A COMPARATIVE STUDY OF DIFFERENT CLASSIFIERS FOR BLOCKBUSTER MOVIES

MASIH MIKHAK

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

A COMPARATIVE STUDY OF DIFFERENT CLASSIFIERS FOR BLOCKBUSTER MOVIES

MASIH MIKHAK

DISSERTATION SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF COMPUTER SCIENCE

FACULTY OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

UNIVERSITY OF MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: MASIH MIKHAK

Matric No: WGA150038

Name of Degree: MASTER OF COMPUTER SCIENCE

Title of Dissertation: A COMPARATIVE STUDY OF DIFFERENT CLASSIFIERS FOR BLOCKBUSTER MOVIES

Field of Study: MACHINE LEARNING

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date:

Subscribed and solemnly declared before,

Witness's Signature Date:

Name:

Designation:

ABSTRACT

The purpose of this study is to investigate the blockbuster movie prediction using machine learning techniques. Predicting blockbuster movie success has been proven to be challenging and difficult and there is no effective research to prove which machine learning classifier and feature combination would provide most accurate prediction. Statistical analysis as well as operational researches are used to achieve best possible prediction, considering the strength and weakness of all these researches highlight the objectives of this research.

This research aims to present and compare which different machine learning classifiers produce most accurate prediction result and determine which feature combination can predict with high accuracy. In this study, four machine leaning classifiers and combination of features were applied to over 400 movies to find out which classifier gives the highest accuracy.

The result of this research represent that the K-NN classifier provide higher accuracy which by comparing the result by actual movies success in Box-office it clearly emphasis the contribution of this research.

KEYWORDS: Blockbuster Prediction, Machine Learning techniques, KNN, Features

ABSTRAK

Tujuan kajian ini adalah untuk menyiasat ramalan filem blockbuster menggunakan teknik pembelajaran mesin. Prediksi kejayaan filem blockbuster telah terbukti mencabar dan sukar dan tidak ada penyelidikan yang berkesan untuk membuktikan pengkelasan pembelajaran dan kombinasi ciri mesin akan memberikan ramalan yang paling tepat. Menggunakan analisis statistik dan juga penyelidikan operasi digunakan untuk meramalkan kemungkinan terbaik kekuatan dan kelemahan semua penyelidikan ini menyerlahkan objektif penyelidikan ini.

Penyelidikan ini bertujuan untuk membentangkan dan membandingkan mana yang berbeza pengeluar pembelajaran mesin menghasilkan hasil ramalan yang paling tepat dan menentukan kombinasi ciri yang boleh meramalkan dengan ketepatan yang tinggi. Dalam kajian ini, empat pengelas yang bersandar mesin dan gabungan ciri-ciri telah digunakan untuk lebih 400 filem untuk mengetahui pengelas mana yang memberikan ketepatan tertinggi

Hasil daripada eksperimen ini menunjukkan bahawa pengelas K-NN memberikan ketepatan tertinggi yang dengan membandingkan hasil kejayaan filem sebenar di Box-office ia dengan jelas menekankan sumbangan penyelidikan ini

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Dr. Ram Gopal Raj for the continued support of my dissertation, for his motivation, enthusiasm, and guidance from the beginning stages of this study as well as providing me with valuable experience throughout the work.

My deepest thanks go to my beloved family, for their endless love and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

K-NN : K-Nearest

CHAPTER 1: INTRODUCTION

1.1 Background

The film industry is a multi-billion dollar business. In 2015, the U.S. and Canada observed total Box office gains topping $11.1 Billion (Peukert, Claussen, & Kretschmer, 2017). Nevertheless, the financial success of movie is actually uncertain, with "Blockbuster" and "flops" released practically each year. While experts have performed the work of predicting movie success using various alternatives, they attempted to anticipate Box office income, or cinema admissions. However, from an investor's standpoint, they could need to be as reassured as can be done that his/her investment will in the long run business lead to dividends. This research identifies a movie's success as its success and tries to forecast such success within an automated way to raised support movie shareholders' decisions.

The process of production for any movie start using the development phase, including the engineering of the script and screenplay. Following, the potential film makes its way to the phase of production, the main factor to success. For the duration of this particular phase, the film-making group is amassed, filming areas are determined, and opportunities are garnished, among additional decisions. Then, the film moves to the actual production phase, by which filming occurs. The post-production stage involves the insertion associated with after-effects and editing. The final phase is distribution (Eliashberg, Hui, & Zhang, 2007). To aid the investment decisions of the movie, the prediction of profitability needs to be provided before the real production phrase. In this particular research, we are thinking about predicting a movie’s monetary

Success during its preproduction expression. As a result, we can just leverage data that's available at the moment.

Predictions made before or after the official discharge (the final phase in movie construction) might have more data to use and obtain more accurate results, however they are too late for investors to create any meaningful decision. (Eliashberg et al., 2007). This research proposes the method to use machine learning techniques to supply early predictions of film profitability. Based on historic data, the system use important characteristics for every movie, including "who" will be active in the movie, "How Much" is the budget, and the complement between these features after that it uses various machine learning techniques to predict the success from the movie with different requirements for profitability.

Study shown that using machine learning can be used in movie success prediction with better accuracy compare to other methods such as human expertise .The machine learning aims to determine the value of an unknown sample through learning from the already known datasets example of such machine learning techniques are K-NN, Naïve Bayes and many others. (J. Singh, Singh, & Singh, 2017)

Both feature and classifier play an essential role in identifying the success associated with Blockbuster prediction accuracy. Good choice of features and classifiers has a tendency to improve the accuracy of prediction. While wrong choice of features and has a tendency to decrease the accuracy pf prediction (Blum & Langley, 1997).

1.2 Problem statement

Predicting blockbuster movie success has been proven to be difficult and challenging and because of complexity of such this prediction there are many uncertain predictions which cannot be easily interpreted. Many researches have been done to predict blockbuster movie success but there is only some effective research to prove which feature and machine learning classifier would produce high prediction accuracy and also there is only few effective researches to indicate which feature and classifier combination would produce high blockbuster prediction accuracy and since the prediction percentage is now around 80% there is a need for improvement since it is up to 100%.

1.3 Aims and Objectives of research

This research is aimed at determining feature and classifier combination that would provide the best blockbuster movie prediction with highest possible accuracy.

The main objectives are as follows:

- To investigate the suitable machine learning approaches or classifier for blockbuster movie prediction.
- To determine which feature combination can predict with high accuracy compare to existing prediction accuracies.
- To evaluate the selected classifier feature combination in terms of accuracy and prediction

1.4 Methodology

This section briefly discuss the methodology that being used in this research which provide step by step all processes being done to perform blockbuster movie prediction.

1.4.1 Literature Review

The Most critical phase of every research is to have better understanding about gaps and limitations of research, to aim this purpose several related literatures were reviewed to achieve this purpose.

1.4.2 Dataset

For this research 400 movies were analyzed which all of these movies and their related information were extracted from IMDB website.

1.4.3 Machine Learning Classifiers

In this research 4 different classifiers are used to be applied on data

- Naïve Bayes
- K-NN
- Random Forest (RF)
- Decision Tree

1.4.4 Cross Validation

Using cross validation the initial data set is partitioned in to random subsets of equal dimension (Schneider, 1997). On can be used since the validation data and the remaining subsets are used as training information. This procedure is actually then repeated as. Numerous times as desired, the outcomes are then taken an average of and presented since the validated result. See figure (1.1) for any visual representation. This validation technique on all of the results mentioned below with 5 folds.

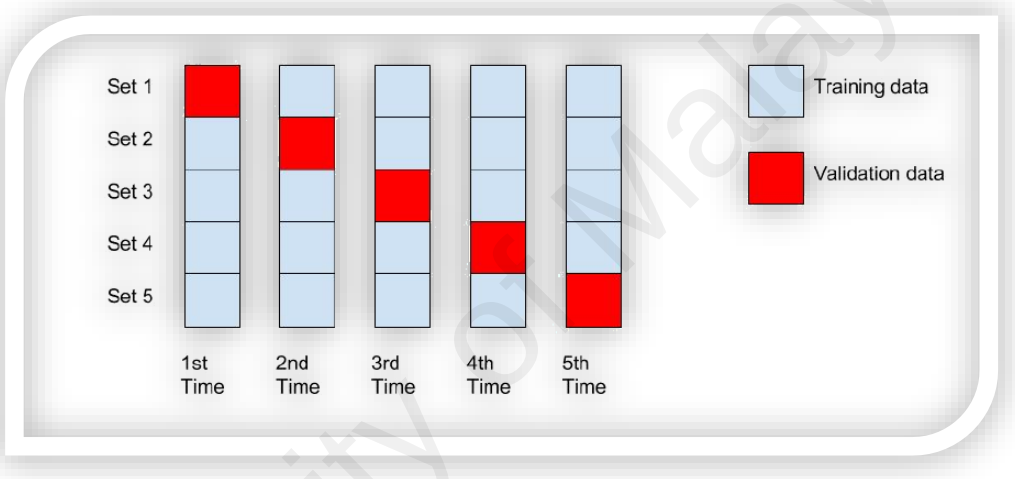


Figure 1.1: Visual representation of 5 fold cross validation

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter is going to discuss the concepts related to results of movie success predictions outcomes with different machine learning classifiers extracted from different literatures. The chapter emphasis different machine learning techniques used in movie prediction and their relationship this chapter also consider different features applied in different researches and consider their strengths and weaknesses.

2.2 Background

Success of any movie is based upon how that movie has been justified, there have been many studies done by people to prioritized box office revenue.(Apala et al., 2013), some research tried to predict the gross of movies depending on regression models using movies data (Sawhney & Eliashberg, 1996), some other researches categorize either failure or success based upon the revenue and then in order to make a prediction apply binary classification but relying on the revenue cannot be a strong measurement for getting desired prediction. Further some people came up with a model using pre-released data as their features(Lee, Park, Kim, & Choi, 2016). In many cases very few features are considered and it make the proposed model work poorly. Although some researches uses some NLP applications to apply in sentiment analysis and then by using movies reviews test their own domain.(Lash & Zhao, 2016)

2.3 Blockbuster movie Prediction Techniques

Related works related to box-office can be divided in two categories 1.the studies that focus on features that affect success of movies 2. Studies that predict viewer preference for movies.

2.3.1 Prediction based on Machine Learning Classifiers

(Hsu, Shen, & Xie, 2014) used three classification techniques for user ratings prediction of movies: neural networks, multiple linear regression and linear combination. The researchers compared these three techniques based on the average errors of each technique. The result for neural network 69% mean of error, the linear combination 65% and multiple linear combination by 57%. (Latif & Afzal, 2016) used multiple attributes (Awards, opening weekends, meta-scores, budgets) in order to classify those movies into different groups (Terrible, Poor, Average, and Excellent based upon their IMDB ratings. In this research authors use different classifiers: Multilayer Perception, Logistic Regression, naïve Bayes, part and Simple Logistic with accuracy of 79.07%, 84.15%, 79.52% and 79.66% respectively. In another research (Nithin, Pranav, Sarath, Lijiya, & Applications, 2014) use Logistic Regression, Linear Regression and support Vector classifiers to predict gross revenue and IMDB ratings of movies. The movies data was extracted from Wikipedia, IMDB and Rotten Tomatoes, then all in formation integrated in one database. The Linear Regression classifier to deploy stochastic gradient for calculating the gross, which is the sum of all attributes used, multiplied each by some weight. The Logistic regression is used by splitting the target variable into different groups of equal sizes and based on histogram representation of movies revenue. The result represents the Linear Regression has the highest accuracy (50.7%) and Logistic regression has the lowest accuracy (39%), the SVM classifier obtained the worst result for the both accuracy (39%). In (Saraee, White, & Eccleston, 2004) the purpose of study is to analyze movie parameters and their relation with rating , to achieve this purpose

author perform relevance analysis to find factors that contribute most to highly rated movies, they find out the relation between film year and it's rating by applying clustering technique. Finally, the researcher classify the rating of upcoming movies by a universal classifier query. The authors used IMDB to extract data. After the relevance analysis many attributes are eliminated from the studied set. The ratings are classified into four different categories: terrible, poor, average and excellent. Based on results the most important factors that affect the movie rating are: director (55%), budget (28%), and actors (90%).

2.3.2 Prediction based on Features used

There are some other researches that use textural features, (Oghina, Breuss, Tsagkias, & de Rijke, 2012) aimed to predict movie rating using two features: textural features and surface extracted from comments ad tweets. The surface feature is the number of online activities around a movie (quantitative) and textural feature refers to meaning of those features (qualitative), for this research the authors use a linear regression model. They implement it multiple times using different features the best result is from combination of tweeter features with users like/dislike on Facebook. The result from this research shows that this combination gives (42%) mean absolute error and a (52%) root squared mean error.

There are researches that use collaborating filtering system (Tomar & Verma) which use previous rating attempts. That is done to be able to guarantee the correctness of movie ranking. The next thing is to evaluate anonymous users to investigate their personality similarity to known users (collaborative filtering). This analysis brings about group the users predicated on their movie type interest. Next, the author operate a similarity computation to anticipate a user's score utilizing a group of similar users. Finally, they use a fuzzy inference process to increase the results of collaborative filtering when

suggestion is unavailable. The authors try to filtering users who scored more than the mean quantity of movies rated each day or they ranked over 25 percent of total users evaluations. The authors assess the similarity index and estimation. They use the collaborative filtering only then incorporate it with the fuzzy system. Results show a recognizable improvement of the prediction square main mistake and the mean dependability with all the collaborative filtering and the fuzzy system mixed, over using the collaborative filtering together.

2.3.3 Prediction based on Statistical models

(Reddy, Kasat, & Jain, 2012) aim to be able to have prediction on box office opening by using the hype analysis. The most important aspect hype analysis is that the success is completely depend on the income it made in its opening weekend and also before its release how much hype it gets from people. To begin the use a web crawler to find number of tweets pertaining to a movie. The tweets are collected through hour basis. In order to perform hype measurement there are three factors. Factor number one is to calculate "No of relevant tweets per second." Second one is "Find the number of distinct users who have posted the tweets". Third factor is "Calculate the reach of a tweet". Here reach of a tweet means that some different person's tweets have different value. Suppose if a well-known actor or director posted a positive tweet for a movie is more valuable than a tweet posted by an average person. In order to calculate the reach of tweet for calculating the reach of a tweet they count the follower of a particular user. They calculated No of relevant tweets per second, Second factor is "Find the and Calculate the reach of a tweet as hype factor by taking the average value of these three factors for each movie. Their analysis based on hype factor, number of screens the movie is going to be released and the average price of all tickets per screen per show. The total model is very simple calculations and they just counted the number of tweets related to a movie, but they don't use any kind of language processing to know if the tweet is positive or negative. A neural network had been used

in the prediction of financial success of a box office movie before releasing the movie in theaters (Sharda & Delen, 2006). This forecasting had been converted into a classification problem categorized in 9 classes. The model was represented with very few features. 6

At (Sharda & Delen, 2006), it was attempted to increase movie gross prediction with using News Analysis where quantitative data are made by Lydia (high-speed processing system for data analysis for gathering and evaluating data In news). That included with two different models (regression and *k*-nearest neighbor models). But they would be considered movie with high budget. This Model can be unsuccessful is a common word being used as name and it cannot predict with absence of news related to that movie. M.H Latif, H. Afzal (Sharda & Delen, 2006) that used just the IMDB database as their key resource and the collected data was not clean. Again their data was unpredictable and extremely noisy as they stated. So they used Central Inclination as a typical for filling lacking values for different attributes. K. Jonas, N. Stefan, S. Daniel, F. Kai use sentiment and relational network evaluation for prediction their hypothesis was predicated on level and positivity evaluation of IMDb's sub forum Oscar Buzz. That they had considered movie critics as the influencer and their predictive point of view. They used carrier of expression which gave incorrect effect when some words were used for negative means. There was no category of award and only concerned with the award for best movie, director, actors/actress and supporting actors/actress. In some instances, success prediction of your movie were made through neural network examination (Rhee, Zulkernine, & Ieee, 2016). Some analysts made prediction predicated on social media, communal network and hype evaluation where they determined positivity and quantity of reviews related to a specific movie. Additionally few people got predicted Box Office movies' success predicated on Twitter tweets and YouTube reviews. In both circumstance, the correctness of prediction will be doubtful and can neglect to give appropriate final result. A small domain is not a decent idea for measurement. In previous

works, most studies were predicated on attributes which were either available before the release or before the release of any movie. Even though some of the researchers considered both types of characteristics but in that case hardly any qualities were counted. The possibility of having better success in prediction goes higher with more attribute involved.

