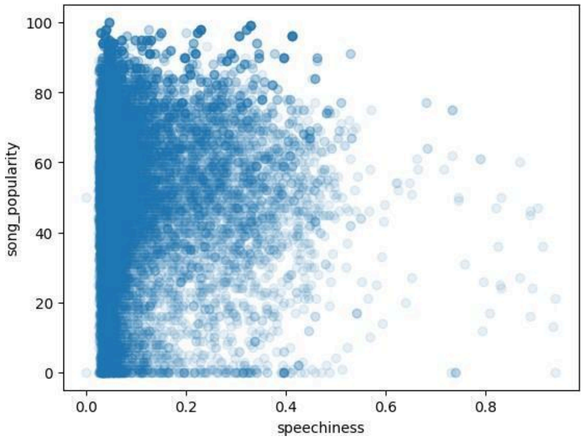


Figure 9. Relationship Between Song Popularity and Speechiness

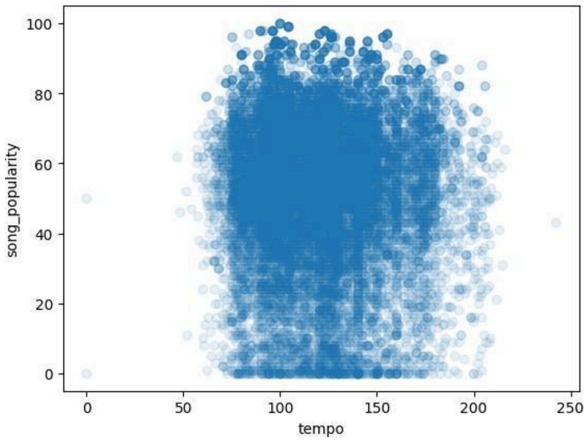


Figure 10. Relationship Between Song Popularity and Tempo

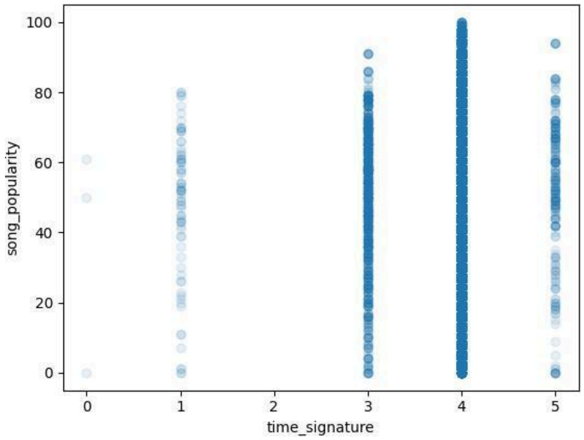


Figure 11. Relationship Between Song Popularity and Time Signature

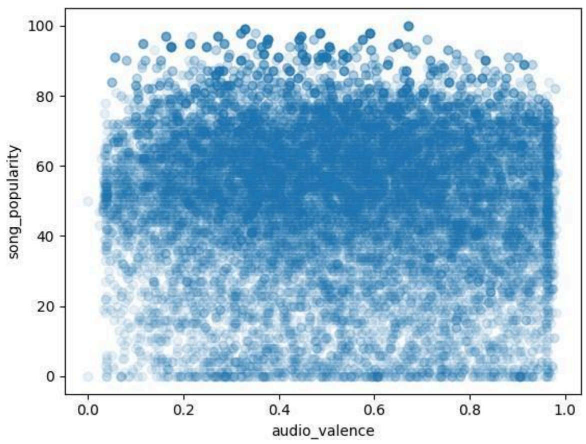


Figure 12. Audio Valence vs Time Signature

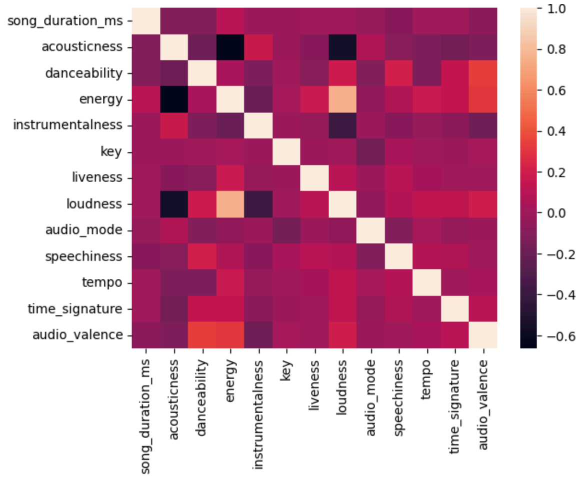


Figure 13. Heatmap of Features

Observations from Figure 13 are:

1. Loudness is inversely correlated with the number of instruments.
2. Loudness is positively correlated with energy.
3. Loudness is inversely correlated with acousticness.
4. Acoustic Ness is inversely correlated with energy.

6. Methodology and Model description

6.1. Data Collection

We acquired two main data files, specifically 'song data.csv' and 'song info.csv,' with details about these files provided in the 'Data Description' section above.

6.2. Data Integration:

We combined these two datasets using the Python pandas library to create a consolidated dataset, which we titled 'our dataset.csv'.

6.3. Data Pre-processing

We performed initial data preprocessing, which included the following steps:

- **Checking for Null or Missing Values:** We systematically reviewed the dataset for any missing or null values.
- **Distribution Analysis:** To gain insights into the data distribution, we generated a count versus song popularity distribution curve.

