Engineering Universe for scientific Research and management
ISSN (Online): 2319-3069
Vol. XIII Issue VI
June 2021

COMPARATIVE ANALYSIS OF PREDICTING MOVIES SUCCESS USING MACHINE-LEARNING TECHNIQUE

Jyoti Jagriti[1], Prof. Chetan Agarwal[2], Prof. Priyanka Parihar[3]
Department of Computer Science Engineering RITS, Bhopal, (MP), India
[1]Jyotijagrti510@gmaill.com, [2]chetan.agrawal12@gmail.com, [3]priyanka.parihar28113@gmail.com

ABSTRACT: The movie success factors depend on the critics, storyline, hero's, music etc. To predict the movie success various data mining and machine learning techniques such as Guassian NB, Multinomial NB, Bernoulli NB, KNeighnors Classifier, Decision Tree, Logistic regression has been developed but, in this work, we use random forest classifier for the prediction of movie success with reduced cost and schedule. The random forest classifier selects the dataset randomly from the available dataset and the generate the decision tree of the selected dataset and then apply the voting on the prediction results and whose score and accuracy will be maximum that will indicates the success of movie. For the sample of IMDb dataset, we use online resource of kaggle and the experimental results is generated from the widely used machine learning programming language Python which helps in the analysis of the proposed methodology. The performance of proposed methodology is measured using the parameters such as Score, accuracy, precision, recall value, F1 score, mean absolute error and mean square error. The comparative analysis of the proposed methodology is done among the existing approach Guassian NB, Multinomial NB, Bernoulli NB, KNeighnors Classifier, Decision Tree, Logistic regression. The score and accuracy value of our proposed methodology is 70% while other is less. Similarly, the F1 score, precision and recall value of proposed methodology is 69%, 66% and 66% while the Bernoulli NB, Multinomial NB and logistic regression is comparatively very less. Similarly, the comparative analysis of proposed and existing approach is done using mean absolute error and mean square error and the value is 32% and 36%, which is very less. These results of proposed methodology improve the success rate of movie success.

Keywords: Random Forest, Logistic Regression, Prediction, IMDB, Gaussian NB, Multinomial NB, Precision, Python

1. INTRODUCTION

Movie industry is a huge sector for investment but larger business sectors have more complexity and it is hard to choose how to invest. Big investments come with bigger risks. The CEO of Motion Picture Association of America (MPAA) J. Valenti mentioned that 'No one can tell you how a movie is going to do in the marketplace. Not until the film opens in darkened theatre and sparks fly up between the screen and the audience' [1]. As movie industry is growing too fast day by day, there are now a huge amount of data available on the internet, which makes it an interesting field for data analysis. Predicting a movie success is a very complex task to do. The definition of a movie success is relative, some movies are called successful based on its worldwide gross income, and some movies may not shine in business part but can be called successful for good critic's review and popularity. A movie revenue depends on various components such as cast acting in a movie, budget for the making of the movie, film critics review, rating for the movie, release year of the movie, etc. Because of these multiple components there is no formula that helps us to provide analysis for predicting how much

revenue a particular movie will be generating. However, by analyzing the revenues generated by previous movies; a model can be built which can help us predict the expected revenue for a particular movie. Such a prediction could be very useful for the movie studios which will be producing the movie so they can decide on different expenses like artist compensations, advertising of the movie, promotions in various cities, etc. accordingly. Moreover, it allows investors to predict an expected return-on-investment (ROI). In addition, it will be useful for many movie theatres to estimate the revenues they would generate from screening a particular movie.

There are many elements impacting the movies of a film, for instance, number of screens for the motion picture, publicizing, time, actors, directors, budget, genre and number of motion pictures that are released in specific duration or time frame and even within past years, months and days. Motion pictures or movies have turned into a vital piece of our lives as manner for passion, compassion, enthusiasm and entertainment [2]. Films have likewise been a noteworthy medium for culture trade between various nations and districts and are therefore an irreplaceable resource for the world. Given this, the motion picture industry has turned into a business and it has enormous market benefit and potential. As an outcome, the information and research about the motion picture industry is getting to be noticeably more profound. Capacity to precisely foresee the movies potential returns over investment based on total cost of ownership for a motion pictures will enable the film line decides the publicity cost and time of demonstrating the motion picture to expand the benefit and returns to investment made therein. The issue of foreseeing the movies gross of a releasing film has been broadly handled in the past from a measurable perspective.