Table 2. 1: summery of relevant literatures

Steamer	Objectives	Data Collection	Limitation
(Latif & Afzal, 2016)	Use different attributes to categorize movies into different categories (excellent, average, poor)	Using IMDB as main resource for data collection.	The author use small number of attributes for this research which cannot be an accurate prediction.
(Nithin et al., 2014)	Prediction is based upon Logistic Regression with stochastic gradient, Linear Regression and support Vector classifiers	Wikipedia, IMDB and Rotten Tomatoes	The Results taken from this research shows that using suggested techniques can't achieves high accuracy
(Lee, K., et al., 2016)	To presents a model for predicting box-office performances of movies	Using survey and research on favorite movies asked from viewers and extracting related data from IMDB	Not paying much attention to the explanation of how the model's features are related to its outcome
(Apala et al., 2013)	To predict the box office performance using data extracted from different social media	official trailers of movies with prerelease dates in selected month	produce much less accurate prediction due to untrusted data
(Mestyan, Yasseri, & Kertesz, 2013)	The main objective is come up with a model to predict financial success of any movies depend on data extracted from activity of online users.	Wikipedia activity records	Limited number of movies may not be suitable based on proposed algorithm

CHAPTER 3: PROPOSED METHODOLOGY FOR BLOCKBUSTER MOVIE PREDICTION

3.1 Introduction

This chapter describe proposed methodology for predicting the Blockbuster movie using machine learning techniques. This chapter also reviews prediction challenges and discuss the reason for importance of this prediction using different feature combination. It also describe the process of gathering data and the way data are analyzed using different classifiers and evaluate their importance for classification task. The workflow of research is given in Figure 3.0.

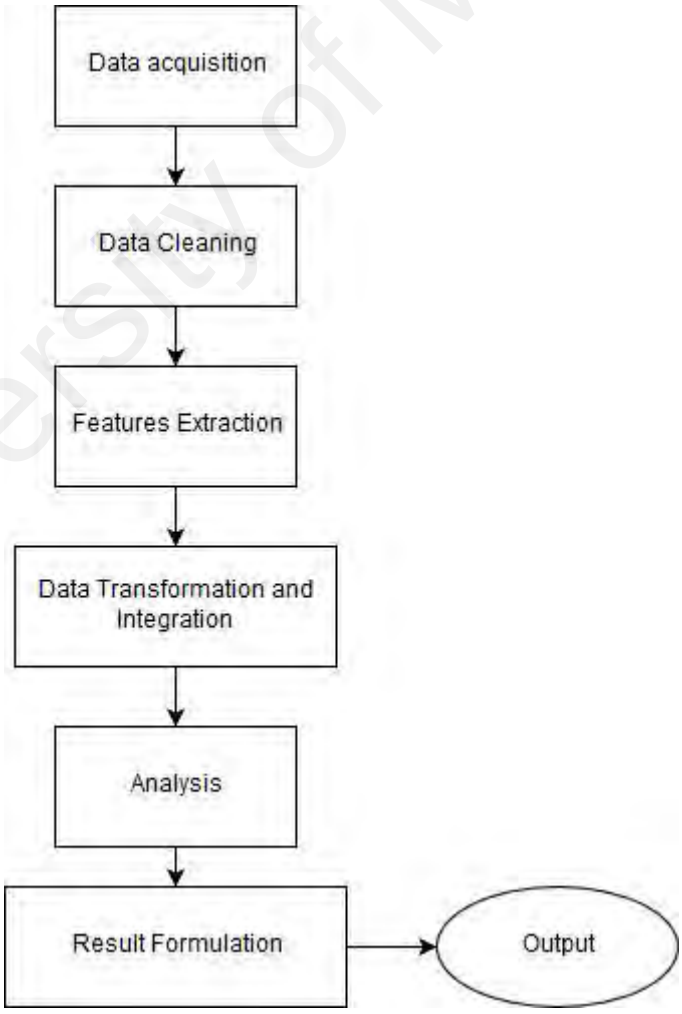


Figure 3.0 Research Workflow

3.2 Prediction Challenges

The ability for early movie success prediction is quite important for adopting and personalizing information for movie production industries, regarding this issue the prediction of popular content is not a simple problem. The imbalanced nature of data makes the problem even more challenging because there is huge difference between the number of movies that become blockbuster and those that become Flop. Choosing and finding the movie features that are suitable to predict with higher accuracy is also can be considered as one of those challenges.

3.3 Model Architecture

Our recommended approach to blockbuster prediction is based on feature-based classification model(V. K. Singh, Piryani, Uddin, Waila, & Ieee, 2013) in which we choose and extract number of features form movies and classify them as Blockbuster/Flop classes. In this section researcher describe the overall architecture of proposed system.

Figure 3.1 illustrate our proposed model architecture, the data collection information and the result extracted from selected features are described in chapter 4. The proposed model apply different features from movies and then different ML approaches used to train selected classifiers, then used to predict whether a movie is Blockbuster or Flop.

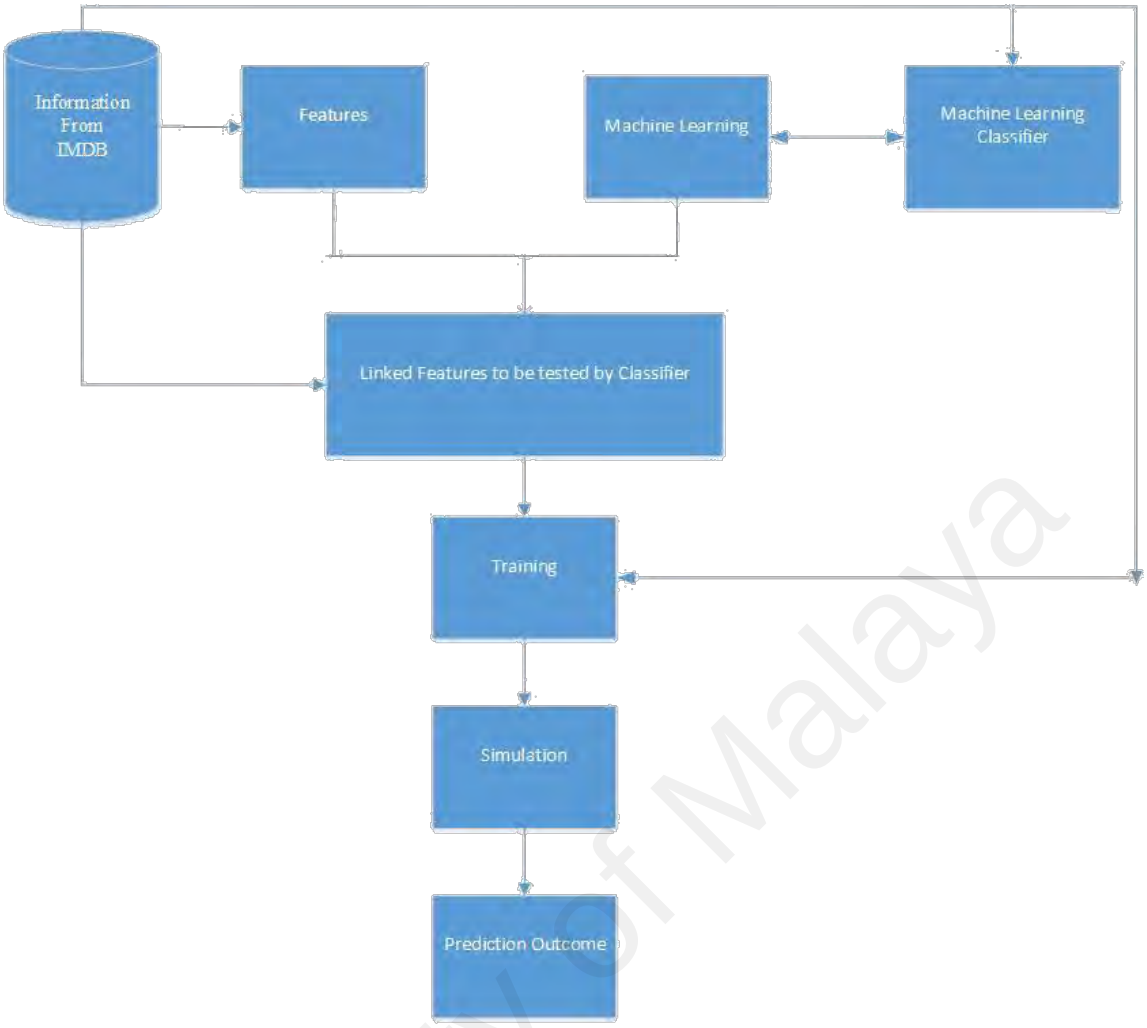


Figure 3. 1: Representation of Proposed Methodology

The two most important facts for learning based systems are selection of appropriate classifier and features which are extracted from the data. In following sections we evaluate and describe our selected classifiers and features in detail

3.4 Feature Description

Movie-based Parameters are those that have direct relation to the movie itself, including who are on the cast and what the movie is about. The most popular feature for cast members is a movie's casts actors, the popularity of actors can be measured by their star powers that have gotten from their popularity. Star powers of actors have been captured by actor earnings(Parimi & Caragea, 2013) it was agreed that higher star powers are helpful for a movie's success. However, no research has explored the profitability of actors. As it costs a great amount of money to cast a famous actor, we believe an actor's record of profitability will be a better Indicator of a movie's profitability than his/her record in generating revenues.

The role of directors in a movie's financial success is often overlooked or downplayed while some research has investigated the individual success of directors and have actually tried to connect directors' star powers to movies' financial success(Lutter, 2014) , few studies. Contrary to these select past studies, we believe that both actors and directors are crucial for films success. As directors, particularly, play important roles in movie productions this research will examine the effect of directors on movie profitability, in addition to actors. On one hand, many of the measurements for teamwork were simplistic and problematic. For example, an actor's experience was based solely on the number of previous movie appearances, without considering what types of movies she has contributed to, and thus has more experience in.

While the success of a film depends a lot on the story as well as the cast of the movie, the budget also plays an instrumental role, having big investment on movie production shows that the movie has acquired decent resources in different part of its production and it can be considered as one of best parameters to use in this research.

3.5 Classification Methods

The classification of data can be carried out using several different classifiers. Regarding to this fact the machine learning has proposed several featured-based classifiers that can be applied in variety of applications (Ghosh, Olewnik, & Lewis, 2018). With regards to the features and aspect of the info, several classifiers can be utilized for a prediction job. In statistical machine learning, a classifier can be either generative or discriminative. A generative classifier attempts to anticipate a probabilistic distribution for each course of data and assign an anonymous test to the category with highest likelihood. Alternatively, discriminative approaches make an effort to depict a curve which best discriminates the info points in several classes. With regards to the nature of the info, features and desired performance and complexity the latest models of can learn. Within this section we will describe the classifiers that people used and the reason why for with them. Within the next chapter we will experimentally show the performance of every method and introduce the perfect model for our problem.

In the following subsection the researcher is going to discuss all the classifiers and features description methods. It is important to highlight that different parameters used in these section is based on default parameters in RapidMiner tool.

3.5.1 Naïve Bayes Classifier

A Naive Bayes classifier is a straightforward generative classifier predicated on the application of the Bayes' theorem with strong assumption-ions that the features are highly independent. Quite simply, a Naive Bayes classifier assumes that the occurrence or lack of a specific feature is unrelated to the presence or lack of other feature, given the school adjustable. Despite their naïve design and seemingly oversimplified assumptions, Naive Bayes' classifiers have worked quite nicely in many complicated real-world situations

such for word classification (Frank & Bouckaert, 2006)spam diagnosis, sentiment classification and with opinion mining.

The Naive Bayes model works perfectly in the issues where the features are independent. Inside our tweet classification problem as you will notice later in this chapter, almost all of the feature are impartial and the Naive Bayes classifier is potentially an effective classifier. The Bayes' classifier calculates the likelihood of an object owned by each one of the classes. Given a class label C for a tweet (popular or non-popular) a tweet which is represented with a feature vector x (x_1,...,x_f). In the Bayes' guideline we can compute class posterior possibility P(c|X) the following:

$$P(c|X) = P(c \mid x_{1,\dots,}x_f) = \frac{P(c)P(x1,\dots,xf\,|c)}{P(x1,\dots,xf\,)}$$

3.5.2 Distance-based Classifiers

The K-Nearest Neighbor (K-NN) classifier is another popular and simple classifier which is potentially well suited for our problem. K-NN is a kind of instance-based learning, or sluggish learning, where the function is merely approximated locally and everything computation is deferred until classification. The k-NN algorithm is among the easiest of most machine learning algorithms. K-nearest neighbor (K-NN) algorithm is a discriminative classification algorithm that assigns query data to the school to which almost all of its k-nearest neighbors belong. A Euclidean distance solution is utilized to get the K-Nearest neighbors from the test pattern from a couple of known classifications (Witten and Frank, 2005). A downside of the essential "majority voting" classification occurs when the course syndication is skewed. Frequently class will dominate the prediction of the new example, because this tends to be common amongst the k nearest neighbors because of their large number (Coomans & Massart, 1982)

3.5.3 Random Forest

The Random forest technique, suggested by (Breiman, 2017), leverages multiple decision trees to anticipate a results. Its result depends upon the prediction that appears the frequently in each one of the individual decision trees and shrubs. Multiple trees and shrubs, or an ensemble of trees, may be used to mitigate the instability of an individual decision tree. An ensemble of trees is established with random examples selected from the suggestions training data. The circumstances excluded with each arbitrary sample can be viewed as "out-of-bag" and used as test examples for calculating out-of-bag prediction exactness.

Tree ensembles decrease over fitting with a couple of diverse trees and shrubs that have a tendency to converge when the place is sufficiently large. By arbitrarily restricting the characteristics used to create the trees, traits that would usually not need been chosen within a decision tree can bring about the breakthrough of cross-attribute correlations and habits that otherwise could have been skipped. It has the potential to boost global prediction and accuracy.

3.6 Features

A crucial factor when creating a prediction model is to signify samples with a good group of features. Good features should be interesting and really should have discriminative power. Which means that the features can discriminate between your movies that is recognition and the ones which do not. The features can be either discrete meaning they can have a value from a couple of defined worth, or they could be continuous meaning the features have a continuing value. A lot of the features that people extracted because of this work are indie.

A sequential Forward Selection (SFS) techniques is used for feature selection, the reason for choosing this technique is that it's low computational burden of (Doak, 1992), The SFS technique works based on greedy search algorithm that determines the features by first starting from an empty set then add a single feature that increment the value of chosen objective function in the superset in sequence to the subset, the following pseudo code discuss this process

1. Feature set initialization
    i. F_0 = (Hsu et al.); i = 0;
2. Select the next best Feature
    ii. X = arg max [□ (f_i + x)]
    iii. Where $x \neq F_i$
3. Update the Feature set
    iv. $F_{i+1} = F_i + x$
4. While i < d
    v. i = i + 1
    vi. Go to step 2

3.7 Data Sampling

For this research we investigated on different resources to obtain reliable data for our research, after going through different resources we found the IMDB website the best place to gather necessary data which can help us throughout this research, for our research we extracted over 400 movies from 2012-2016, and we choose 7 different feature of selected movies to apply machine learning classifiers on them.

3.8 Software Requirements

Software	Usage
1. Windows 10	Operating System
2. Microsoft Word 2013	Text Editor
3. Microsoft Excel data and use them in simulator	Used to collect
4. Microsoft Visio 2013 figures and charts	Used to create
5. Rapid Miner 6.0	Experimentation
6. EndNote 7.0 citation	Referencing and

Table 3. 1: list of software used in this research

CHAPTER 4: RESULT AND EVALUATION

This chapter begins with describing how to import datasets into the simulator used to analyze data. In this research numbers of experiments are conducted in order to getting better observation regarding determination of better prediction accuracy. These experiments are done based on Appling different features and using these features in different fashions by combing them or using them individually in order to understand which combination would produce better prediction accuracy. to achieve this purpose 400 movies in total were analyzed these movies are divided in two classes, Blockbuster and Flop , to have a better prediction we analyzed data through 5 steps, at first step 100 movies were analyzed after that 150 movies, 200 movies , 300 movies and at the end 400 movies were analyzed .