6.4. Categorization of Songs

To effectively categorize the songs, we used the describe function to calculate percentiles and scores. We divided the songs into four distinct categories based on their popularity scores:

- **Potential Masterpiece:** This category includes songs with exceptionally high scores, indicating a strong potential for success.
- **Popular:** These songs have already reached a significant level of popularity.
- **Mildly Popular:** This category encompasses songs with moderate popularity levels.
- **Potential Flop:** Songs in this group have scores that suggest a lower likelihood of success.

For the classification of data into these categories, we computed the percentage of songs in each:

- Percentage of songs classified as potential masterpieces: 10.96%
- Percentage of songs deemed popular: 14.57%
- Percentage of songs categorized as mildly popular: 24.48%
- Percentage of songs considered potential flops: 49.97%

6.5. Outlier Detection:

We utilized a boxplot to identify outliers within the data. As shown in Figures 14 and 15, many sample points in the dataset significantly diverge from the central tendency, suggesting the existence of outliers. Consequently, we need to establish a method for their removal.

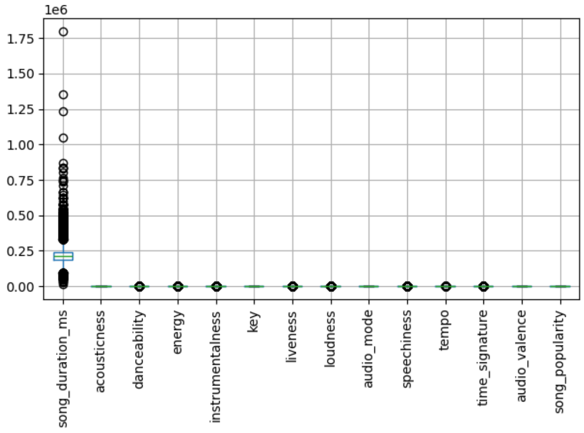


Figure 14. Boxplot of features (I)

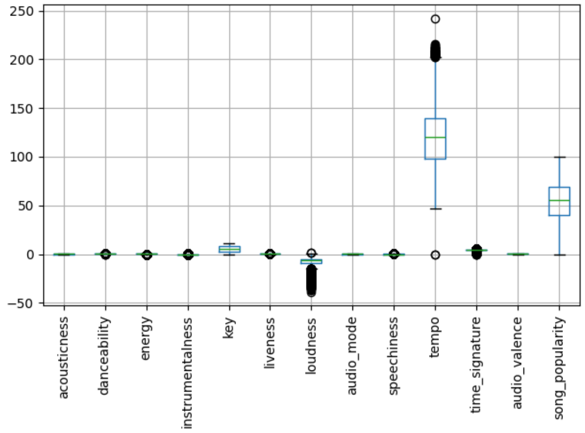


Figure 15. Boxplot of features (II)