There are many elements impacting the movies of a film, for instance, number of screens for the motion picture, publicizing, time, actors, directors, budget, genre and number of motion pictures that are released in specific duration or time frame and even within past years, months and days. Under this scheme, we are centering to build up a strategy in light of affiliation using machine learning to upgrade the forecast of success of the movie. In this article, random forest classifier is used for prediction of movie success and this classifier is compared with the existing machine learning techniques. With the increasing in huge amount of data a good data analysis is required and machine learning approach is one of them. Nowadays this approach is extensively used for data analysis purpose. The comparative analysis of proposed and existing technique is done using various performance measuring parameters such as Precision, recall, F1 score, mean square error, mean absolute error, root mean square error etc. and it is analyzed that our proposed approach outperforms than the existing approach. It means our approach is much better in the prediction of movie success rate.

2. RELATED WORK

Kumar, and Kumar (2018) proposed framework predicts the achievement of a motion picture in light of its gainfulness by utilizing chronicled information from different sources. Utilizing informal community examination and content mining methods, the framework naturally separates a few gatherings of highlights, including "who" are on the best composition (actor and director) what a film is about, "when" a motion picture will be released, and in addition "semi variety" highlights that match "who" with "what", and "when" with "what". Examination comes about with motion pictures amid years' time frame demonstrated that the framework beats benchmark techniques by a substantial edge in anticipating motion picture productivity. Novel highlights we proposed likewise made extraordinary commitments to the expectation. Moreover, to planning a choice emotionally supportive network with reasonable utilities, our investigation of key factors for motion picture productivity may likewise have suggestions for hypothetical research on group execution and the achievement of imaginative work [2].

Latif and Afzal (2016) used IMDB for our experimentation. They created dataset and then transformed it and applied machine learning approaches to build efficient models that can predict the movies popularity. Performing data mining on IMDB is a hard task because of so many attributes related to a movie and all in different dimensions with lots of noisy data and missing fields. After performing

classification, they have found out that their best results are achieved through simple logistic and logistic regression at around 84 %. The attributes that contributed the most to information are metascore and number of votes for each movie, Oscar awards won by the movies and the number of screens the movie is going to be screened [3].

Chaudhari et al. (2016) developed a tool, which can predict the success of movie being a hit or flop. As this factor is important for everyone involved in the movie, for example: If a movie is flop, it exacerbates the image of actor or director. The tool will use searching algorithms and then use of bespoke system to predict the percentage of success of movie which is yet to be released. Their analysis of the data collected from various resources like IMDb, Kaggle. They gather a series of interesting facts and relationships using a variety of data mining techniques such as Bayes Classification Algorithm, Decision Tree etc. Subsequently, a classifier is learned and used to classify new movies with respect to their predicted box-office collection. Experimental results showed that the proposed approach improved the classification accuracy as compared to a fully independent setting. In particular, they discovered the rate of success with respect to various parameters such as language, country, budget, Facebook likes of the actors and actresses etc and focus on relevant details such as the relationship between the budget of the movie and rating of the movie, language and rating, facebook likes and rating etc. The data mining techniques used will enable us to uncover information which will both confirm or disproved common assumptions about movies, and also allow us to predict the success of a future film given select information about the film before its release [4].

Meenakshi et al., (2018) developed a system based upon data mining techniques that may help in predicting the success of a movie in advance thereby reducing certain level of uncertainty. An attempt is made to predict the past as well as the future of movie for the purpose of business certainty or simply a theoretical condition in which decision making the success of the movie is without risk, because the decision maker has all the information about the exact outcome of the decision, before he or she makes the decision. With over two million spectators a day and films exported to over 100 countries, the impact of Bollywood film industry is formidable. They gathered a series of interesting facts and relationships using a variety of data mining techniques [7]. In particular, they concentrated on attributes relevant to the success prediction of movies, such as whether any particular actors or actresses are likely to help a movie to succeed. They additionally reported on the techniques used, giving their implementation and utility. Additionally, they found some attention-grabbing facts, such as the budget of a movie isn't any indication of how well-rated it'll be, there's a downward trend within the quality of films over time, and also the director and actors/actresses involved in the movie [5].

Quader et al. (2017) proposed a decision support system for movie investment sector using machine learning techniques. This research helps investors associated with this business for avoiding investment risks. The system predicted an approximate success rate of a movie based on its profitability by analyzing historical data from different sources like IMDb, Rotten Tomatoes, Box Office Mojo and Metacritic. Using Support Vector Machine (SVM), Neural Network and Natural Language Processing the system predicts a movie box office profit based on some pre-released features and post-released features. They showed Neural Network gives an accuracy of 84.1% for pre-released features and 89.27% for all features while SVM has 83.44% and 88.87% accuracy for pre-released features and all features respectively when one away prediction is considered. Moreover, they figured out that budget, IMDb votes and no. of screens are the most important features which play a vital role while predicting a movie’s box-office success [6].

