Evaluation On Dhallywood Movies Based On Machine Learning

Sharmin Akter
Student Id:012181001

A Thesis

in

The Department

of

Computer Science and Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Science in Computer Science and Engineering

United International University

Dhaka, Bangladesh

October 2020

Approval Certificate

This thesis titled "**Evaluation On Dhallywood Movies Based On Machine Learning**" submitted by **Sharmin Akter**, Student ID:**012181001**, has been accepted as Satisfactory in fulfillment of the requirement for the degree of Master of Science in Computer Science and Engineering on 15 October, 2020.

Board of Examiners

1.

_____ Supervisor

Dr. Mohammad Nurul Huda
Professor & Director - MSCSE Program
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

2.

_____ Head Examiner

Dr. Swakkhar Shatabda
Associate Professor & Undergraduate Program Coordinator
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

3.

_____ Examiner-I

Mr. Suman Ahmmed
Assistant Professor
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

4.

_____ Examiner-II

Rubaiya Rahtin Khan
Assistant Professor
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

5.

_____ Ex-Officio

Prof. Dr. Salekul Islam
Professor and Head of the Dept.
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

Declaration

This is to certify that the work entitled "**Evaluation On Dhallywood Movies Based On Machine Learning** " is the outcome of the research carried out by me under the supervision of Dr. Mohammad Nurul Huda**,** Professor & Director, MSCSE Program.

_____

Sharmin Akter
012181001
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

In my capacity as supervisor of the candidate's thesis, I certify that the above statements are true to the best of my knowledge.

_____

Dr. Mohammad Nurul Huda
Professor & Director - MSCSE Program
Department of Computer Science and Engineering
United International University
Dhaka, Bangladesh

Abstract

The film is an exciting source of investment for passionate movie makers. The profitable nature of motion picture industry attracts movie creators to involve with it. In a such scenario, this is very important to evaluate movie status to find relevant features of a movie that make it successful. Machine learning is a popular trend for analyzing movie data. In our proposed research, we have tried to evaluate the status of Dhallywood movie based on three different class classifiers such as: Binary class (Hit-Yes, No), Triple class (Excellent, Good, Bad), and Four class (Excellent, Very Good, Good, Bad). The method of analyzing data has been described in details. Collection of Dhallywood movie data is the main challenge of this research work. The collected data have analyzed in a different way to set a target variable, which has improved the accuracy of models. The collected data have analyzed using five ML algorithms, and each algorithm applied three times for three different groups of class. Then the analytical results have compared to find out the best algorithm. From the comparative analysis it is found that accuracy of Triple class classification is higher than that of Binary and Four class classification. In addition to that among all applied algorithms, Random Forest provides the highest accuracy which is near about 85%. This research provides a new approach to set target variable classes based on Wikipedia data, news, actor actress biography, and viewer response on YouTube for a particular movie. We have selected this approach because the Dhallywood movie rating is not accurate on IMDb for all movies due to lack of budget and revenue data.

Acknowledgement

I would like to express my gratitude to all those who have helped me to complete my MSc in Computer Science and Engineering Thesis.

I am especially grateful to my supervisor Dr. Mohammad Nurul Huda, Professor and MSCSE Director, Department of Computer Science and Engineering at United International University, for his continuous guidance, advice, effort and veritable suggestion throughout the research.

I am also grateful to Head Examiner Dr. Swakkhar Shatabda, Associate Professor, Department of Computer Science and Engineering at United International University, for his suggestion and valuable review for both conference and thesis book.

I would like to thank, Mr. Suman Ahmmed, Assistant Professor, Department of Computer Science and Engineering at United International University for reviewing my thesis book and his valuable comments.

I also like to thank Rubaiya Rahtin Khan, Assistant Professor, Department of Computer Science and Engineering at United International University for her valuable time to review my thesis book.

Lastly, I would like to express my sincere appreciation to all of them who have willingly helped me out with their abilities.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

This chapter presents an overview of the motivation, purpose and problem statements. This also chapter focuses on different research challenges that faced during this research work. The objectives and contributions of this thesis have been presented in the last part of this chapter. The chapter ends with the organization of the thesis.

1.1 Motivation

The film is always the most cherished source of entertainment. There is a risk factor associated with making any film. The factor is whether the movie can fulfill the expectations of the audience or not. It also involves economic factors. A system is badly necessary there to estimate and predict the revenue that will be earned by a movie in accordance with the huge investment required for a movie [1]. The severity of the problem increases when we realize there is a multitude of factors that impacts the revenue of the movie. The story behind Dhallywood movie success is not similar to those of Hollywood and Bollywood. Hence a Dhallywood movie cannot be evaluated similarly as a movie of Hollywood and Bollywood. Those movies receive greater impact than Dhallywood movies by publishing a movie trailer or item song before the movie is released as the movie trailer has a great impact on its success [2][3]. Wikipedia activities and information helps the researchers to understand the movie popularity [4]. User review, critic review, and news analysis are other ways to understand movie status [5][6]. The Bangladesh film industry is losing its attraction day by day and the yearly movie production rate has been decreased. As a result, this is very crucial moment for the Bangladesh film industry. A proper evaluation system should be adopted with Dhallywood to save this sector. It has been reported that many researchers of other countries worked on their movies to enhance their film industry, besides they have a well-organized movie database [7]-[12]. Motivated from the outcome of their research, we have devoted our research to develop a well-organized database for Dhallywood.

1.2 Purpose

Previous studies proved that it is possible to predict the success of a movie using attributes such as budget, rating, actors rating, revenue and so on. The goal of this thesis is to examine the accuracy and performance of a better movie evaluation system, using a greater data set with some new features and methods those are previously not used with machine learning. In our work, we have focused on Wikipedia movie information, the biography of actor-actress, sometimes YouTube viewer response to analyze Dhallywood movie status instead of directly predict success. Our purpose is to find factors, that are playing a very important role to make a movie excellent.

1.3 Objective and Research Challenges

Main objective of this study is to collect Bangladeshi movie data from different sources. No work has been done previously on Dhallywood movies in this approach. Previous studies focused mainly on Hollywood and Bollywood movies. The main advantage of Hollywood and Bollywood movies are the availability of data[6]-[8]. Therefore, this research work focuses on data collection and analyze data using machine learning.

The key objective of our research is as follows.

- To prepare a dataset for Dhallywood movies.
- To develop a system to predict movie success/status based on actor-actress and genre.
- To predict the influence of screening in different festivals, joint production, and foreign actor on its success.
- To apply some machine learning algorithms and compare results.

The film industry is a massive sector to invest and earn money. Every year several movies are released in Hollywood, Bollywood, and Dhallywood (Bangladesh) as well. But a very few Bangladeshi or Dhallywood movies become super hit, blockbuster or all times blockbuster. Rigorous researches have been carried out previously on the Hollywood and Bollywood movies, in the domain of online review mining and success prediction. By applying different methods various research groups are exploring the different ways to predict success.

Considering our key objective of features, we face some challenges as follows

- Due to unorganized and proper data availability many challenges have faced to drive analysis on Dhallywood movies.
- It is a tricking and challenging task to collect data on Bangladeshi/ Dhallywood movies.
- Revenue, budget, and rating is the primary parameter of a movie success but Dhallywood movie related any source does not contain these data thoroughly for all movies.
- To improve accuracy, again and again we extract features and prepare datasets.

There were some research questions as follows.

- How do you start collecting data?
- Which factors are related to Dhallywood movie success?
- Will Our dataset features give good accuracy of the model?
- Will the dataset fit to the model?
- How we can improve our model accuracy?

1.4 Thesis contribution

The contributions of this thesis work are summarized as follows.

- There were a lot of problems making a dataset of Bangladeshi movies. Finally, we make a dataset.
- Here we apply different techniques to extract features, that are never done by any research before.
- We study on Bangladeshi movies previous studies do not focus on this.

1.5 Organization of thesis

The organization of the thesis is as follows.

- **Chapter 2**

  In the literature review chapter, we will discuss previous work that predicts movie success using different machine learning tools and techniques. Also focus on the dataset features and data collection techniques. A summary table will signify their notable works briefly.

- **Chapter 3**

  The methodology chapter will present the outline of the methodology of this thesis work, such as data collection, feature selection, feature extraction, data preprocessing, and dataset description. The data collection section 3.2 will focus on the techniques of data collection. There are several methods for feature selection and extraction. Chapter 3 will discuss the applied methods.

- **Chapter 4**

  The experimental analysis and result chapter will focus on our model's accuracy, comparison of models result, and model validation. We will apply five machine learning algorithms to analyze and compare their accuracy three times using different classes of target variables.

- **Chapter 5**

  Conclusion and future work chapter will discuss our overall thesis work. This chapter will also include the thesis outcome and future work.

Chapter 2

Literature Review

In the recent era, machine learning is the most popular and useful technique to convert data into the correct format. In the modern age, different internet sources are the origin of data. The data are mixed nature, complex structure, and large those are tough to analyze. Which data format will compatible with the proposed work is possible to determine through the machine learning techniques. This chapter will discuss machine learning and its challenges to apply machine learning on movie related data analysis.

2.1 Machine Learning in Movie Data Analysis

Machine learning is a method of analyzing a set of data. Machine learning is a fast-growing trend to identify data patterns. A lot of work has been done previously on movie data to analyze reasons for success and flop. Several parameters influence on the success of a movie such as revenue, budget and so on. All past research on movies include data from movie relevant internet sources and try to find a better outcome by applying different ML methods and algorithms.

2.1.1 Movie Success Prediction Through Machine Learning

In the early days, many people prioritized gross box office revenue as a parameter of movie success [13][14]. There is another parameter of the movie which is very significant. Sometimes a movie may not earn revenue but yet it can achieve a huge positive response from the audience. Previous researches used to predict movie success or movie rating with available information and most of them used to focus only on one or two variables as a success parameter. Movie status prediction through IMDb rating is a popular approach nowadays. Wardak Ruheen Bristi, Zakia Zaman, and Nishat Sultana developed a model that can predict IMDb rating [15]. In their work, they define four output classes using ratings.

Subramaniyaswamy V., Vignesh Vaibhav M., Vishnu Prasad R., Logesh R in their work classified data into 3 classes low budget, medium budget, and big-budget film[2]. Garima Verma and Hemraj Verma were done a study on Bollywood movie success by classifying

data into two class hit or flop[7]. In their research, they included music rating with IMDb rating and used different machine learning models. They applied SVM and multiple regression to predict box office success.

A mathematical model was developed by Javaria Ahmad, Prakash Duraisamy, Amr Yousef and Bill Buckles to predict the success and failure of the upcoming movies based on several attributes [8]. There were some criteria in calculating movie success which includes budget, actors, director, producer, set locations, story writer, movie release day, competing for movie releases at the same time, music, release location, and target audience. This system can be used to predict the success or failure of upcoming movies by the moviemakers as well as by the audience.

Muhammad Hassan Latif and Hammad Afza wrote a paper, in their paper they focused on movie rating prediction using machine learning [9]. They extracted data from IMDb and divides rating into four classes. A research was done by L. Zhou where movie review mining used machine learning and semantic orientation [10]. This paper result showed better findings than the earlier studies and also proved the movie review mining is a more challenging application than many other types of review mining.

Rijul Dhir and Anand Raj were provided a quite efficient approach to predict IMDb score on IMDb Movie Dataset [16]. They were tried to unveil the important factors influencing the score of IMDb Movie Data. The exploratory analysis found that the number of voted users, number of critics for reviews, the number of Facebook likes, duration of the movie and gross income of movies affect the IMDb score strongly. After building the five models they found out that the Random Forest represents the movie features more accurately.

Nahid Quader and Dipankar Chaki developed a system that can predict 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[17]. Using Support Vector Machine (SVM), Neural Network, and Natural Language Processing the system predicts a movie box office profit based on some pre-released and post-release features. A larger number of reviews from IMDb and Rotten Tomato are inserted in the dataset. This paper 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. They figure

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**.** Authors in their paper measured movie success with two parameters: Gross box office collection, Critics rating. Since these two parameters judge the success of movies on different independent and unrelated levels which were able to increase accuracy. Here they also showed the interrelation between these classical factors. If a particular actor or actress worked with a particular production house, their films perform well in the box office.

Anand Bhave, Himanshu Kulkarni, Vinay Biramane, and Pranali Kosamkar were analyed that along with the classification factor, social media feedback improves accuracy (FB, Twitter, YouTube) [18]. A study combines temporal abstraction (TA) with data mining techniques to find some important rules to adjust marketing strategies [19]. The framework had three main modules, a data pre-processing module, a data attribute selection module, and a temporal abstraction module. Most of the basic information was gathered from IMDb, such as the number of comments and stars. Information was gathered for other attributes from Amazon and Box Office Mojo.

P. Nagamma, H. R. Pruthvi, K. K. Nisha, and N. H. Shwetha were validating feature effectiveness using clustering and sentiment Classification by involving text preprocessing, text transformation [20]. Here researchers applied different data mining techniques for a better outcome and used fuzzy clustering for sentiment classification also apply SVM for final sentiment classification. Result also found that using the TF-IDF approach of feature selection gave an accuracy of 10% more than using the 14 key words method for box-office prediction.

A. Samad, H. Basari, B. Hussin, I. G. Pramudya, and J. Zeniarja, in their paper applied opinion mining refers to the application of computational linguistics, natural language processing, and text mining to identify or classify whether the movie is good or not based on message opinion[21].They used Support Vector Machine (SVM) classifier to analyzed and understand the data patterns. This research concerns binary classification which was classified into two classes: positive and negative. The positive class showed good message opinion and the negative class showed the bad message opinion of certain movies. They were validated the accuracy level of SVM applying 10-Fold cross validation and confusion matrix.

R. Niraj and J. Singh illustrated the user-generated and professional critic also play a vital role in movie success[6]. Their study contributes to the literature by bringing a new measure of valence in the UGR/movie literature from personal and group psychology literature. In 2015, M. Lash et al discussed in their work about three factors 'who', 'what' and 'when'[22]. It is possible to predict the profit of a movie in the early stage of production by analyzing factors: 'who'-its actor, actress directors, and social network, 'what'-genre and rating, as well as 'when'- when a movie will be released.

Marton Mestyan, Taha Yasseri, and Janos Kertesz in their work considered the activity level of editor, Wikipedia view and number of pages views by readers [4]. They defined various variables and applied the logistic regression model for data analysis. Athira M D and Laksmi K S used an ensemble classification method to film reviews as well as film-related meta-data to generate more precise outcomes [23]. They focused on prediction using the best k features which improved the accuracy of success prediction. Where they described if two or more classifiers gave the same outcome, the max voting method will predict film as a success, failure, or neutral. When the outcomes were combined, the max voting gives an accuracy of 90%.

