Track-Hit : Music Hit Predictor

Group 4

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

CSE343/ECE343 — Machine Learning Monsoon 2021

Track-Hit : Music Hit Predictor

Akhilesh Reddy
2019230

Gitansh Raj Satija
2019241

Vinay Pandey
2019288

Yatharth Taneja
2019346

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY **DELHI**

Motivation

Music is a great stress buster, helps in relaxation, and elevates our mood. Songs have different lyrics and sentiments attached, unique audio features which leave an impact on the audience. While some songs become a sensation, some fail to leave their mark. With so many new artists coming up, releasing a top-of-the-chart song is all the more difficult.

There have been different studies to predict the success of a song based on its audio features, lyrics, and emotion. Some researchers have also tried to study the artist’s impact, releasing date, and other metadata surrounding the song

Motivation

Our focus is only to analyze and study the impact of acoustic features on song popularity when the song’s artist is not known. This will be helpful, especially for newer artists who do not have any previous hits and want to predict the success of a song based on its physical characteristics. The model can also be helpful for music producers and investors looking to find potential talent

We also present a new dataset containing information only on Indian Songs; analyze the difference in the impact of acoustic features on song popularity in India as opposed to globally

Literature review

The paper Dance Hit Song Prediction explores the possibility of predicting dance hit songs by using basic musical features and capturing advanced features that model the temporal aspects. The research in the paper categorized different characteristics of a song into three different features, namely, meta-information(like artist name, artist location), basic analyzer features(like acoustics, danceability, energy), and temporal features (timbre and beat difference). A total of five models, namely, Decision Trees, Ripper Ruleset, Naive Bayes, Logistic Regression, and Support Vector Machines, were built for each dataset using diverse classification techniques and tested through 10-fold cross-validation.

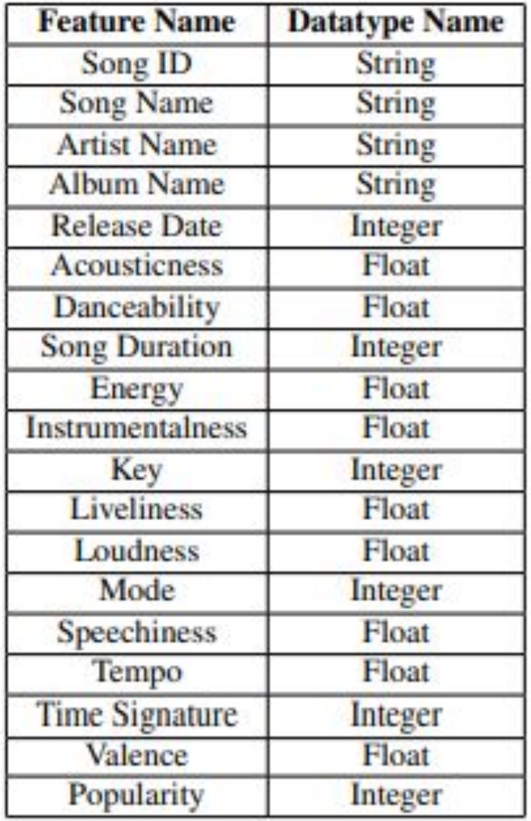
Literature review

Musical trends and predictability of success in contemporary songs in and out of the top charts takes into account the acoustic features as well as the correlation between the songs success and the artists associated by considering songs released in the UK between 1985 and 2015 to understand the dynamics of success and study various multi-decadal trends. A random forest has been used to predict success. The paper deals with both the ‘musical’ factors and the ‘socio-economic’ factors such as popularity (‘superstar’) status and the demographics of the artist(s), the level of promotion, the label/company and other such meta information, and their correlation with the song’s success.