3. MACHINE LEARNING CLASSIFICATION

Machine Learning is a perception which consents the machine to acquire from examples and experience, and that moreover deprived of being overtly programmed. Thus, instead of you scripting the code, what you do is you feed data to the generic technique, and the technique/ machine builds the logic based on the given data.[13,8] It permits the computers system or the machines to construct data-driven decisions rather than being explicitly programmed for carrying out a certain task. These programs or techniques are designed in a

way that they learn and improve over time when are exposed to new data.

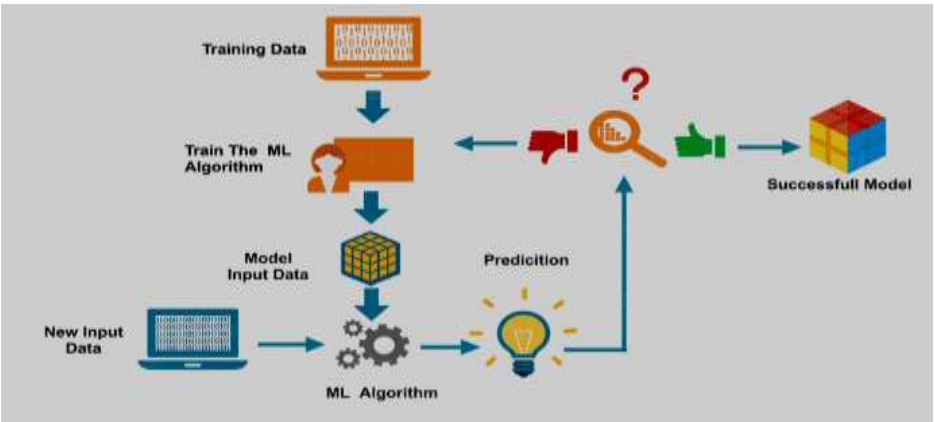


Fig.1: Working steps of Machine learning technique

Machine Learning Technique is trained using a training data set to create a model. When new input data is introduced to the ML technique, it makes a prediction based on the model. The prediction is evaluated for accuracy and if the accuracy is acceptable, the Machine Learning technique is deployed. If the accuracy is not acceptable, the Machine Learning technique is trained again and again with an augmented training data set.[8] This is just a very high-level example as there are many factors and other steps involved.

3.1 Types of Machine Leaning Techniques

There are three important types of Machine Learning Techniques such as supervised learning, unsupervised learning and reinforcement learning, which we are discussing in detail:

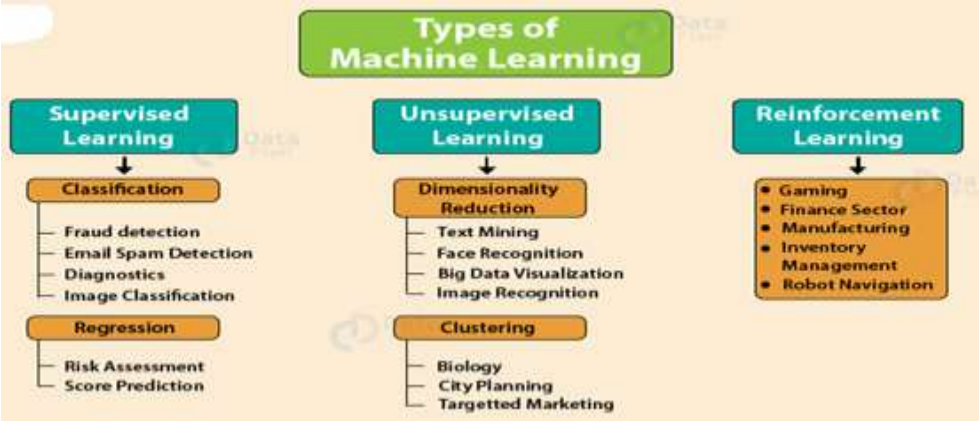


Fig.2: Classification of Machine Learning Techniques

3.1.1 Supervised Learning

Supervised Learning is the most popular paradigm for performing machine learning operations. It is widely used for data where there is a precise mapping between input-output data. The dataset, in this case, is labeled, meaning that the algorithm identifies the features explicitly and carries out predictions or classification accordingly. [9] As the training period progresses, the algorithm is able to identify the relationships between the two variables such that we can predict a new outcome. Resulting Supervised learning algorithms are task-oriented. As we provide it with more and more examples, it is able to learn more properly so that it can undertake the task and yield us the output more accurately. Some of the algorithms that come under supervised learning are as follows:

Linear regression, random forest, support vector machine, artificial intelligence [10], etc. There are two main types of supervised learning problems: they are classification that involves predicting a class label and regression that involves predicting a numerical value [10].

