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 Boxoffice 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

Abst	ract		iii
Abst	rak		iv
Ackı	nowledg	ements	v
Tabl	e of Cor	itents	vi
List	of Figur	es	viii
List	of Table	rs	X
		ools and Abbreviations	
CHA	APTER	1: INTRODUCTION	1
1.1	Backg	round	1
1.2		m statement	
1.3	Aims a	and Objectives of research	3
1.4	Metho	dology	4
	1.4.1	Literature Review	4
	1.4.2	Dataset	4
	1.4.3	Machine Learning Classifiers	4
	1.4.4	Cross Validation	5
CHA	APTER	2: LITERATURE REVIEW	6
2.1	Introdu	iction	6
2.2	Backg	round	6
2.3	Blockb	ouster movie Prediction Techniques	7
	2.3.1	Prediction based on Machine Learning Classifiers	7
	2.3.2	Prediction based on Features used	8
	2.3.3	Prediction based on Statistical models	9

CHAPTER 3: PROPOSED METHODOLOGY FOR BLOCKBUSTER MOVIE

PREDICTION 13

3.1	Introdu	iction	13
3.2	Predict	ion Challenges	14
3.3	Model	Architecture	14
3.4	Feature	e Description	16
3.5	Classif	ication Methods	17
	3.5.1	Naïve Bayes Classifier	17
	3.5.2	Distance-based Classifiers	18
	3.5.3	Random Forest	19
3.6	Feature	es	19
3.7	Data S	ampling	20
3.8	Softwa	re Requirements	21
СНА	PTER	4: RESULT AND EVALUATION	22

4.1	Importing Datasets and feature Assigning	.22
4.2	Research Result	.29
	4.2.1 Naïve Bayes	. 30

REFERENCES 53

LIST OF FIGURES

Figure 4. 1 Open Rapid miner new project	23
Figure 4. 2 import dataset	23
Figure 4. 3 Displaying movie info and features	24
Figure 4. 4 attribute annotation	25
Figure 4. 5 selecting features and attributes	26
Figure 4. 6 Input datasets to simulator	27
Figure 4. 7 Contents of testing Dataset	27
Figure 4. 8 Annotation Option to select or deselect different options	28
Figure 4. 9 : Naïve Bayes Result based on 100 movies	30
Figure 4. 10: Naïve Bay Result based on 150 movies	31
Figure 4. 11 : Naïve Bay Result based on 200 movies	32
Figure 4. 12 : Naïve BayResult based on 300 movies	33
Figure 4. 13 : Naïve Bay Result based on 400 movies	34
Figure 4. 14 : Random Forest Result based on 100 movies	35
Figure 4. 15 : Random Forest Result based on 150 movies	36
Figure 4. 16 : Random Forest Result based on 200 movies	37
Figure 4. 17: Random Forest Result based on 300 movies	38
Figure 4. 18: Random Forest Result based on 400 movies	39
Figure 4. 19 : K-NN Result based on 100 movies	40
Figure 4. 20 : K-NN Result based on 150 movies	41
Figure 4. 21: K-NN Result based on 200 movies	42
Figure 4. 22 : K-NN Result based on 300 movies	43
Figure 4. 23: K-NN Result based on 400 movies	44

Figure 4. 24 : Decision Tree Result based on 100 movies	.45
Figure 4. 25 : Decision Tree Result based on 150 movies	.46
Figure 4. 26 : Decision Tree Result based on 200 movies	.47
Figure 4. 27: Decision Tree Result based on 200 movies	.48
Figure 4. 28 : Decision Tree Result based on 300 movies	.49
Figure 4. 29 : Decision Tree Result based on 400 movies	. 50

LIST OF TABLES

Table 4. 1 Result based on all features combinations	
Table 4. 2 : Random Forest Result based all features combinations	
Table 4. 3 : K-NN Result based all features combinations	44
Table 4. 4 : Decision Tree Result based all features combinations	51

LIST OF SYMBOLS AND ABBREVIATIONS

K-NN : K-Nearest

university Malaya

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.

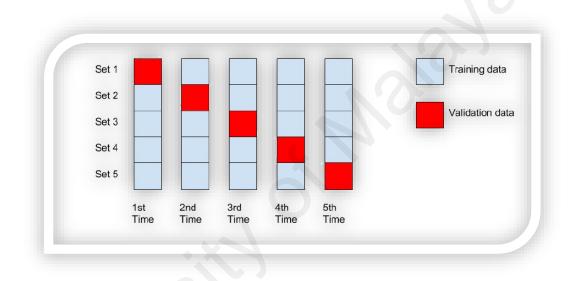


Figure 1.1: Visual representation of 5 fold cross validation

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)	PredictionisbaseduponLogisticRegressionwithstochasticgradient,LinearRegressionand supportVectorclassifiers	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.