6.6. Model description

We applied multiple models to the dataset, primarily to obtain accuracies on the non-preprocessed data and gain insights. The models we utilized are those covered in the lectures to date. For all models, we employed K-Fold cross-validation with K set to 10 and reported the average validation accuracy.

6.6.1 Logistic Regression

Solver: ’lbfgs’
Maximum Iterations: 10000
Regularization (C): 1.0
Penalty: ’l2’
Accuracy: 52.%

Naive Bayes

We used Gaussian Naive Bayes. There are no hyperparameters in case of Naive Bayes.
Accuracy: 35.66%

6.6.2 Decision Tree

Criterion: ’gini’
Accuracy:62.63%

6.6.3 Random Forest

Number of Estimator(trees):1000
Criterion: ’gini’
Accuracy: 77.01%

6.6.4 SVM

C: 10
gamma: 0.01
kernel: RBF
Accuracy:53.02%

7. Result and Analysis

We are obtaining notably good accuracies even without preprocessing the data. Random Forest have emerged as the top-performing models for this dataset, with their accuracies outlined in sections 6.6.3.

Certain features, such as song duration and liveliness, offer significant insights into the popularity of songs, while others, like audio mode, are less effective for making predictions.

This implies that by preprocessing the dataset, performing feature engineering, eliminating outliers, and fine tuning hyperparameters, we can further enhance model accuracies.

8. Conclusion and Future Plan

The overall conclusion of the analysis indicates that it is possible to predict a song's likelihood of becoming popular using Spotify's data with a reasonably high degree of accuracy.

Additionally, the exploratory data analysis suggests that the data requires some preprocessing. Since the machine learning models were employed to obtain preliminary results, they should be fine-tuned with the optimal parameters.

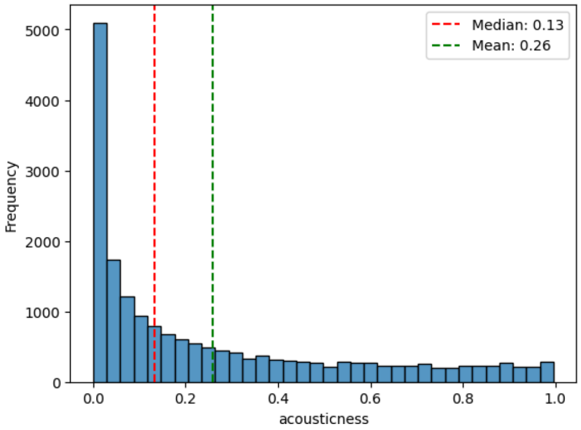
In the future, we plan to implement outlier removal, perform feature engineering, and tune the hyperparameters of the machine learning models.

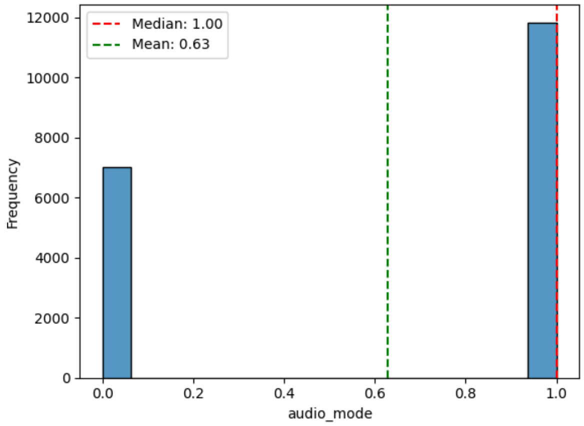
Additionally, we will utilize ensemble techniques to enhance prediction accuracy. We will also evaluate, analyze, and compare various machine learning models trained on this dataset to deepen our understanding. If we discover any useful insights, we aim to integrate them into the model to potentially improve its performance