Literature review

Hit Song Prediction for Pop Music By SIAMESE CNN with ranking loss explores predicting the popularity via training the model based on hit scores and relative ranking through a convolutional neural network instead of regression/classification models. Using euclidean loss and pairwise ranking loss, they were able to devise relative ranking among songs, and by using a sampling method called A/B, they were able to predict results with high accuracy. The Siamese architecture optimizes the model parameters by jointly considering rating and ranking-based loss functions. The proposed model is a multi-objective Siamese CNN which optimizes both the mean squared error (MSE) in predicting (i.e. rating) the hit song scores for both songs and the pairwise ranking loss in deciding which one of the two is likely to have a higher score

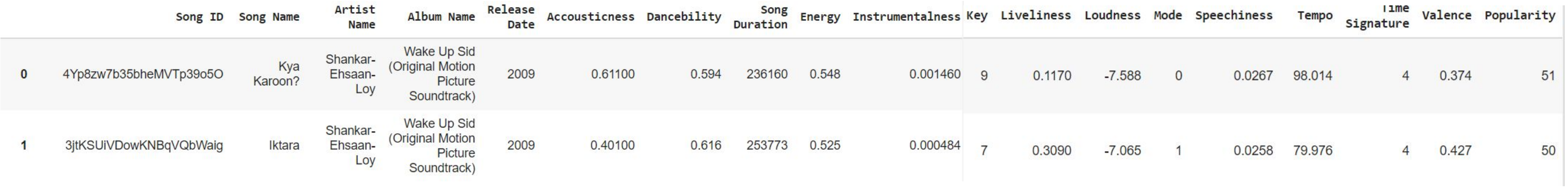
Dataset description :- Attributes



Feature Name	Datatype Name
Song ID	String
Song Name	String
Artist Name	String
Album Name	String
Release Date	Integer
Acousticness	Float
Danceability	Float
Song Duration	Integer
Energy	Float
Instrumentalness	Float
Key	Integer
Liveliness	Float
Loudness	Float
Mode	Integer
Speechiness	Float
Tempo	Float
Time Signature	Integer
Valence	Float
Popularity	Integer

The dataset contains some metadata and various audio features about the songs. The Meta-Informative features include track id, track name, artist name and album name. The Acoustic features include the measure of acousticness, danceability, energy, loudness among others.

Dataset description :- Attributes



	Song ID	Song Name	Artist Name	Album Name	Release Date	Accousticness	Dancebility	Song Duration	Energy	Instrumentalness	Key	Liveliness	Loudness	Mode	Speechiness	Tempo	Time Signature	Valence	Popularity
0	4Yp8zw7b35bheMVTp39o5O	Kya Karoon?	Shankar-Ehsaan-Loy	Wake Up Sid (Original Motion Picture Soundtrack)	2009	0.61100	0.594	236160	0.548	0.001460	9	0.1170	-7.588	0	0.0267	98.014	4	0.374	51
1	3jtKSUiVDowKNBqVQbWaig	Iktara	Shankar-Ehsaan-Loy	Wake Up Sid (Original Motion Picture Soundtrack)	2009	0.40100	0.616	253773	0.525	0.000484	7	0.3090	-7.065	1	0.0258	79.976	4	0.427	50

Dataset description :- Preprocessing

For preparing the dataset for the model, we removed all the features(fields) which were unique to the song (such as Song ID, Song Name) and information related to the artist.

We analyzed the dataset and realized a huge chunk (around 30%) of the songs in the dataset had 0 popularity.

Further, we visualized the distribution in popularity of songs against all the features and removed all the extreme outliers like songs of duration more than 8 minutes.

We normalized the audio features using min-max normalization on the column.

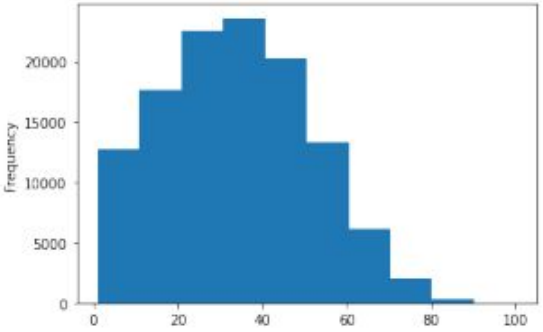
We were left with 118,795 different songs which were composed by 14,222 different artists.