2.1.2 Summary from The Previous Study

Table 2.1: Summary Table from the Previous Study

Author Name	**Dataset Collection**	**Feature Selection**	**Features**	**Classifier and validation method**
Subramani yaswamy [2]	BoxOfficeMojo and Wikipedia	Linear regression	Budget, revenue, actor, actress, ROI, trailer view, Wikipedia view.	SVM and Multiple regression
Marton Mestyan [4]	Wikipedia article, news, BoxOfficeMojo	Correlation analysis.	Activity level of view, Wikipedia view, newsreader, actor, actress, rating	Logistic regression

R. Niraj and J. Singh [6]	Raw data from different sources like UGR comments, revenue data etc		Genre, revenue, director star, actor star, professional review, volume of UGR comments, UGR positive comment	Regression analysis
J. Ahmad, et al [8]	Bollywood movie dataset	Correlation analysis.	Movie Name, Year of release, Genres (Drama, Action, Romance, Comedy), Other, Directors , Music directors, Producers , Language Hindi	Mathematical model(own developed)
Muhammad Hassan Latif and Hammad Afza[9]	IMDb	Information gain	Name, Genre, Budget, User Votes, Metascore, Opening week business, Awards, Number of Screens,Rating	Logistic regression, Simple Logistic, Multilayer Perceptron, J48, Naive Bayes, ROC
Pimwadee Chaovalit and L. Zhou [10]	Corpus, clean dataset in movie reviews domain (http://www.cs.cornell.edu/people/pabo/movie-review-data/),IMDb.		Ratings, no of reviews.	Confusion matrix, 3-fold validation,
Warda Ruheen Bristi	IMDb		Title, studio, actor, actress, rating, genre, year, country	Bagging, Naïve Bayes, Random

,Zakia Zaman and Nishat Sultana[15]				Forest, IBK, 10-fold cross validation
R. Dhir and A. Raj [16]	IMDb, Twitter tweets, YouTube	Correlation analysis.	Movie titles, directors, genres, countries of origin, and the Facebook popularity of the top three actors	Random forest, KNN, Gradient Boost, SVM, Ada boost.
N.Quader [17]	IMDb, Rotten tomatoes, Box office mojo, Metacritic(dataset contains 755 movies released in between 2012 to 2015)	Correlation analysis.	Total 15 features are used in our proposed model. We take in consideration of tomato critics' meter, tomato critics' rating, tomato audiences' score and tomato audiences' rating from Rotten Tomatoes, Meta score of Metacritic and IMDb rating from IMDb for a particular movie, number of viewers rated the movies. The multiplied value of rating and number of users who rated are used as a single feature.	MLP,SVM, NN
A. Bhave [18]	IMDb,social media, Youtube, Twitter	Linear regression	Cast, Producers, Directors, Genre, movie revenue, movie production budget.	Multivariant linear regression.
L. Chen, C. Chi, and L. Huang [19]	Text data from IMDb,	Temporal Abstraction module	overall (OA), screenplay (ST), special effects (SE),	5-fold cross-validation

	Amazon,BoxOfficeMojo		director (PDR), and actors and actresses (PAC)	
P. Nagamma [20]	IMDb	Validating Feature effectiveness by clustering	User review	SVM, Autoregression model
A. Samad [21]	Raw data from twitter	Case normalization, tokenization, Stemming, Generate n-grams.		SVM, SVM-SPO,10-fold cross validation, confusion matrix.
M. Lash [22]	A movie archive website, BoxOfficeMojo	Correlation analysis.	Genre, actor actress, plot, release time, social media response	Logistic regression
Garima Verma and Hemraj Verma[24]	IMDb, bollymoviereview, planetbollywood,boxofficeindia	Linear regression	Number of screens, rating, music rating, actor, actress, genre.	Multiple Regression, Discriminant Analysis, Artificial Neural Network, Logistics regression, Multinomial regression

2.2 Challenges in Movie Dataset Collection for Analysis

The most challenging part of this research work is data collection. The data that is required for predicting and analyzing movie success and failure are scattered on the different

sources. Actually, a research outcome depends on the proper dataset and way of thinking. Some authors use dataset collected from popular movie websites like IMDb, Rotten Tomatoes, Box office mojo, and Metacritic. These datasets have a definite number of common types of features like genre, actor, actress, rating, budget, user review, gross and so on. But these limited features most of the cases are not enough to analyze success or failure of a movie. Most of the researchers collect data from multiple sources according to their research requirements. In 2018, Rijul Dhir and Anand Raj collected data from twitter tweets and YouTube comments [16]. In their research perspective, twitter and YouTube data were important for movie popularity analysis. Another work is done by Anand Bhave in 2015, where user data was collected from social media besides YouTube and Twitter[18]. L. Chen, C. Chi, and L. Huang in their work use text data scraped from IMDb consumer reviews that can help them to find factors that affect movie sales [19]. Pimwadee Chaovalit and L. Zhou collected a corpus dataset that includes positive and negative both types of reviews collected from IMDb[10]. P. Nagamma et al in their research scrape IMDb data in a different way reviews a movie directly collected from saved HTML source code where HTML source code contains multiple information like the review, rating, usefulness along the date[20]. Rakesh Niraj and Jagdip Singh in 2015 done research with a dataset of user and critic reviews [6]. They have collected data from the specialized source. A. Samad et al focus on opinion mining of movie review they used a dataset of raw twitter data [21].

Another research was done by R. Niraj and J. Singh their data came from different specialized sources like raw data from different sources like UGR comments, revenue data and so on [6]. M. Lash focused on who, what, and when the factor of a movie and they collect data from a movie archive as their requirements [22]. Some researchers used data generated in Wikipedia, YouTube, and also scraped data from BoxOfficeMojo[2]. Wikipedia data pasted into Google Sheets by enabling a simple script then exported into Excel.

In 2019, Garima Verma and Hemraj Verma scrape data from a variety of sources like imdb.com, bollymoviereviewz.com, planetbollywood.com, boxofficeindia.com, and bollywoodhungama.com through scraper extension in crome[7]. Another study where researchers in used Wikipedia articles, news and BoxOfficeMojo data for movie success analysis [4].

2.3 Challenges in Choosing Proper ML Algorithm and Method

In machine learning there are several algorithms for predictive modeling. The most widely used predictive models are decision trees, regression, and neural networks[15][25]. The suitable machine learning model depends on the category of problem, understanding the nature of data, data processing, training set size and so on. Support Vector Machine, Neural Network, Logistic Regression, K-Nearest Neighbor, and Random Forest are the most popular machine learning models to predict movie success [2][11] [16][17]. Table1 shows the summary of different works based on machine learning. Some researcher applied their own models to get better accuracy.

2.4 Related Works

Most of the researchers gathered data from different sources to extract the best features [4][21][22][31]. Under the umbrella of traditional factors, it is acknowledged that researchers have laid focus on some common features like genre, rating, time of release, and actors [8][16]. Several works had been done on movie success prediction, using different machine learning models based on movie rating and also compared accuracy of the models [2][7][9][12][15][16][24]-[30]. Wikipedia information-based rating, users and critics review impact also analyzed by some research [5][6][32][33]. The actor, actress, villain and director have an influence on movie success [34]. Some studies compared those features with other features that depict the popularity of the movie as a whole [18][22]. Many researchers have used correlation feature selection [8][17] and where another study compares some feature selection techniques [35]. Hence in this study, we will find out the most relevant feature which could influence the success rate of movies. We will also focus on Wikipedia information user and critic review. In light of the above work factors playing an important role to make a Bangladeshi movie popular will be investigated and compared using ML models along with some feature selection methods. Then models will be validated by other performance measures such as correlation, cross-validation and ROC curve.

Chapter 3

Methodology and Methods

This chapter presents research methodology and experimental design of this research work. In addition to that different machine learning models have been described in this chapter. These models have been applied to evaluate Dhallywood movie data.

3.1 Outline of Methodology

Bangladeshi movie dataset is not available on Kaggle, UCI machine learning, Socrata, Google public dataset, or any dataset website. Therefore, it was challenging to complete the whole work in one cycle. We have prepared a dataset for our research as there were no well-organized dataset available. Before preparing the dataset, we have analyzed the taste and choice of Bangladeshi people and some other related factors of the movie. After brainstorming we have found some features then scraped data from different sources. The dataset has been processed to make it useable for the machine learning model. However, the result was not satisfactory in the first stage. Therefore the dataset is reconstructed with some new features. Our proposed model validated after achieving a satisfactory result. The methodology is divided into 8 phases as follows:

- Brainstorm feature
- Create feature
- Data fetching and dataset preparation
- Data processing
- Feature extraction and feature selection
- Machine learning model
- Result
- Model validation

The whole process is shown in the following diagram Figure 3.1.

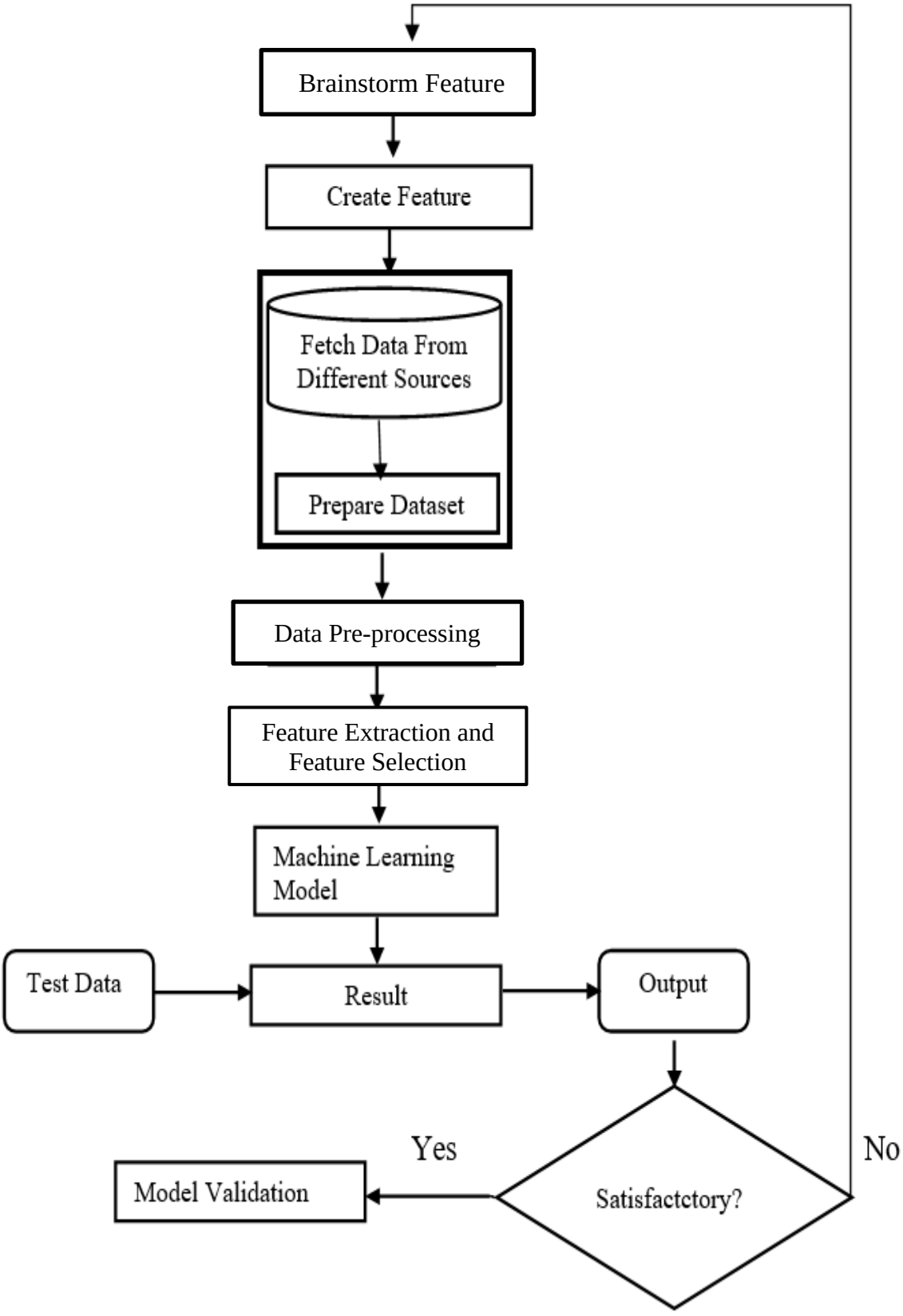


Figure 3.1: Outline of Methodology

3.2 Brainstorm Features

We have prepared and updated our dataset several times by creating new features and rearranging existing features to improve the accuracy of the models. The first dataset is shown in Figure 3.1, data was collected from bmdb. We get a very poor result from this dataset.

Title	Movie type	Color/Bn	Digital	35 mm	Length	Language	Director	Producer	Year	Production house	Story Writer	Actors	Rating
Chakor	Drama,Roma	Color	0	1	141	bangla	Montazur Rahman Al	Jahangir A	1992	Shonamoi Films	Aslam	Diti,Elias Kanchon	7
Khoma	Drama,Roma	Color	0	1	123	bangla	Malek Afsari	Sheknur A	1992	Rosey Films	Malek Afsari	Shabana,Alomgir	6.3
Nay Juddho	Action	Color	0	1	129	bangla	Montazur Rahman Al	Bonolota	1991	Bonolota cholochett	Montazur Rahmar	Shuchorita,Elias Kanch	4.4
Kothipuron	Drama,Roma	Color	0	1	127	bangla	Malek Afsari	Sheknur A	1989	Rosey Films	Ahad Khan	Rozina,Alomgir	8.6
Ranga Vabi	Drama,Roma	Color	0	1	149	bangla	Motin Rahman	Shabana	1989	S S Productions	Iqbal Kasmeri	Shabana,Alomgir	8.5

Figure 3.2: Dataset 1 Collected From bmdb.co.

3.3 Create Feature

Proper user rating, review, budget, gross income, meta score, these are very important information about a movie. In case of Hollywood and Bollywood movies all the data are readily available. However, some issues exist due to unavailability of all data in Dhallywood movies. Therefore, of using dataset presented in Figure 3.2 obtained result was not satisfactory. Hence some new features have been added to the prepared dataset to improve the obtained results. In Section 3.4 data collection sources and dataset preparation process has been discussed in detail.

3.4 Data Collection and Dataset Preparation

Data has been collected form the webpages of bmdb.co, prothomalo.com, ntvbd.com, bdnews24, Wikipedia and IMDb. Redundant and unnecessary data have been removed from the prepared dataset. Data has been collected based on different attributes like the movie title, genre, movie length, award, actor, actress, villain, number of release country, revenue, producer, production house, director, Joint venture, foreign actor, language. For my research, I have collected Dhallywood movie data from 1972- 2019. Data collection and dataset preparation is the most important and analytical part. This phase has been performed through three steps.

- **Step1**: Data have been collected only from bmdb.co [39]. The language of this website is Bangla. So, data have been inserted into the datasheet manually. Many attributes of data were missing and all the available data not correct even. Movie genre information was also confusing as these were different from Wikipedia. Most of the movie rating in bmdb was 0 out of 10. Therefore the rating of the movies were reported from IMDb. Sample dataset after completing step 1 is shown in Figure 3.2. This dataset contains only 14 features.
- **Step 2**: In this step, a new dataset has been prepared again because in the previous dataset we found some problems like genre information, also different name spelling of the same actor, writer, and director. In step two, we collect Dhallywood movie data from IMDb [38] by using parsehub data scraper. This software considers each data as a span if one data is missing it includes self-generated unauthorized value which hampers the results.

Figure 3.3: Data Scrape Using Parsehub From IMDb

title	year	genre	rating	director	actor1	actor2	actor3	actor4	votes	writer1	writer2
Mrittika Maya	2013	Drama, Family	8.1	Gazi Rakayet	Raisul Islam Asa	Titas Zia	Shormy Mala	Lutfur Rahman G	48	Gazi Rakayet	Gazi Rakayet
Meherjaan	2011	Drama, History	2.9	Rubaiyat Hossai	Jaya Bachchan	Victor Banerjee	Omar Rahim	Shayna Amin	667	Rubaiyat Hossain	
Surja Dighal Bari	1979	Drama	8.3	Sheikh Niamat A	Dolly Anowar	Rawshan Zamil	Sitara Begum	Sufia	252	Abu Ishaque	Masihuddin Sha
Aynabaji	2016	Crime, Mystery,	9.2	Amitabh Reza Cl	Chanchal Chowc	Masuma Rahma	Bijori Barkatullah	Partho Barua	18,956	Syed Gaosul Ala	Anam Biswas
Dahan	1985	Drama	8.5	Sheikh Niamat A	Humayun Faridi	Bobita	Dolly Anowar	Rawshan Zamil	58	Sheikh Niamat A	Sheikh Niamat A
Ghani	2006	Drama	8.3	Kazi Morshed	Raisul Islam Asa	Dolly Johur	Masum Aziz	Naznin Hasan Cl	21	Kazi Morshed	
Haajar Bachhar Dhc	2005	Comedy, Drama,	8	Suchanda	Riaz	Sharmin Zoha Sl	A.T.M. Shamsuz	Shahnur	677	Zahir Raihan	Aziz Misir
Swatta	2017	Drama	7.7	Hashibur Reza K	Shakib Khan	Paoli Dam	Jayanto Chattopa	Don	258	Sohani Hossain	Ferdous Hasan
Gangajatra	2009	Drama	7.2	Syed Wahiduzza	Ferdous Ahmed	Poppy	Shimla	Shahidul Alam S	10	Syed Wahiduzza	Hasan Ahmed
Guerrilla	2011	Drama, History,	8.2	Nasiruddin Yousi	Jaya Ahsan	Ferdous Ahmed	Shampa Reza	Ahmed Rubel	2,385	Syed Shamsul H	Nasiruddin Yous
Television	2012	Drama	8.2	Mostofa Sarwar I	Shahir Huda Rur	Chanchal Chowc	Nusrat Imrose Ti	Mosharraf Karim	4,242	Mostofa Sarwar I	Anisul Haque

Figure 3.4: Dataset 2 Represents Data Collected From IMDb Including Title, Year, Genre, Actor, Actress(part1)

user_review	critic_review	award	Hit	joint_venture	release_country	foreign_actor	screenings_in_festivals
0	1	15	yes	no	3	no	yes
30	78	15	yes	no	1	yes	yes
1	0	14	yes	no	1	no	yes
96	7	13	yes	no	4	no	yes
0	1	13	yes	no	1	no	no
0	1	13	yes	no	1	no	no
5	0	13	yes	no	1	no	yes
4	4	13	yes	no	1	no	yes
0	1	13	no	no	2	no	yes
13	4	12	yes	no	1	no	yes
11	4	12	yes	no	2	no	yes
1	38	12	no	no	5	no	yes

Figure 3.5: Dataset 2 Represents Data Collected From IMDb Including User_review, Critic-review, Joint_venture (part2)

- In dataset 2, some information has collected from Wikipedia. As a part of our analysis, we set a target variable which shows a movie hit or flop (yes/no). The target variables are set based on rating, Wikipedia information, and different news article. Total 20 features included in dataset two those are, title, year, genre, rating, actor1, actor2, actor3, actor4, vote, writer1, writer2, director, number of user review, critic review, award, joint venture, release country, foreign actor, number of release country and screenings in festivals.
- Step 3: In this step, we modified our previous dataset for better accuracy. Three new features are included namely actor, actress, and villain. Two more target variables such as hit category -3(excellent, good, bad) and hit category –4(excellent, very good, good, bad) are added in the dataset. While categorizing movie status in three and four categories reaction of the viewer on YouTube for a movie is analyzed for more precise information about it.