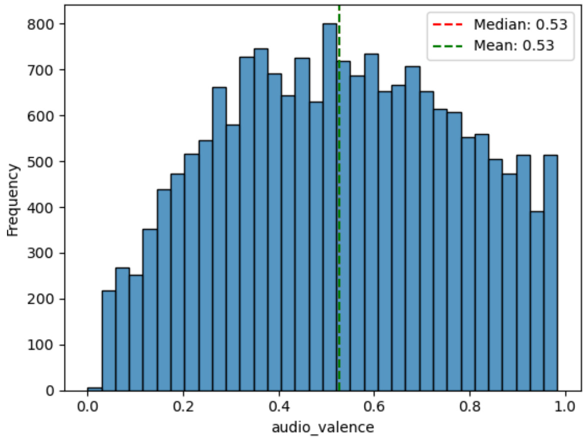
9. References

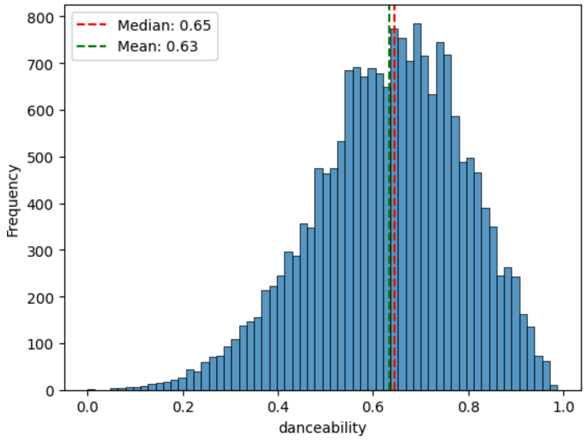
1. https://www.kaggle.com/datasets/edalrami/19000-spotify-songs
2. https://towardsdatascience.com/song-popularity-predictor-1ef69735e380
3. https://developer.spotify.com/documentation/web-api
4. https://www.kaggle.com/code/amansorout/spotify-song-popularity-classification
5. https://www.researchgate.net/publication/341420234_Predicting_Music_Popularity_on_Streaming_Platforms/link/5f0d0bd8a6fdcca32ae97ccc/download
6. http://cs229.stanford.edu/proj2015/140_report.pdf
7. https://www.researchgate.net/publication/370712842_Popularity_Prediction_of_Music_by_Machine_Learning_Models