Indian Data set

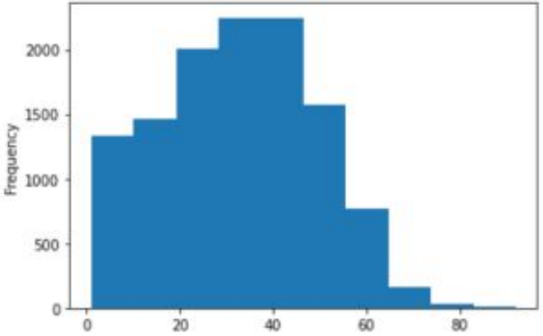
We collected data on songs made by Indian Artists by querying track information based on ’Bollywood’, ’India Indie’, ’Tollywood’ genres and by fetching online playlists consisting of songs by Indian Artists.

We also collected data from our original dataset, whose artists were Indian. Following preprocessing techniques similar to our original dataset, we were able to extract 11,844 different songs by 1,394 different artists

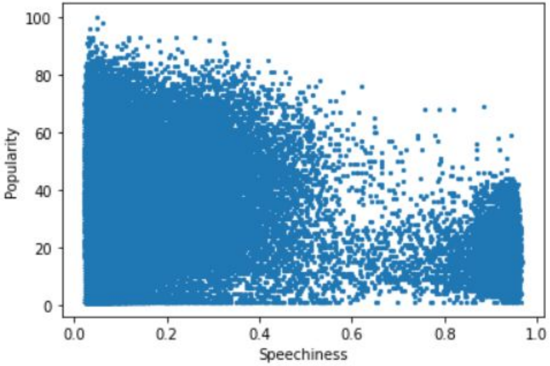
Dataset description :- Visualisation

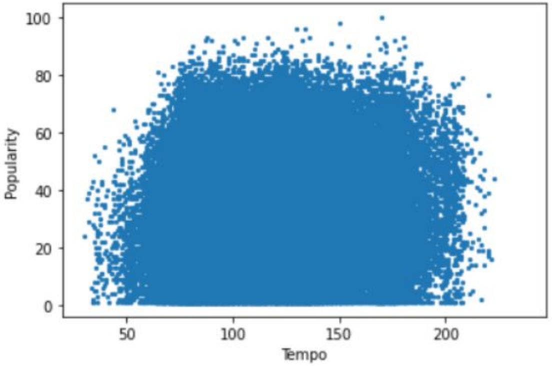


(a) Global Dataset

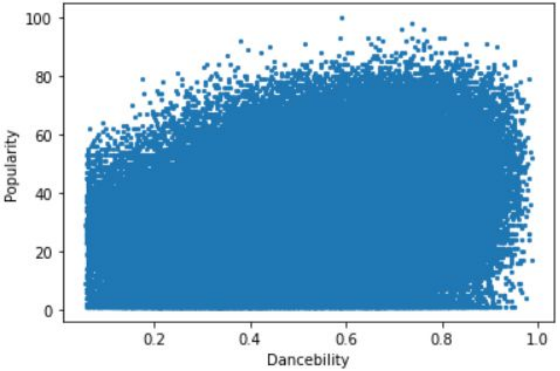


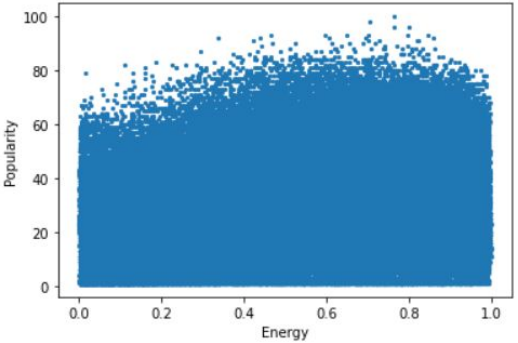
(b) Indian Dataset

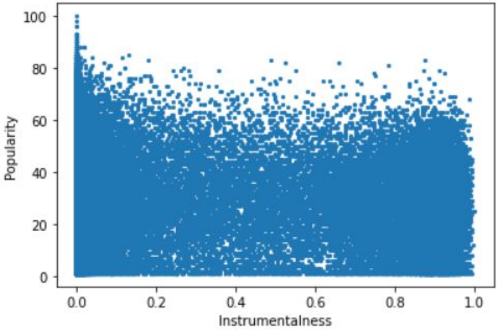


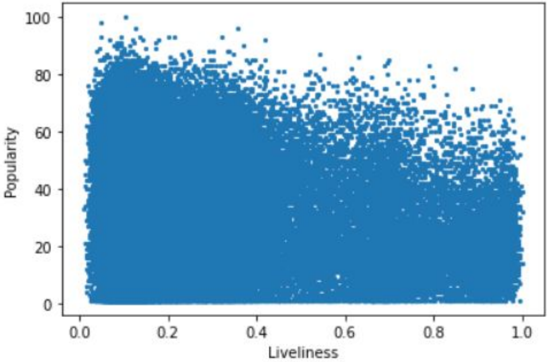