Actor	Actress	Villain	Hit	Hit category-4	Hit category-3
Anando	Shabana	Golam Mustafa	yes	very good	good
Victor Banerjee	Jaya Bachchan	Humayun Faridi	yes	very good	good
Ferdous Ahmed	Poppy	Shahidul Alam S	no	good	good
Rahul Bose	Shahana Goswa	Rahul Bose	no	good	good
Raisul Islam Asa	Naila Azad Nup	Pijush Bandyopa	no	good	good
Tauquir Ahmed	Afsana Mimi		no	good	good
Shakib Khan	Apu Biswas	Misa Sawdagar	no	good	good
Raisul Islam Asa	Rokeya Prachy	Masum Aziz	no	good	good
Jewel Jahur	Shimla	Mamunur Rashic	no	good	good
Sohel Rana	Aruna Biswas	Miju Ahmed	no	good	good

Figure 3.6: New Categories Hit Category-4 and Hit Category-3 Create from Existing Features.

Table 3.1: Dataset Features Description

Feature	Description
Title	Name of the movie
Genre	Category of the movie (drama, romantic, history, war, action and so on)
Year	Movie released year
Director	Director name
Actor1	Leading male or female actor of the movie
Actor2	Second leading male or female actor of the movie
Actor 3	Third leading male or female actor of the movie
Actor 4	Fourth leading male or female actor of the movie
Actor	Leading male actor of the movie
Actress	Leading female actor of the movie
Villain	Leading negative role of the movie
Votes	User votes for a movie
Writer 1	Story writer of the movie
Writer 2	Dialogue writer of the movie
User review	Number of user comments
Critic review	Number of critic comments
Rating	Rating given by viewer out of 10
Meta score	Score given by meta critics
Award	Number of awards achieved by a movie
Foreign actor	Foreign actor participation in a movie
Joint venture	Movie jointly directed by more than one country
Number of release county	Number of released countries
Screenings in different festival	Movie exhibited in different festival or not

Finally, there are 23 features in our dataset with three target variables. These target variables are Hit, Hit category-3, and Hit category-4. We have faced so many contradictory and puzzling situations while setting target classes. These variables are set by analyzing multiple sources like: Wikipedia information, viewer response on YouTube, actor actress biography, user rating, and votes for a movie. The quality of movie can be judged from the

information available on IMDb and Wikipedia. However, to estimate the quality of a movie we had to analyze lot of data available scatteredly in different sources.

3.5 Data Pre-processing

Data processing make a dataset machine readable. Our collected dataset contains raw data, before applying machine learning model we need to convert the dataset into machine understandable format.

3.5.1 Data Cleaning

In this stage redundant and unnecessary data have been removed. While scraping data, miss span produce some garbage value which value insert into dataset, we remove those garbage value from the dataset.

3.5.2 Missing Value Handling

Sometimes our dataset contains null values, value marked with '?' and some other symbols. As null value is not possible to handle nor convert into other values so those instances were removed. If any instance contains '?' or value, then these values have been converted into numeric values by the encoding method.

3.5.3 Splitting Multiple Genres into Individual Column

Genre column contains multiple genres for the same movie. Therefore, individual columns for each genre category has been generated. Suppose, a movie is a combo of 2, 3 or 4 genres. To find correlation and which genre movie is most popular it is necessary to split genres. Sample code is given in appendix section A.1.

g_ Adventure	g_ Comedy	g_ Crime	g_ Drama	g_ Family	g_ Fantasy	g_ History	g_ Horror	g_ Music	g_ Musical	g_ Mystery	g_ News	g_ Romance	g_ Sci-Fi	g_ Sport	g_ Thriller	g_ War	g_Action	g_Adventure	g_Animation	g_Biography	g_Comedy	g_Crime
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3.7: Table After Splitting Multiple Genres into Individual Column

3.5.4 Convert Some Column into Binary Value

Some columns have been converted into binary values these are joint venture, no of release country, foreign actor and screening in different festivals. In primary stage these values were ‘yes’ and ‘no’. Quantitative values ‘0’and ‘1’ have been set ‘yes’ and ‘no’ respectively. Sample code is given in appendix section A.2.

joint venture	no of release country	foreign actor	Screenings in different festivals
1	1	1	1
0	3	0	1
0	1	0	1
0	1	0	0
1	1	0	0

Figure 3.8: Conversion of Some Columns into Binary Values

3.5.5 Change Datatype

Some features datatype has been changed in this step. These changes were made to covert some features into required datatype. For this purpose, some columns were converted into string. Sample code is given in appendix section A.3.

3.5.6 Encoding Categorical Data

For categorical features and instances those instances have been converted into numeric values using the encoding process applying Scikit-learn library. Label encoder is an encoding process that converts any object (text) type data into an integer (numeric) by labeling the value between 0 to n classes. Label encoding is a property of Scikit-learn library. For using the label encoding process it is necessary to import “sklearn.preprocessing **import** Label Encoder”. Though another encoding process ’One-Hot’ encoder also exists which encodes object values into binary. So, ‘One-Hot’ encoder will be applicable for solving any binary class classification problem.

3.5.7 Feature Scaling

Feature scaling is very important before feature selection. It is necessary to normalize or standardize features to compare the data within a particular range. Sometimes, it also helps

to make the calculation faster within the algorithm. Feature scaling is required for KNN which performed by using sklearn, StandardScaler().

3.6 Feature Extraction and Feature Selection

Irrelevant features add noise, which reduces the accuracy of the model. To minimize effects of noise and high dimensionality as well as to increase accuracy of the model, some form of dimension reduction is necessary for data mining. Feature selection and extraction are two approaches of dimension reduction.

3.6.1 Feature Extraction

Data have been extracted from IMDb, Wikipedia [40], Bangladesh box office [41], and many other internet sources. In this part, some more features have been created from existing features to improve accuracy. The dataset preparation steps have been discussed in section 3.4. First of all, the name of four leading actors have been collected actor2, actor3, and actor4 reduce model accuracy. Then these four features have been rearranged into three namely actor, actress, and villain. It has been found that newly created features increase the accuracy of our model.

E	F	G	H	I	J	K
actor1	actor2	actor3	actor4	Actor	Actress	Villein
Shabana	Anando	Farid Ali	Golam Mustafa	Anando	Shabana	Golam Mustafa
Jaya Bachchan	Victor Banerjee	Omar Rahim	Humayun Faridi	Victor Banerjee	Jaya Bachchan	Humayun Faridi
Ferdous Ahmed	Poppy	Shimla	Shahidul Alam S	Ferdous Ahmed	Poppy	Shahidul Alam S
Raisul Islam Asa	Jayanto Chattopa	Mamunur Rashic	Pijush Bandyopa	Raisul Islam Asa	Naila Azad Nup	Pijush Bandyopa
Momtazuddin Ah	Afsana Mimi	Tauquir Ahmed	Rawshan Zamil	Tauquir Ahmed	Afsana Mimi	Rawshan Zamil
Shakib Khan	Apu Biswas	Rumana	Misa Sawdagar	Shakib Khan	Apu Biswas	Misa Sawdagar
Rokeya Prachy	Raisul Islam Asa	Masum Aziz	Maruf Sheikh	Raisul Islam Asa	Rokeya Prachy	Masum Aziz

Figure 3.9: Feature Creation from an Existing Feature by Converting Actor1, Actor2, Actor3 and Actor4 into Actor, Actress and Villain.

3.6.2 Feature Selection

There are several methods for feature selection among them four method have been applied and compared in this research work. These are Filter method, Wrapper method, Trial and Error method and Hybrid method.

Correlation: Correlation is a statistical method to analyze the degree of relationships of two numerically measured variables. A positive correlation means that two or more variables have a strong relationship with each other moving in a common direction, while a negative correlation means that the variables are moving in the opposite direction. The correlation technique is appropriate for certain kinds of data. It is used for quantifiable data in which numbers are representing something informative and meaningful. Using, the correlation analysis we can find the affinity between all the variables with each other in the dataset. To remove the unnecessary features using Filter feature selection method correlation analysis have been performed among all the parameters. The feature meta score has only one value in the entire column so it has been removed.

The different features correlation has been analyzed, which are shown in Figure10 and Figure 11. Correlation heatmap colors indicate positive and negative relations. Darker color indicates a negative correlation and lighter color indicates a positive correlation. Table 3. shows our final features with description.

We have found a positive correlation as follows:

- Joint venture with Foreign actor
- Votes with user review
- Number of release country with g_Sci-Fi
- g_Action with g_Drama
- User review with g_Mystery

Negative correlations as follows

- Screening in different festivals with g_Action
- g_Comedy with g_Action
- g_Action with g_Family

The movie title and movie releasing year are not important features to analyze a movie. Hence these two parameters have been removed from our dataset.

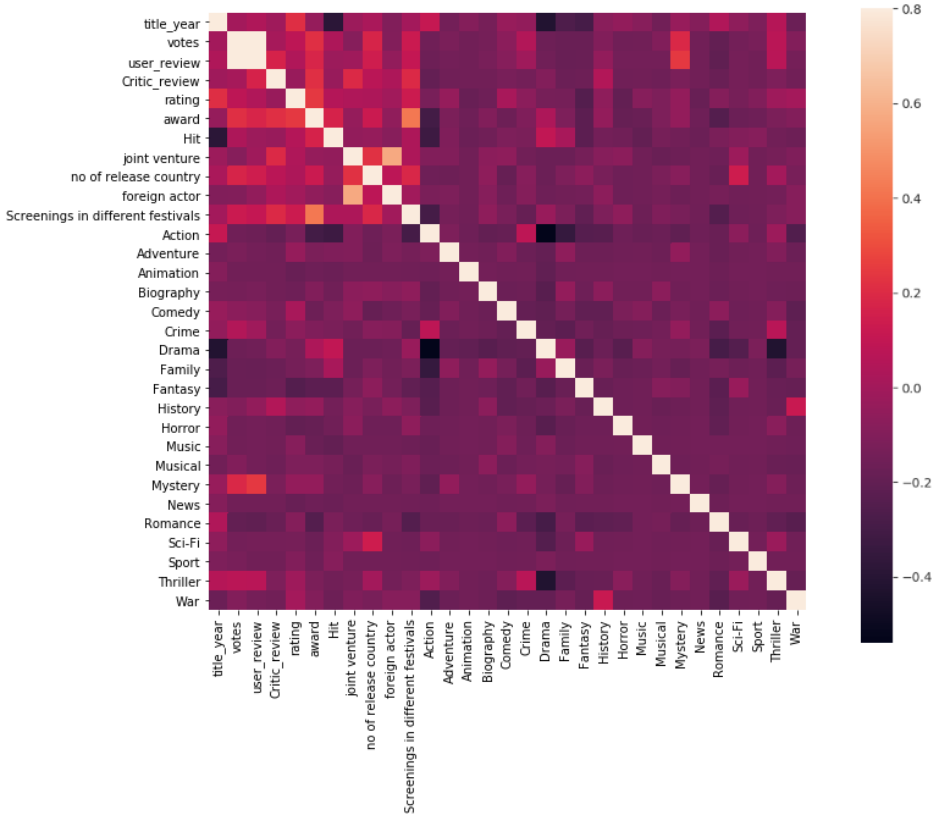


Figure 3.10 : Correlation Heatmap Before Dropping Unnecessary Features.

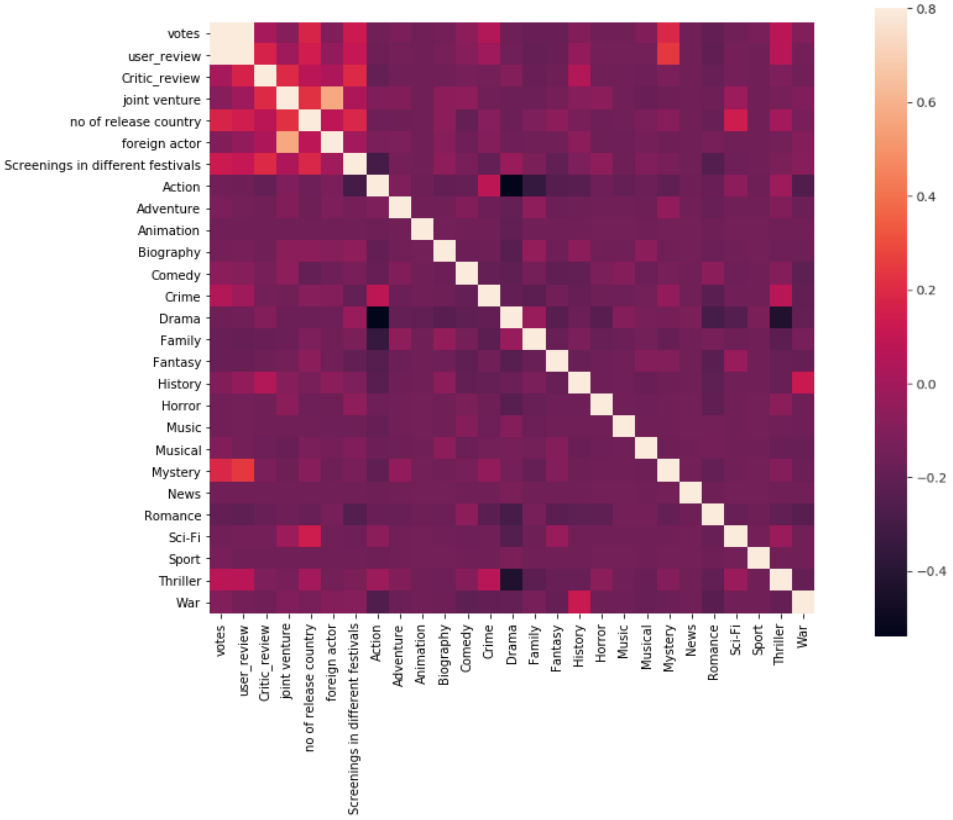


Figure 3.11: Correlation Heatmap of Features After Removing the Unnecessary Column

Trial and Error: Trial and Error method has been applied to identify important features for selected machine learning models. This method feed some features to machine learning model to evaluate their performance. From the evaluation it is decided whether the feature kept or removed to enhance accuracy. Trial and Error method allows to recursively add or eliminate features. As a result, this method can be more accurate than Filtering. Proposed methodology also shows the effect of frequent changes in features affects the accuracy of a model. Orange is a data mining platform to conduct data analysis and visualization which has been applied in this research work. By applying this method, the features actor and actress have been found as important features.