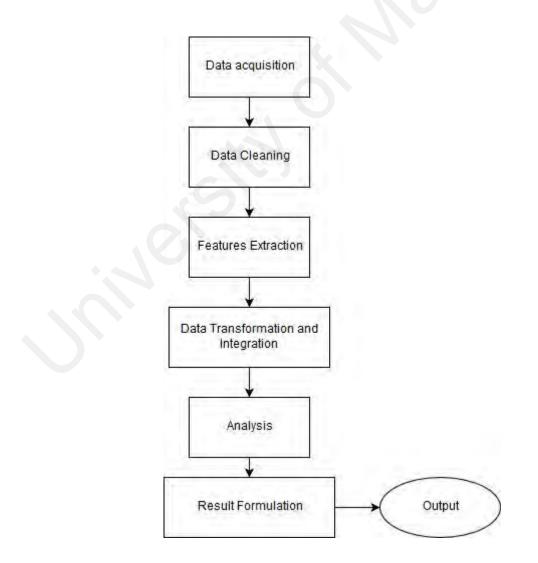


Figure 3.0 Research Workflow

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.

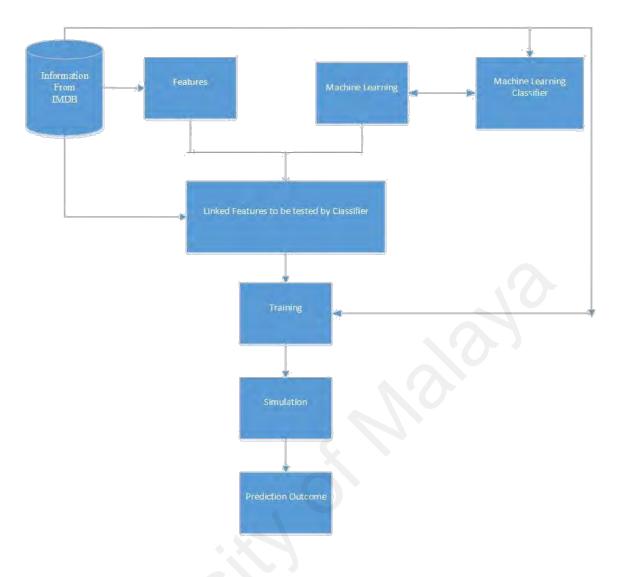


Figure 3. 1: Representation of Proposed Methodology

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:

$$P(c|X) = P(c | x_{1,...,x_f}) = \frac{P(c)P(x_{1,...,x_f}|c)}{P(x_{1,...,x_f})}$$

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. $X = \arg \max [\Box (f_i + x)]$
 - 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	Used to collect
data and use them in simulator	
4. Microsoft Visio 2013	Used to create
figures and charts	
5. Rapid Miner 6.0	Experimentation
6. EndNote 7.0	Referencing and
citation	

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:

Interpretation of the second secon	A.		- ¤ ×
Open, save and print processes	Viters Design Results		G Chuestions? +
Becent Processes	Process	Parameters	
Save Process Ctri+S 1Data = +	🕼 Process 100% 🖉 🖉 🖉 🖉	Process	
Save Process as	Insure	logyerbosity	init e G
Import Process.	lie	logfile	a (a)
Laport Process		resultive	
PrintEgport Image		random seed	2001 0
d Ept			
		send mail	usver + 0
		encoding	SYSTEM * (1)
Onerstore			
openatora		Thide advanced parameters	
Data Access (45)		Change compatibility (7.4 000)	
Blending (77)		Help ×	
Cleansing (26) Modeling (129)			
 Scoring (9) 		RapidMiner Studio Core	
Validation (29) Usity (85)		Synopsis	
Extensions (17)		The root operator which is the outer	most operator of every process.
		Description	
		root operator of the process. This op	te operator of this class, and it must be the erator provides a set of parameters that are
		of global relevance to the process like random number generator.	logging and initialization parameters of the
	Leverage the Wisdom of Crowds to get operator recommendations based on your process design!	Parameters	
Get more operators from the Marketplace.	Adviste Wisdom of Crowds	logverbosity (optional)	

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:

👌 📒 + 🖬 = 🚺	· .		(9	Design	Results				Questions?
Repository		Process					Parameters 🛛 🖂		
🕒 Add Data	10 ÷	Import Data - Select th	e data location.				× 355		
Samples							ty.	int.	*
DB DB				Select the data lo	auon.		-		
Local Repository stants									
Cloud Repository encomedee		Te Documents				* - 5 存金	1 A.		
		Bookmarks	File Name	Size	Туре	Last Modified			
		+ - Last Directory	Adobe		File Folder	May 5, 2015	A led	2001	
			Any DVD Cloner Platinum		File Folder	Nov 6, 2015			
			Audio Notetaker Recordings		File Folder	Oct 5, 2016		never	
			Awosoft		File Folder	Apr 29, 2015			
			Corel VideoStudio Pro Custom Office Templates		File Folder File Folder	May 5, 2015 Sep 24, 2014		SYSTEM	*
			DonationCoder		File Folder	Jan 14, 2017			
			FeedbackHub		File Folder	Dec 23, 2017			
			My EndNote Library Data		File Folder	Apr 14, 2016			
			NetBeansProjects		File Folder	May 3, 2017			
			Simpo PDF to Word		File Folder	Feb 1, 2015	10		
			Sony PMB		File Folder	Sep 23, 2014			
			Sound recordings		File Folder	Mar 9, 2016			
perators ×			Thesis Data		File Folder	Jun 1, 2017			
			ViberDownloads		File Folder	Sep 6, 2015	advanced parameters		
			Virtual Machines		File Folder	Feb 12, 2017	ge compatibility (7 4.000)		
Data Access (45)			Visual Studio 2012		File Folder	Jan 3, 2018	AL LOUIS PROPERTY DE ALCONS		
Blending (77)			Xilisoft		File Folder	Feb 1, 2015			
			1 1 pkt	129 KB	Cisco Packet Tracer	Nov 7, 2016	-26		
Cleansing (26)			My EndNote Library eni	42 KB 13 KB	EndNote Library EndNote Library	Jan 19, 2018 Jun 1, 2017			
Modeling (129)			Untitled1.htm	1.KB	HTM File	Sep 22, 2017	ocess		
Scoring (9)			a province and	1 PM	PILM F HU	560 22 2017	pidMiller Storage Core		
Validation (29)									
Utility (85)		All Files					operator which is the out	er most operator of every proce	55.
Extensions (17)						Please select a f			
							ion		
						- Previous -	Gancel cess must contain exactly	one operator of this class, and it	
						The women's the	issue upportator of the process. This of		
							of global relevance to the process l		
							random number generator.		
		Leverage the Wisdom of Cr	owds to get operator recommendation	ns based on your process design	l.				
				Activate Wisdom of Crowd			Parameters		
Get more operators from the Marketplace									

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

🔪 📒 🖌 ト ラ ピー	 • 						Viewa:	Design	R	sults								Questions?
Repository		Prec	Process												Parame	eters		
😋 Add Data	Import Data - Select the cells to import.													×	355			
Samples									ls to imp							ty.	init	*
DB			_	_	_	_	Selec	ct me ce	is to imp	on.				_				
Cloud Repository Manual																		
same to promy construction		Sheet BluckBuster Cell range M20 M20			s	Seleci Ali 🧹 Define header row			12									
			A	8	c	D	E	F	G	н	T	a.	ĸ	ι	M	Hed.	2001	
		1	director	movie_title	duration	actor_1	actor_2	Earning	genres	num_vol_	actor_3_	plot_key_	num_us_	language	co ^		never	
		2	Zack Sny.	Batman v S.	183.000	Henry C	Lauren	4071972	Action(Ad.	371639	Alan D. P.	based o	3018.000	English	Us		nevei	
		3	Anthony	Captain Am	147,000	Robert D_	Scarlett J	3630242	Action Ad.	272670	Chris Ev.	based o	1022.000	English	US		SYSTEM	
		4	Justin Lin	Star Trek B.	122,000	Sofia Bo	Melissa	3626451	Adion(Ad	53607.000	Lydia Wil	hatredis	432.000	English	US			
		5		The Legend.	110.000	Christop	Alexande	3626451	Action(Ad	42372.000	Casper	africa ca	239.000	English	US			
		6		X-Men: Apo	144.000	Jennifer	Michael	3302490	Action(Ad.	148379	Tye Sher	mutantis	622.000	English	US			
		100,000	David Ayer	Suicide Squ	123.000	Will Smith	Robin At	3235055	Action Ad	118992	Ike Barin	based.o	971.000	English	UE			
		8	Jon Favr.	The Jungle	106.000	Scatlett J	Bill Murray	1510871	Adventur	106072	Garry Sh	fireljungl.	398.000	English	UF			
		9		Independen	120.000	Vivica A	Sela Ward	1549850	ActionJAd	58137.000		alien balt	520,000	English	US			
perators		10		Godzilła Re	120.000	Mark Chi	Shin'ya T	1435234	Action[Ad.	374.000	Atsuko M	bloodigo	13.000	Japanese		the second se		
		11	James B	Alice Throu WarcraftÁ	113.000	Johnny Dominic	Alan Ric.	1260888	Adventur	21352 000	Ruth Ne	based o	131.000 781.000	English	US	advanced parameters		
Data Access (45)		13		The Legend	110.000	Christop.	Alexande	1240517	ActionIAd	42372.000		africajca	239.000	English	US	ge compatibility (7.4.000)		
Blending (77)		14		Ghostbuste	116.000	Ed Begle	Kate Mc.	1240517	ActionIC	69757.000		ghostigh	1211.000	English	US			
Cleansing (26)		15		Kung Fu Pa	95.000	JK Sim_	Angelina	1180996	ActionIAd	64322.000		chinajku_	145.000	English	US	*		
Modeling (129)		16		Gods of Equ.	126.000	Gerard B	Elodie Y.	1180996	Action Ad.	51892.000		ancient e	273.000	English	US	ocess		
Scoring (9)		17		The BFGÅ	117.000	Mark Rul	Penelop.	1085218	Adventur	12572.000	Rafe Spall	cannibali	105.000	English	UF	oldMoser Sturtila Core		
Validation (29)		18	Hideaki	Godzilla Re	120.000	Mark Chi	Shin'ya T_	1072251	Action Ad.	374.000	Atsuko M	bloodigo_	13.000	Japanese	Ja	the surveyore		
Utility (85)		19	Paul Gre.	Jason Bour	123.000	Matt Da.	Riz Ahmed	1023155	Action[Th.	40123.000	Ato Essa	black op_	297.000	English	UF	and the second second		
Extensions (17)		20	Cedric N_	The Hunts.	120.000	Chris He_	Sam Cla	1023101	Action)Ad.	37750.000	Charlize	based o	134.000	English	USY	operator which is the outer n	lost operator of every p	rocess.
			<				-								5	ion		
		+ Previous+ Next 💥											xt 🗶	Cancel	cel sess must contain exactly one operator of this class, and it must be the rator of the process. This operator provides a set of parameters that are			
															ofgloba	I relevance to the process like		
		Leverage the Wisdom of Crowds to get operator recommendations based on your process design!												random	number generator.			
													Param	eters				