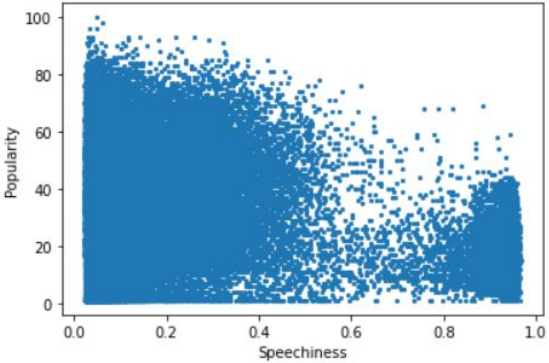
Dataset description :- Visualisation











Dataset description :- Feature Selection

Fisher's Score

It is a supervised algorithm that returns the rank of each feature based on the fisher's score. This rank can be used for feature selection among different variables. The higher the rank of the variable, the more useful is the feature in predicting the target variable.

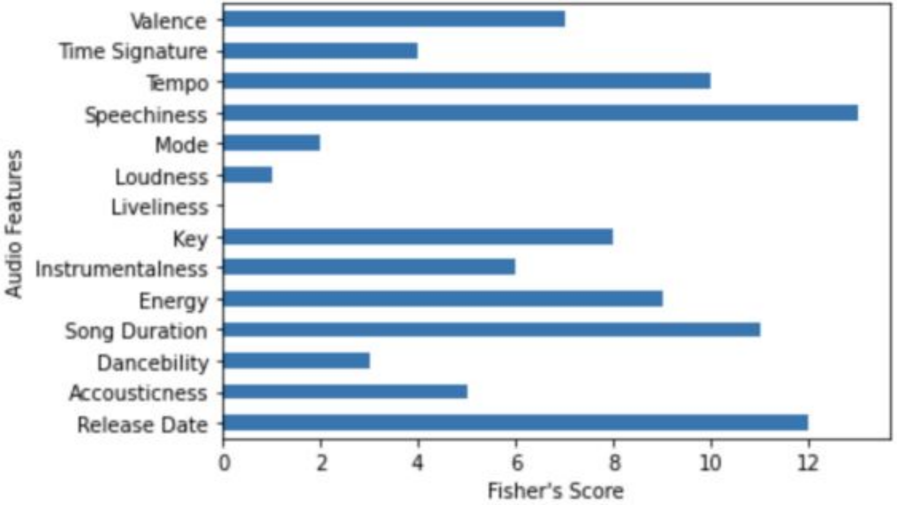


Figure 2: Fisher's Score

Dataset description :- Feature Selection

Information Gain

Information gain for each variable is calculated in the context of the target variable and is used for feature selection. It is calculated by subtracting the weighted entropy for each variable from the original entropy. The higher the information gain, the greater is the decrease in entropy.