4.1 Importing Datasets and feature Assigning

To analyze data using RapidMiner, there are different steps which should be done in order to import dataset and features, which discussed in below:

Step 1: First open RapidMiner Software and from the main window select New Process as shown in Figure 4.1:

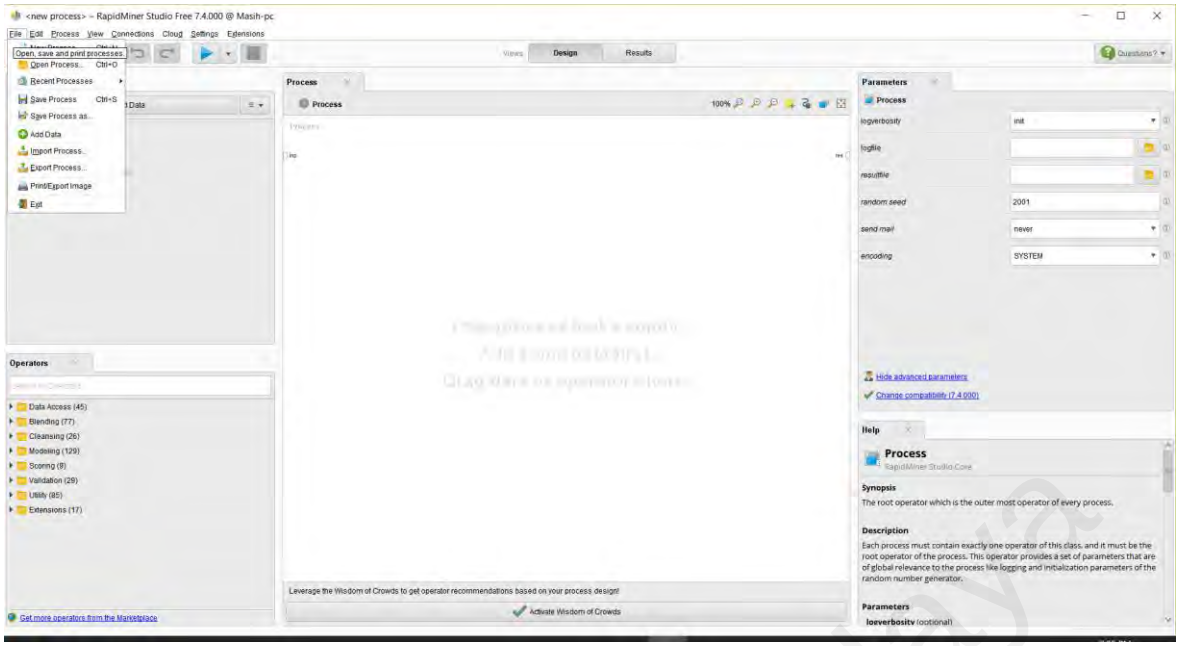


Figure 4. 1 Open Rapid Miner new project

Step 2: In this step we import datasets to RapidMiner Simulator, in this step from File menu we choose add data in order to import training data into simulator, Figure 4.2 shows this step:

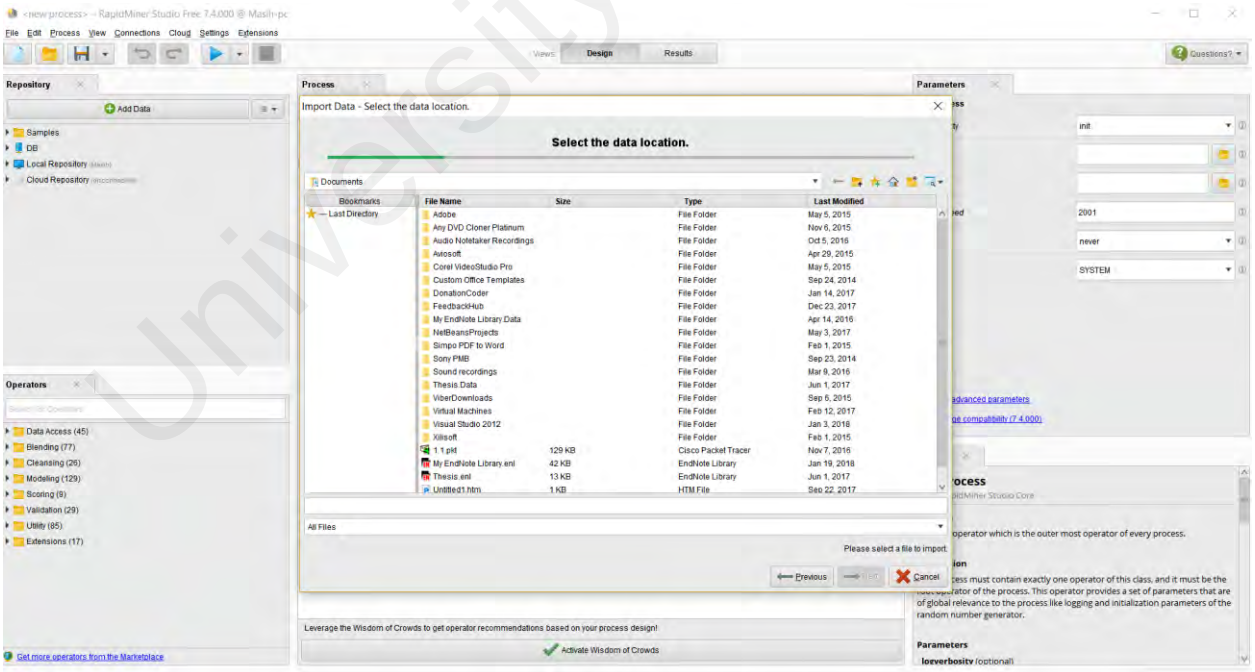


Figure 4. 2 Dataset Import

Step 3: After choosing the file from the file Explorer windows the dialog box will be shown as in Figure 4.3

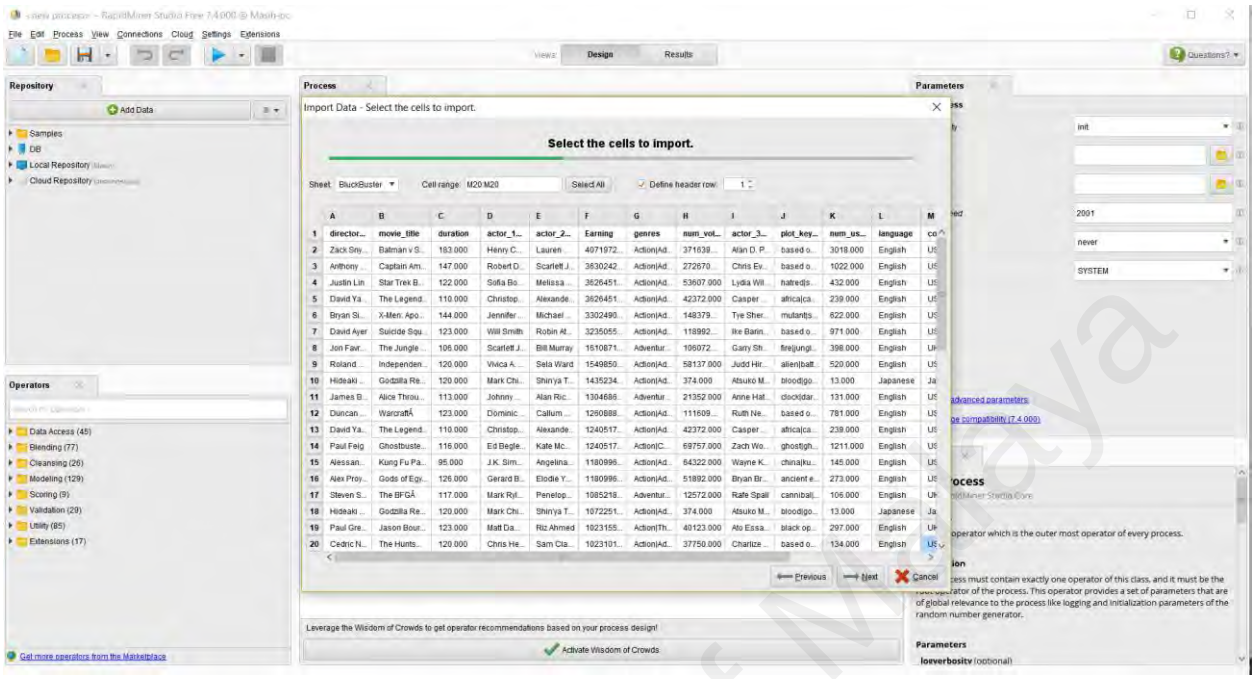


Figure 4. 3 Displaying movie info and features

Step 4: RapidMiner Simulator provide attribute annotation which make it possible to comment a specific record so it won't show during runtime. Figure 4.4 show this step clearly

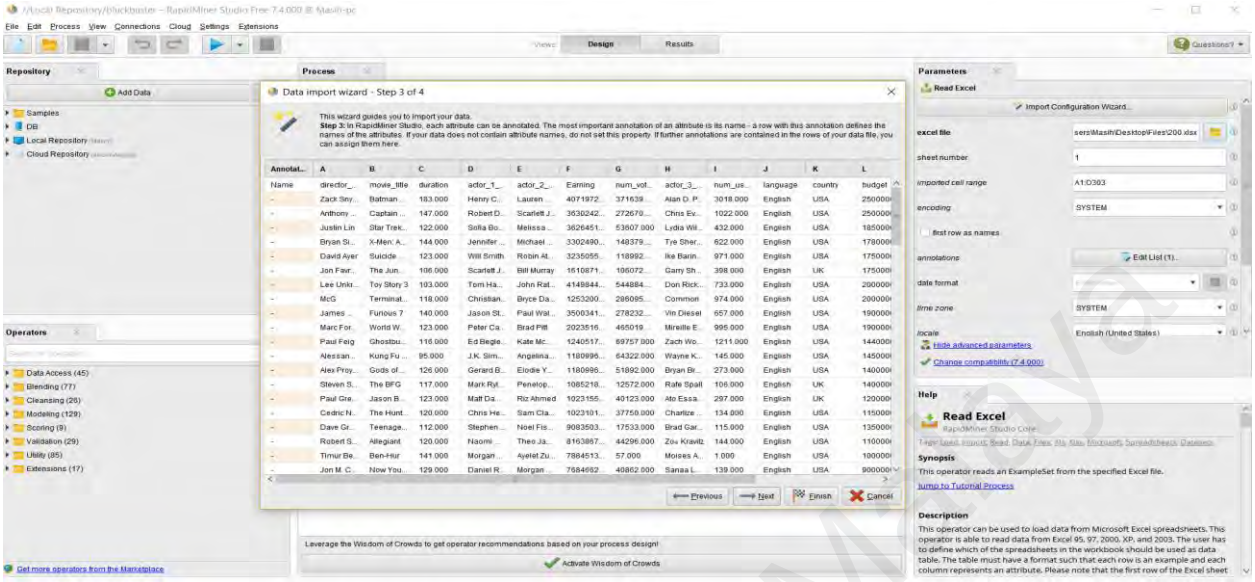


Figure 4. 4 attribute annotation

Step 5: After clicking the next button we will be directed to the next step witch is we are able to choose appropriate datatype for each feature there are many options provided in this step for example we can select or deselect specific feature, choosing the ID attribute and also class label, we choose type attribute salable and press next button Figure 4.5 represents this step.

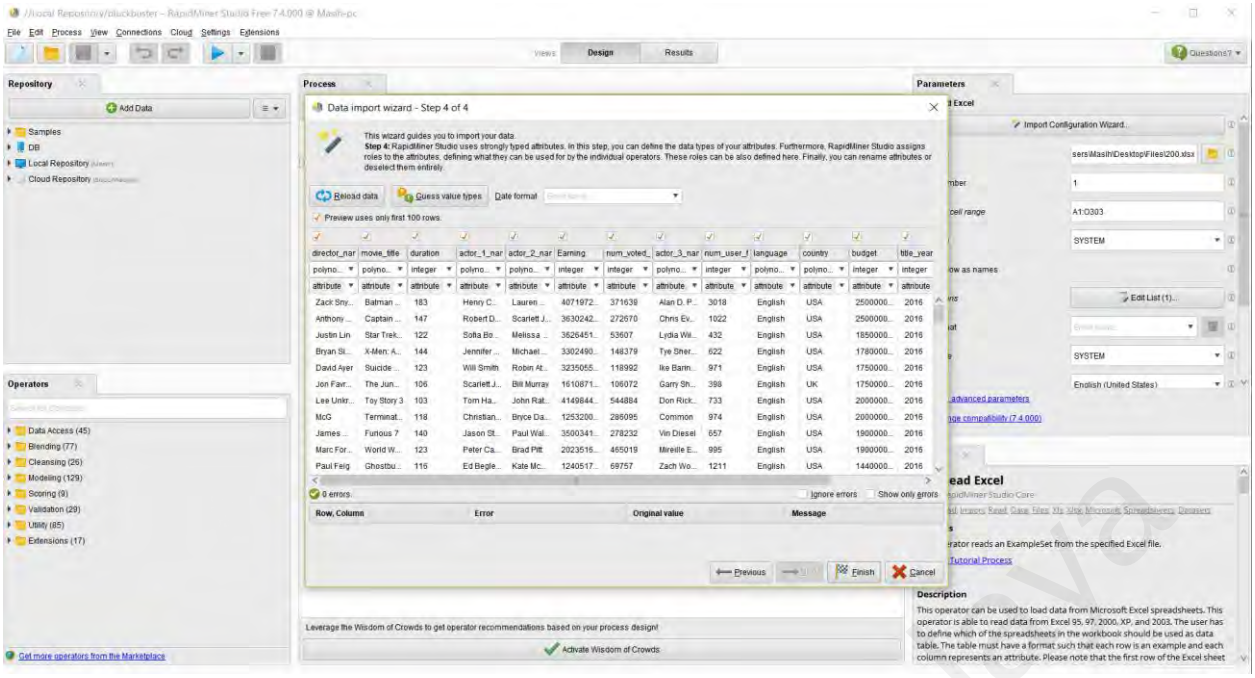


Figure 4. 5 Selecting features and attributes

To analyze datasets using testing dataset we should import data and repeat steps 2 to step 5, in this step class label is not selected in order to use datasets for evaluating and testing.

Step 1: The first is to import datasets to simulator using file menu and then selecting Import Data as in the Figure 4.7

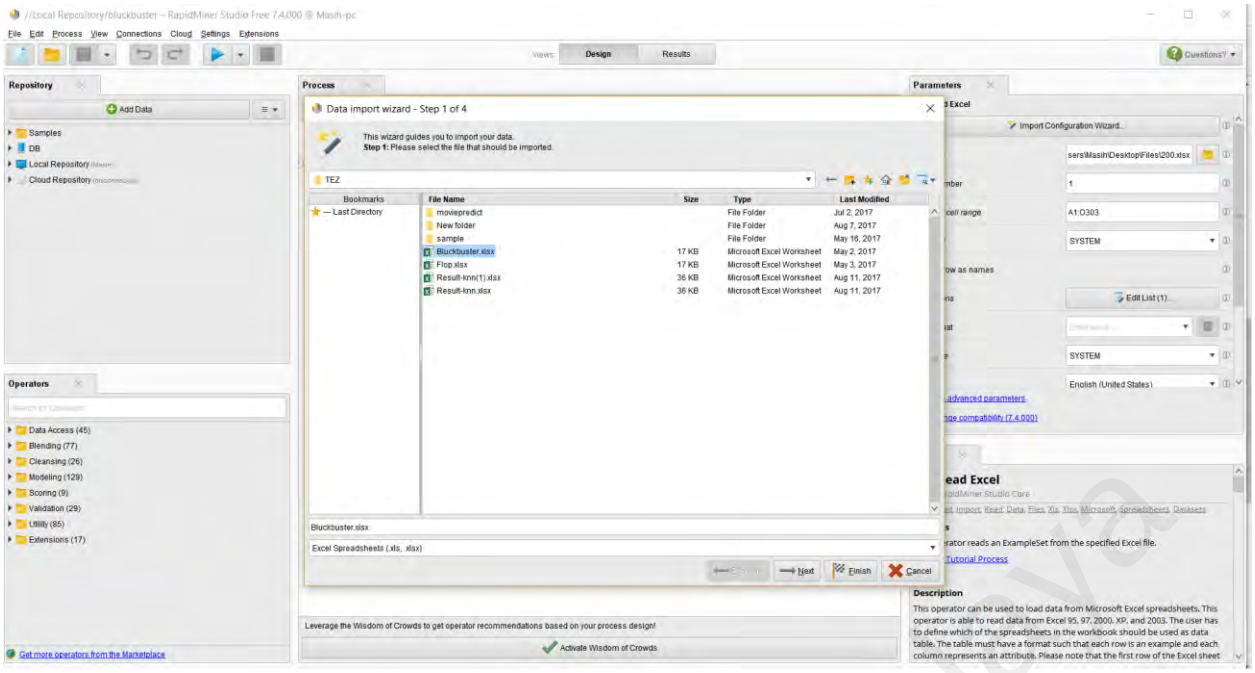


Figure 4. 6 Input datasets to simulator

Step 2: After choosing the datasets the dialog box as in the Figure 4.8 should be appear