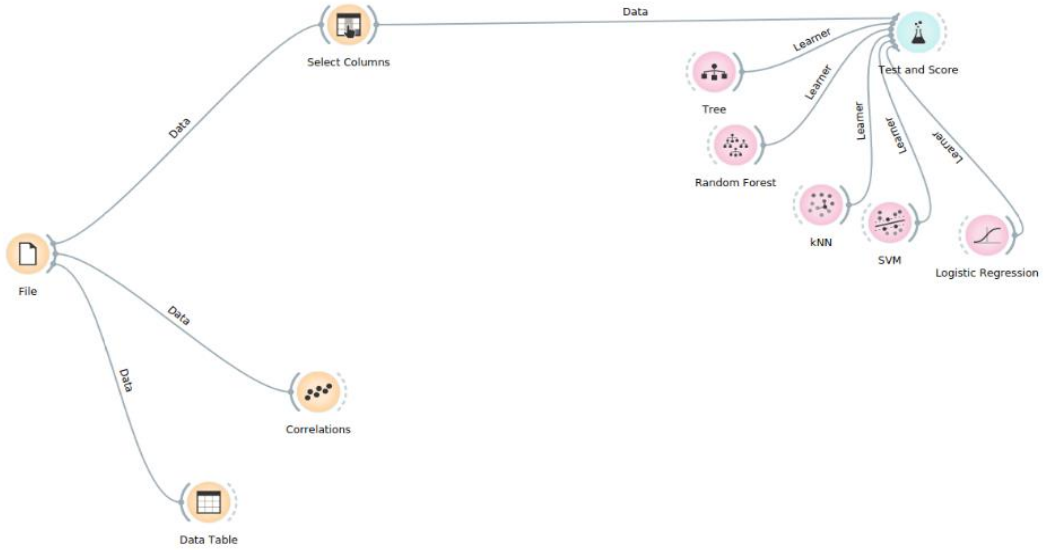
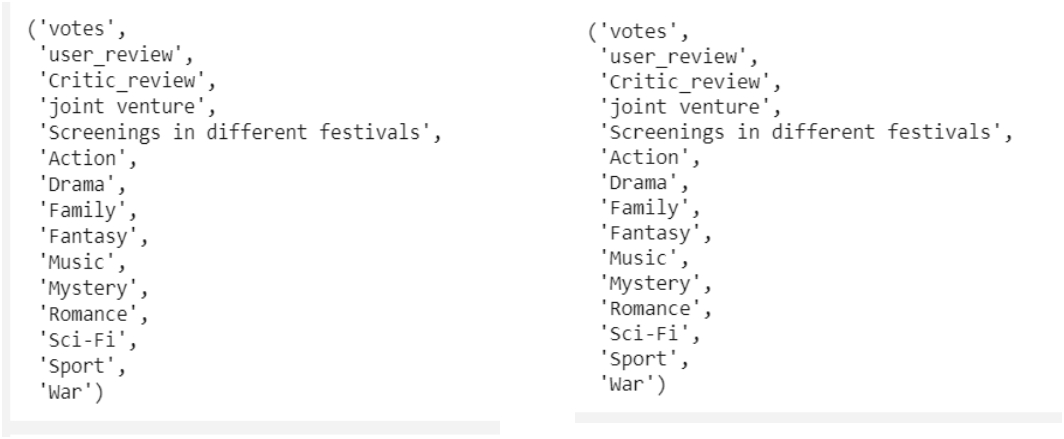


Figure 3.12: Orange Setup for Our Data Analysis to Find Out Most Important Features Using Trial and Error Method

Forward Selection and Backward Elimination: In the Forward Selection method, the model starts working with an empty feature. It is an iterative model and each iteration model keeps adding new features from the dataset till it gives the best performance. In the Backward Elimination method, the model starts working with all the features of the dataset and removes the least significant features from the dataset. Forward Selection and Backward Elimination both are Wrapper methods. Both the methods provide the same features in output.



(a) (b)

Figure 3.13: Figure (a) Shows the Best Features Calculated by Forward Selection Method. Figure (b) Shows the Best Features Calculated by Backward Elimination Method.

Comparison among the Feature selection methods:

All the features have been analyzed applying four different methods. Wrapper method is widely recognized, which is better than the Filter method for feature selection [36]. So, instead of comparing the Wrapper and Filter method, comparison among the selected models in different ways. First of all, Forward Selection and Backward Elimination and Trial & Error method have been analyzed combinedly. In the next step correlation has been incorporated with these three. Based on two ways of feature analysis, it has been found that the second one is more effective for the prepared dataset.

Table 3.2: Dataset Features After Feature Selection Using Forward Selection, Backward Elimination with Trial and Error Method Combinedly (Method 1)

Feature	**Description**
Genre	Category of the movie (drama, romantic, history, war, action and so on)
Director	Director name
Actor	Leading male actor of the movie
Actress	Leading female actor of the movie
Villain	Leading negative role of the movie
Votes	User votes for a movie
Writer 1	Story writer of the movie

User review	Number of user comments
Critic review	Number of critic comments
Joint venture	Movie jointly directed by more than one country
Screenings in different festival	Movie exhibited in different festival or not

Table 3.3: Dataset Features After Feature Selection Using Correlation, Forward Selection, Backward Elimination with Trial and Error Method Combinedly (Method 2)

Feature	**Description**
Genre	Category of the movie (drama, romantic, history, war, action and so on)
Director	Director name
Actor	Leading male actor of the movie
Actress	Leading female actor of the movie
Villain	Leading negative role of the movie
Votes	User votes for a movie
Writer 1	Story writer of the movie
User review	Number of user comments
Critic review	Number of critic comments
Foreign actor	Foreign actor participation in a movie
Joint venture	Movie jointly directed by more than one country
Number of release county	Number of released countries
Screenings in different festival	Movie exhibited in different festival or not

3.7 Machine Learning Model

After finalizing the features we have applied five algorithms as follows:

- Support vector machine (SVM): SVM is a popular machine learning algorithm for pattern recognition and classification. In this work, three different classes have applied namely: binary class, triple class, and four class. This model is selected as it is best for binary classification and also can produce a good result for binary classification [26].
- Random Forest: Random Forest is a tree-based machine learning algorithm. It consists of a large number of decision trees and a very powerful predictive model [22]. The Random Forest can find very complex dependencies but not fit for noisy

data. This algorithm can produce better accuracy because instead of searching the most important feature directly, it searches from a random subset of features. Random inputs and random features creation ensure a good result in classification [37]. This is the purpose to select Random Forest for this work.

- Decision Tree: Decision Tree is also a tree-based algorithm for regression and classification model [11]. It is a flowchart like structure in which each node acts as an attribute test, each branch works like a test outcome, and each leaf node operates as a class label. The contrast between the Decision Tree and the Random Forest is the decision tree is built on an entire dataset with all features, whereas random forest randomly selects feature and make subsets. The first tree-based model (Random Forest) shows good accuracy that why another tree-based model for our work to compare with other models.
- Logistic Regression: Logistic Regression is useful for categorical data therefore this has been selected. Logistic Regression performs well with classification problems, as well as for small datasets [26]. Dhallywood movie dataset is small therefore this model is selected for comparison.
- K-Nearest Neighbor (KNN): KNN is one of the simplest and widely used machine learning prediction algorithms. For a large dataset it works slowly but for a small dataset it predicts very fast. In the KNN algorithm, the parameter k is used to select k number of nearest neighbors, then calculate the distance of nearest neighbors. It finds the category which has the highest number of neighbors considering the query points. This model is selected for its faster operation for smaller dataset.

For each algorithm we calculate three times with different classes of target variables.

- Binary class classification – Yes or No
- Triple class classification – Excellent, Good, Bad.
- Four class classification - Excellent, Very Good, Good, Bad.

3.8 Validation

We use three model validation techniques that are as follows. Model detailed discussion is given in the next chapter.

- Confusion Matrix: The matrix is NxN; we use 3 different values of N according to our different classes, which are 2, 3 and 4.
- 10-Fold Cross Validation.
- ROC Curve.

Chapter 4

Experimental Analysis and Results

This chapter presents experimental process and results. The experiment has been performed using different machine learning models such as: SVM, KNN, Decision Tree, Logistic Regression, and Random Forest. After that these results have been validated using 10-Fold Cross Validation, Confusion Matrix and ROC curve.

4.1 Experimental Analysis with Different Class

The goal of this research work is to evaluate the Dhallywood movie status using machine learning. In this research work each model has been applied for three different classes. In Chapter 3 Section 3.7, these categories have been discussed.

4.1.1 Support Vector Machine (SVM)

Three different class classification results obtained using Forward Selection, Backward Elimination with Trial and Error method are as follows. (Method 1)

- Binary class classification – Yes or No

  Accuracy – 68.12%

- Triple class classification – Excellent, Good, Bad

  Accuracy – 70.93%

- Four class classification - Excellent, Very Good, Good, Bad

  Accuracy – 46.82%

Three different class classification results obtained using correlation, Forward Selection, Backward Elimination with Trial and Error method are as follows. (Method 2)

- Binary class classification – Yes or No

  Accuracy – 68.20%

- Triple class classification – Excellent, Good, Bad

  Accuracy – 70.93%

- Four class classification - Excellent, Very Good, Good, Bad

Accuracy – 47.39%

4.1.2 K Nearest Neighbor (KNN)

Three different class classification results obtained using Forward Selection, Backward Elimination with Trial and Error method are as follows. (Method 1)

- Binary class classification – Yes or No

  Accuracy – 67.63%

- Triple class classification – Excellent, Good, Bad

  Accuracy – 70.93%

- Four class classification - Excellent, Very Good, Good, Bad

  Accuracy – 42.77%

Three different class classification results obtained using correlation, Forward Selection, Backward Elimination with Trial and Error method are as follows. (Method 2)

- Binary class classification – Yes or No

  Accuracy – 67.63%

- Triple class classification – Excellent, Good, Bad

  Accuracy –70.97 %

- Four class classification - Excellent, Very Good, Good, Bad

  Accuracy – 43.93%

4.1.3 Logistic Regression

Three different class classification results obtained using Forward Selection, Backward Elimination with Trial and Error method are as follows. (Method 1)

- Binary class classification – Yes or No

  Accuracy – 63.003%

- Triple class classification – Excellent, Good, Bad

  Accuracy – 73.23%

- Four class classification - Excellent, Very Good, Good, Bad

  Accuracy – 45.66%

Three different class classification results obtained using correlation, Forward Selection, Backward Elimination with Trial and Error method are as follows. (Method 2)

- Binary class classification – Yes or No

  Accuracy – 65%

- Triple class classification – Excellent, Good, Bad

  Accuracy – 73%

- Four class classification - Excellent, Very Good, Good, Bad

  Accuracy – 47%

4.1.4 Decision Tree

Three different class classification results obtained using Forward Selection, Backward Elimination with Trial and Error method are as follows. (Method 1)

- Binary class classification – Yes or No

  Accuracy – 72.25%

- Triple class classification – Excellent, Good, Bad

  Accuracy – 78.19%

- Four class classification - Excellent, Very Good, Good, Bad

  Accuracy – 47.40%

Three different class classification results obtained using correlation, Forward Selection, Backward Elimination with Trial and Error method are as follows. (Method 2)

- Binary class classification – Yes or No

  Accuracy – 75%

- Triple class classification – Excellent, Good, Bad

Accuracy – 80%

- Four class classification - Excellent, Very Good, Good, Bad

  Accuracy – 49%

4.1.5 Random Forest

Three different class classification results obtained using Forward Selection, Backward Elimination with Trial and Error method are as follows. (Method 1)

- Binary class classification – Yes or No

  Accuracy – 68.78%

- Triple class classification – Excellent, Good, Bad

  Accuracy – 82.55%

- Four class classification - Excellent, Very Good, Good, Bad

  Accuracy – 43.95%

Three different class classification results obtained using correlation, Forward Selection, Backward Elimination with Trial and Error method are as follows. (Method 2)

- Binary class classification – Yes or No

  Accuracy – 71%

- Triple class classification – Excellent, Good, Bad

  Accuracy – 85%

- Four class classification - Excellent, Very Good, Good, Bad

  Accuracy – 47%

4.2 10-Fold Cross-Validation

Cross-validation is a way to ensure perfection of a model with bias and variance value, by running the whole process of testing and training multiple times. To check the bias of all models the mean of 10 values have been calculated. Low bias indicates more perfection of a model whereas high bias indicates imperfection. The standard deviation shows the variance of a model. If the value is high the performance of the model varies a lot and it

tends to imperfection. The sample code is used for all models is given in Appendix section A.4. A comparison of bias-variance among all models have been shown in Table 4.2.

4.2.1 Support Vector Machine (SVM)

Bias and variance have been calculated using Forward Selection, Backward Elimination with Trial and Error feature selection method. (Method 1)

Binary class classification

- Mean : 0.6440372670807453
- Standard deviation : 0.04013275318663973

Triple class classification

- Mean : 0.7480519480519481
- Standard deviation : 0.019437181229994507

Four class classification

- Mean : 0.39519668737060043
- Standard deviation : 0.06602086806909442

Bias and variance have been calculated using correlation, Forward Selection, Backward Elimination with Trial and Error feature selection method. (Method 2)

Binary class classification

- Mean : 0.6267908902691512
- Standard deviation : 0.06638988103800655

Triple class classification

- Mean : 0.7519480519480519
- Standard deviation : 0.0213002850218918

Four class classification

- Mean : 0.3995238095238095
- Standard deviation: 0.06408354470302617

4.2.2 K Nearest Neighbor (KNN)

Bias and variance have been calculated using Forward Selection, Backward Elimination with Trial and Error feature selection method. (Method 1)

Binary class classification

- Mean : 0.6439958592132505
- Standard deviation : 0.02906838934005273

Triple class classification

- Mean : 0.764935064935065
- Standard deviation : 0.04626367046897892

Four class classification

- Mean : 0.40828157349896477
- Standard deviation : 0.07451575572142108

Bias and variance have been calculated using correlation, Forward Selection, Backward Elimination with Trial and Error feature selection method. (Method 2)

Binary class classification

- Mean : 0.6454451345755693
- Standard deviation : 0.02907870995843102

Triple class classification

- Mean : 0.48571428571428565
- Standard deviation : 0.0631441339280167

Four class classification

- Mean : 0.39668737060041404
- Standard deviation : 0.07732967757697784

4.2.3 Logistic Regression

Bias and variance have been calculated using Forward Selection, Backward Elimination with Trial and Error feature selection method. (Method 1)

Binary class classification

- Mean : 0.6499585921325053
- Standard deviation : 0.06174370221064117

Triple class classification

- Mean : 0.762337662337662
- Standard deviation : 0.03020702168730652

Four class classification

- Mean : 0.42231457800511507
- Standard deviation : 0.055636743611095836

Bias and variance have been calculated using correlation, Forward Selection, Backward Elimination with Trial and Error feature selection method. (Method 2)

Binary class classification

- Mean : 0.6513871635610766
- Standard deviation : 0.05684877969841414

Triple class classification

- Mean : 0.7597402597402598
- Standard deviation : 0.031943828249997

Four class classification

- Mean : 0.39956521739130435
- Standard deviation : 0.0573226422821773

4.2.4 Decision Tree

Bias and variance have been calculated using Forward Selection, Backward Elimination with Trial and Error feature selection method. (Method 1)

Binary class classification

- Mean : 0.646935817805383
- Standard deviation : 0.026138509315950204

Triple class classification

- Mean : 0.8197278911564627
- Standard deviation : 0.04433567787190713

Four class classification

- Mean : 0.42409937888198757
- Standard deviation : 0.0507609790804009

Bias and variance have been calculated using correlation, Forward Selection, Backward Elimination with Trial and Error feature selection method. (Method 2)

Binary class classification

- Mean : 0.6223395445134575
- Standard deviation : 0.03196682677347736

Triple class classification

- Mean : 0.7948051948051947
- Standard deviation : 0.05773015784647498

Four class classification

- Mean : 0.4081366459627329
- Standard deviation : 0.05157865183420961

4.2.5 Random Forest

Bias and variance have been calculated using Forward Selection, Backward Elimination with Trial and Error feature selection method. (Method 1)

Binary class classification

- Mean : 0.6656521739130434
- Standard deviation : 0.031164812475802967

Triple class classification

- Mean : 0.8183356117566645
- Standard deviation : 0.062259585406755864

Four class classification

- Mean : 0.40660455486542446
- Standard deviation : 0.056225148777777186

Bias and variance have been calculated using correlation, Forward Selection, Backward Elimination with Trial and Error feature selection method. (Method 2)

Binary class classification

- Mean : 0.6599585921325051
- Standard deviation : 0.036883138398081924

Triple class classification

- Mean : 0.8194805194805195
- Standard deviation : 0.040842039463678904

Four class classification

- Mean : 0.4327329192546584
- Standard deviation : 0.057205118809511096

4.3 Confusion Matrix

Confusion matrix is a powerful tool of machine learning for analyzing binary and multiple class classifier performance. It is a table that is used by a classification model/classifier to mark out the achievement of known true value on a set of test data. For measuring accuracy, precision, recall, and f1 score, confusion matrix is needed immensely.

		Actual Value: Positive (1)	Actual Value: Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 4.1: Confusion Matrix

True Positive (TP): When the machine predicts the positive value, that means both predicted and actual class is positive/yes (1) or class are labeled correctly by classifier, it is called true positive.

True Negative (TN): When the machine predicts the negative value, that means both predicted and actual class are negative/no (0) or a classifier can label the class correctly where value is negative is called true negative.

False Positive (FP): Machine predicts the value is positive/yes (1) but its actual value is negative/no (0) that is classifier label the class incorrectly as positive is called false positive.

False Negative (FN): Machine predicts the value is negative/no (0) but its actual value is positive/yes (1) that is classifier label the class incorrectly as positive is called false positive.