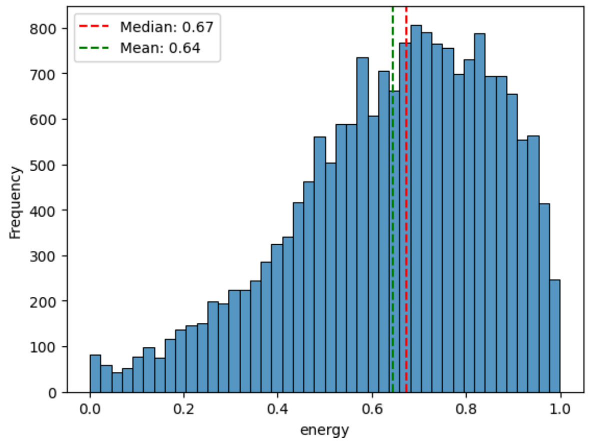
10. Some other relevant graphs

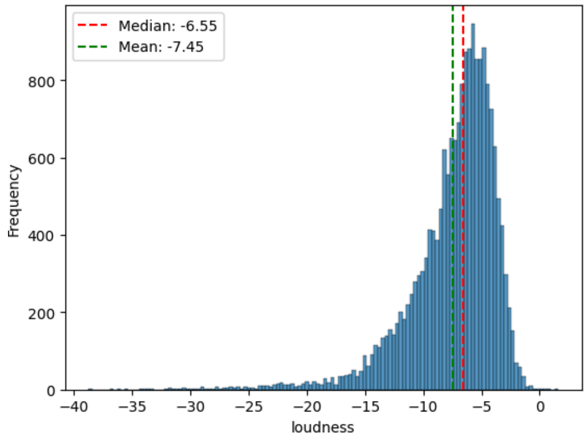


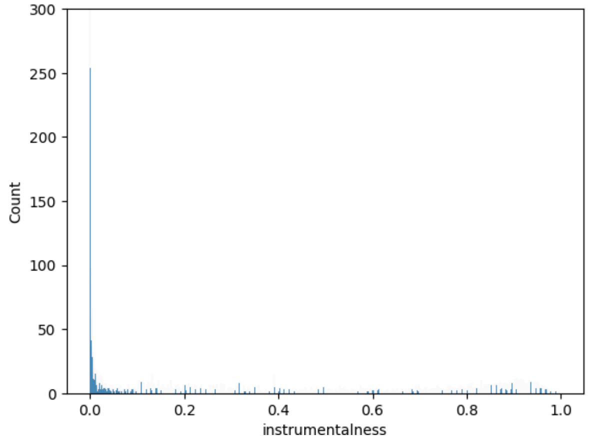


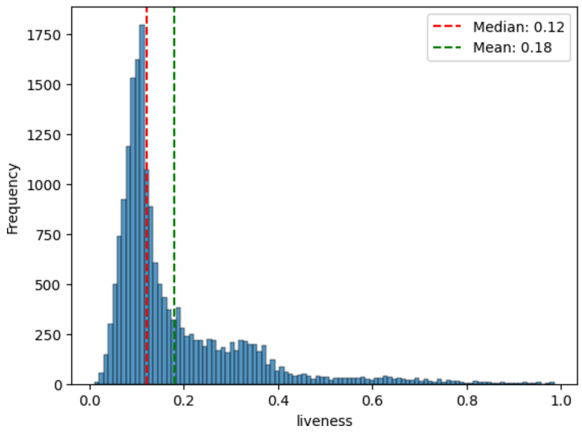


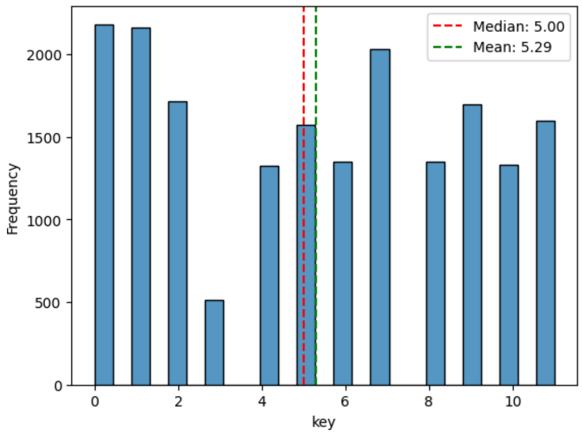


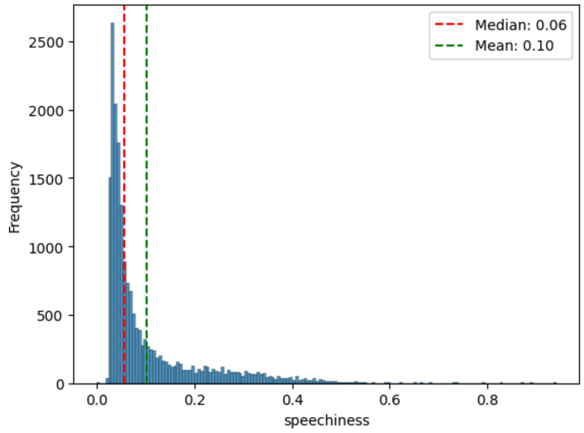

Median: 0.67
Mean: 0.64
Frequency
0
100
200
300
400
500
600
700
800
0.0
0.2
0.4
0.6
0.8
1.0
energy