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

	• •					Viewo	Dest	pei -	Results						Gan Can	estions?	
epository 🛸	Pr	ocess												Parameters 🛸			
C Add Data	🐠 Data ii	mport wizar	ort wizard - Step 3 of 4 X										×	Read Excel			
Samples		This wirard	guides you to	ininor trour	oten									> Import C	Configuration Wizard	d.	
DE Local Repository Inner Cloud Repository Inner	1	Step 3: in R	apidMiner Stu le attributes. If	sio, each an	ribute can de a les not contain			excel file	sers/Masih/Desktop/Files/200.dsx								
	Annotat	A	8	с	D	E	F	G	н	1	J	к	L	sheetnumber	1		
	Name	director	movie_title	duration	actor_1_	actor_2	Earning	num_vot_	actor_3	num_us_	language	country	budget 🖄	imported cell range	A1:D303		
		Zack Sny	Batman	183.000	Henry C	Lauren	4071972	371639	Alan D. P.	3018,000	English	USA	250000		SYSTEM		
		Anthony	Captain	147.000	Robert D.	Scarlett J_	3630242	272670	Chris Ev.	1022.000	English	USA	250000	encoding	SYSTEM		
		Justin Lin	Star Trek	122.000	Solia Bo.	Melissa_	3626451	53607.000	Lydia Wil.	432.000	English	USA	185000	first row as names			
	-	Bryan Si David Aver	X-Men: A., Suicide	144.000	Jennifer	Michael	3302490 3235055	148379	Tye Sher like Barin	622.000 971.000	English	USA	178000				
		Jon Favr.	The Jun	106.000	Will Smith Scarlett J	Robin AL	1610871	106072	Garry Sh.	398 000	English	USA	175000	annolations	East List (1).		
		Lee Unkr	Toy Story 3	103.000	Tom Ha	John Rat.	4149844	544884	Don Rick.	733.000	English	USA	200000	date format	-	100	
	-	McG	Terminat	118.000	Christian.	Bryce Da	1253200	286095	Common	974.000	English	USA	200000				
	-	James _	Furious 7	140.000	Jason St.	Paul Wal	3500341	278232	Vin Diesel	657.000	English	USA	190000	lime zone	SYSTEM	.*	
erators 8	-	Marc For.	World W.	123 000	Peter Ca.	Brad Pitt	2023516	465019	Mireille E	995.000	English	USA	190000	locale	Enclish (United States)		
		Paul Feig	Ghostbu	116.000	Ed Begle.	Kate Mc	1240517	59757 D00	Zach Wo	1211,000	English	ABU	144000	Tinde advanced parameters	and the second s		
	-	Nessan	Kung Fu	95.000	J.K. Sim	Angelina	1180996	64322.000	Wayne K	145.000	English	USA	145000	Change compatibility (7.4.000)			
Data Access (45)	1.0	Alex Proy	Gods of_	126.000	Gerard B	Elodie Y_	1180996	51892.000	Bryan Br_	273.000	English	USA	140000				
Ellending (77)	-	Steven S	The BFG	117.000	Mark Ryt	Penetop	1085218	12572.000	Rafe Spall	106.000	English	UK	140000	Help			
Cleansing (28)	-	Paul Gre	Jason B	123.000	Matt Da.	Riz Ahmed		40123.000	Ato Essa	297.000 134.000	English	UK	120000	and a farmer			
Modeling (129) Scoring (9)	12	Cedric N. Dave Gr	Teenage	112,000	Chris He_ Stephen_	Sam Cla	1023101	17533.000	Charlize	115.000	English	USA	135000	Read Excel RapidMiner Studio Cole			
Validation (29)		Robert S.	Allegiant	120.000	Naomi	Theo Ja	8163867	44296.000	Zo4 Kravitz	144.000	English	USA	110000		03 film Microsoft Spreidtheath Dates		
Ubility (85)	3,	Timur Be	Ben-Hur	141.000	Morgan	Avelet Zu	7884513	57 000	Moises A.	1.000	English	USA	100000	Synopsis	nos lente morrente surgenerente e man	-	
Extensions (17)		Jon M. C.	Now You	129.000	Daniel R	Morgan	7684662	40862.000	Sanaa L	139.000	English	USA	9000000	This operator reads an ExampleSet.	from the specified Excel file.		
	<												2	Jump to Tutonal Process			
									- Erev	ious -	Next R	Einish	X Cancel				
														Description			
															lata from Microsoft Excel spreadsheets		
	L	everage the W	is dom of Crow	ds to get op	erator recomm	endations ba	sed on your pr	ocess designi							Excel 95, 97, 2000, XP, and 2003. The use s in the workbook should be used as di		