Accuracy: When we train a model, we measure its accuracy either trained accuracy or prediction accuracy. Trained accuracy represents that our trained model is being trained correctly by measuring its ratio. Predicted accuracy represents the predicted observation of the total observation or known value. On the other hand, accuracy is the percentage of the test set which is classified correctly.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad \ldots\ldots (4.1)$$

Precision: Precision is the measurement of the exactness of the actual class. That is predicted result is being consistent when the measurement is repeated or how well a result can be determined when it is measured.

$$\text{precision} = \frac{TP}{TP+FP} \quad \ldots\ldots(4.2)$$

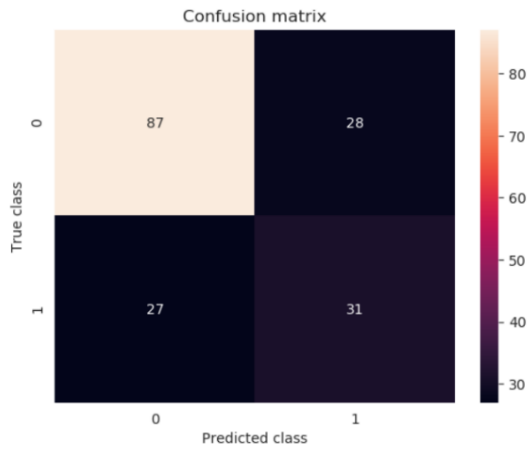
Recall: Recall is the measurement of completeness of the actual class. When we predict on test set then the ratio of positive observation on all the observation of actual positive is called recall.

$$\text{recall} = \frac{TP}{TP+FN} \quad \ldots\ldots(4.3)$$

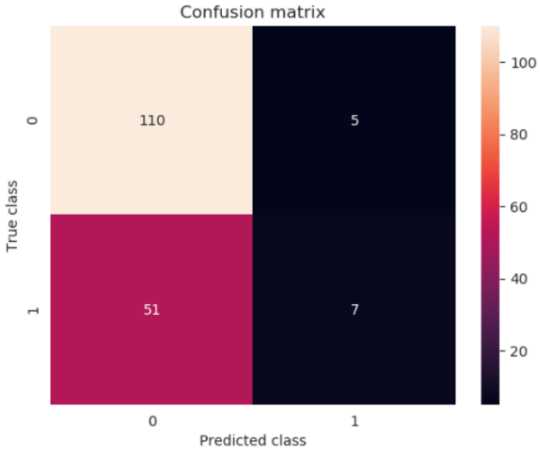
F Score: F1 score / F measurement is the combination of both precision and recall that it is taken both false positives and false negatives into accounts.

$$\text{F} - \text{score} = \frac{2\times precision\times recall}{precision+recall} \quad \ldots\ldots(4.4)$$

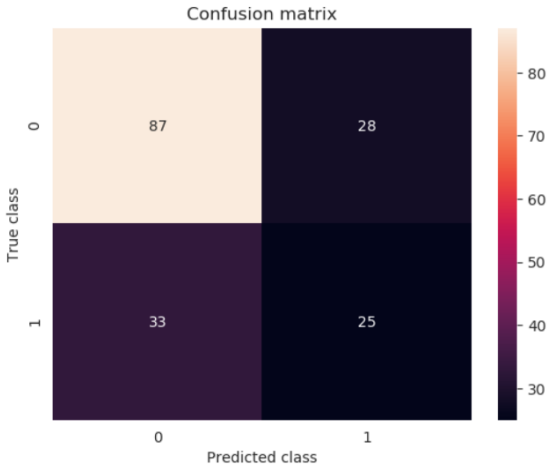
4.3.1 Confusion Matrix for Binary Class



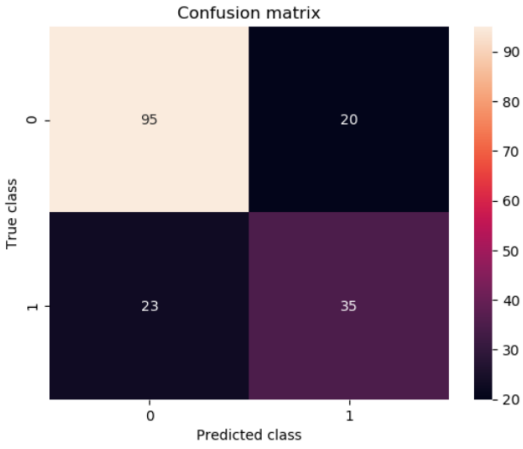
(a) SVM



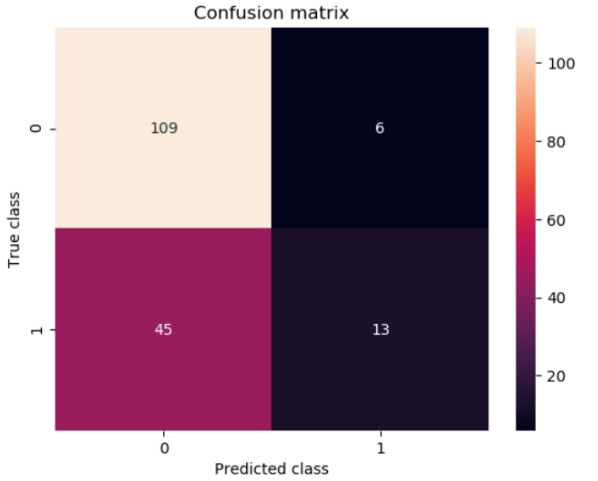
(b) KNN



(c) Logistic Regression



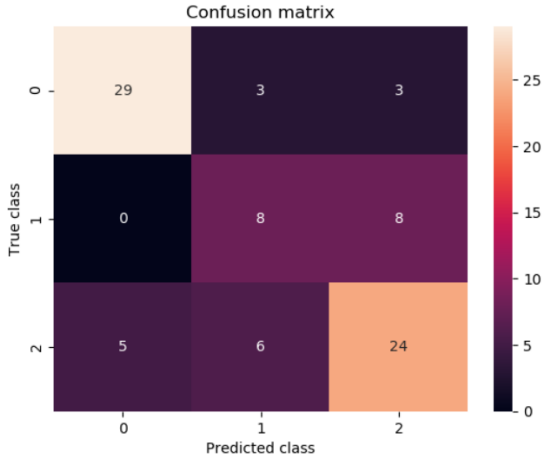
(d) Decision Tree



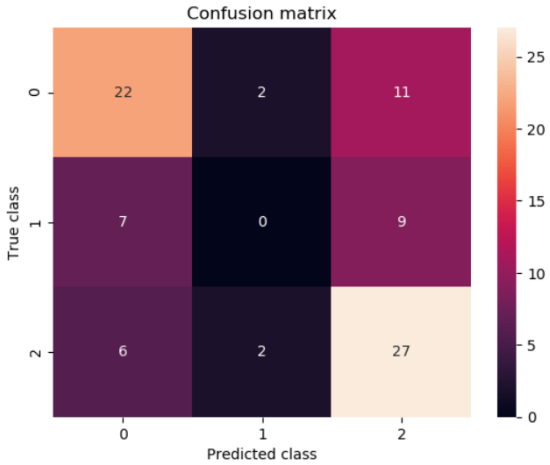
(e) Random Forest

Figure 4.2 : Confusion Matrix of Five ML Models for Binary Class Classification

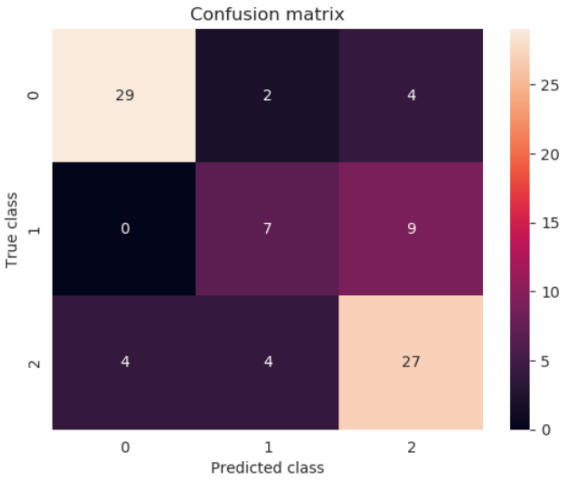
4.3.2 Confusion Matrix for Triple Class



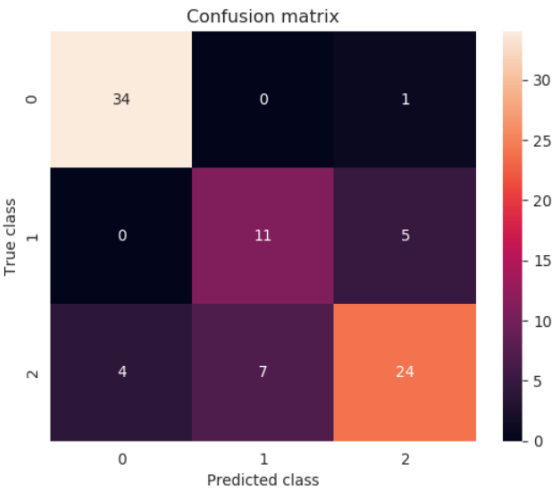
(a) SVM



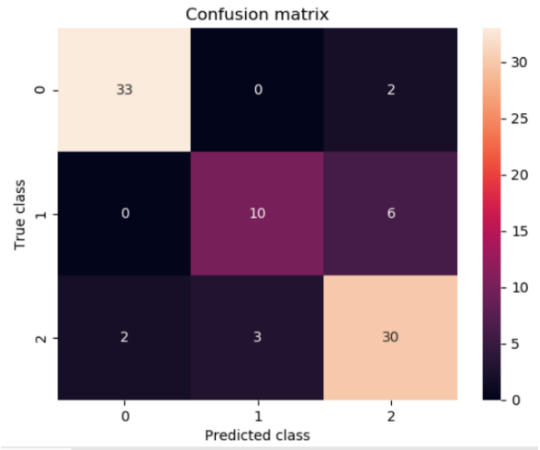
(b) KNN



(c) Logistic Regression



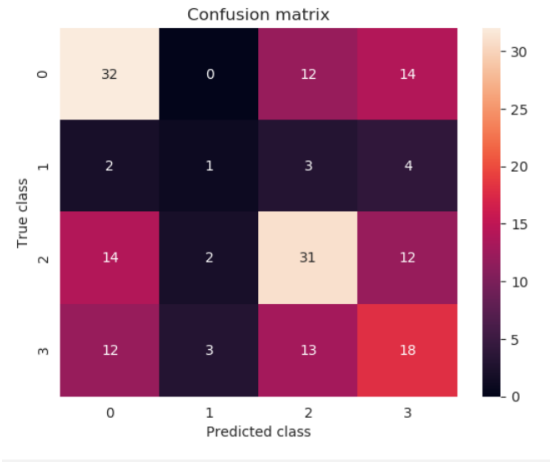
(d) Decision Tree



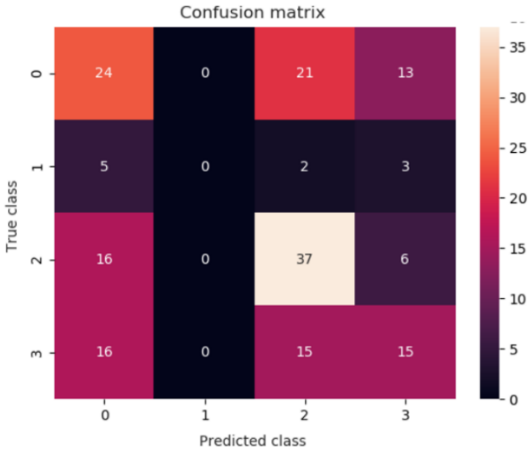
(e) Random Forest

Figure 4.3: Confusion Matrix of Five ML Models for Triple Class Classification

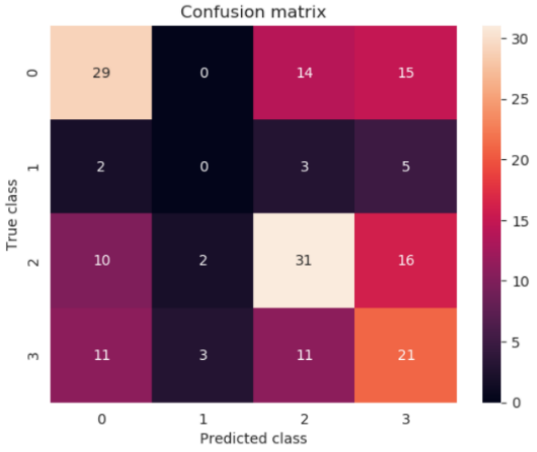
4.3.3 Confusion Matrix for Four Class



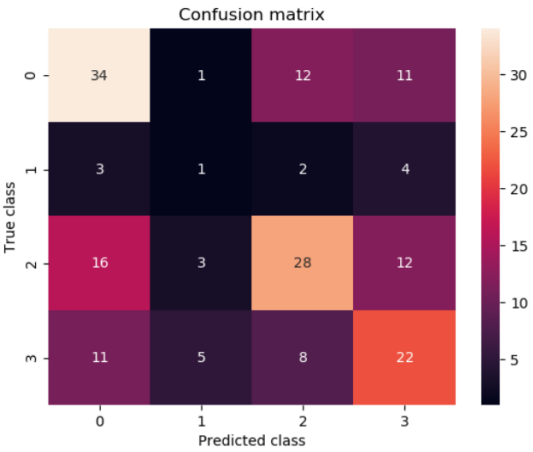
(a) SVM



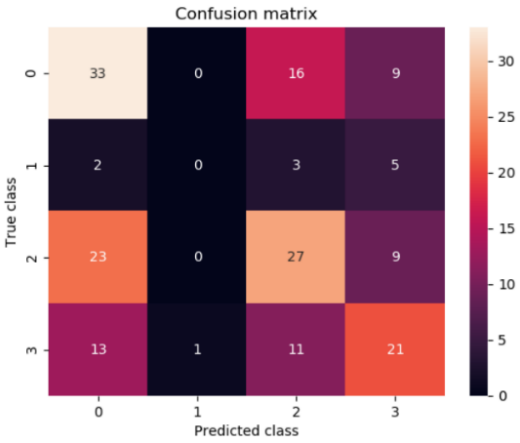
(b) KNN



(c) Logistic Regression



(d) Decision Tree



(e) Random Forest

Figure 4.4: Confusion Matrix of Five ML Models for Four Class Classification

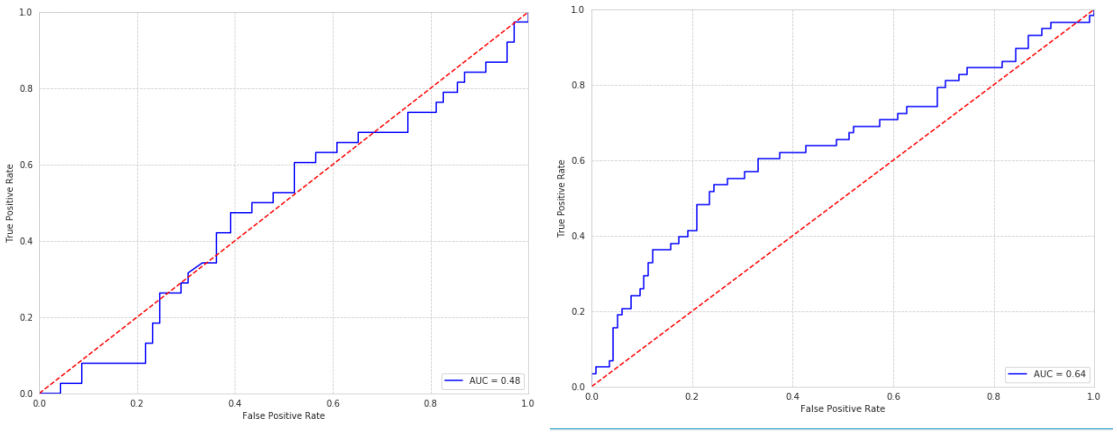
The confusion matrix illustrates the performance of a classification model that can give an actual idea of a model. Accuracy can hide the exact details of a model because it misleads when each class has unequal data. Whereas the confusion matrix can reveal the correct information by calculating correct and incorrect predictions of each class. In section 4.1 it is shown that the accuracy of five supervised ML models with three separate class classifications. Accuracy works well when balanced data is available for each class. If data is imbalanced another performance measure for data is required. For the analytical purpose in some cases false negative is less concern where precision is better. If false negative is a high concern, then recall is better. Imbalanced data sometimes tends to high recall and sometimes tends to high precision. The harmonic mean of precision and recall (F1score) is also an effective measure for imbalanced data. More FP indicates less precision and more FN indicates less recall. Both these values affect the performance measures of a model. Four class classifier performance is very poor than other classes because of imbalance data and more false predictions.