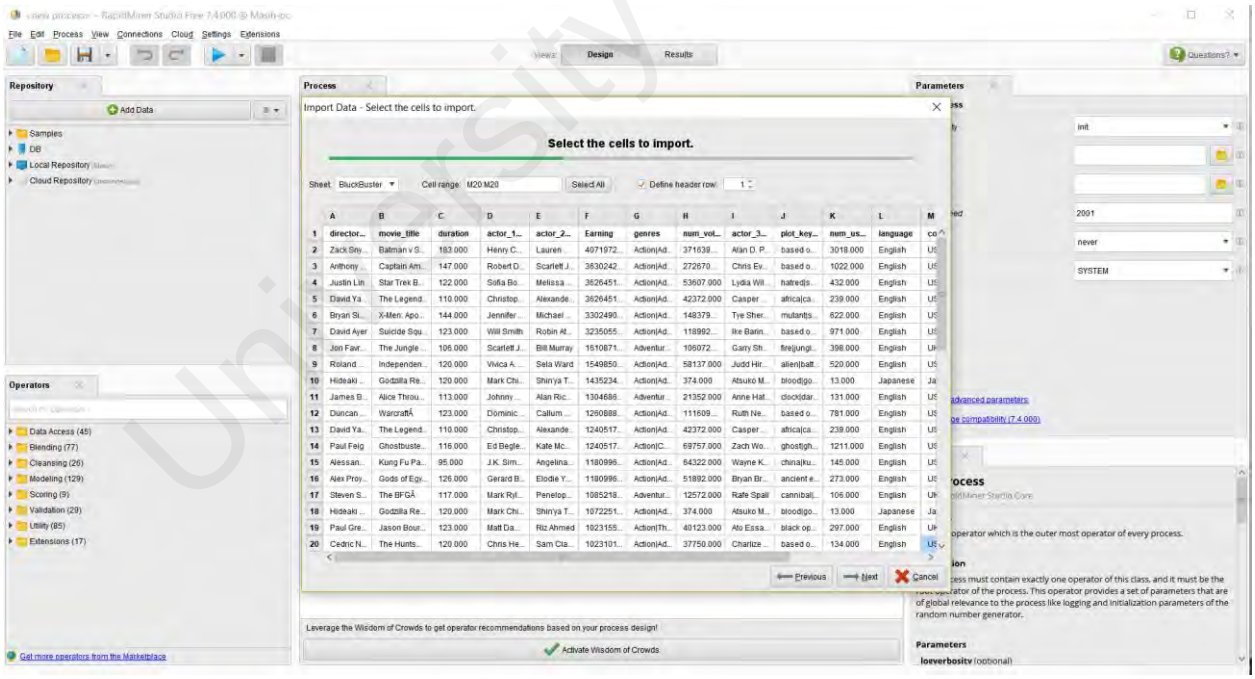


Figure 4. 7 Contents of testing Dataset

Step 4: RapidMiner Simulator provides attribute annotation which make it possible to comment a specific record so it won't show during runtime. Figure 4.9 show this step clearly.

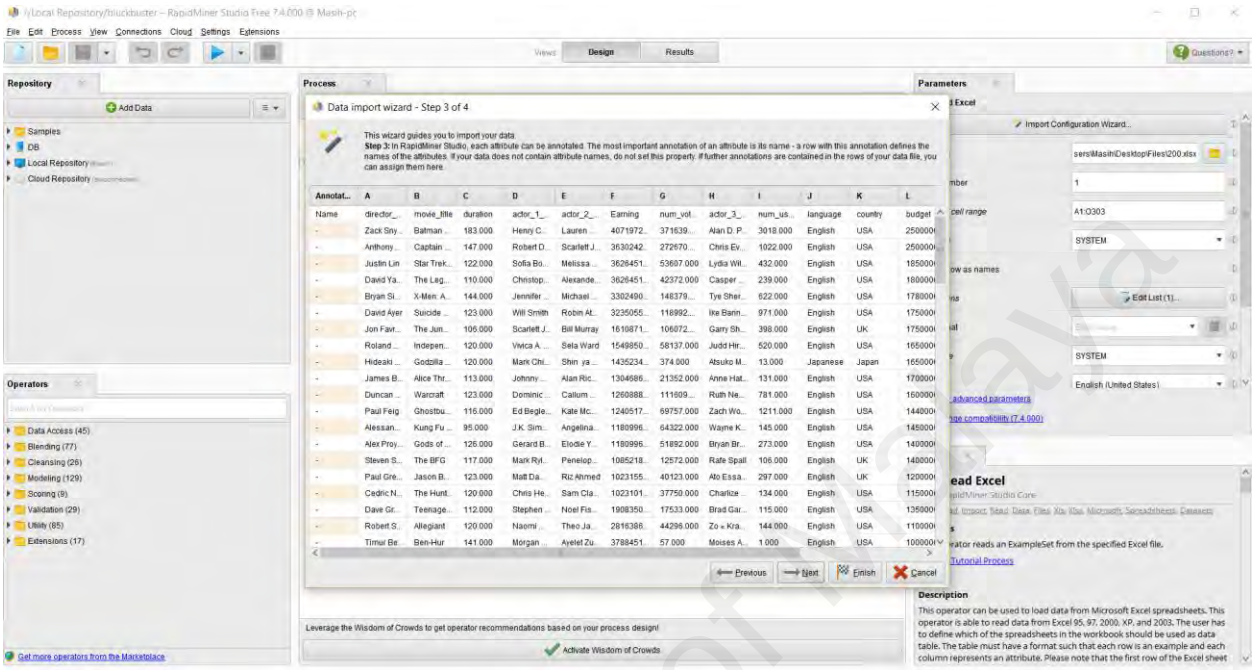
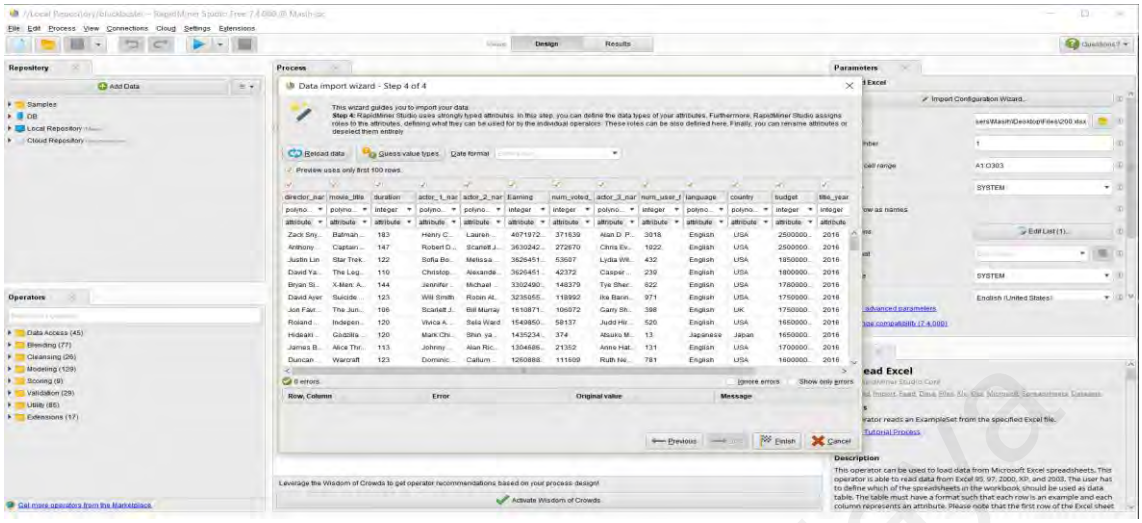


Figure 4. 8 Annotation Option to select or deselect different options

Step5: In this step attribute data type can be selected, the difference is that no label is being selected because the goal is to test the system. The Figure 4.10 presents this step.



4.2 Research Result

This section is describing and comparing results taken from different experiments that is done on movie datasets using Rapid Miner software, for this experiment and to make sure the results are accurate and reliable we applied classifiers on selected features using 100 movies and then repeated the experiment with adding more movies, the reason is to know how adding more movies will affect the accuracy percentage.

For this research total number of 400 movies are gathered and at the first stage 100 movies are analyzed then we repeated the same process by adding 100 more movies and we continue this process by adding 100 movies until we get to 400 movies, after applying 400 movies on our features and getting the results we get to the point that with adding more movies there is no change on accuracy percentage for all classifiers.

For this experiments we have used a number of movie features in which can be much effected on our desired prediction, for this experiment we applied Director, First actor, second actor, third actor, IMDB rate, movie budget to apply our Machine Learning classifiers on these features, the reason for choosing these features is that all these features are the most important factors for consideration of movie success in box-office.

These experiments aims to justify how choosing different attributes and classifiers can help researcher to have most effective prediction on blockbuster movie.

For the first experiment 100 movies are being considered and we applied 4 classifiers on these small datasets, at first we applied each classifiers on dataset based on one feature and after applying all features we did the same process using feature combination. The Below Figures represents the comparison between results extracted from applying each classifier on dataset.

4.2.1 Naïve Bayes

Figure 4.9 represents the results taken from applying naïve Bayes classifier on different number of datasets.

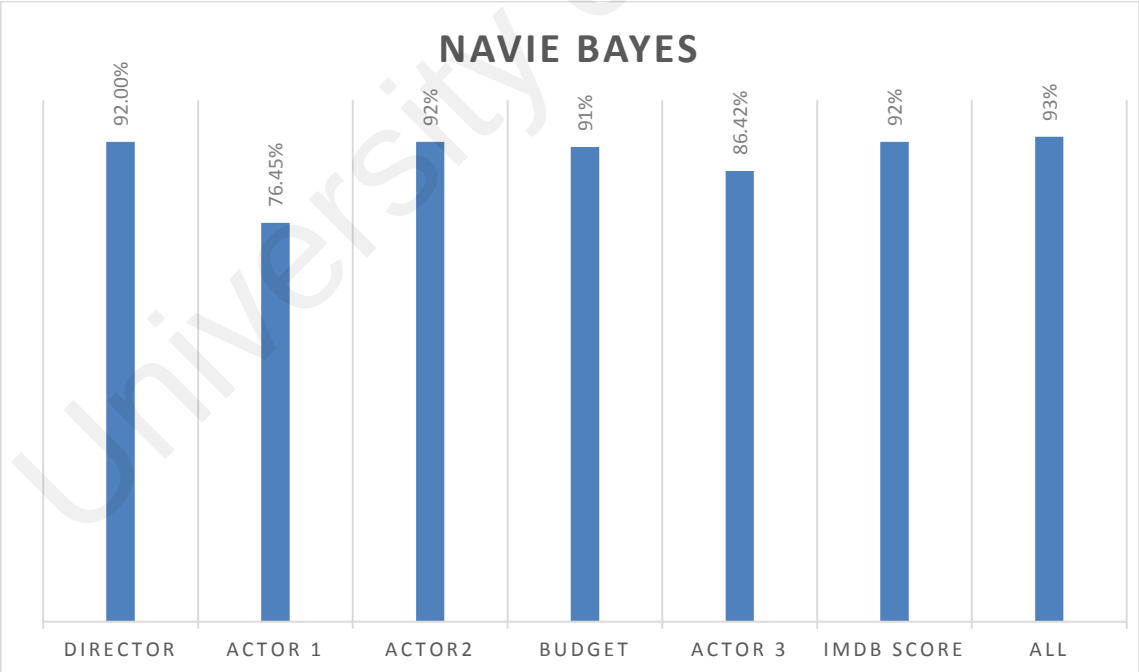


Figure 4. 9 : Naïve Bayes Result based on 100 movies

In this figure Naïve Bayes classifier giving high accuracy percentage on Director, IMDB score and on combination of features, but the prediction based on Actor 1 is giving lowest accuracy compare to others features.

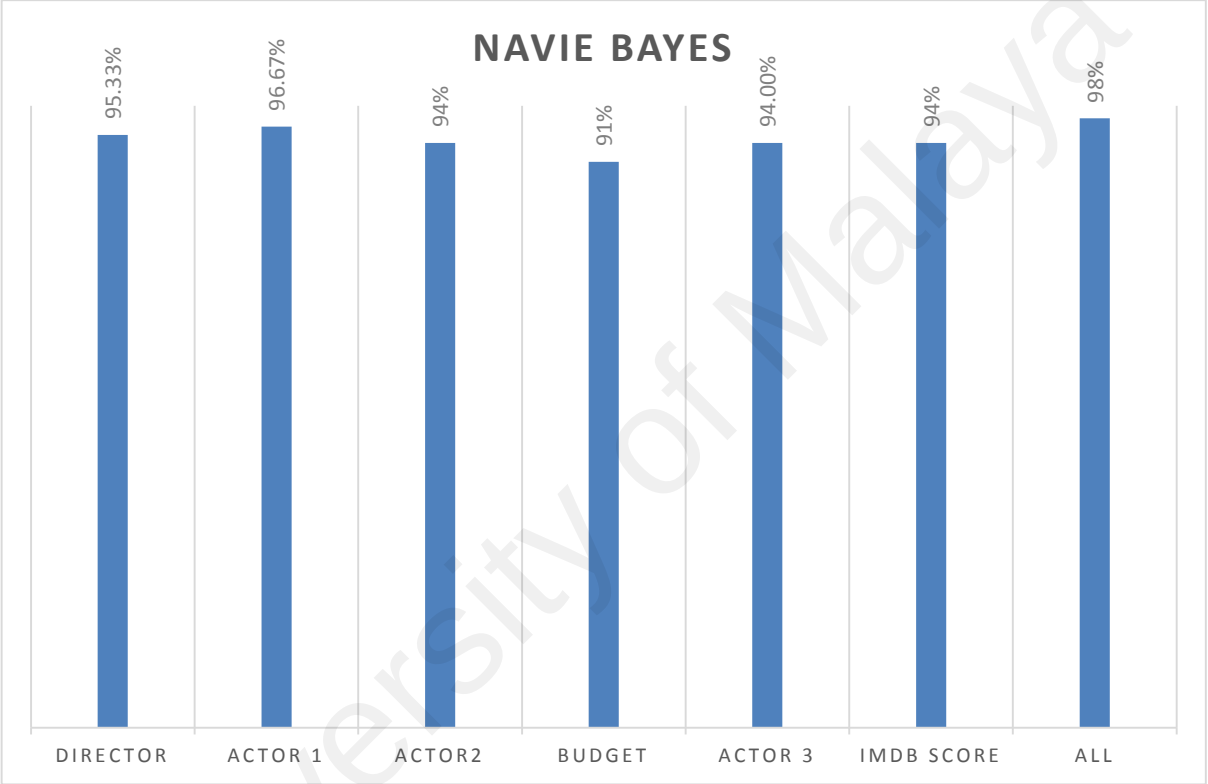


Figure 4. 10: Naïve Bay Result based on 150 movies

Figure 4.10 presents the results taken from applying 150 movies, comparing this figure with previous figure shows a noticeable change in prediction based on Actor 1 feature following with other features which has better result comparing their equivalents from previous figure.