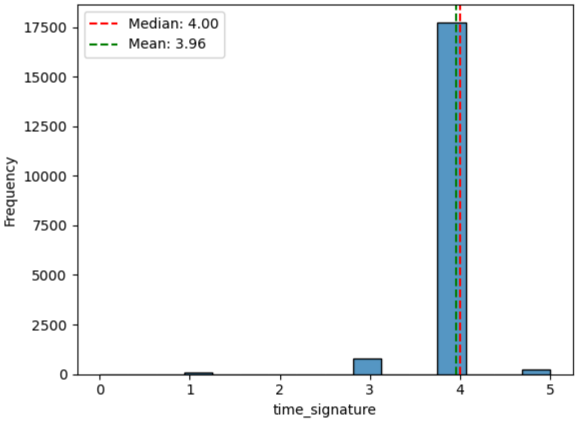

Median: -6.55
Mean: -7.45
Frequency
0
200
400
600
800
−40
−35
−30
−25
−20
−15
−10
−5
0
loudness

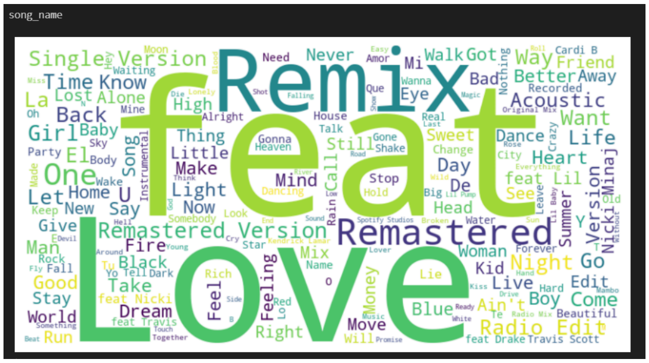

Count
0
50
100
150
200
250
300
0.0
0.2
0.4
0.6
0.8
1.0
instrumentalness

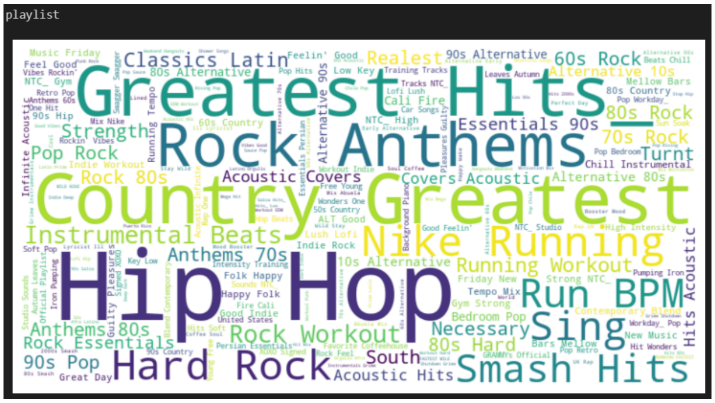

Median: 0.12
Mean: 0.18
Frequency
0
250
500
750
1000
1250
1500
1750
0.0
0.2
0.4
0.6
0.8
1.0
liveness