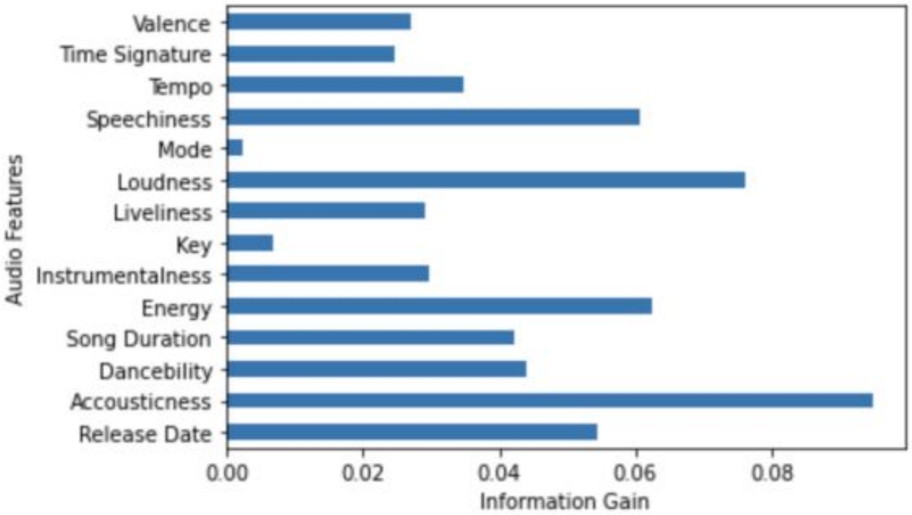
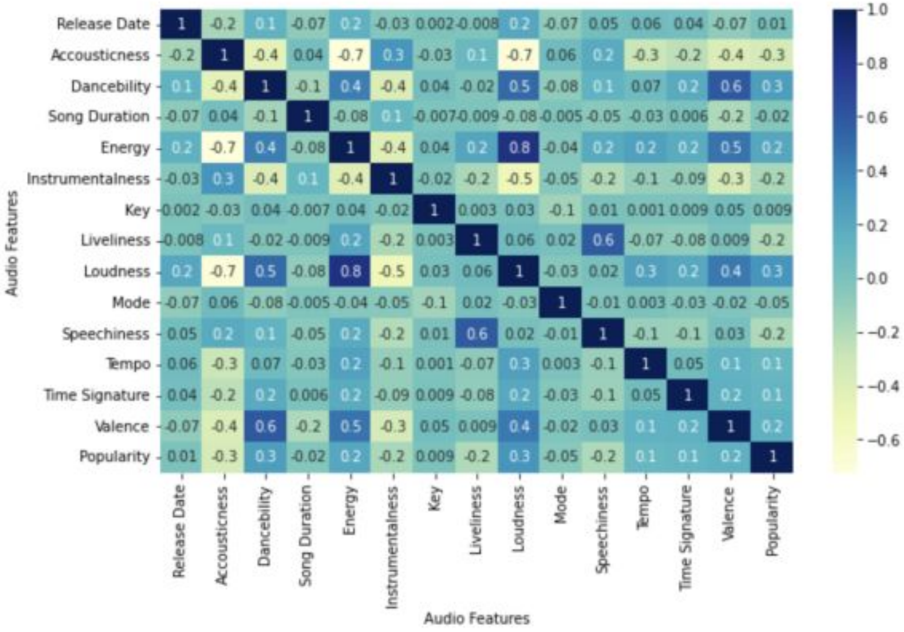


Figure 3: Information Gain

Dataset description :- Feature Selection

Correlation Coefficient

It is a measure of the linear relationship between 2 or more variables. It helps in predicting a variable based on the value of another variable. It helps in deciding the features which are largely correlated with each other and can be dropped after determining their correlation with the target variable and therefore help in feature selection.