- Classification: Supervised learning problem that involves predicting a class label.
- Regression: Supervised learning problem that involves predicting a numerical label.

Both classification and regression problems may have one or more input variables and input variables may be any data type, such as numerical or categorical [45].

3.1.2 Unsupervised Learning

Unsupervised machine learning holds the advantage of being able to work with unlabeled data. This means that human labor is not required to make the dataset machine-readable, allowing much larger datasets to be worked on by the program. The model learns through observation and finds structures in the data. Once the model is given a dataset, it automatically finds patterns and relationships in the dataset by creating clusters in it [11]. In supervised learning, the labels allow the algorithm to find the exact nature of the relationship between any two data points. However, unsupervised learning does not have labels to work off of, resulting in the creation of hidden structures. Relationships between data points are perceived by the algorithm in an abstract manner, with no input required from human beings. The creation of these hidden structures is what makes unsupervised learning algorithms versatile. Instead of a defined and set problem statement, unsupervised learning algorithms can adapt to the data by dynamically changing hidden structures.[11] This offers more post-deployment development than supervised learning algorithms. What it cannot do is add labels to the cluster, like it cannot say this a group of apples or mangoes, but it will separate all the apples from mangoes. Suppose we presented images of apples, bananas and mangoes to the model, so what it does, based on some patterns and relationships it creates clusters and divides the dataset into those clusters. Now if a new data is fed to the model, it adds it to one of the created clusters. The example of unsupervised learning is k-mean clustering, principle component analysis, SVD, FP-growth etc [16]. There are many types of unsupervised learning, although there are two main problems that are often encountered by a practitioner: they are clustering that involves finding groups in the data and density estimation that involves summarizing the distribution of data [11].

- Clustering: Unsupervised learning problem that involves finding groups in data.
- Density Estimation: Unsupervised learning problem that involves summarizing the distribution of data.

3.1.3 Reinforcement Learning

Reinforcement learning directly takes inspiration from how human beings learn from data in their lives. It features an algorithm that improves upon itself and learns from new situations using a trial-and-error method. Favorable outputs are encouraged or 'reinforced', and non-favorable outputs are discouraged or 'punished'. Based on the psychological concept of conditioning, reinforcement learning works by putting the algorithm in a work environment with an interpreter and a reward system. In every iteration of the algorithm, the output result is given to the interpreter, which decides whether the outcome is favorable or not [12]. In case of the program finding the correct solution, the interpreter reinforces the solution by providing a reward to the algorithm. If the outcome is not favorable, the algorithm is forced to reiterate until it finds a better result. In most cases, the reward system is directly tied to the effectiveness of the result [12]. In typical reinforcement learning use-cases, such as finding the shortest route between two points on a map, the solution is not an absolute value. Instead, it takes on a score of effectiveness, expressed in a percentage value. The higher this percentage value is, the more reward is given to the algorithm. Thus, the program is trained to give the best possible solution for the best possible reward [17]. This simple feedback reward is known as a reinforcement signal.

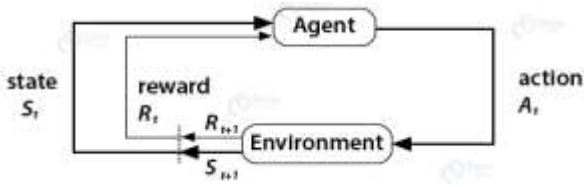


Fig. 3: Example of reinforcement learning

4. PROPOSED WORK

Most of the earlier work are predicting the IMDB Score. Means they are using different attribute to predict the IMDB Score hence they taken this as a Regression Problem. But our main focus is to predict the success rate. So, we divide the whole Range of IMDB in five different categories so that we can take it as Classification Problem and hence we can also increase the past scores. (To get a good Score in Regressor requires proper dataset but in classification a good score can easily raise). So, we had classified the movies in a category followed as:

Table 1: Range of IMDb rating

IMDB Rating	Score (My system of scoring)
0-2	0
2-4	1
4-6	2
6-8	3
8 and so on	4

This is all about the workflow as it's all divided in the following steps which made this easy.