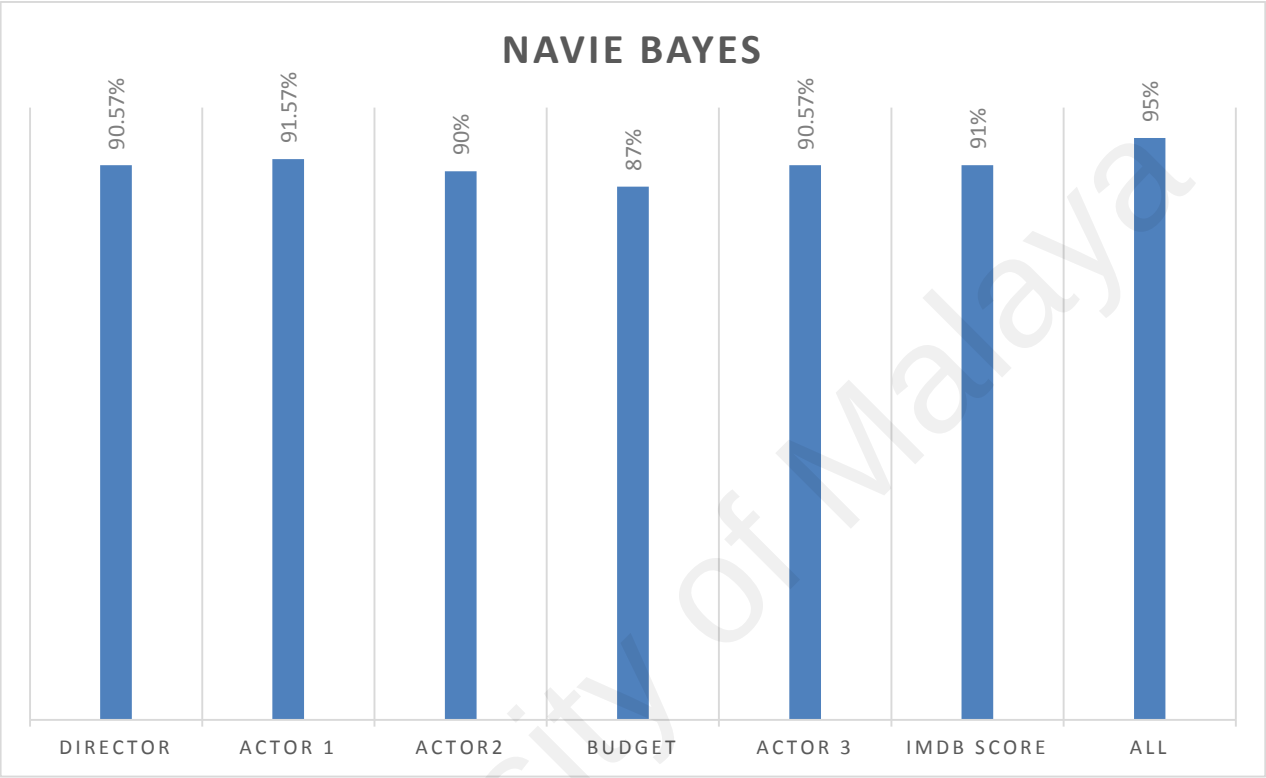


Figure 4. 11 : Naïve Bay Result based on 200 movies

With adding more movies to out experiment we get interesting results , as given figure presents, the results has started to decrease , this shows we need to continue adding more movies to see how adding more movies will affect the accuracy.

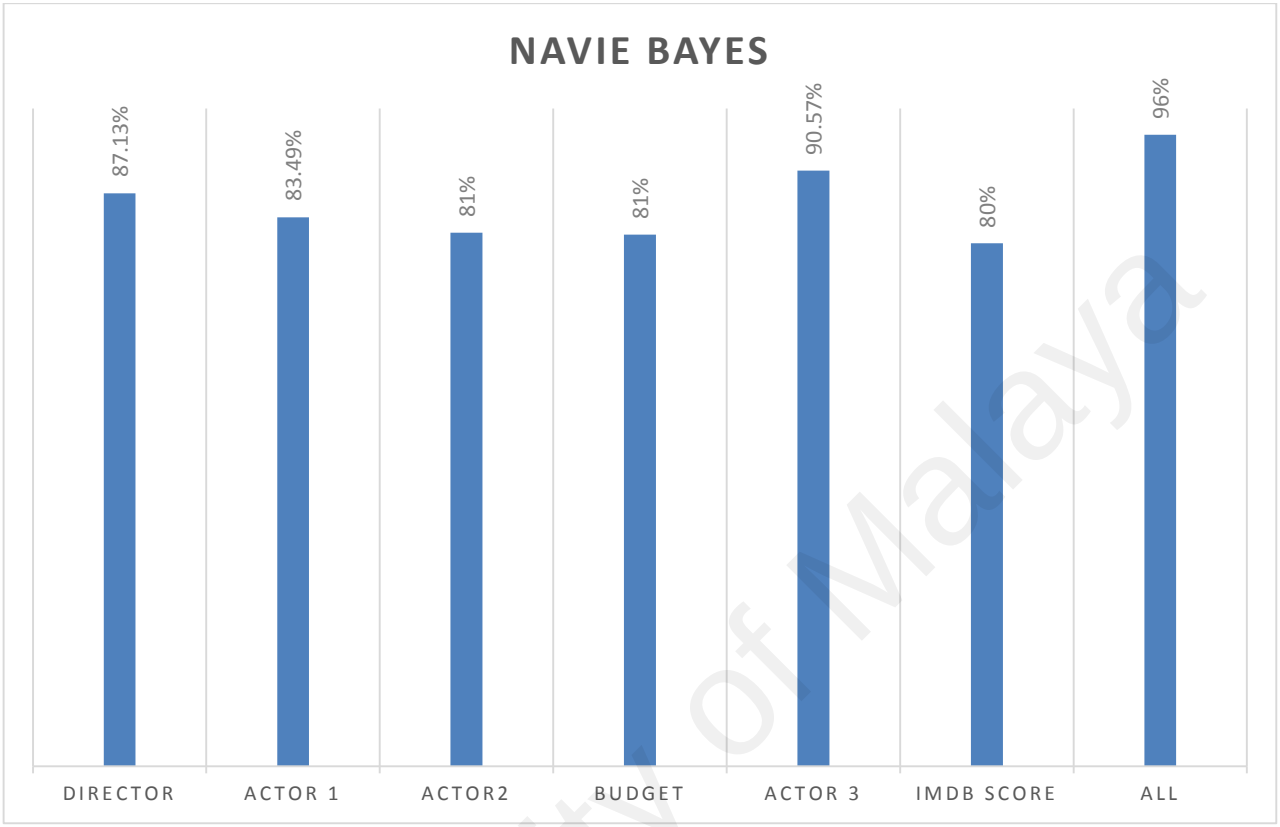


Figure 4. 12 : Naïve Bayes Result based on 300 movies

Figure 4.13 clearly shows that adding more movies to our analysis will lead to deduction in prediction results on each feature, but the prediction based on feature combination is giving higher percentage.

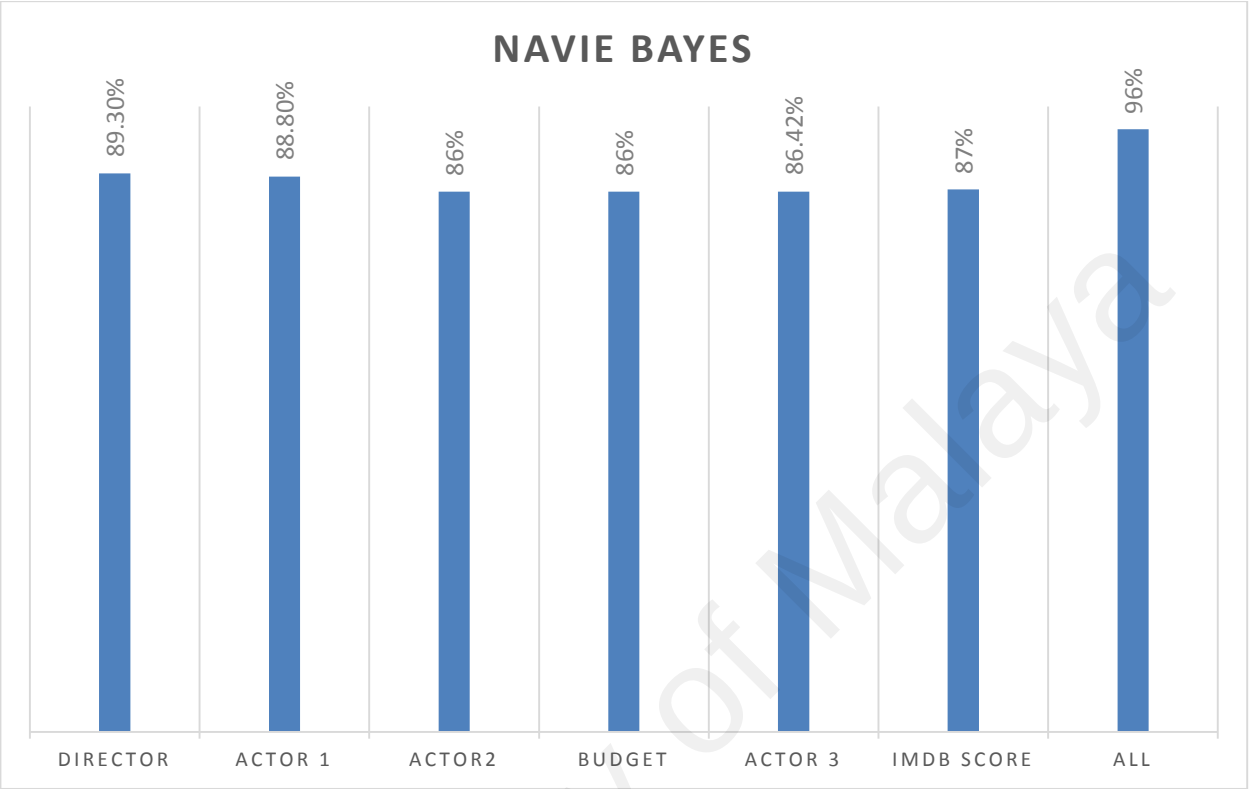


Figure 4. 13 : Naïve Bay Result based on 400 movies

Conducting this experiment using 400 movies shows that adding more movies doesn't have much effect on increasing accuracy percentage result for Naïve Bayes classifier, using all features keep giving high accuracy.

	true Blockbuster	true Flop	class precision
pred. Blockbuster	172	13	92.97%
pred. Flop	4	113	96.58%
class recall	97.73%	89.68%	

Table 4. 1 Result based on all features combinations

Considering Table 4.1 for this experiment, in the first row classifier predicted (172) blockbuster movies correctly with 92.97% accuracy. The Naïve Bayes classifier could achieve 94.37% accuracy in overall.

Random Forest

The second classifier which is used for this research is Random Forest, below figures represents the results taken from applying this classifier on our datasets.

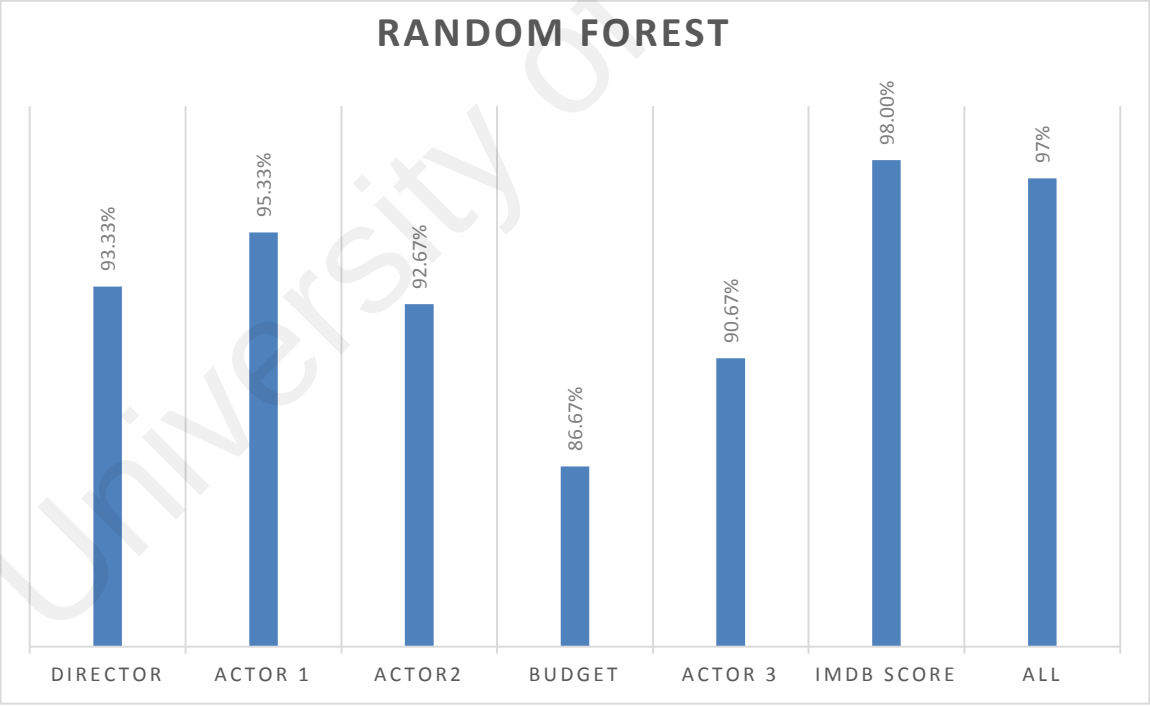


Figure 4. 14 : Random Forest Result based on 100 movies

For the first experiment we applied our classifier on 100 movies, the result showing that prediction based on IMDB Score giving higher accuracy comparing with other features, the lowest accuracy is based on budget.

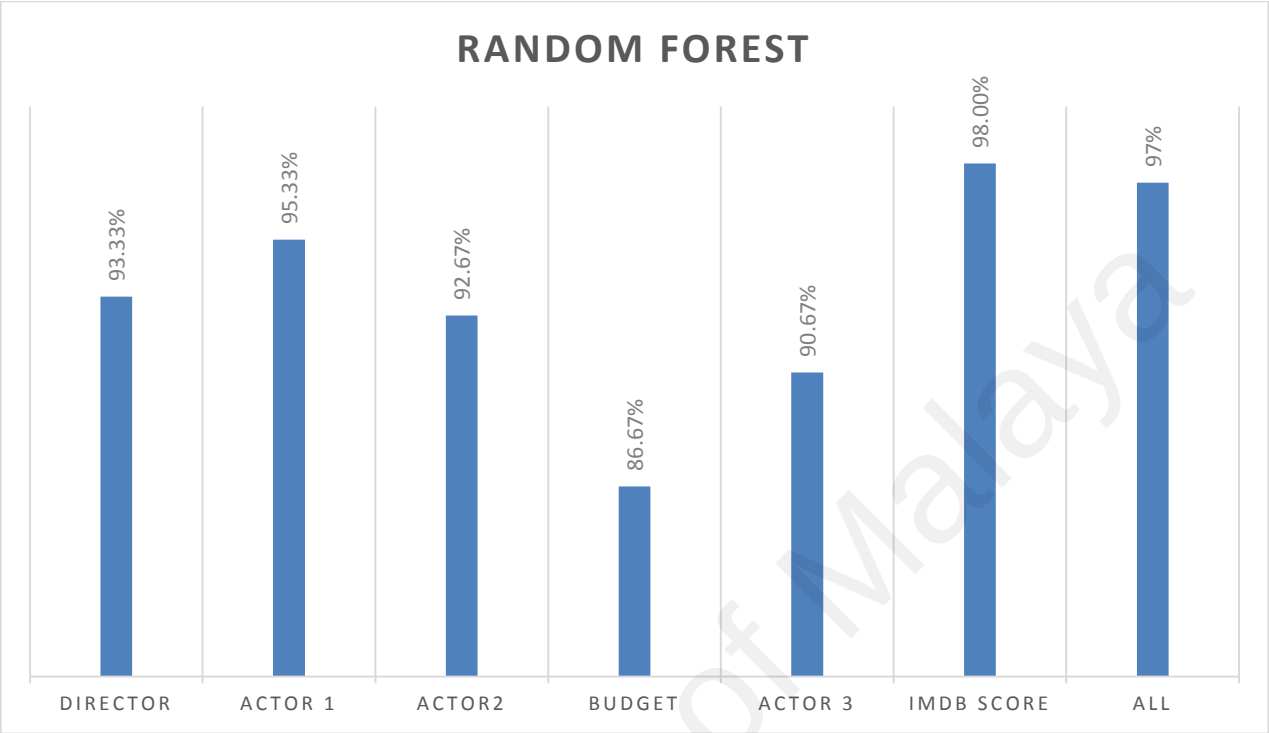


Figure 4. 15 : Random Forest Result based on 150 movies

Next experiment is done using 150 movies, interestingly the results are exactly similar to results taken from 100 movies experiment so researcher keep adding more data to dataset to see how adding more movies will affect these results.