4.4 ROC Curve

A ROC (Receiver Operator Characteristic) curve is a graphical representation that shows the diagnostic ability of a binary classifier. It is possible to create a ROC curve for N number of classes using the “One Vs All” methodology. For example, our target variable has three classes named Excellent, Good, and Bad. We will have one ROC for Excellent classified against Good and Bad, another ROC for Good classified against Excellent and Bad, and the third one of Bad classified against Good and Excellent. A ROC

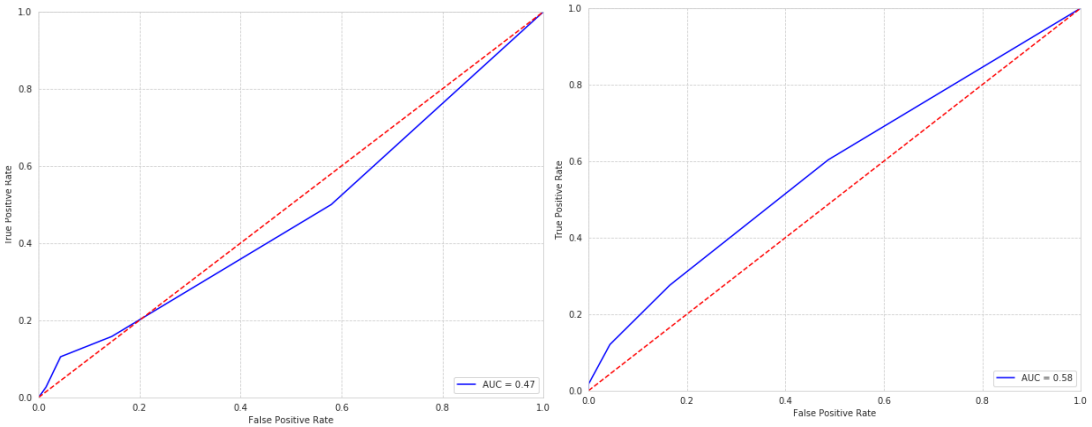
curve represents how a model deals with true positive rate and false positive rate, where TPR has been plotted as a function of FPR. One common approach to summarize the performance of a classifier is to calculate AUC (Area Under the ROC Curve). AUC score 5.0 indicates no distinction between positive and negative class, below 5.0 indicates poor performance, a range 0.7 to 0.8 is considered acceptable, a range 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding.

4.4.1 ROC Curve for Binary Class Classification



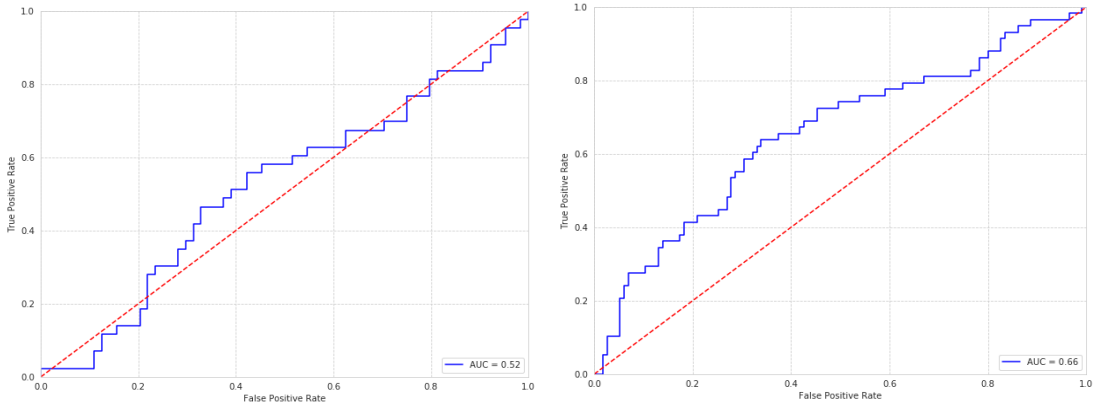
(a) Before Feature Selection (b) After Feature Selection

Figure 4.5: ROC Curve of SVM Binary Class Classification



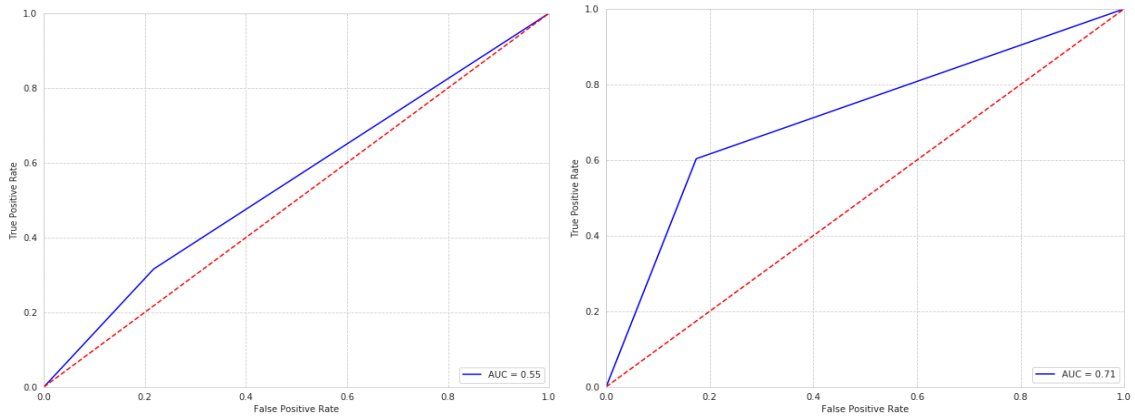
(a) Before Feature Selection (b) After Feature Selection

Figure 4.6: ROC Curve of KNN Binary Class Classification



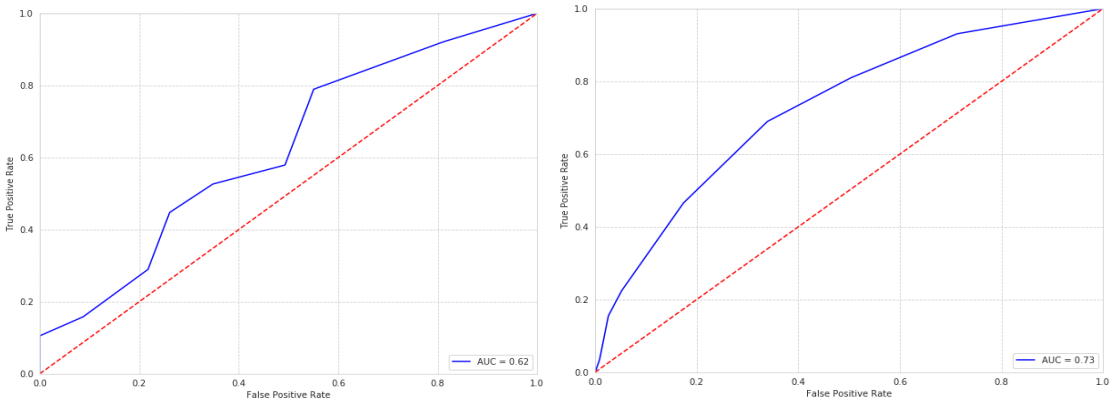
(a) Before Feature Selection (b) After Feature Selection

Figure 4.7: ROC Curve of Logistic Regression Binary Class Classification



(a) Before Feature Selection (b) After Feature Selection

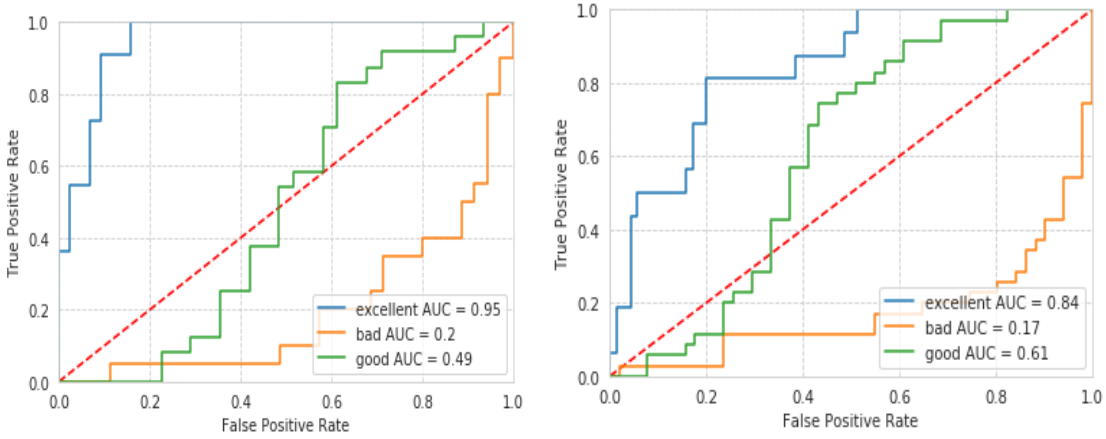
Figure 4.8: ROC Curve of Decision Tree Binary Class Classification



(a) Before Feature Selection (b) After Feature Selection

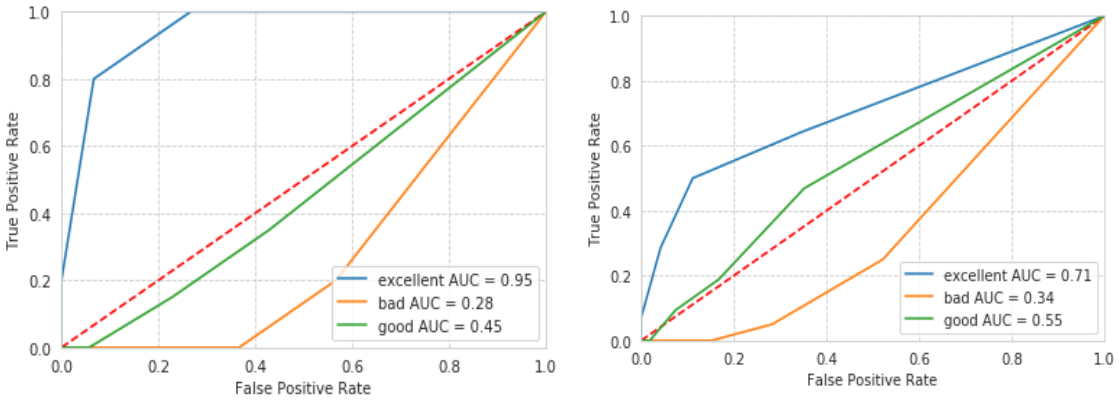
Figure 4.9: ROC Curve of Random Forest Binary Class Classification

4.4.2 ROC Curve for Triple Class Classification



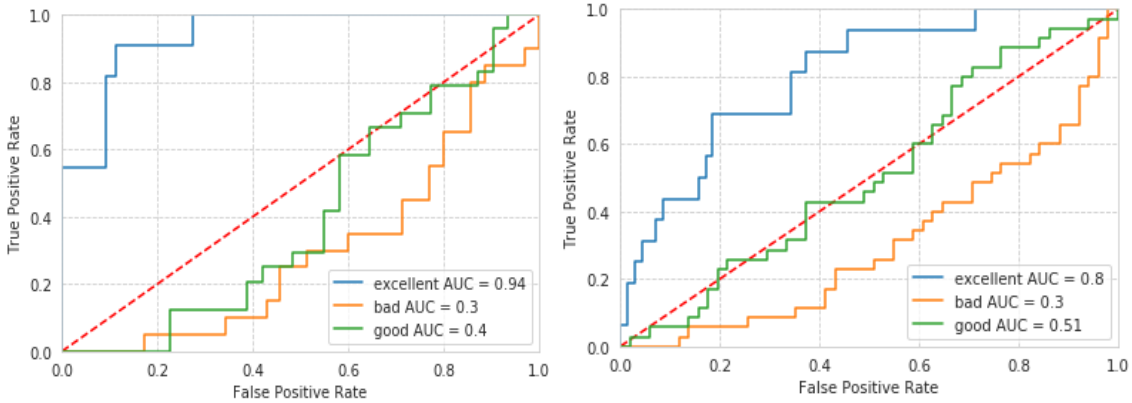
(a) Before Feature Selection (b) After Feature Selection

Figure 4.10: ROC Curve of SVM Triple Class Classification



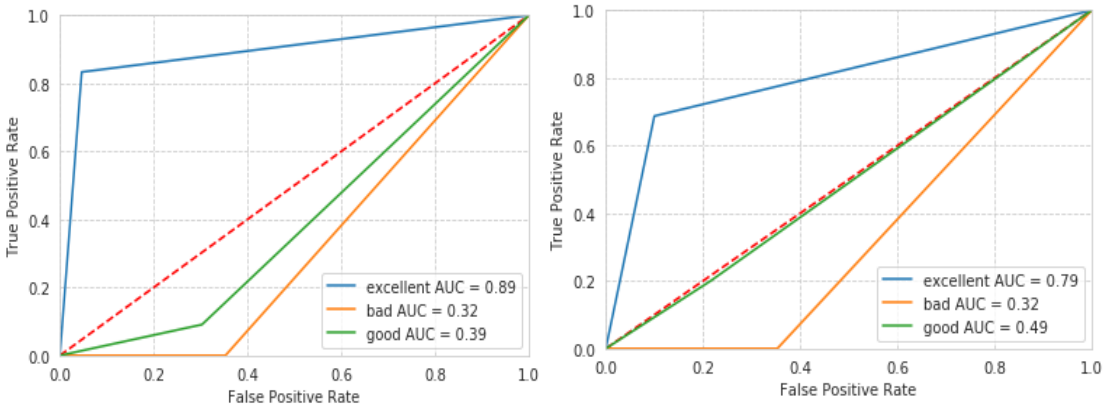
(a) Before Feature Selection (b) After Feature Selection

Figure 4.11: ROC Curve of KNN Triple Class Classification



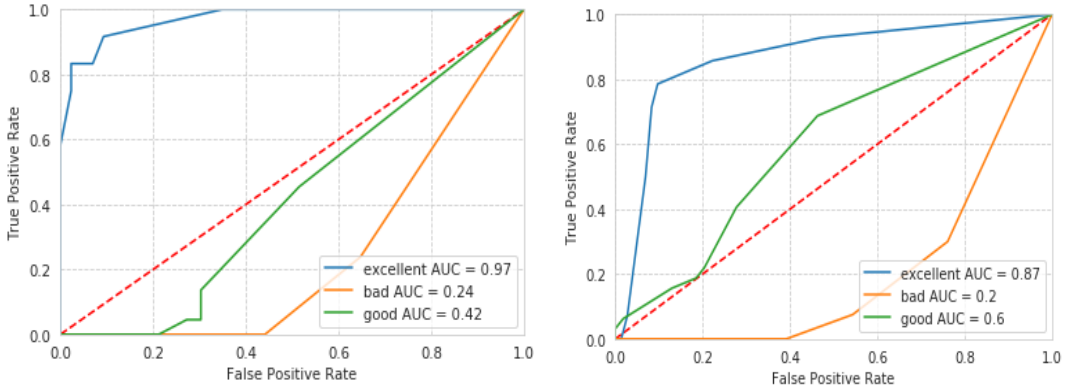
(a) Before Feature Selection (b) After Feature Selection

Figure 4.12: ROC Curve of Logistic Regression Triple Class Classification



(a) Before Feature Selection (b) After Feature Selection

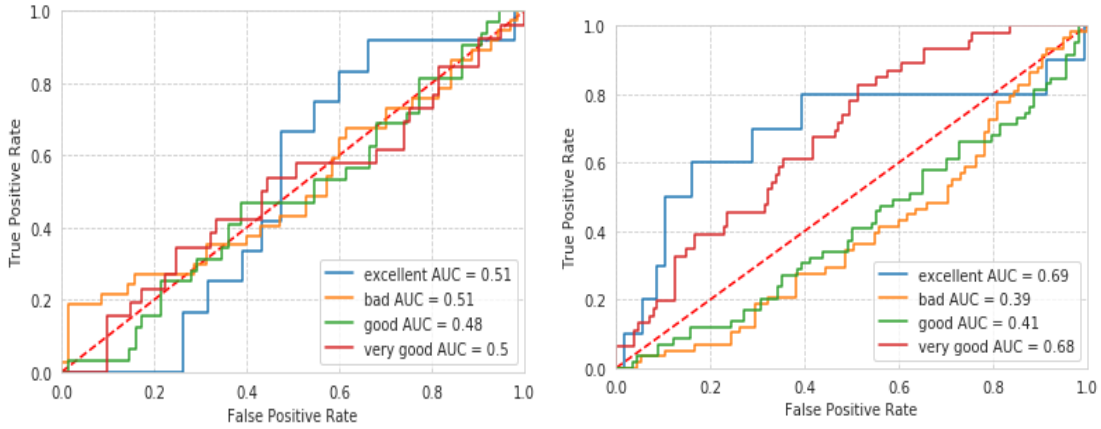
Figure 4.13: ROC Curve of Decision Tree Triple Class Classification