Fig.4: Work flow of Movie prediction

To predict the movies success rate, our methodology follows the subsequent steps which is discussing below:

Step1: Collecting database

In the first step we need dataset to work upon. As we are familiar with a website named as Kaggle, which is the best place for all kinds of datasets. So, talking about our aim we need a dataset which comprises of every single detail about the movie. Hence, we have the official; dataset of IMDB movies which is collection of every detail of around 4000 +n movies. Discussing about the attributes or the details of each movie, as we have in following attributes.

Table 2: Attributes of Movie dataset

color	director_name	num_critic_for_reviews	duration
director_facebook_likes	actor_3_facebook_likes	actor_2_name	actor_1_facebook_likes
gross	genres	actor_1_name	movie_title
num_voted_users	cast_total_facebook_likes	actor_3_name	facenumber_in_poster
plot_keywords	movie_imdb_link	num_user_for_reviews	language
country	content_rating	budget	title_year
actor_2_facebook_likes	imdb_score	aspect_ratio	movie_facebook_likes

So, with the help of this dataset we will try to implement a Classification model which can easily predict the Movie success rate as 0 with least and so on**.**

Step2: Cleaning Database

In this step we already had collected the database so this is the time where we need to filter or clean the dataset. So in this we take Care of few things as following. As cleaning the data is first task in ML & DS workflow. Without this we will face many issues in exploring the required terms. As cleaning just not actually means to clean the data. It exactly means filtering and modifying your data such that it is easy to explore, understand and model.

1. First task is to remove all the Nan or Empty values.
2. Then we need to handle the missing value.
3. Handling the Outliers

So, in this process we clean the data and make it ready for the Training purpose.

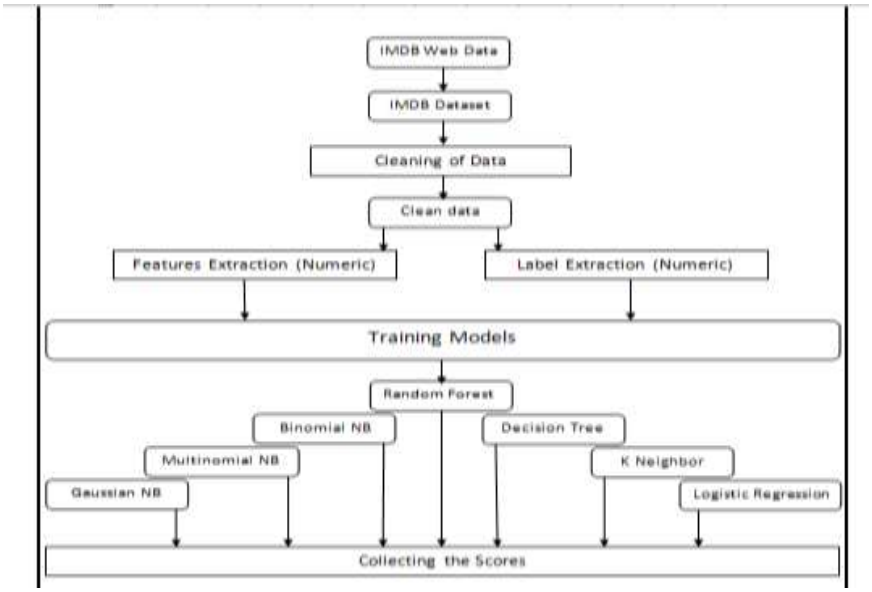


Fig.5 Data flow diagram of proposed work

Step3: Picking Features (Necessary)

In this step, we need to the features or we can say that we need to select the columns which we need to feed in the following model.

As in given dataset we have three types of data

1. Numerical data
2. Categorical Data
3. Composite data

As all the classifiers are best suited for the Numerical Values so we will be going to select all the numerical columns for the training purpose.

Step4: Training Different Models

From the last step we have training dataset and resulting IMDB scores now we need train different model or we can say that we need to train different classifiers so that prediction can be taken out. In this we are using following Models.

1. Random Forest Regressor
2. Logistic regression
3. Decision tree
4. K Neighbors Classifier
5. Gaussian Naïve Bayes
6. Multinomial Naïve Bayes
7. Binomial Naïve Bayes

Here we have used a few of best classifiers.

Step5: Printing all the Scores

In this step we printed all the scores of all used classifiers hence get to know that Random Forest is the best regressor.