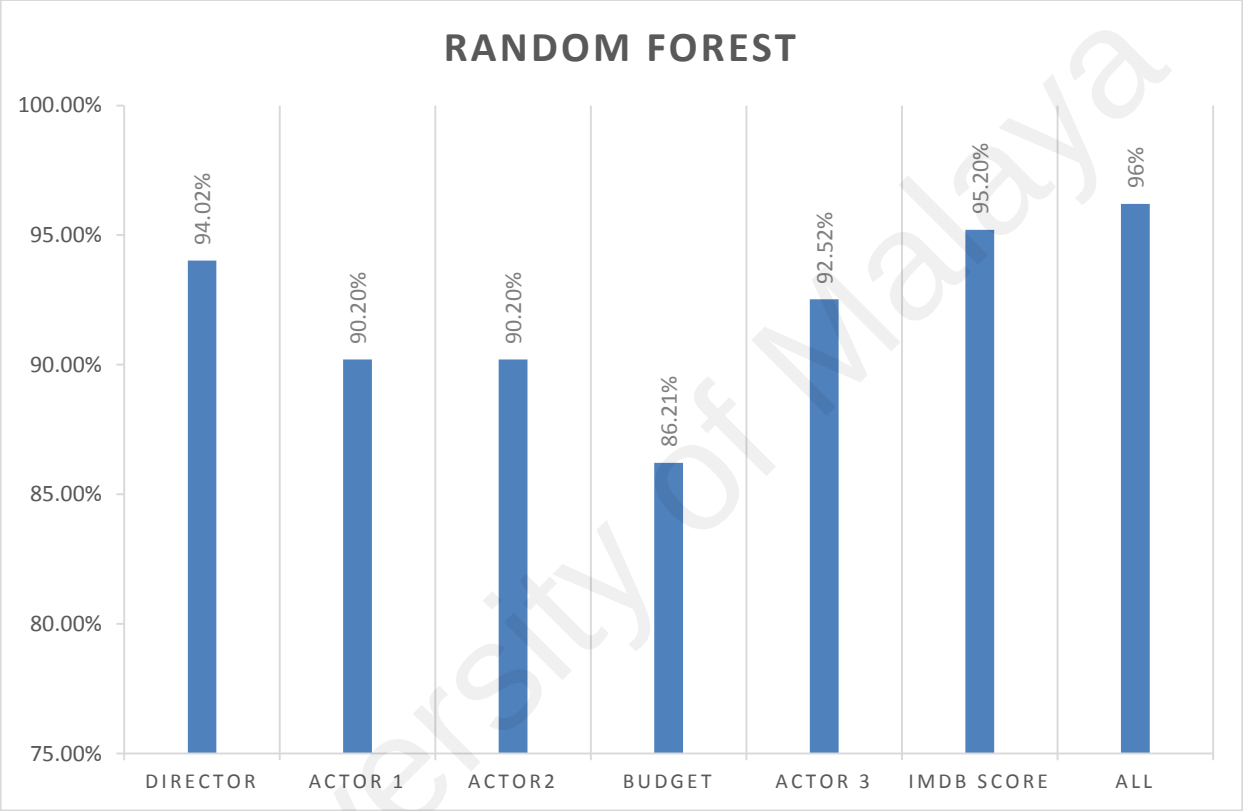


Figure 4. 16 : Random Forest Result based on 200 movies

By adding more movies to our experiment and using 200 movies we can see the results are changing, in this experiment Actor1 and IMDB having bigger deduction in the results and highest accuracy belong to features combinations.

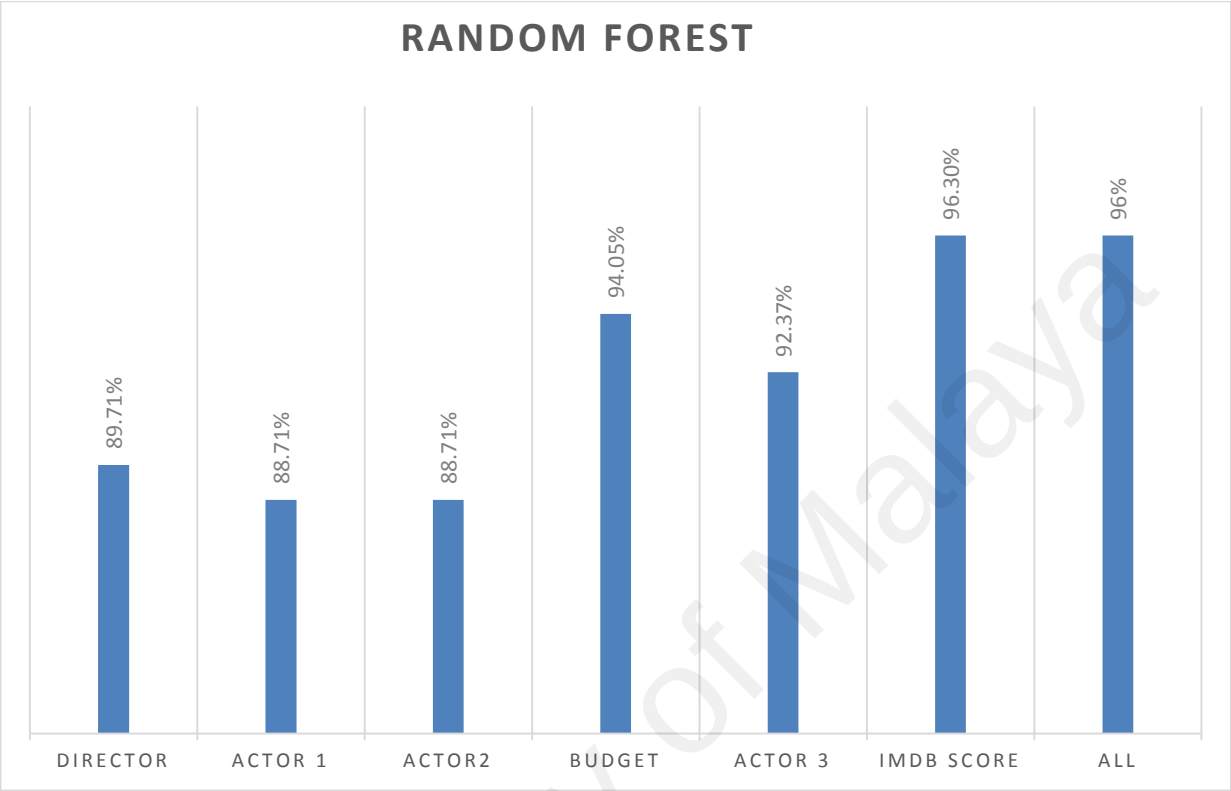


Figure 4. 17: Random Forest Result based on 300 movies

As the figure shows by using 300 movies, except the Director and Actor1 features all other features have better accuracy comparing with previous experiment.

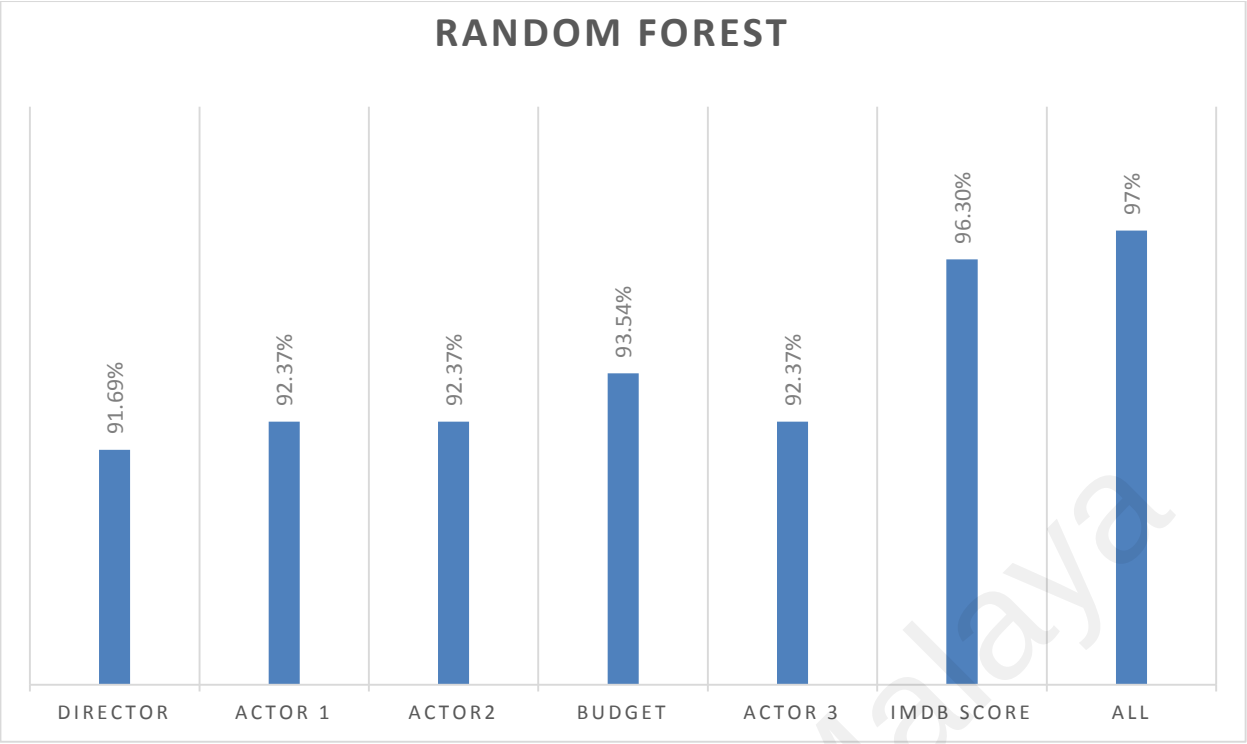


Figure 4. 18: Random Forest Result based on 400 movies

Figure 4.19 clearly states that with using 400 movies to perform this experiment mostly all the features having more accurate percentage and similar to previous results, feature combination giving higher accuracy.

	true Blockbuster	true Flop	class precision
pred. Blockbuster	174	15	92.06%
pred. Flop	2	111	98.23%
class recall	98.86%	88.10%	

Table 4. 2 : Random Forest Result based all features combinations

Considering the Table 4.2 regarding this experiment, in the first row the classifier correctly predicted 174 blockbuster movies with accuracy of 92.06%, and in the second row of this table, the classifier predicted 111 flop with accuracy of 98.23%, this classifier could achieve the overall prediction of 94.37%.

K-nearest

Figure 4.19 represent result taken from applying K-NN classifier on out dataset, in this experiment similar to previous ones we apply different number of data to see how different number of movies will affect prediction accuracy.

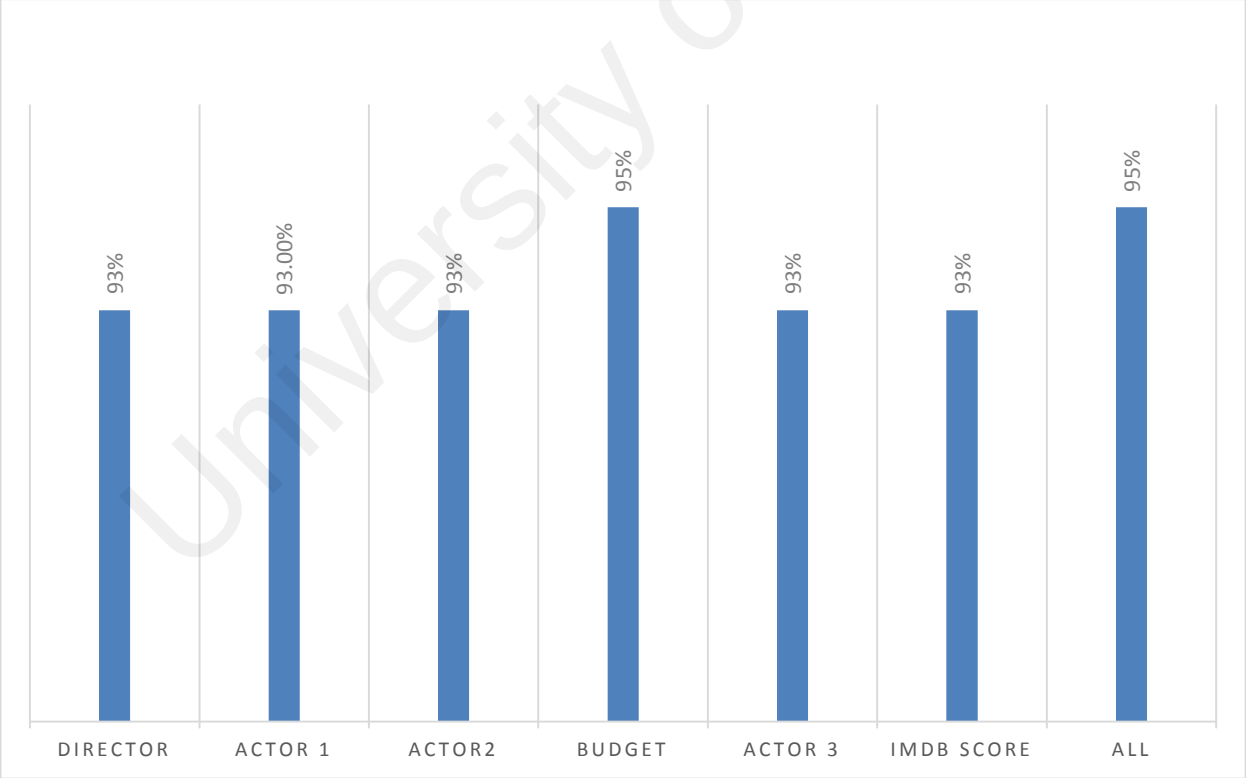


Figure 4. 19 : K-NN Result based on 100 movies

Above figure shows the results based on applying K-NN classifier on 100 movies, as it is clear all of features have high accuracy. The Budget and all Feature combinations have the highest prediction with 95% accuracy. To see how these numbers can be trusted we keep adding more movies to see the change in these predictions.

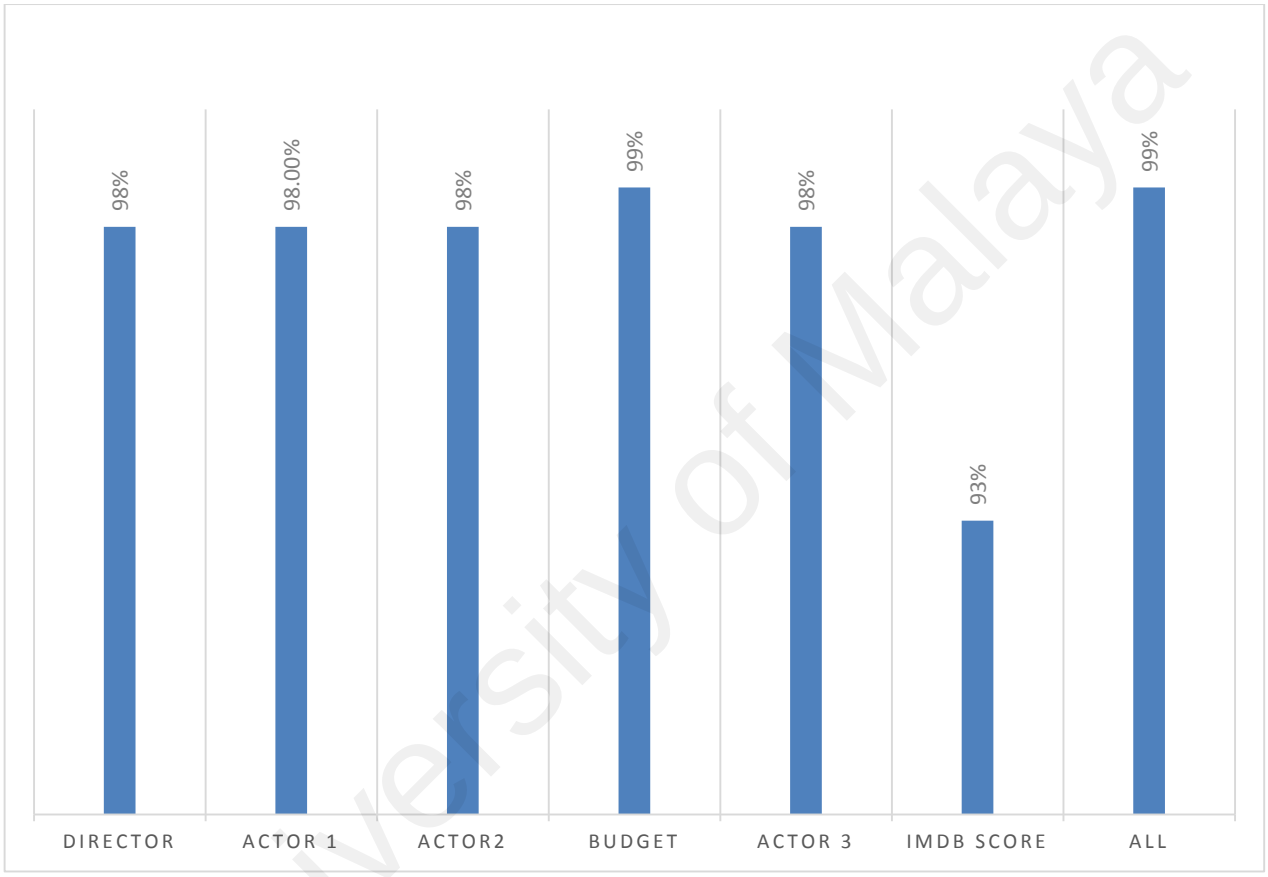


Figure 4. 20 : K-NN Result based on 150 movies

Figure 4.20 represents that after adding more data to the dataset the accuracy percentage increase for all the feature, by comparing the results we can realize that except IMDB score feature all other feature have high percentage of prediction.