(a) Before Feature Selection (b) After Feature Selection

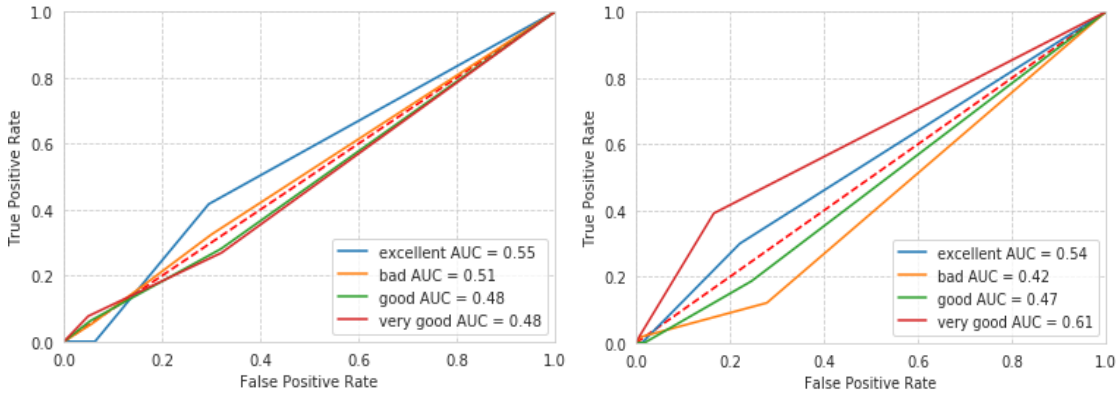
Figure 4.14: ROC Curve of Random Forest Triple Class Classification

4.4.3 ROC Curve for Four Class Classification



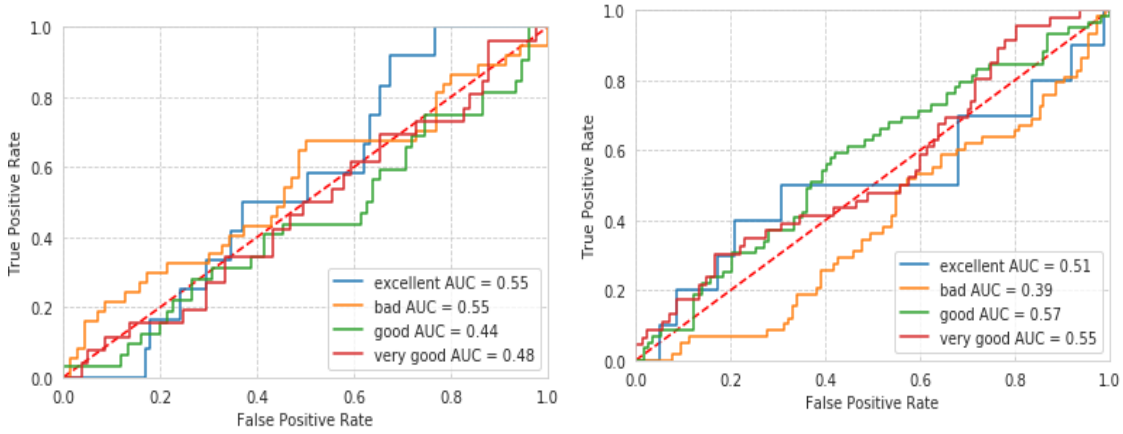
(a) Before Feature Selection (b) After Feature Selection

Figure 4.15: ROC Curve of SVM Four Classification



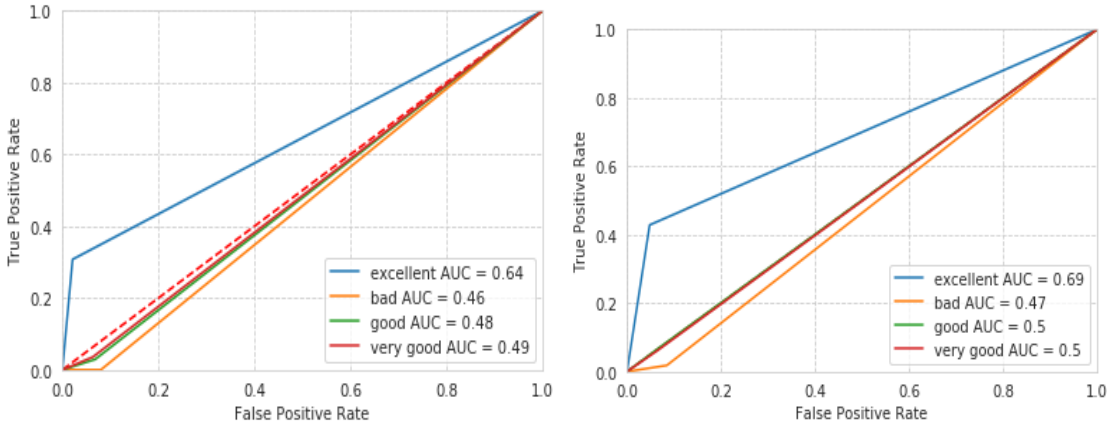
(a) Before Feature Selection (b) After Feature Selection

Figure 4.16: ROC Curve of KNN Four Class Classification



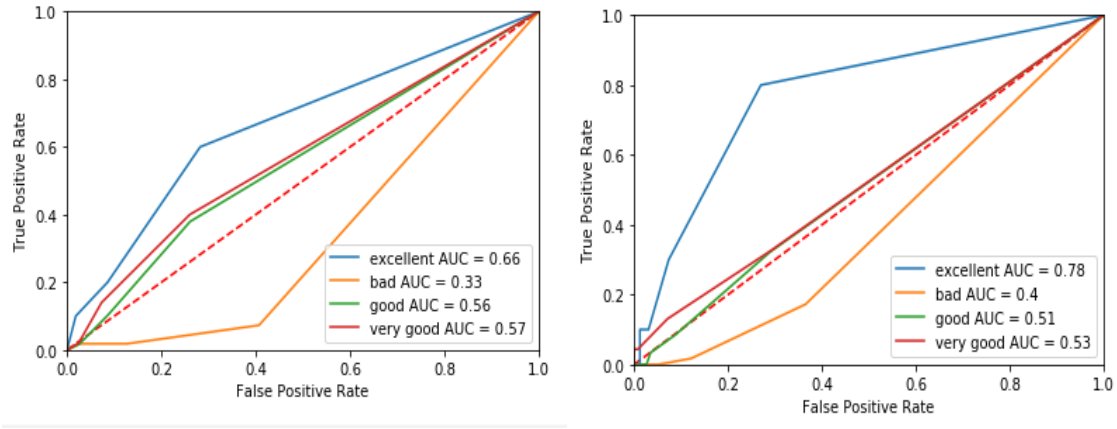
(a) Before Feature Selection (b) After Feature Selection

Figure 4.17: ROC Curve of Logistic Regression Four Class Classification



(a) Before Feature Selection (b) After Feature Selection

Figure 4.18: ROC Curve of Decision Tree Four Class Classification



(a) Before Feature Selection (b) After Feature Selection

Figure 4.19: ROC Curve of Random Forest Four Class Classification

AUC is a performance measure of ML models. Figure 4.5 to Figure 4.19 presents the ROC curve for each model twice one is before feature selection another is after feature selection. The ROC curve of binary class classification model shows better performance after feature selection. However, for another two classes (triple and four) the AUC score is not balanced.

4.4 Results Analysis and Comparison

To determine the dependency of two categorical values the Chi-square test is used. The dependency of four categorical values with three different classes have been determined using this test. The obtained result shown in Table 4.1. From the observation of data presented in Table 4.1 it is found that actor-actress have high dependency with target classes. Here the value of alpha is considered as 0.10. If the p-value is lower than alpha, it indicates an interdependence between two variables. Whereas higher alpha value means the hypothesis of interdependence is incorrect.

Table 4.1: Categorical Features P-value Using Chi2 Method

Class	**Categorical Feature (P-value)**			
	Actor	Actress	Director	Villain
Binary	0.000601	0.0100591	0.1714873	0.218612
Triple	0.000127	0.0002978	0.0008706	0.011534
Four	0.000064	0.00000008	0.0030367	0.003046

Table 4.2: Actor and Actress Impact On Model Accuracy

ML models	Classification	Accuracy (%) Without Actor and Actress (Method 1)	Accuracy (%) With Actor and Actress (Method 1)	Accuracy (%) Without Actor and Actress (Method 2)	Accuracy (%) With Actor and Actress (Method 2)
SVM	Binary Class	63.97%	68.12%	61.74%	68.20%
	Triple Class	65.18%	70.93%	63.23%	70.93%
	Four Class	34.26%	46.82%	34.56%	47.39%
KNN	Binary Class	66.21%	67.63%	66.44%	67.63%
	Triple Class	68.11%	70.93%	68.01%	70.97%
	Four Class	36.51%	42.77%	42.11%	43.93%
Logistic Regression	Binary Class	62.01%	63.003%	59.97%	64.73%
	Triple Class	69.03%	73.23%	62.92%	73.25%
	Four Class	33.33%	45.66%	36.51%	46.82%
Decision Tree	Binary Class	70.68%	72.25%	71.92%	75.14%
	Triple Class	72.36%	78.19%	73.19%	80.23%
	Four Class	37.60%	47.40%	38.20%	49.13%
Random Forest	Binary Class	66.31%	68.78%	66.53%	70.52%
	Triple Class	79.11%	82.55%	79.38%	84.88%
	Four Class	38.76%	43.93%	38.83%	46.82%

Table 4.3: Accuracy(%), Mean and Standard Deviation of Five ML Algorithms With Three Different Class Using 10-Fold Cross Validation

ML models	**Classification**	**Method 1**			**Method 2**		
		Accuracy	**Mean**	**Std**	**Accuracy**	**Mean**	**Std**
SVM	Binary Class	68.12%	0.6440	0.0401	68.20%	0.6267	0.0663
	Triple Class	70.93%	0.7480	0.0191	70.93%	0.7519	0.0213
	Four Class	46.82%	0.3951	0.0660	47.39%	0.3995	0.0640
KNN	Binary Class	67.63%	0.6439	0.0290	67.63%	0.6454	0.0290
	Triple Class	70.93%	0.7649	0.0462	70.97%	0.4857	0.0631
	Four Class	42.77%	0.4082	0.0745	43.93%	0.3966	0.0773
Logistic Regression	Binary Class	63.003%	0.6499	0.0617	64.73%	0.6513	0.0568
	Triple Class	73.23%	0.7623	0.0302	73.25%	0.7597	0.0319
	Four Class	45.66%	0.4223	0.0556	46.82%	0.3995	0.0573
Decision Tree	Binary Class	72.25%	0.6469	0.0261	75.14%	0.6223	0.0319
	Triple Class	78.19%	0.8197	0.0443	80.23%	0.7948	0.0577
	Four Class	47.40%	0.4240	0.0507	49.13%	0.4081	0.0515
Random Forest	Binary Class	68.78%	0.6656	0.0311	70.52%	0.65995	0.0368
	Triple Class	82.55%	0.8183	0.0622	84.88%	0.8194	0.0408
	Four Class	43.93%	0.4066	0.0562	46.82%	0.4327	0.0572

Table 4.4: Different ML Models Classification Results with Accuracy, Precision, Recall and F-score Using Forward Selection, Backward Elimination with Trial and Error Feature Selection Method (Method 1)

ML models	**Classification**	**Accuracy (%)**	**Precision** (weighted average)	**Recall** (weighted average)	**F-Score** (weighted average)
SVM	Binary Class	68.12%	64.08	63.70	63.86
	Triple Class	70.93%	68.24	68.09	68.16
	Four Class	46.82%	39.57	38.67	38.84
KNN	Binary Class	67.63%	63.32	53.86	49.85
	Triple Class	70.93%	66.11	63.74	64.10
	Four Class	42.77%	31.22	33.20	31.97
Logistic Regression	Binary Class	63.003%	59.02	58.51	58.68
	Triple Class	73.23%	71.50	68.86	69.66
	Four Class	45.66%	35.48	35.62	35.52
Decision Tree	Binary Class	72.25%	68.71	68.29	68.31
	Triple Class	78.19%	78.65	77.58	78.03
	Four Class	47.40%	39.67%	39.57%	39.54%
Random Forest	Binary Class	68.78%	66.73	55.58	52.64
	Triple Class	82.55%	82.53	73.83	76.27
	Four Class	43.93%	40.60	36.32	36.82

Table 4.5: Different ML Models Classification Results with Accuracy, Precision, Recall and F-score Using Correlation, Forward Selection, Backward Elimination with Trial and Error Feature Selection Method (Method 2)

ML models	**Classification**	**Accuracy (%)**	**Precision** (weighted average)	**Recall** (weighted average)	**F-Score** (weighted average)
SVM	Binary Class	68.20%	64.42	64.55	64.48
	Triple Class	70.93%	66.97	67.14	67.04
	Four Class	47.39%	40.01	39.21	39.39
KNN	Binary Class	67.63%	63.33	53.86	49.04
	Triple Class	70.97%	70.15	68.01	68.06
	Four Class	43.93%	32.30	34.02	32.92
Logistic Regression	Binary Class	64.73%	59.83	62.76	59.37
	Triple Class	73.25%	69.74	67.91	68.52
	Four Class	46.82%	36.28	37.02	36.51
Decision Tree	Binary Class	75.14%	72.07	73.01	71.42
	Triple Class	80.23%	76.86	78.15	77.23
	Four Class	49.13%	41.005	40.97	40.85
Random Forest	Binary Class	70.52%	69.60	58.59	57.40
	Triple Class	84.88%	83.39	80.84	81.81
	Four Class	46.82%	35.39	37.08	36.09

The classification results are summarized in Table 4.2-4.5 using two different methods of feature selection. Actor and actress both features have a huge impact on model accuracy. The accuracy achieved by models with or without actor-actress, for both feature selection method is shown in Table 4.2. From Table 4.2 it is revealed that, the impact of actor and actress is extreme on four-class classification. Figure 4.20 and Figure 4.21 interprets the relation between actor & rating and actress & rating.

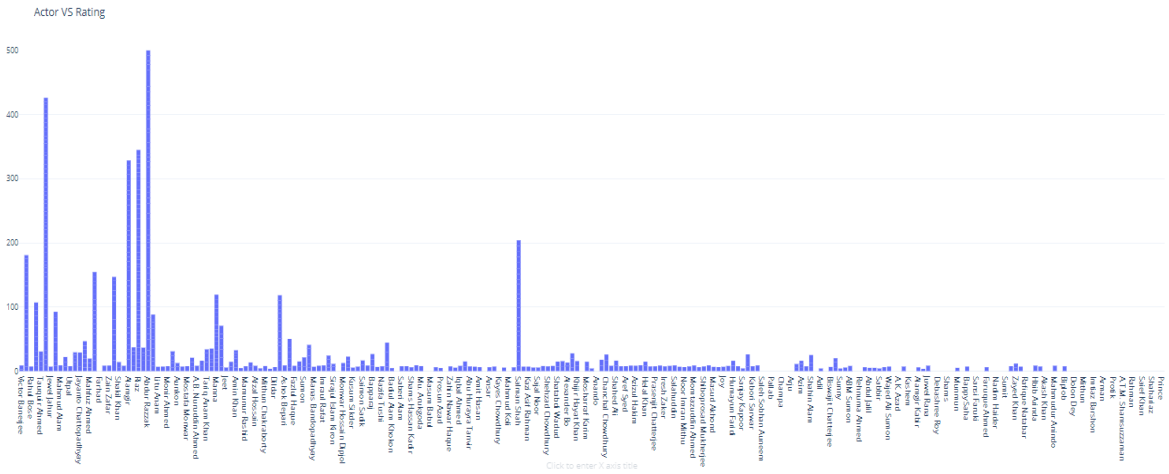


Figure 4.20: Actor Vs Rating Relationship Graph

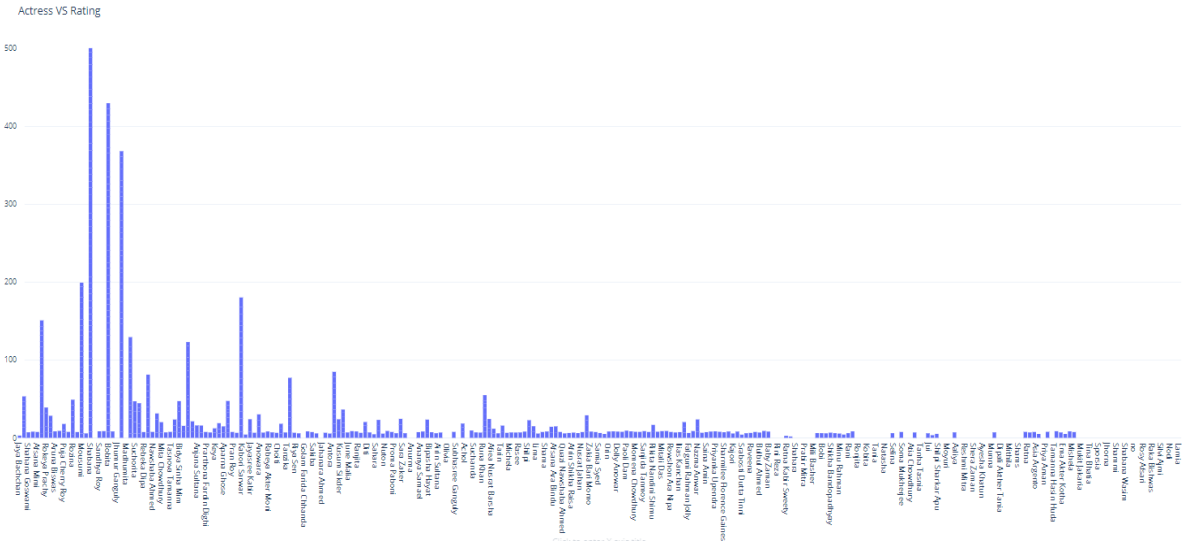


Figure 4.21: Actress Vs Rating Relationship Graph.