Step6: Result

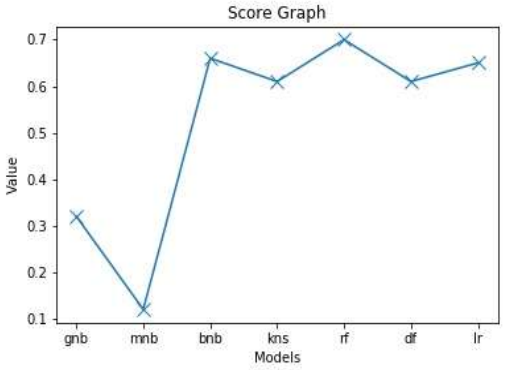
As a result, we final implemented a classifier which can predict the Success rate of any IMDB movie.

5. RESULT ANALYSIS

In this section of the dissertation we perform the result analysis on different measuring parameters like score, accuracy, precision, recall, f1-measure, mean absolute error and mean square error and comparison is done between the proposed methodology (random forest classifier), logistics regression and K neighbors' classifiers.

5.1 Comparison of Score

The score parameter is used prove the rating score to the movie and the comparative analysis of this parameter is done among different machine learning such as GuassianNB, MultinomialNB, BernoulliNB, KNeighnorsClassifier, Decision Tree, Logistic regression and our proposed method (random forest). The simulation results of our proposed method and existing method is shown in table 3 and it is 70% which is much more about the other exiting approach. The analysis is done using the comparison graph shown in figure 6 and it is found that our proposed method has higher value than the others. It means that the proposed method is more success in the prediction of movie success or hit.



Graph 6: Comparison of Score parameters

Table 3: Comparative analysis of score parameter between Random forest and existing method

Comparison of Score			
S. No.	Name of Classifier	Short Name	Score
1	GaussianNB	GNB	0.32
2	MultinomialNB	MNB	0.12
3	BernoulliNB	BNB	0.66
4	KNeighborsClassifier	KNS Model	0.61
5	Random Forest (Proposed Method)	RF	0.7
6	Decision Tree	DT	0.61
7	Logistic regression	LR	0.65

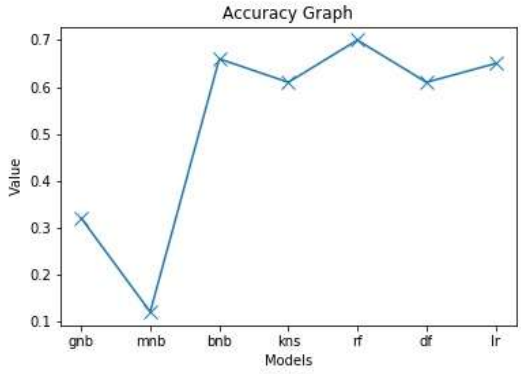
5.2 Comparison of Accuracy

This section presents the comparison of accuracy parameter to show the prediction accuracy of movie and the comparative analysis of this parameter is done among different machine learning such as Gaussian NB, Multinomial NB, Bernoulli NB, KNeighnors Classifier, Decision Tree, Logistic regression and our proposed method (random forest). The simulation results of our proposed method and existing method is shown in table 4 and it is 70% which is much more about the other exiting approach. The analysis is done using the comparison graph shown in figure 7 and it is found that our proposed method has higher value than the others. In this the value of accuracy is equivalent to score parameter. If score of the movie will high accuracy of the movie prediction will high. And it is analyzed that the proposed method is more success in the prediction of movie success or hit.

Table 4: Comparative analysis of accuracy parameter between Random forest and existing method

Comparison of Accuracy			
S. No.	Name of Classifier	Short Name	Accuracy
1	Gaussian NB	GNB	0.32
2	Multinomial NB	MNB	0.12
3	Bernoulli NB	BNB	0.66
4	KNeighbors Classifier	KNS Model	0.61

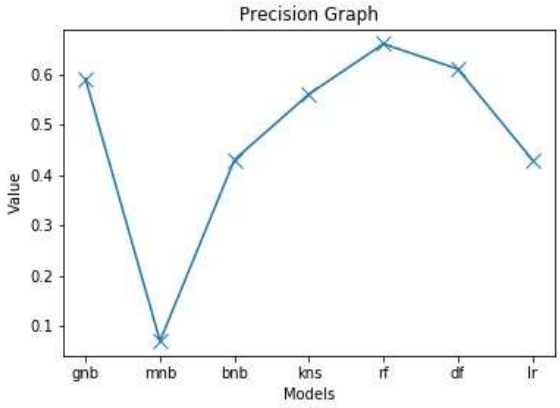
5	Random Forest (Proposed Method)	RF	0.7
6	Decision Tree	DT	0.61
7	Logistic regression	LR	0.65



Graph 7 : Comparison of accuracy parameters

5.3.3 Comparison of Precision Score

This section presents the comparison of precision score parameter to show the prediction accuracy of movie and the comparative analysis of this parameter is done among different machine learning such as Gaussian NB, Multinomial NB, Bernoulli NB, KNeighnors Classifier, Decision Tree, Logistic regression and our proposed method (random forest). The simulation results of our proposed method and existing method is shown in table 5 and it is 66% which is much more about the other exiting approach. The analysis of precision parameter is done using the comparison graph shown in figure 8 and it is found that our proposed method has higher value than the others. Due to the higher precision value it is analyzed that the proposed method is more success in the prediction of movie success or hit.