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.

2 📒 🗏 ・ ちぐ 🕨	-					View	s De	sign	Results							📢 Qu	uestions7 •
Repository		Process												Paran	ieters 🗶		
G Add Data	= +	🕘 Data i	mport wiza	ard - Step 4	of 4									×	I Excel		
Samples DB Local Repository Junety		1	Step 4: Ra roles to th	pidMiner Stu	to import your d dio uses strong efining what the	ly typed attrib									> Import C	onliguration Wizard.	0
Cloud Repository data measure		C) Reioa	id data	Guess va	live types D	ite format									mber	1	00
		Prestew	uses only fin	at 100 rows.											cell range	A1:0303	60
		director_nar	movie_title	duration	actor_1_nar	actor_2_na	Earning.	Num_voted	actor_3_na	(num_user_)	(language	e country	budget	title_year	· · · · · · · · · · · · · · · · · · ·	SYSTEM	• 3
		polyno *	polyno.	integer	* polyno. *	polyno. *	integer *	integer *	polyno	integer *	polyno *	polyno *	integer ·	integer attribute	ow as names		œ
		Zack Sny	Batman .	183	Henry C.	Lauren	4071972	371639	Alan D. P.	3018	English	USA	2500000	2016 ^	vis	📝 Edit List (1)	0
		Anthony Justin Lin	Captain Star Trek.	147	Robert D Softa Bo	Scarlett J. Melissa	3630242 3626451	272670 53607	Chris Ev.	1022 432	English English	USA USA	2500000	2016 2016	kat	én en 🔻	H a
		Bryan Si David Aver	X-Men: A. Suicide	144	Jennifer Will Smith	Michael	3302490 3235055	148379	Tye Sher like Barin	622 971	English English	USA.	1780000	2016		SYSTEM	* 0
erators		Jon Favr	The Jun.	106	Scarlett J	Bill Murray	1610871_	106072	Garry Sh	398	English	ик	1750000	2016		English (United States)	* 3
		Lee Unkr McG	Toy Story Terminat	3 103 . 118	Tom Ha., Christian	John Rat. Bryce Da.	4149844	544884 285095	Don Rick. Common	733 974	English English	USA.	2000000	2016	advanced parameters		
Data Access (45) Blending (77)		James	Futious 7	140	Jason St	Paul Wal	3500341	278232	Vin Diesel		English	USA	1900000	2016			
Cleansing (26)		Marc For Paul Feig	World W	123	Ed Begle	Brad Pitt Kate Mc	2023516 1240517	465019 69757	Mireille E Zach Wo	995	English	USA. USA	1900000	2016			
Modeling (129) Scoring (9)		< O errors:	-			-			-		-	Ignore en	ors Sho	> only errors	ead Excel		
Validation (29)		Row, Colur	na		Error			Ong	jinal value			Wessage				On other Microsoft, Spreadtherers, Danas	ea.
Utility (85) Extensions (17)										← Be	vious —		Einish	Cancel	s trator reads an ExampleSet <u>Jutorial Process</u>	from the specified Excel file.	
															perator can be used to load d	ata from Microsoft Excel spreadsheet: xcel 95, 97, 2000, XP, and 2003. The us	
		reverage me	Leverage the Wisdom of Crowds to get operator recommendations based on your process design! to define												in the workbook should be used as d such that each row is an example and		