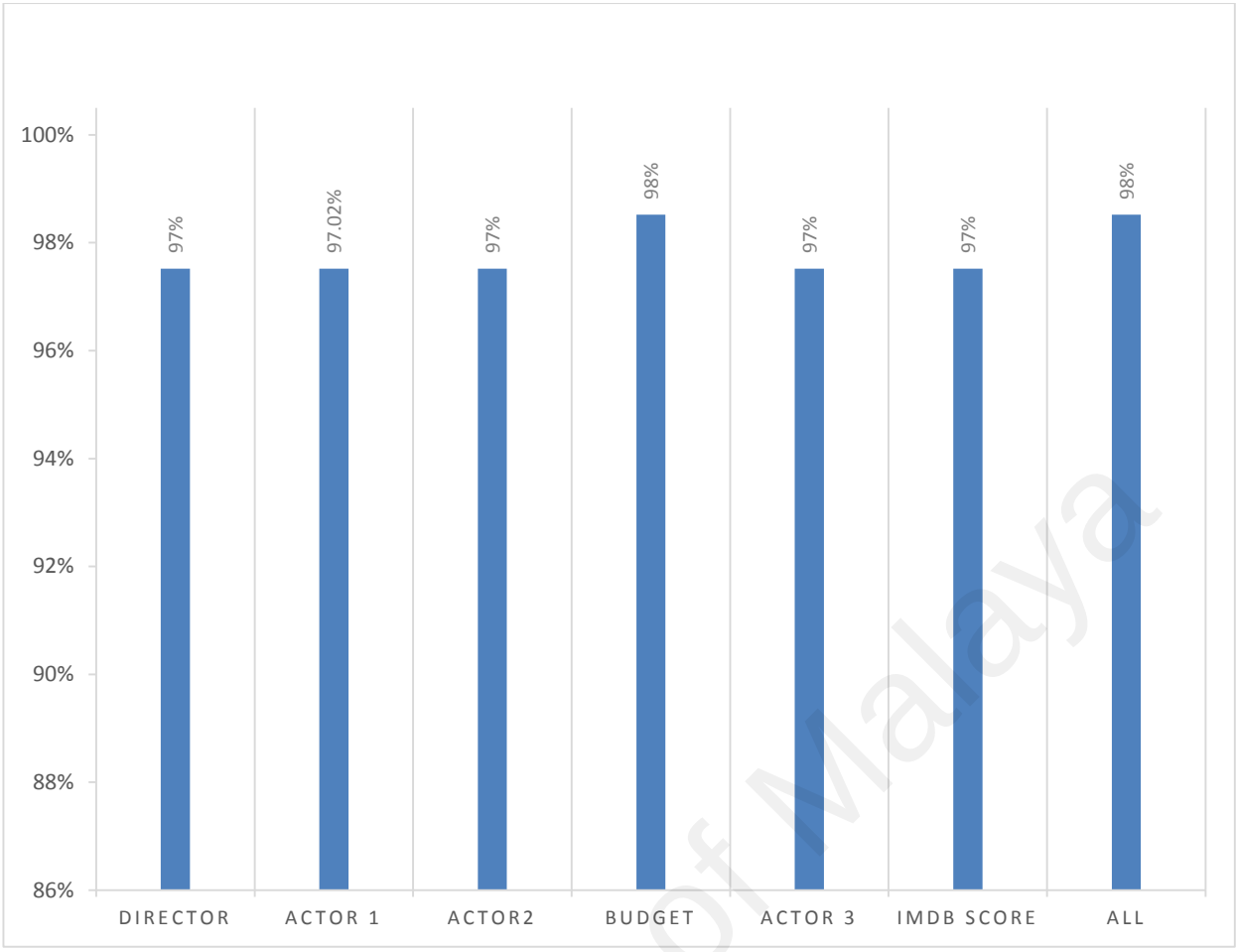


Figure 4. 21: K-NN Result based on 200 movies

Figure 4.21 shows that after adding 50 more movies to our experiment the accuracy percentage decrease, by comparing the extracted from this experiment with previous results we can see in almost all features we have less accuracy.

Figure 4. 22 : K-NN Result based on 300 movies

Figure 4.22 represent that adding more movies to our experiment is leading to deduction in all results.

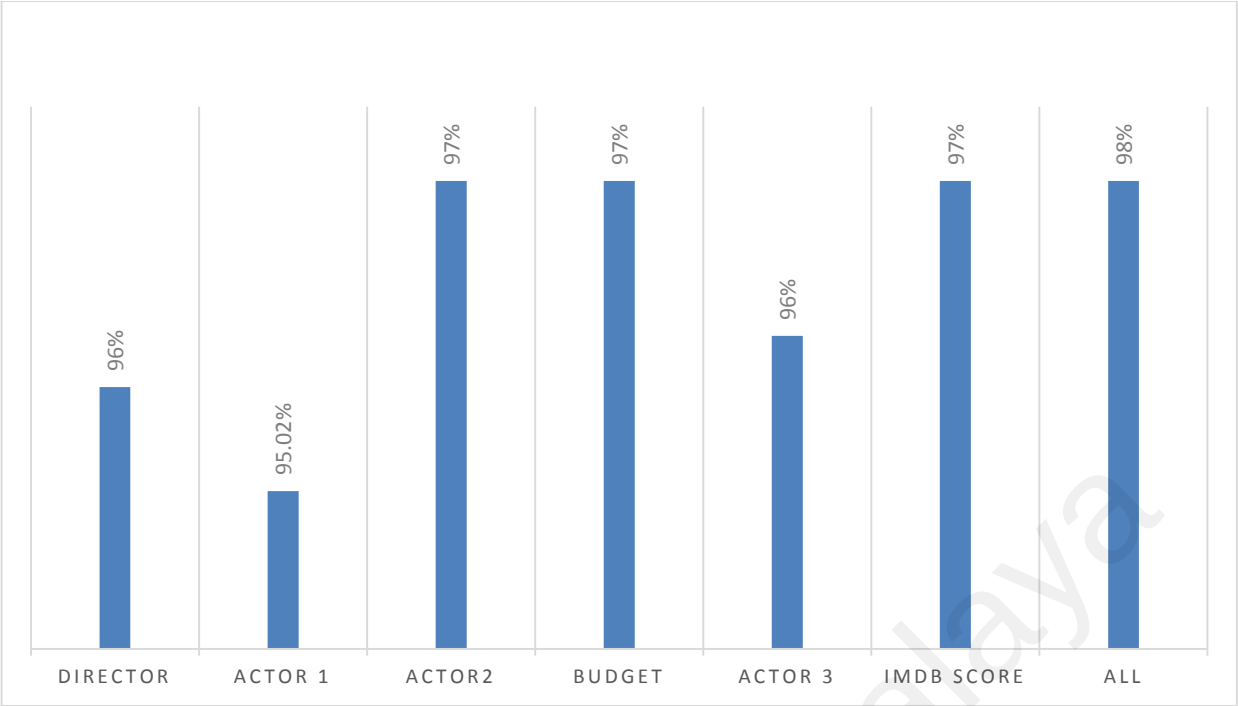


Figure 4. 23: K-NN Result based on 400 movies

The last experiment shows that after adding more data, the result do not change and most of the features have the same result similar to previous experiment, this figure represents that the prediction based on all feature gives the highest accuracy result compare to other features.at this point we can realize that K-NN classifier has the better prediction compare with last two classifier so far.

	true Blockbuster	true Flop	class precision
pred. Blockbuster	173	3	98.30%
pred. Flop	3	123	97.62%
class recall	98.30%	97.62%	

Table 4. 3 : K-NN Result based all features combinations

By Considering the Table 4.3, it can be observed that the K-NN classifier was correctly predicted 173 blockbuster movies with accuracy of 98.30%, in the second row this classifier predicted 123 flop movies by accuracy of 97.62%, and the overall prediction accuracy is 98%.

Decision Tree

The last classifier that used in this research in Decision tree, similar to other classifiers the researcher applied it on different number of movies to achieve more accurate prediction.

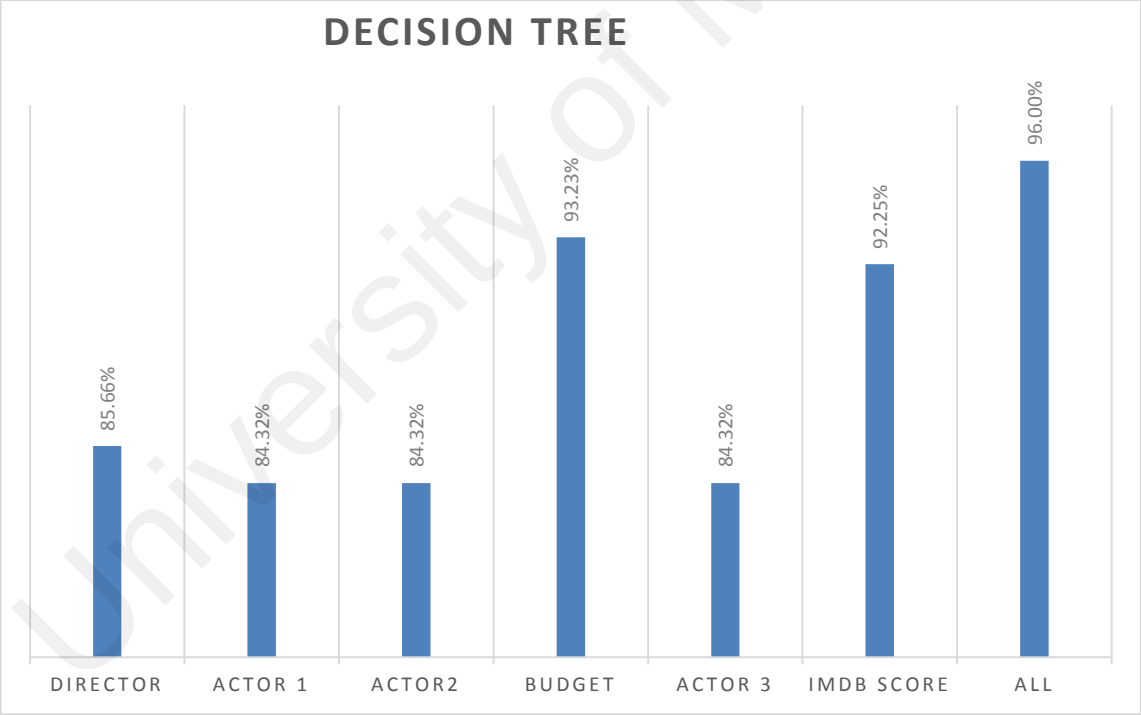


Figure 4. 24 : Decision Tree Result based on 100 movies

Figure 4.24 represents the results from applying the Decision Tree classifier on 100 movies, this figure shows that the lowest prediction belongs to Actors features and the highest belongs to feature combinations.

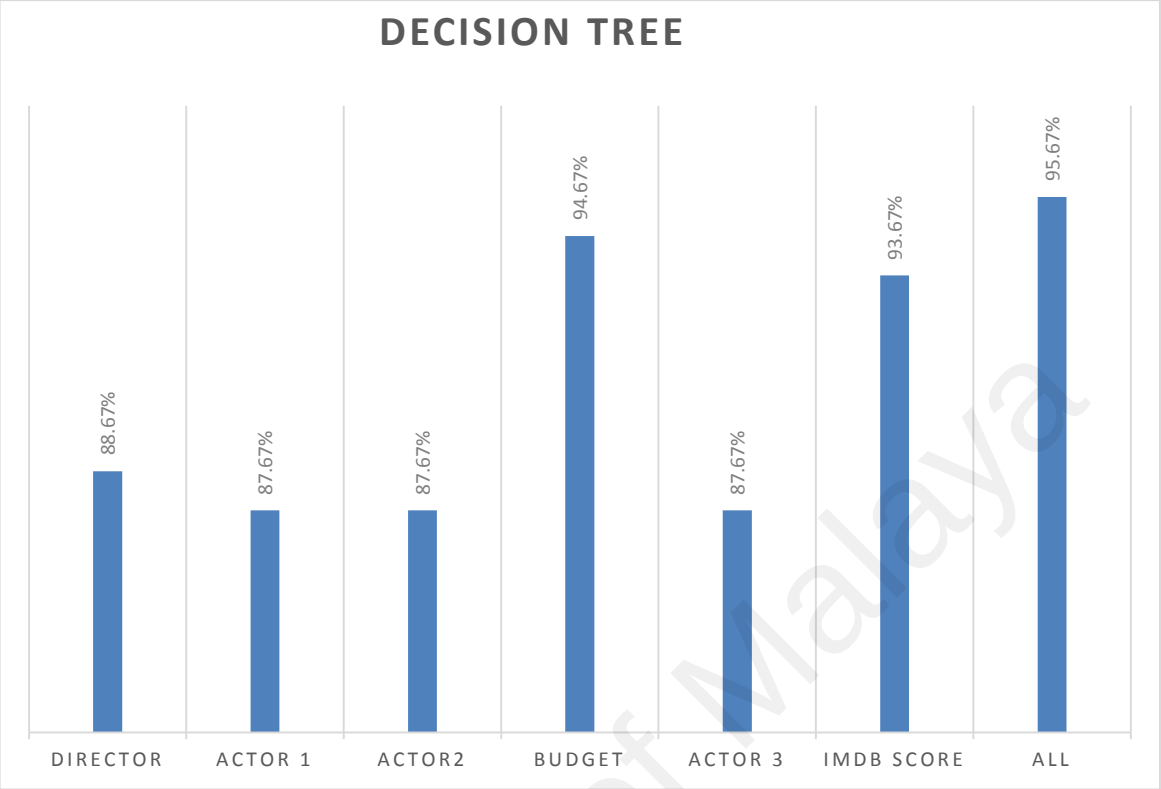


Figure 4. 25 : Decision Tree Result based on 150 movies

By adding 50 movies to dataset, the prediction percentage increase in all features, the numbers might not be noticeable but we can understand that by adding more data our prediction becomes more accurate.

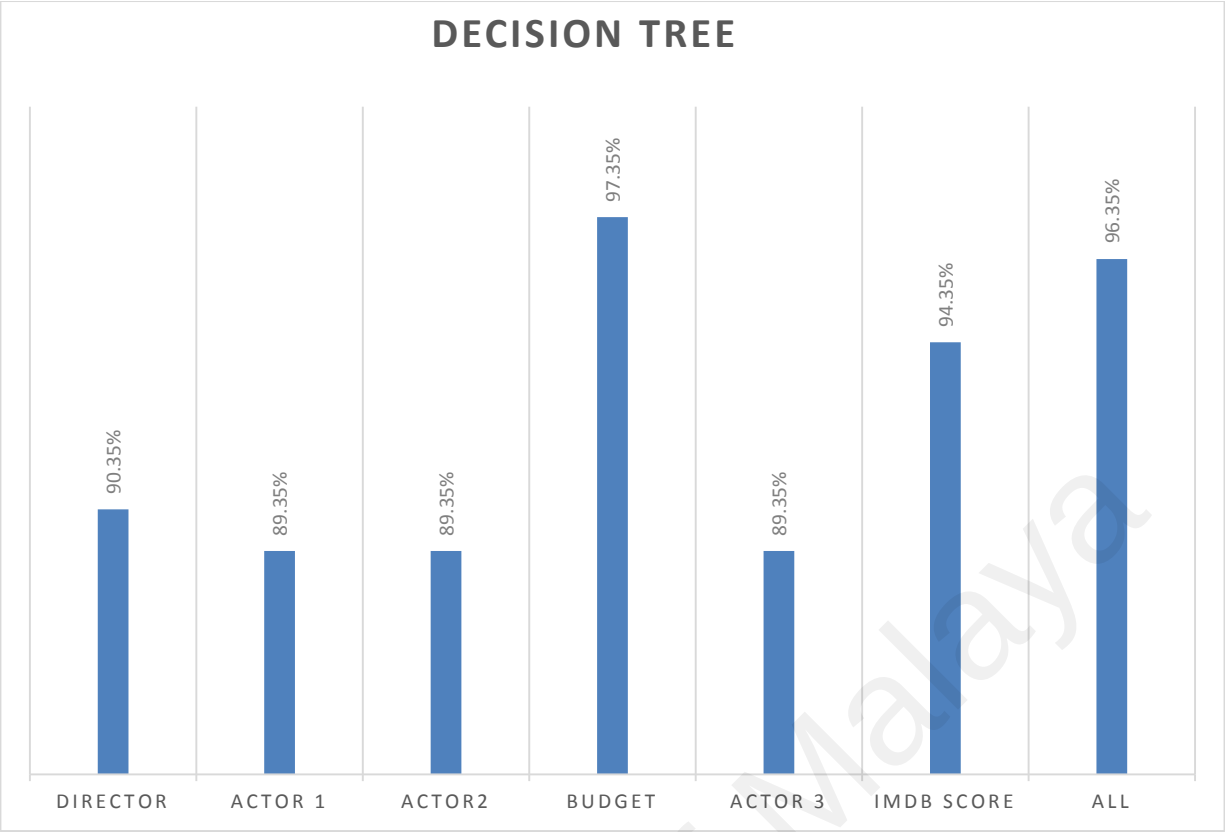


Figure 4. 26 : Decision Tree Result based on 200 movies

Figure 4.26 shows that by adding more data the prediction results keep raising, in this figure the prediction based on Budget gives the highest prediction percentage and then feature combination has the next big number compare with other features.