Graph 8 : Comparison of Precision Score

Table 5: Comparative analysis of Precision parameter between Random forest and existing method

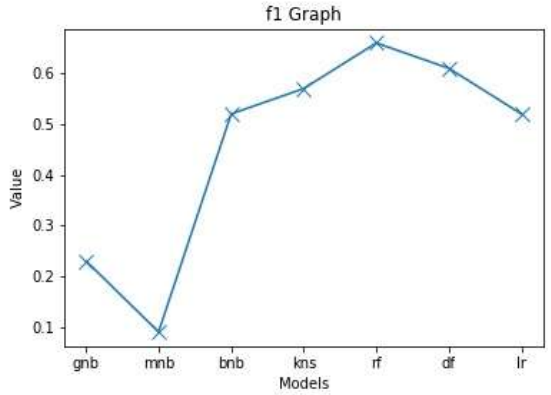
Comparison of Precision Score			
S. No.	Name of Classifier	Short Name	Precision Score
1	Gaussian NB	GNB	0.59
2	Multinomial NB	MNB	0.07
3	Bernoulli NB	BNB	0.43
4	KNeighbors Classifier	KNS Model	0.56
5	Random Forest (Proposed Method)	RF	0.66
6	Decision Tree	DT	0.61
7	Logistic regression	LR	0.43

5.3.4 Comparison of F1 Score

This section presents the comparison of F1 score parameter to show the prediction accuracy of movie and the comparative analysis of this parameter is done among different machine learning such as Gaussian NB, Multinomial NB, Bernoulli NB, KNeighnors Classifier, Decision Tree, Logistic regression and our proposed method (random forest). The simulation results of our proposed method and existing method is shown in table 6 and it is 66% which is much more about the other exiting approach. The analysis of F1 score parameter is done using the comparison graph shown in figure 9 and it is found that our proposed method has higher value than the others. Due to the higher F1 score value it is analyzed that the proposed method is more success in the prediction of movie success or hit.

Table 6: Comparative analysis of F1 score parameter between Random forest and existing method

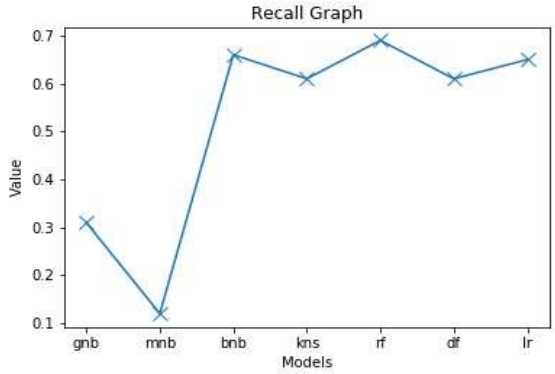
Comparison of F1 Score			
S. No.	Name of Classifier	Short Name	F1 Score
1	Gaussian NB	GNB	0.23
2	Multinomial NB	MNB	0.09
3	Bernoulli NB	BNB	0.52
4	KNeighbors Classifier	KNS Model	0.57
5	Random Forest (Proposed Method)	RF	0.66
6	Decision Tree	DT	0.61
7	Logistic regression	LR	0.52



Graph 9 : Comparison of F1 score parameters

5.5 Comparison of Recall Score

This section presents the comparison of Recall score parameter to show the prediction accuracy of movie and the comparative analysis of this parameter is done among different machine learning such as GuassianNB, MultinomialNB, BernoulliNB, KNeighnorsClassifier, Decision Tree, Logistic regression and our proposed method (random forest). The simulation results of our proposed method and existing method is shown in table 7 and it is 69% which is much more about the other exiting approach. The analysis of recall score parameter is done using the comparison graph shown in figure 10 and it is found that our proposed method has higher value than the others. Due to the higher recall value it is analyzed that the proposed method is more success in the prediction of movie success or hit.