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

🐠 //Local Repository/bluckbuster – RapidMine/ Studio Free 7.4J	000 @ Masin-pc						- 🗆 🕫
Elle Edit Brocess View Connections Cloud Settings Extensions		Views:	Design Results				Q Cuestions? •
Repository	Process				Para	meters 🔀	
😋 Add Data 🛛 🖃 🕶	Data import wizar	d - Step 1 of 4			×	1 Excel	
Samples	This wizard	guides you to import your data.				> Import C	Configuration Wizard. (D)
DB Local Repository dataset	Step 1: Plea	se select the file that should be imported.					sers/Masih/Desktop/Files/200.xtsx 📘 🛈
Cloud Repository (and and and and and and and and and and	E TEZ					mber	1 0
	Bookmarks	File Name moviepredict New folder	Size	Type File Folder File Folder	Last Modified Jul 2, 2017 Aug 7, 2017	 cell range 	A1:0303 @
		sample Bluckbuster xisx	17 KB	File Folder Microsoft Excel Worksheet	May 16, 2017 May 2, 2017		SYSTEM + D
		Result-knn(1) xlax	17 KB 36 KB	Microsoft Excel Worksheet Microsoft Excel Worksheet	Aug 11, 2017	ow as names	0
		Result-knn.xlsx	36 KB	Microsoft Excel Worksheet	Aug 11, 2017	173	Edit List (1)
		1000				kat .	Emittande
							SYSTEM • D
Operators ×						advanced parameters	Enclish (United States) 🔹 🗇
						tige compatibility (7.4.000)	
Data Access (45) Blending (77)						and the second s	
Cleansing (26)						and the second s	8
Modeling (129) Scoring (9)						ead Excel	
Validation (29)						at Import Head Data Files.	XIs Xiss Microsoft Spreadsheets Desisters
Unity (85) Extensions (17)	Bluckbuster.sisx					s	from the specified Excel file.
	Excel Spreadsheets (.xls, .	dsx)			Einish X Gancel	Tutonal Process	non die speureo excerne.
	1					ription	data from Microsoft Excel spreadsheets. This
	Leverage the Wisdom of Cro	wds to get operator recommendations based	d on your process design!		oper	ator is able to read data from l	Excel 95, 97, 2000, XP, and 2003. The user has s in the workbook should be used as data
Get more operators from the Marketplace		V A	clivate Wisdom of Crowds		table	The table must have a forma	t such that each row is an example and each ease note that the first row of the Excel sheet v