	Release Date	Accousticness	Dancebility	Song Duration	Energy	Instrumentalness	Key	Liveliness	Loudness	Mode	Speechiness	Tempo	Time Signature	Valence	Popularity
Release Date	1	-0.2	0.1	-0.07	0.2	-0.03	0.002	-0.008	0.2	-0.07	0.05	0.06	0.04	-0.07	0.01
Accousticness	-0.2	1	-0.4	0.04	-0.7	0.3	-0.03	0.1	-0.7	0.06	0.2	-0.3	-0.2	-0.4	-0.3
Dancebility	0.1	-0.4	1	-0.1	0.4	-0.4	0.04	-0.02	0.5	-0.08	0.1	0.07	0.2	0.6	0.3
Song Duration	-0.07	0.04	-0.1	1	-0.08	0.1	-0.007	-0.009	-0.08	-0.005	-0.05	-0.03	0.006	-0.2	-0.02
Energy	0.2	-0.7	0.4	-0.08	1	-0.4	0.04	0.2	0.8	-0.04	0.2	0.2	0.2	0.5	0.2
Instrumentalness	-0.03	0.3	-0.4	0.1	-0.4	1	-0.02	-0.2	-0.5	-0.05	-0.2	-0.1	-0.09	-0.3	-0.2
Key	0.002	-0.03	0.04	-0.007	0.04	-0.02	1	0.003	0.03	-0.1	0.01	0.001	0.009	0.05	0.009
Liveliness	-0.008	0.1	-0.02	-0.009	0.2	-0.2	0.003	1	0.06	0.02	0.6	-0.07	-0.08	0.009	-0.2
Loudness	0.2	-0.7	0.5	-0.08	0.8	-0.5	0.03	0.06	1	-0.03	0.02	0.3	0.2	0.4	0.3
Mode	-0.07	0.06	-0.08	-0.005	-0.04	-0.05	-0.1	0.02	-0.03	1	-0.01	0.003	-0.03	-0.02	-0.05
Speechiness	0.05	0.2	0.1	-0.05	0.2	-0.2	0.01	0.6	0.02	-0.01	1	-0.1	-0.1	0.03	-0.2
Tempo	0.06	-0.3	0.07	-0.03	0.2	-0.1	0.001	-0.07	0.3	0.003	-0.1	1	0.05	0.1	0.1
Time Signature	0.04	-0.2	0.2	0.006	0.2	-0.09	0.009	-0.08	0.2	-0.03	-0.1	0.05	1	0.2	0.1
Valence	-0.07	-0.4	0.6	-0.2	0.5	-0.3	0.05	0.009	0.4	-0.02	0.03	0.1	0.2	1	0.2
Popularity	0.01	-0.3	0.3	-0.02	0.2	-0.2	0.009	-0.2	0.3	-0.05	-0.2	0.1	0.1	0.2	1

Figure 4: Correlation Coefficient

Dataset description :- Feature Selection Inference

After observing the results of all the above techniques we arrived at the conclusion to drop the features ‘Mode’, ‘Key’ and ‘Time Signature’. The reason is that the features ‘Mode’, ‘Key’ and ‘Time Signature’ are providing relatively lower information as compared to other features

Methodology

The main objective is to classify songs into popular or unpopular sets. The secondary objective is to also assign a popularity score to songs on the basis of musical factors.The initial steps involve preprocessing the data using data visualization and feature selection techniques.After the preprocessing stage, different models have been tried for both classification and regression.

Methodology

Regression

Applying Regression techniques on the data to assign a popularity score to songs on the basis of musical factors. The regression models trains on the training set and tries to establish a relationship between the features of a song and their popularity. The model then predicts a popularity score on the testing data. The models have been trained on a training dataset of size 95000+ global songs.

We make use of Linear, Lasso, and Ridge Regression models along with various other regressors like XGBoost, AdaBoost, Decision Tree and Random Forest Regressors.

Methodology

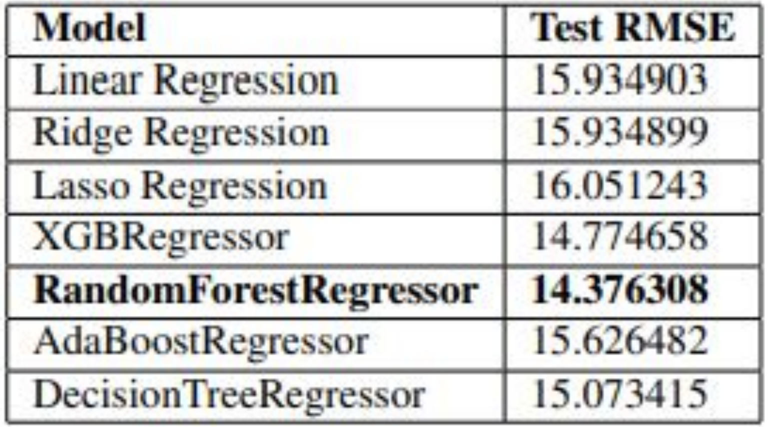
Classification

We applied binary classification to segregate data into popular and non-popular songs. Songs have been classified into a binary category of successful/non successful based on their popularity score.