Table 4.3, represents the mean and standard deviation (bias and variance) values of ML models. Table 4.4 and Table 4.5 shows the accuracy, precision, recall, and f-score of all classifiers and their comparisons with binary, triple, and four class classification. Among

all classifiers Random Forest shows the best result for triple class classification for both feature selection methods. Firstly, two classes were set for evaluating a movie. After that four classes were set to recognize the status of a movie more accurately. But accuracy was lower than the previous one due to the small size of the data in each class.

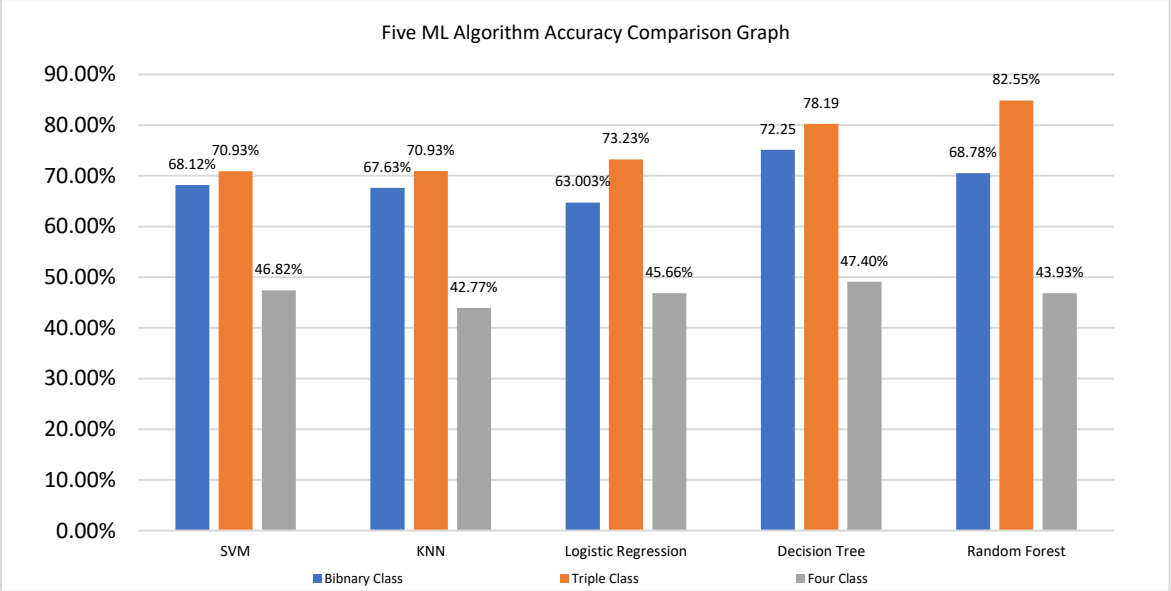


Figure 4.22: Five ML Algorithm Accuracy Comparison Chart (Using Feature Selection Method 1)

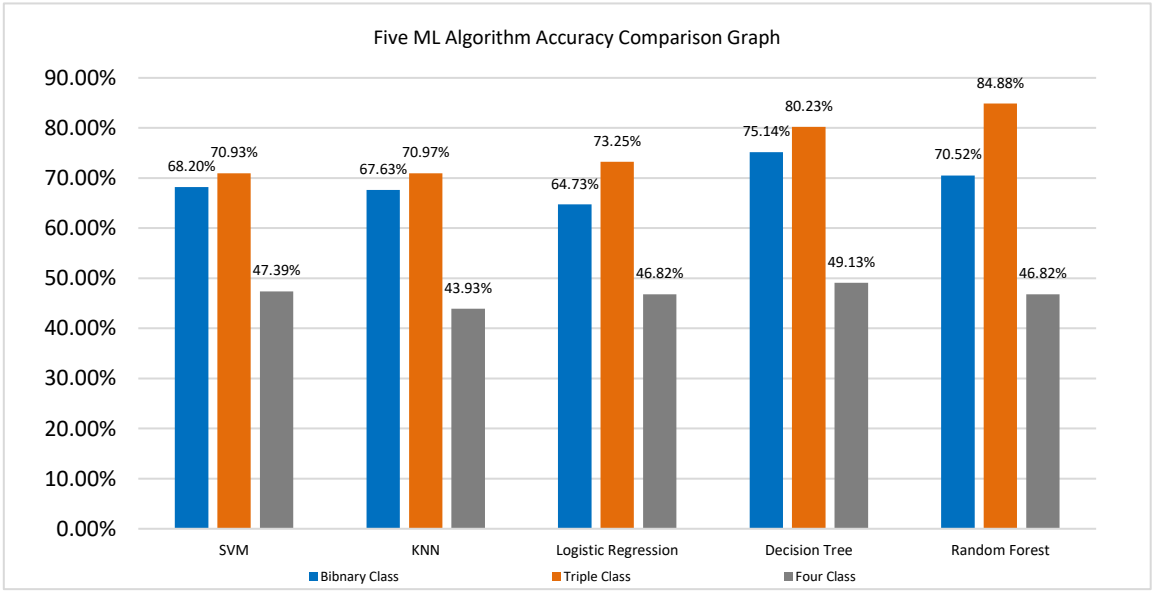


Figure 4.23: Five ML Algorithm Accuracy Comparison Chart (Using Feature Selection Method 2)

The AUC score is always greater than 0.5 for Excellent class shown in section 4.4. For Very Good and Good class AUC is around 0.5 and for Bad class score is always lower than 0.5. Bad class shows a very poor AUC score because of more false predictions.

There is a problem with the classes Very Good and Good. For example, one director has made two movies with the same actor, actress, villain, and writer but one is good another is very good, then model becomes confused at the time of prediction so accuracy becomes low. To overcome this problem the movies have been categorized into three classes. ROC area of each algorithm shows, the excellent class score is better than good and bad class. Five ML model comparison graph is presented, in Figure 4.22 and Figure 4.23 those show accuracies with three different classes.

Chapter 5

Conclusion and Future Work

In chapter presents the concluding remarks and future direction of this research work considering its limitations.

5.1 Discussion

The main purpose of this thesis is to evaluate Dhallywood movies and make a proper dataset that will help to increase the accuracy of our applied model. Here the dataset has been updated until the accuracy of the model reaches to a satisfactory level. A rigorous analysis has been conducted based on Wikipedia information, news article information, the actor and actress biography of a movie to find out its status (super hit, hit and flop). Considering these we can find out which movie has played significant role of on the career of actors or actresses. Collected data is both categorical and numerical, here apply five supervised machine learning algorithms have been applied. In the first phase of data collection, data was collected from bmdb.co. Due to lack of proper information again data scraped from IMDb. In first attempt, three different classes based on IMDb rating did not provide a satisfactory result. Hence information from some other sources were considered for better understanding of the exact status of the movie. Two important success parameters budget and gross income were not obtainable for all movies so these features were excluded from the early stage of our research. In Chapter 4 section 4., two different approaches for feature selection were analyzed. From the analysis method 2 ensures more accuracy than method 1. Finally using 10-fold cross-validation and confusion matrix performances of all models were evaluated. In addition to that, ROC curves have plotted as another performance measure where AUC score shows each class performance. This study discovers actor-actress impact on movie status, if we exclude these two features, model accuracy and performance both fall below than previous. Each algorithm has been applied three times and finally a good accuracy was achieved for Random Forest three-class classification.

5.2 Conclusion

This research intends to assess the status of Dhallywood movies. In this research one of the challenges was preparation of dataset due to scarcity of data. To solve this problem different approaches were applied to manage data to achieve a better outcome. From the comparison of target classes with rating it was found that a very little affinity exists between rating and movie status. Therefore, this research focused on informative data analysis also. From our analysis the effect of different features on movie status have been determined using machine learning. Our research also found the reason behind the popularity of the Dhallywood movie mostly depends on actors and actresses. Among all five algorithms with three different class classification Tree-based algorithm offered the best performance.

5.3 Future Work

This research work sets the ground for future researchers to further investigate and potentially exploit other facts and features to measure the movies popularity of Dhallywood movies. Movie success does not depend only on features related to movies. The number of audiences plays a major role in its success. Movie audiences depend on many features like political conditions and economic stability of a country. If proper rating and revenue data are found in future accuracy of our research can be improved. In addition to that this work can be improved further by increasing both the number of movies and features in the dataset. We would like to include other social media sources as movie data, such as Twitter and Facebook.

In future researchers can apply some other machine learning models with different feature selection methods and compare results with our present work. Future study can also include CNN and ANN to ensure better performance of our prediction model.

References

[1] A. N. N. Approach, “Predicting Movie Box Office Profitability,” pp. 665–670, 2016.

[2] V. Subramaniyaswamy, V. V. M, V. P. R, and R. Logesh, “Predicting Movie Box Office Success using Multiple Regression and SVM,” *2017 Int. Conf. Intell. Sustain. Syst.*, no. Iciss, pp. 182–186, 2017.

[3] A. Mahata, N. Saini, and S. Saharawat, “Intelligent Movie Recommender System Using Machine Learning,” pp. 94–110.

[4] T. Yasseri, “Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data,” vol. 8, no. 8, 2013.

[5] I. W. I. C. Acm, W. Zhang, and S. Skiena, “on Web Intelligence Improving Movie Gross Prediction Through News Analysis,” 2009.

[6] R. Niraj and J. Singh, “Impact of user-generated and professional critics reviews on Bollywood movie success,” *Australas. Mark. J.*, pp. 1–9, 2015.

[7] G. Verma and H. Verma, “Predicting Bollywood Movies Success Using Machine Learning Technique,” *2019 Amity Int. Conf. Artif. Intell.*, pp. 102–105, 2016.

[8] J. Ahmad, P. Duraisamy, A. Yousef, and B. Buckles, “Movie Success Prediction Using Data Mining,” pp. 2015–2018, 2017.

[9] M. H. Latif and H. Afzal, “Prediction of Movies popularity Using Machine Learning Techniques,” vol. 16, no. 8, pp. 127–131, 2016.

[10] L. Zhou, “Movie Review Mining : a Comparison between Supervised and Unsupervised,” vol. 00, no. C, pp. 1–9, 2005.

[11] A. Kanitkar, “Bollywood Movie Success Prediction using Machine Learning Algorithms,” *2018 3rd Int. Conf. Circuits, Control. Commun. Comput.*, pp. 1–4, 2018.

[12] F. B. Moghaddam and C. Trattner, “Predicting Movie Popularity and Ratings with Visual Features,” *2019 14th Int. Work. Semant. Soc. Media Adapt. Pers.*, pp. 1–6, 2019.

[13] “A shortened version of this paper appeared in,” vol. 13, no. 3, pp. 15–24, 2000.

[14] S. Gopinath, “Performance,” no. May 2015, 2013.

[15] W. R. Bristi, “Predicting IMDb Rating of Movies by Machine Learning Techniques,” *2019 10th Int. Conf. Comput. Commun. Netw. Technol.*, pp. 1–5, 2019.

[16] R. Dhir and A. Raj, “Movie Success Prediction using Machine Learning Algorithms and their Comparison,” *2018 First Int. Conf. Secur. Cyber Comput. Commun.*, pp. 385–390, 2018.

[17] N. Quader, “A Machine Learning Approach to Predict Movie Box-Office Success,” pp. 22–24, 2017.

[18] A. Bhave, “Role of Different Factors in Predicting Movie Success,” vol. 00, no. c, 2015.

[19] L. Chen, C. Chi, and L. Huang, “Exploring contextual factors from consumer reviews affecting movie sales : an opinion mining approach,” *Electron. Commer. Res.*, no. 0123456789, 2019.

[20] P. Nagamma, H. R. Pruthvi, K. K. Nisha, and N. H. Shwetha, “An Improved Sentiment Analysis Of Online Movie Reviews Based On Clustering For Box-Office Prediction,” pp. 933–937, 2015.

[21] A. Samad, H. Basari, B. Hussin, I. G. Pramudya, and J. Zeniarja, “Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization,” *Procedia Eng.*, vol. 53, pp. 453–462, 2013.

[22] M. Lash, S. Fu, S. Wang, and K. Zhao, “Early Prediction of Movie Success — What , Who , and When,” vol. 1, pp. 345–349.

[23] Athira M D and Laksmi K S ,“Movie success prediction using ensemble classifier,” pp. 20–24, 2020.

[24] H. Verma and G. Verma, “Prediction Model for Bollywood Movie Success : A Comparative Analysis of Performance of Supervised,” *Rev. Socionetwork Strateg.*, no. 0123456789, 2019.

[25] R. Aditya, I. Journal, R. A. Abarja, and A. Wibowo, “Movie Rating Prediction using Convolutional Neural,” vol. 8, no. 5, 2020.

[26] N. Quader, O. Gani, and D. Chaki, “Performance Evaluation of Seven Machine Learning Classification Techniques for Movie Box Office Success Prediction,” no. December, pp. 7–9, 2017.

[27] L. Jiang and Z. Wang, “Predicting Box Office and Audience Rating of Chinese Films using Machine Learning.”

[28] S. Basu, *on Opinion Mining and Artificial Neural*. Springer Singapore.

[29] R. Ahuja, “Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor,” *2019 9th Int. Conf. Cloud Comput. Data Sci. Eng.*, pp. 263–268, 2019.

[30] B. D. Ruths, “Social media for large studies of behavior,” pp. 1–2.

[31] M. Ahmed, M. Jahangir, and H. Afzal, “Using Crowd-source based features from social media and Conventional features to predict the movies popularity,” 2015.

[32] A. A. A. L, “A Recommender System for Wiki Pages : Usage Based Rating Approach,” pp. 553–558, 2012.

[33] K. Lee and J. Park, “Predicting movie success with machine learning techniques : ways to improve accuracy,” *Inf. Syst. Front.*, 2016.

[34] S. Gaenssle, O. Budzinski, and D. Astakhova, “Conquering the Box Office: Factors influencing Success of International Movies in Russia,” 2018.

[35] Anuj Sharma and Shubhamoy Dey, "A comparative study of feature selection and machine learning techniques for sentiment analysis",*Proceedings of the 2012 ACM Research in Applied Computation Symposium Association for Computing Machinery*, New York, NY, USA, 1–7, 2012.

[36] L. Talavera, “An evaluation of filter and wrapper methods for feature selection in categorical clustering.”

[37] L. E. O. Breiman, “Random Forests,” pp. 5–32, 2001.

[38] "Internet movie database," IMDb.com, [Online]. Available: http://imdb.com. [Last Accessed 12-05- 2020 at 02.00 pm].

[39] “Bangla Movie Database”, [Online] Available: https://bmdb.co/ [Last Accessed on 29-01-2020 at 12.00 am].

[40] “Wikipedia”,[Online] Available: https://www.wikipedia.org/ [Last Accessed on 24-05-2020 at 09.00 pm].

[41] “Bangladesh Movie box office”,[Online] Available https://bdboxoffice.com/ [Last Accessed on 29-05- 2020 at 08.00 pm]

Appendix A

A.1 Splitting Multiple Genres into Individual Column

```
cleaned = df.set_index('title').genre.str.split(',', expand=True).stack()
cleaneddf = pd.get_dummies(cleaned, prefix='g').groupby(level=0).sum()
df = pd.merge(df, cleaneddf, on='title')
df.head()
```

A.2 Convert Some Column into Binary Value

```
df['joint venture'] = (df['joint venture'] == 'yes').astype(int)
df['foreign actor'] = (df['foreign actor'] == 'yes').astype(int)
df['Screenings in different festivals'] = (df['Screenings in different festivals'] == 'yes').asty
pe(int)
df.head()
```

A.3 Change Datatype

```
df['Hit'] = (df['Hit'] == 'yes').astype(int)
df['joint venture'] = (df['joint venture'] == 'yes').astype(int)
df['foreign actor'] = (df['foreign actor'] == 'yes').astype(int)
df['Screenings in different festivals'] = (df['Screenings in different festivals'] == 'yes').asty
pe(int)
df.head()
```

A.4 10-Fold Cross-Validation

```
from sklearn.model_selection import cross_val_score
accuracies = cross_val_score(estimator=classifier, X=X_train, y=y_train, cv=10)
accuracies
accuracies.mean()
accuracies.std()
```