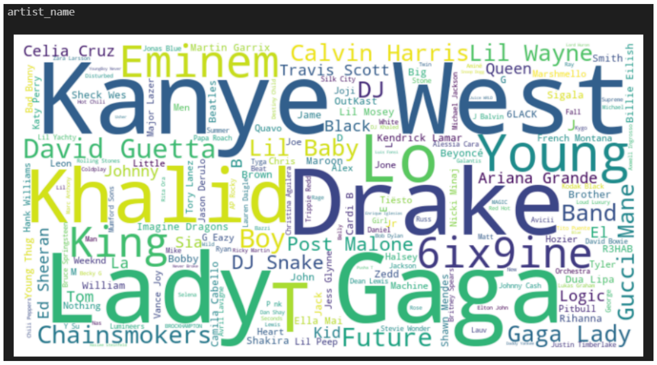

Median: 5.00
Mean: 5.29
Frequency
0
500
1000
1500
2000
0
2
4
6
8
10
key

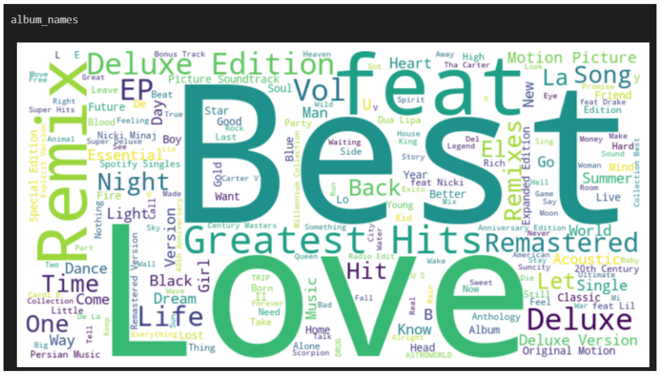

Median: 0.06
Mean: 0.10
Frequency
0
500
1000
1500
2000
2500
0.0
0.2
0.4
0.6
0.8
speechiness


Median: 4.00
Mean: 3.96
Frequency
17500
15000
12500
10000
7500
5000
2500
0
0
1
2
3
4
5
time_signature


song_name
feat
Love
Remix
Remastered
Single Version
Remastered Version
Time
Know
Lost
Alone
Back
Girl
Baby
One
Home
Let
Say
Thing
Little
Make
Light
Now
Mind
Still
Sweet
Day
Head
Dance
Heart
Life
Want
Acoustic
Better
Away
Way
Friend
Walk
Got
Never
Need
Bad
Eye
Mi
Fire
Black
Take
Dream
World
Stay
Good
Man
Give
Fall
Run
Rich
Feel
Feeling
Money
Move
Right
Blue
Ain't
Radio Edit
Night
Go
Kid
Live
Boy
Come
Edit
Beautiful
feat Drake
Travis Scott
Woman
Forever
Summer
Version
Nicki Minaj
Lie
Mix
Stop
De
La
High
Song


playlist
Greatest Hits
Rock Anthems
Country Greatest
Hip Hop
Nike Running
Hard Rock
Smash Hits
Sing
Run BPM
Rock Workout
Running Workout
Instrumental Beats
Classics
Latin
Realest
60s Rock
80s Rock
70s Rock
Essentials 90s
Strength
Pop Rock
Rock 80s
Acoustic Covers
Covers Acoustic
Alternative 80s
Anthems 70s
10s Alternative
Anthems 80s
Rock Essentials
90s Pop
Necessary
80s Hard
South
Acoustic Hits
Alternative 90s
Alternative 10s
Cali Fire
Turnt
Bedroom Pop
Tempo Mix
Music Friday
Feel Good
90s Alternative


artist_name
Kanye West
Drake
Lady Gaga
Khalid
Eminem
Calvin Harris
Lil Wayne
David Guetta
Young
Lil Baby
Travis Scott
Queen
King
DJ Snake
6ix9ine
Ariana Grande
Band
Malone
Post
Chainsmokers
Future
Gaga Lady
Ed Sheeran
Gucci Mane
Logic
Celia Cruz
Imagine Dragons
Kendrick Lamar
Black
Boy
Tom
Johnny
Sia
Lil Mosey
Kid
Lo
Jone
Cardi B
Nicki Minaj
Billie Eilish
Dua Lipa
Pitbull
Rihanna
Shakira


album_names
Best
Love
feat
Remix
Deluxe Edition
Greatest Hits
Remastered
Deluxe
Deluxe Version
Motion Picture
Song
Vol
EP
Night
Life
Time
One
Dance
Back
Hit
Let
Single
Remixes
Essential
Black
Dream
Light
Know
Music
Girl
Version
Special Edition
Picture Soundtrack
Good
Man
Star
Go
Live
Summer
Better
Heart
Original Motion
Persian Music
Classic