We have used logistic regression,Gaussian Naive Bayes, K-Nearest Neighbours, Decision Tree, Random Forest, SVM, Neural Network and Gradient Boosting Classifiers in the project

Results/Analysis/Conclusion [Regression]

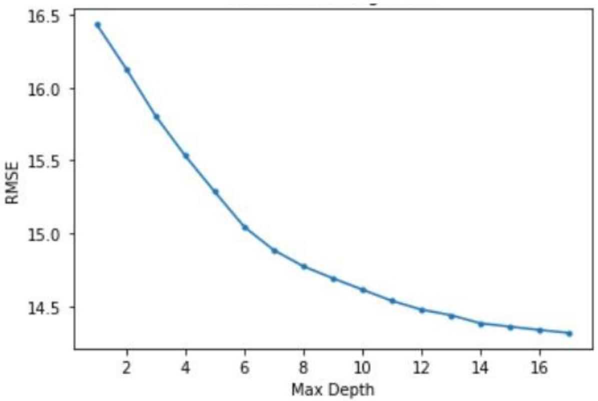
We have received satisfying results comparable to state of art models in the preliminary models that we have trained. We observe that Random Forest Regressor gives us the least RMSE (best performance) of 14.376.



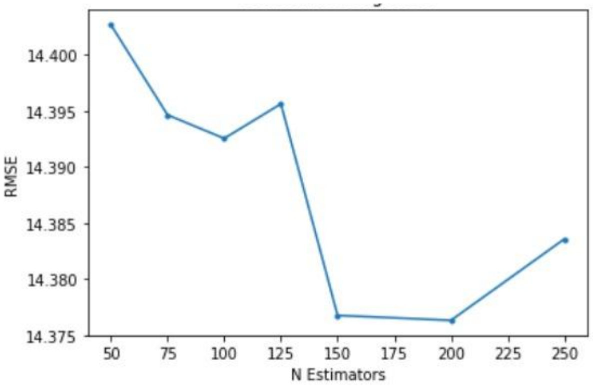
Model	**Test RMSE**
Linear Regression	15.934903
Ridge Regression	15.934899
Lasso Regression	16.051243
XGBRegressor	14.774658
RandomForestRegressor	**14.376308**
AdaBoostRegressor	15.626482
DecisionTreeRegressor	15.073415

Results/Analysis/Conclusion [Regression]

We selected Random Forest Regressor for hyperparameter tuning. We found the best performance to be at max depth=14 with 200 estimators.