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

🗋 📒 H + רו ר 🕨 🕨	- 1						viewa:	Design	Re	sults								Questions?
Repository		Proce	155												Parame	ters		
😋 Add Data		Impo	rt Data - S	elect the cells	to import.			-							×	355		
Samples DB Local Repository Name			_	_			Sele	ct the cel	lls to imp	ort.						ty-	init	
Cloud Repository concerns		Shee	at BluckBu	ster T C	ell range M2	20 M20	s	Relect All	Z Define	header (ow	12							
			A	B	c	D	E	F	G	в	ĩ	i	к	ι	M	hed .	2001	
		1	director	movie_title	duration	actor_1	actor_2	Earning	genres	num_vol_	actor_3_	plot_key_	num_us_	language	co ^		never	
		2	Zack Sny	Batman v S.	183.000	Henry C	Lauren	4071972	Action(Ad.	371639	Alan D. P.	based o	3018.000	English	Us		increa.	
		line.	Anthony	Captain Am	147.000	Robert D_	Scarlett J.	3630242	Action Ad	272670	Chris Ev.	based o	1022.000	English	US		SYSTEM	
			Justin Lin	Star Trek B.	122,000	Sofia Bo	Melissa	3626451	Adion)Ad.	53607.000	Lydia Wil	hatredis	432.000	English	US			
			David Ya	The Legend.	110.000	Christop	Alexande	3626451	Action Ad	42372.000	Casper	africa ca.	239.000	English	US			
			Bryan Si	X-Men: Apo	144.000	Jennifer	Michael	3302490	Action)Ad	148379.	Tye Sher	mutantis	622.000	English	US			
		100.00	David Ayer	Suicide Squ	123.000	Will Smith	Robin At	3235055	Action Ad	118992	Ike Barin	based.o	971.000	English	UE			
		a line of	Jon Favr.	The Jungle	106.000	Scatlett J	Bill Murray	1610871	Adventur	106072	Garry Sh	fireljungl	398.000	English	UF			
		incr-		Independen .	120.000	Vivica A.	Sela Ward	1549850	Action)Ad	58137.000	Judd Hir.	alien balt	520,000	English	US			
perators		1000	Hideaki	Godzilla Re Alice Throu	120.000	Mark Chi	Shin'ya T Alan Ric	1435234	Action(Ad	374.000 21352.000	Atsuko M.	blood(go diackidar	13.000	Japanese English				
			Duncan	WarcraftÅ	123.000	Dominic	Callum	1260888	ActionIAd	111609	Ruth Ne.	based o	781.000	English	US US	advanced parameters		
			David Ya.	The Legend	110.000	Christop.	Alexande.	1240517	Action(Ad	42372.000	Casper_	africa ca	239.000	English	US	pe compatibility (7.4.000)		
Data Access (45)		(received)	Paul Felo	Ghostbuste	116.000	Ed Begle	Kate Mc.	1240517	ActionIC	69757.000		ghostigh	1211.000	English	US			
Cleansing (26)		in the second	Alessan	Kung Fu Pa	95.000	JK Sim_	Angelina	1180996	ActionIAd	64322.000	Wayne K	chinalku_	145.000	English	US	×		
Modeling (129)		16	Alex Prov	Gods of Egy.	126.000	Gerard B	Elodie Y	1180996	ActionIAd	51892.000	Brian Br	ancient e	273.000	English	US			
Scoring (9)			Steven S	The REGA	117.000	Mark Rd_	Penelog.	1085218	Adventur	12572.000	Rafe Spall	cannibali	105.000	English	UF	OCESS addMoserStartin Core		
Validation (29)		1000	Hideaki	Godzilla Re.	120.000	Mark Chi.	Shin'ya T_	1072251	ActionIAd	374.000	Atsuko M	bloodigo	13.000	Japanese				
Utility (85)		19	Paul Gre	Jason Bour	123.000	Matt Da	Riz Ahmed	1023155	ActionITh.	40123.000	Ato Essa	black op	297.000	English	UP			
Extensions (17)		20	Cedric N_	The Hunts.	120.000	Chris He	Sam Cla	1023101	ActionIAd.	37750.000	Charlize	based o.	134.000	English	USU	operator which is the outer	most operator of evi	ery process.
			<					1							5	inn.		
												- Eleviou		xt 🗙	Cancel	cess must contain exactly or		
															ofgloba	ator of the process. This op relevance to the process like number generator.		
		Lever	age the Wise	tom of Crowds 1	o get operator	recommenda	tions based o	n your proces	s design!							ineritari Berrergener		

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.