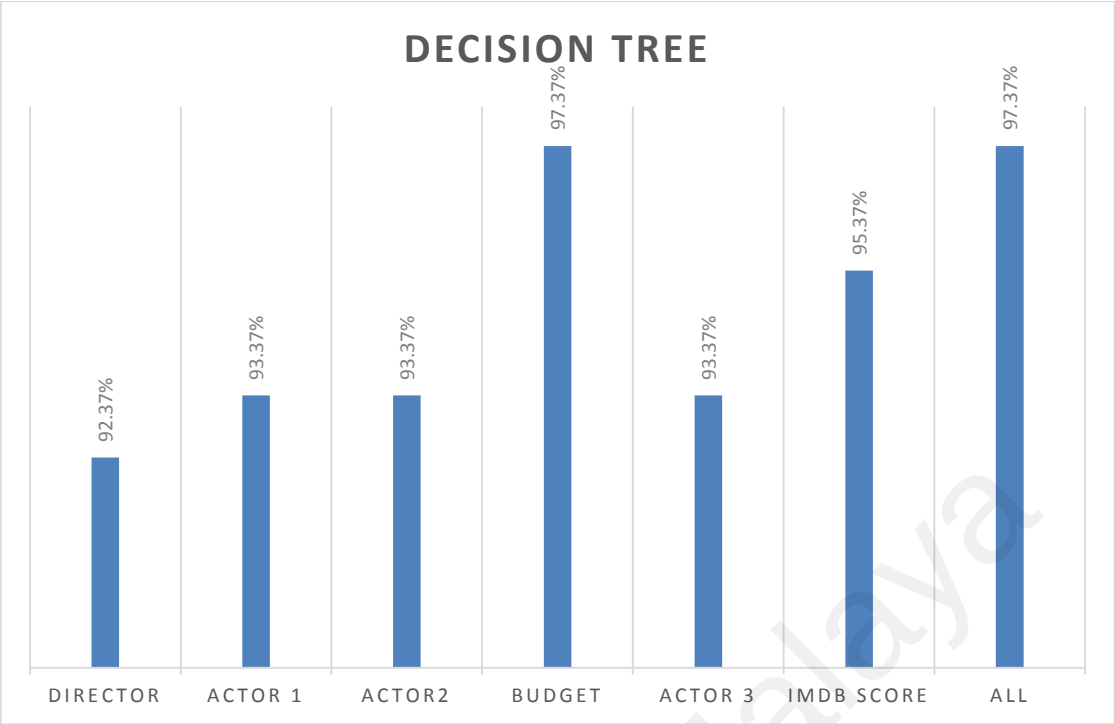


Figure 4. 27: Decision Tree Result based on 200 movies

Figure 4.27 represents the results extracted from applying decision tree classifier on 200 movies, the results show that there is a big changes in some features like Actors features, but the highest prediction percentage belongs to features combination and Budget.

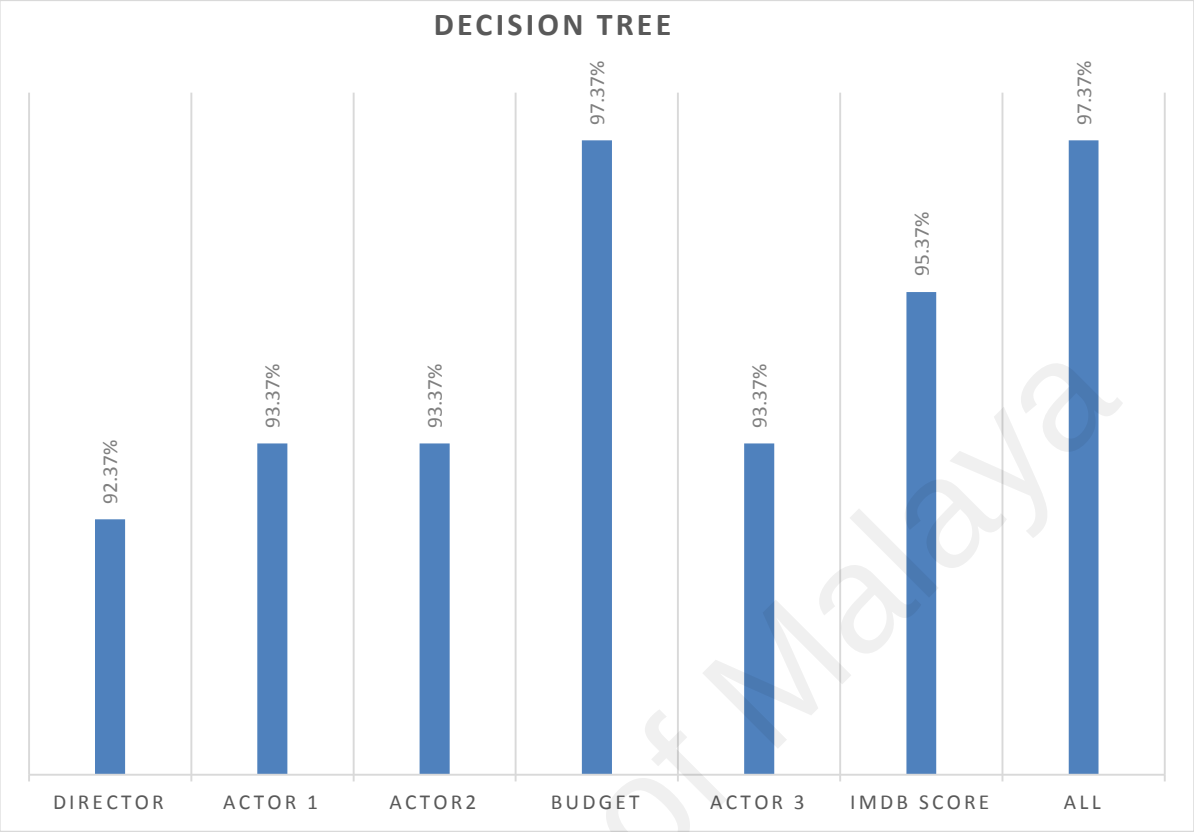


Figure 4. 28 : Decision Tree Result based on 300 movies

The above figure shows that adding 50 more movies to this experiment doesn't have that much effect on the results compare with previous figure and prediction percentages is similar to 200 movies. To make sure if adding more data will affect this result we continue this experiment using 400 movies.

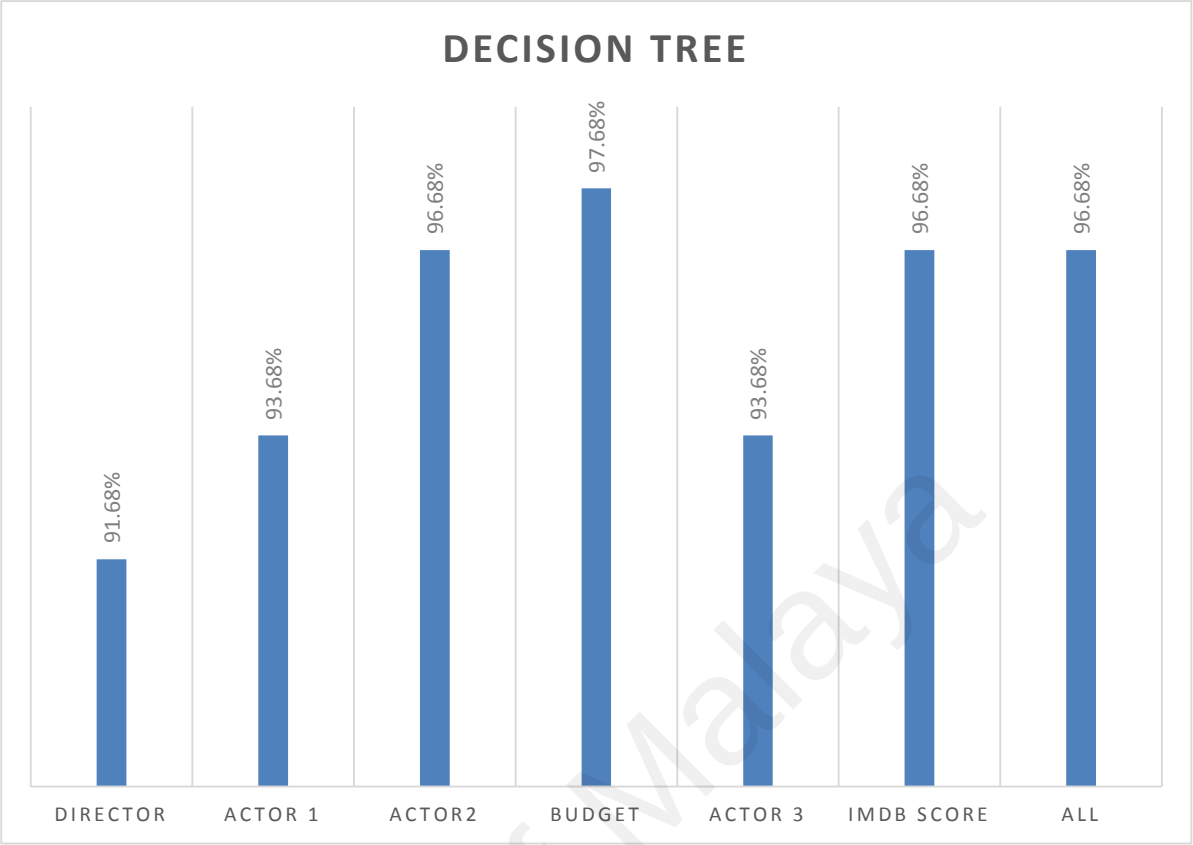


Figure 4. 29 : Decision Tree Result based on 400 movies

Figure 4.29 shows interesting results, for all features there is a rapid growth in the prediction percentage, the budget feature has the best prediction compare with other features.

	True Blockbuster	true Flop	class precision
pred. Blockbuster	174	8	95.60%
pred. Flop	2	118	98.33
class recall	98.86%	93.65%	

Table 4. 4 : Decision Tree Result based all features combinations

By considering the Table 4.4 it can be observed that the Decision Tree classifier correctly predicted 174 blockbuster movies with accuracy of 95.60%, the overall accuracy prediction percentage for this classifier is 96.35%.

By doing this research and using different classifiers following with different features we reached the fact that in all the classifiers, after using the certain number of movies for performing each experiment , the result and accuracy percentage did not change and in some conditions they begin to have deduction in the accuracy. To achieve highest accuracy researcher applied different number of movies to understand how different number of movies will affect the result.

The result shows that K-NN classifier gives the highest accuracy in most of features.

CHAPTER 5: CONCLUSION

This chapter aims to draw a general conclusion and also discusses the future direction for this research. This conclusion also points out some of limitation related to proposed techniques which has been evaluated by the researcher.

This research can be extended from different point of view for future studies. In this research several classifiers are examined separately, the most noticeable extension to this research is to implement methods which combine different features to get the most accurate result. For this purpose different literatures from different researches examined to come up with best possible methods for extracting best possible feature s and classifies which can lead us to most accurate results.

In this research we combined different features and classifiers to predict blockbuster movies, after performing different series of experiments with different numbers data we get that with using 400 movies we can have better prediction on our proposed classifiers and these results shows that K-NN classifier can predict the blockbuster movies with higher accuracy compared with other classifiers. At the end we reached the fact that combination of all selected features along with using K-NN classifier has given the best accuracy percentage.

REFERENCES

Apala, K. R., Jose, M., Motnam, S., Chan, C. C., Liszka, K. J., & de Gregorio, F. (2013). Prediction of Movies Box Office Performance Using Social Media. *2013 Ieee/Acm International Conference on Advances in Social Networks Analysis and Mining (Asonam)*, 1209-1214.

Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence, 97*(1-2), 245-271.

Breiman, L. (2017). *Classification and regression trees*: Routledge.

Coomans, D., & Massart, D. L. (1982). Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-Nearest neighbour classification by using alternative voting rules. *Analytica Chimica Acta, 136*, 15-27.

Doak, J. (1992). CSE-92-18-an evaluation of feature selection methodsand their application to computer security.

Eliashberg, J., Hui, S. K., & Zhang, Z. J. (2007). From story line to box office: A new approach for green-lighting movie scripts. *Management Science, 53*(6), 881-893.

Frank, E., & Bouckaert, R. R. (2006). *Naive bayes for text classification with unbalanced classes.* Paper presented at the European Conference on Principles of Data Mining and Knowledge Discovery.

Ghosh, D., Olewnik, A., & Lewis, K. (2018). Application of Feature-Learning Methods Toward Product Usage Context Identification and Comfort Prediction. *Journal of Computing and Information Science in Engineering, 18*(1), 10. doi:10.1115/1.4037435

Hsu, P. Y., Shen, Y. H., & Xie, X. A. (2014). Predicting Movies User Ratings with Imdb Attributes. In D. Miao, W. Pedrycz, D. Slezak, G. Peters, Q. Hu, & R. Wang (Eds.), *Rough Sets and Knowledge Technology, Rskt 2014* (Vol. 8818, pp. 444-453).

Lash, M. T., & Zhao, K. (2016). Early Predictions of Movie Success: The Who, What, and When of Profitability. *Journal of Management Information Systems, 33*(3), 874-903. doi:10.1080/07421222.2016.1243969

Latif, M. H., & Afzal, H. (2016). Prediction of Movies popularity Using Machine Learning Techniques. *International Journal of Computer Science and Network Security, 16*(8), 127-131.

Lee, K., Park, J., Kim, I., & Choi, Y. (2016). Predicting movie success with machine learning techniques: ways to improve accuracy. *Information Systems Frontiers*. doi:10.1007/s10796-016-9689-z

Lutter, M. (2014). Creative success and network embeddedness: Explaining critical recognition of film directors in Hollywood, 1900–2010.

Mestyan, M., Yasseri, T., & Kertesz, J. (2013). Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. *Plos One, 8*(8). doi:10.1371/journal.pone.0071226

Nithin, V., Pranav, M., Sarath, B., Lijiya, A. J. I. J. o. D. M. T., & Applications. (2014). Predicting Movie Success Based on IMDB Data. *3*, 365-368.

Oghina, A., Breuss, M., Tsagkias, M., & de Rijke, M. (2012). *Predicting imdb movie ratings using social media.* Paper presented at the European Conference on Information Retrieval.

Parimi, R., & Caragea, D. (2013). *Pre-release box-office success prediction for motion pictures.* Paper presented at the International Workshop on Machine Learning and Data Mining in Pattern Recognition.

Peukert, C., Claussen, J., & Kretschmer, T. (2017). Piracy and box office movie revenues: Evidence from megaupload. *International Journal of Industrial Organization, 52*, 188-215.

Reddy, A. S. S., Kasat, P., & Jain, A. (2012). Box-office opening prediction of movies based on hype analysis through data mining. *International Journal of Computer Applications, 56*(1).

Rhee, T. G., Zulkernine, F., & Ieee. (2016). *Predicting Movie Box Office Profitability A Neural Network Approach*.

Saraee, M., White, S., & Eccleston, J. J. T. o. t. W. I. (2004). A data mining approach to analysis and prediction of movie ratings. 343-352.

Sawhney, M. S., & Eliashberg, J. (1996). A parsimonious model for forecasting gross box-office revenues of motion pictures. *Marketing Science, 15*(2), 113-131. doi:10.1287/mksc.15.2.113

Schneider, J. (1997). Cross validation. *A Locally Weighted Learning Tutorial Using Vizier, 1*.

Sharda, R., & Delen, D. (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications, 30*(2), 243-254. doi:10.1016/j.eswa.2005.07.018

Singh, J., Singh, G., & Singh, R. (2017). Optimization of sentiment analysis using machine learning classifiers. *Human-Centric Computing and Information Sciences, 7*(1), 32.

Singh, V. K., Piryani, R., Uddin, A., Waila, P., & Ieee. (2013). *Sentiment Analysis of Movie Reviews A new Feature-based Heuristic for Aspect-level Sentiment Classification*.

Tomar, R., & Verma, C. User propensity analysis for Movie prediction rating based on Collaborative filtering and Fuzzy system.