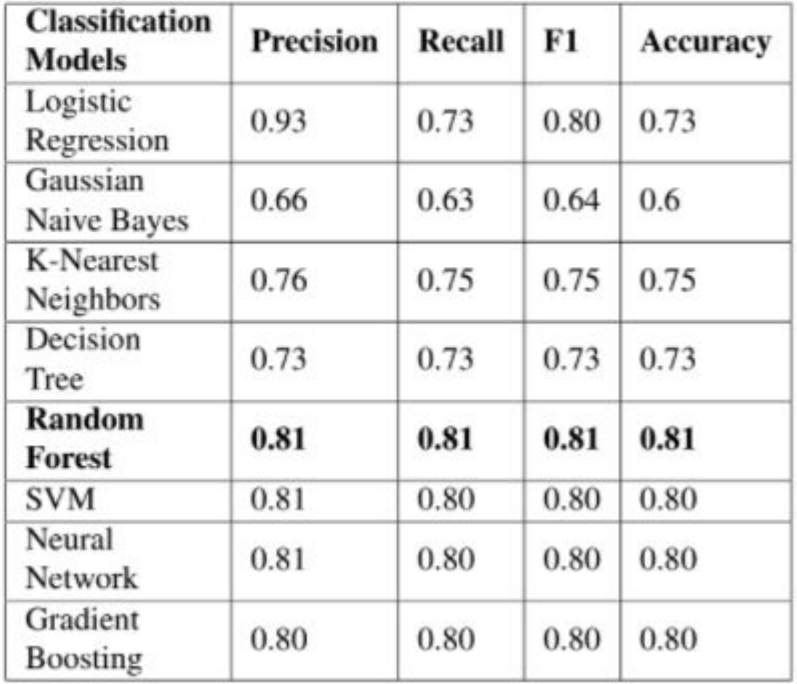


RMSE for Different Max Depths

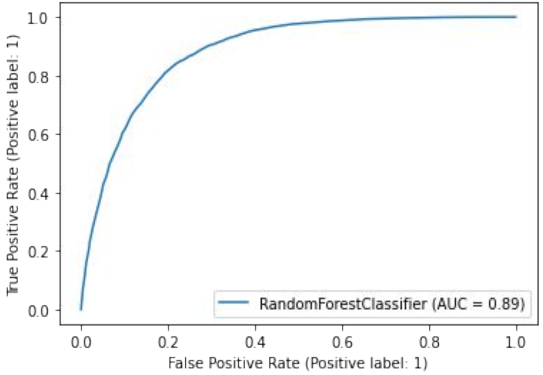


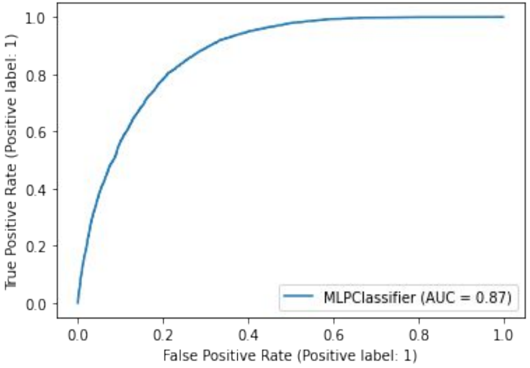
RMSE for Different Number of Estimators at max depth=14

Results/Analysis/Conclusion [Classification]



Classification Models	**Precision**	**Recall**	**F1**	**Accuracy**
Logistic Regression	0.93	0.73	0.80	0.73
Gaussian Naive Bayes	0.66	0.63	0.64	0.6
K-Nearest Neighbors	0.76	0.75	0.75	0.75
Decision Tree	0.73	0.73	0.73	0.73
Random Forest	**0.81**	**0.81**	**0.81**	**0.81**
SVM	0.81	0.80	0.80	0.80
Neural Network	0.81	0.80	0.80	0.80
Gradient Boosting	0.80	0.80	0.80	0.80





Indian Dataset

We observe a notable difference in the performance of model trained on our global dataset when we tested it on Indian Song Dataset. We predict popularity scores on our best Regression Model (Random Forest Regressor) for the Indian Songs and observed an RMSE of 16.29 (considerably higher than that of Global Dataset). On training a separate model using Indian Dataset Only, we get an RMSE of 14.34 indicating that features have different importance in the Indian Dataset. Thus we conclude that on analyzing indigenous songs we can have better predictions.

Timeline

Week 1-2: Dataset Generation ✓

Week 3: Pre-processing of Data and its Visualization. ✓

Week 4-5: Feature extraction, feature selection, and correlation ✓

Week 6-8: Training/Testing different Models (like Logistic Regression, Random Forests, etc.) ✓

Week 9: Analyzing the performance of different models. ✓

Week 10: Fine-Tuning the model followed by deploying the system ✓

Week 11: Final Report Writing + Buffer ✓

Individual Contributions

- **Akhilesh Reddy:** Literature review, Data Extraction and Collection by using available datasets online, Applying models Linear regression , Logistic Regression, Ridge and Lasso regularisation , Decision Tree, Random Forest,K-Nearest Neighbours, and Analysis and inference of the data
- **Gitansh Raj Satija:** Literature review, Data Extraction and Collection by fetching various playlists online, Indian Dataset Generation, Data cleaning and Preprocesssing, and Analysis and inference of the data, Gaussian NB, Regressor Models, Indian Dataset Analysis
- **Vinay Pandey:** Literature review, Data Extraction and Collection : Using spotify api based on year, Feature selection using Information gain , correlation coefficient and mean absolute difference, and Analysis and inference of the data
- **Yatharth Taneja:** Literature review, Data Extraction and Collection : Using spotify api based on genre , Feature selection using l1 regularisation, fisher index, Information gain , correlation coefficient , Applying models like Neural Network, SVM,Gradient Boosting, and Analysis and inference of the data