Graph 10: Comparison of Recall score parameters

Table 7: Comparative analysis of Recall score parameter between Random forest and existing method

Comparison of Recall Score			
S. No.	Name of Classifier	Short Name	Recall Score
1	Gaussian NB	GNB	0.31
2	Multinomial NB	MNB	0.12
3	Bernoulli NB	BNB	0.66
4	KNeighbors Classifier	KNS Model	0.61
5	Random Forest (Proposed Method)	RF	0.69
6	Decision Tree	DT	0.61
7	Logistic regression	LR	0.65

6. CONCLUSION

In this research article, we use machine learning technique for our experimentation which have powerful classification like Gaussian NB, Multinomial NB, Bernoulli NB, KNeighnors Classifier, Decision Tree, Logistic regression and random forest for classification and regression. The experimental analysis of proposed method and existing method is done using the performance measuring parameters like precision, recall, F1 score, accuracy, etc. For the simulation of proposed method and existing method Python language uses which is easy implement and consume less computation time than other language. The result generated for proposed method after simulation for the accuracy and score parameter is 70% which is much more than the existing method. Similarly, the analysis of proposed and existing method is done using precision and recall parameter and the value of proposed methods is 66% which is also more than the existing method. Later the analysis of proposed and exiting method is done using F1 score and the value of F1 score of is 69% and 58% which is about 11% more than the existing method.

We also perform the analysis using the mean absolute error, mean square error and root mean squared error which is 32%, 36% and 61% of proposed method which is very less than the existing method. Based on these parameters we can easily predict the success rate of movies. Though the movie can be predicted using the available IMDb dataset but if data is not available like new hero, movie genre, lead actor's name, and the country where the movie will be originated prediction cannot be done. So, in future work, it is necessary to add these factors also which can improve the prediction rate of movie success. And also use the hybrid technique of machine learning by using the essential features of random forest and some other technique like logistic regression which will improve the accuracy for successful of movie prediction.

REFERENCE

[1]. B. R. Litman & H. Ahn. (1998). Predicting financial success of motion pictures. In B. R. Litman (Ed.), the motion picture mega-industry. Boston, MA: Allyn & Bacon Publishing, Inc.

[2]. Kumar and Kumar, "Predicting Movie Success or Failure using Linear Regression & SVM over Map-Reduce in Hadoop", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 6, Issue 6, June 2018.

[3]. "Prediction of Movies popularity Using Machine Learning Techniques", International Journal of Computer Science and Network Security, VOL.16 No.8, August 2016, pp 127-131.

[4] Chaudhari et al., "A Data Mining Approach to Language Success Prediction of A Feature Film", International Journal of Engineering Sciences & Management Research, 2016, pp. 1-9.

[5] Meenakshi et al., "A Data mining Technique for Analyzing and Predicting the success of Movie", Journal of Physics: Conf. Series 1000 (2018) 012100 doi :10.1088/1742-6596/1000/1/012100. Pp 1-9.

[6] Quader et al., "A Machine Learning Approach to Predict Movie Box-Office Success", 20th International Conference of Computer and Information Technology (ICCIT), 22-24 December, 2017.

[7] G. Diamantopoulos and M. Spann 2005 Performance analysis of CART and C5.0 using sampling techniques Advances in Computer Applications. 11 3, pp. 233–243.

[8] https://www.edureka.co/blog/what-is-machine-learning/

[9] Christopher M. Bishop, "Pattern Recognition and Machine Learning (Information Science and Statistics)", 2006. Page-3

[10] Russel, "Artificial Intelligence: A Modern Approach", January 1, 2015

[11] Hastie et al. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Second Edition (Springer Series in Statistics) , 2016, pp. 28. [online]. Available: https://www.amazon.com/Elements-Statistical-Learning-Prediction-Statistics.

[12] Richard S. Sutton et al., "Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning",2018, second edition, pp.-2. [Online]. Available : https://www.amazon.com/Reinforcement-Learning-Introduction-Adaptive-Computation.

[13] A. Smola and S. Vishwanathan, Introduction to Machine Learning. United Kingdom at the University Press, Cambridge, October 1, 2010.

[14] Online]. Available: www.analyticsvidhya.com.

[15] Sunpreet Kaur, Sonika Jindal "A Survey on Machine Learning Algorithms", International Journal

of Innovative Research in Advanced Engineering (IJIRAE)2016,Issue 11, Volume 3.
[16] W. Gerstner, “Supervised learning for neural networks: a tutorial with JAAv exercises”.
[17] o. R. .s. P. breiman L, friedman J.H., “Classification and regression trees.” Belmont CA Wadsworth International group, 1984. B. C. . U. P.E.tgoff, “Multivariate decision trees: machine learning,” no. 19, 1995, pp. 45–47.