	· 5 C 🕨	•					View	Des	ign	Results							🚱 Qu	estions? •
Repository			Process	× I											Paran	neters		
	😋 Add Data	≡ •	🔹 Data in	mport wizar	rd - Step 3 c	of 4									×	1 Excel		
Samples			This wizard guides you to import your data.											Import C	onfiguration Wizard	D		
DB			1	Step 3: In R	tapidMiner Stu he attributes. If	dio, each attr	ibute can be a			nt annotation o Ihis property if					n defines the ur data file, you		sersiMasihiDesktop/Files/200.xlsx	
and the second second			Annotat	A	8	с	D	E	F	G		1	a.	к	t.	mber	1	-
			Name	director	movie_title	duration	actor_1_	actor_2	Earning	num_vot	actor_3	num_us	languag	e country	budget A	cell range	A1:0303	10
			1.00	Zack Sny _	Batman .	183 000	Henry C	Lauren	4071972	371639	Alan D. P.	3018.000	English	USA	250000		SYSTEM	
			A 12	Anthony	Captain	147.000	Robert D.	Scarlett J	3630242	272670	Chris Ev	1022.000	English	USA	250000		STRIEM	-
			-	Justin Lin	Star Trek	122.000	Sofia Bo.	Melissa	3626451	53607.000	Lydia Wit	432.000	English	USA	185000	ow as names		D.
				David Ya	The Lag	110.000	Christop	Alexande	3626451	42372.000	Casper_	239.000	English	USA	180000			
				Bryan Si David Aver	X-Men: A Suicide	144.000	Jennifer	Robin At	3302490	148379	Tye Sher	622.000 971.000	English	USA	178000	ins	EatList(1).	
			-	Jon Favr	The Jun_	106.000	Scarlett J.	Bill Murrav	1610871	105072	Garry Sh	398.000	English	UK	175000	at		100
				Roland_	Indepen	120.000	Vivica A	Sela Ward	1549850	58137.000	Judd Hir	520.000	English	USA	1650001			
			-	Hideaki	Godzilla_	120.000	Mark Chi	Shin ya_	1435234	374.000	Atsuko M	13.000	Japanes		165000	*	SYSTEM	• 10
perators			-	James B	Alice Thr.	113.000	Johnny	Alan Ric	1304686	21352.000	Anne Hat.	131.000	English	USA	170000		English (United States)	
permore			-	Duncan	Warcraft	123.000	Dominic	Callum	1260888	111609	Ruth Ne	781.000	English	USA	1600001	advanced parameters	Enterent (Onited States)	
			44	Paul Feig	Ghostbu	116.000	Ed Begle	Kate Mc.	1240517_	69757.000	Zach Wo	1211.000	English	USA	144000	age compatibility (7.4.000)		
Data Access (45)			÷.	Alessan	Kung Fu	95.000	J.K. Sim	Angelina	1180996	64322.000	Wayne K_	145.000	English	USA	145000	The second secon		
Blending (77)				Alex Proy	Gods of	125.000	Gerard B	Elodie Y	1180996	51892.000	Bryan Br	273,000	English	USA	140000	The second second		
Cleansing (26)			1.27	Steven S	The BFG	117.000	Mark Ryl.	Penelop.	1085218	12572.000	Rafe Spall	106.000	English	UK	1400001			
Modeling (129)				Paul Gré Cedric N	Jason B.	123.000	Matt Da. Chris He	Riz Ahmed Sam Cla	1023155	40123.000	Ato Essa	297.000	English	UK	120000	ead Excel		
Scoring (9) Validation (29)				Dave Gr.	Teenage	112.000	Stephen	Noel Fis	1908350	17533.000	Brad Gar	115.000	English	USA	135000	ipidMiner Studio Core	(b. Ku, Microsoft, Spread-been, Datas	in the second
Ubity (85)				Robert S.	Allegiant	120.000	Naomi	Theo Ja	2816386	44296.000	Zo = Kra	144 000	English	USA	1100000	an unport peak tiess (ties)	the state subclassic associationers remain	uera
Extensions (17)				Timur Be	Ben-Hur	141.000	Morgan	Ayelet Zu	3788451	57.000	Moises A.	1 000	English	USA	1000001~	rator reads an ExampleSet	from the specified Excel file.	
			<												×	Tutorial Process		
											erev Brev	tous -	+ Next	W Einish	X Cancel			
			-											-	Descr	iption		
			Leverage the l	Nisdom of Cri	owds to get op	erator recom	mendations b	ased on your (process desig	pal					opera	tor is able to read data from E	ata from Microsoft Excel spreadsheets xcel 95, 97, 2000. XP, and 2003. The us	er has
			tode											in the workbook should be used as d such that each row is an example and				

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.

						100		ngn	Results							🧐 Gu	estone?	
Repository		Process	Process Param								noters							
O Add Data	= +	Data in	mport wiza	ard - Step 4	of 4									×	d Excel			
Samples		This witard aildes you to import your data												✓ Import C	onfiguration Wizard.	10		
DB Local Repository Cloud Repository		1	Step 4: RI roles to th	pidMiner Stu	dio uses strong	ly hped attric							pidiliner Studio o can rename a		Inter	sersWassPyDeoxdopWide6\200.xlax		
		Beloa	ad data	Quess vi	alue types D	ate formal												
		· Proview	uses only fin	at 100 rows.											cell range	A1:0303	18	
			4	191	- 21	4	2.	4	1	2	1	2	10	1		SYSTEM		
		director_nar	movia_title	duration		actor_2_na	ar Laming	num_voted.	actor_3_nar	num_user_f	language	country	tudget	\$80_year				
		atticute *	attribute	integer attribute	▼ polyno. ▼		· integar ·	attribute *	polyno_ *	infeger +	potyno *	polyno	 integer * attribute * 		ow as hames			
		Zack Sm.	Batman	183	 attribute Henry C_ 	Lauren	4071972	371639	Alan D.P.,	attribute *	English	USA	2500000 -	attribute 2016	416	EditList(1).		
		Anthony	Claptain	147	Robert D	Scanot J		272670	Chris Ex.	1022	English	USA	2500000	2016				
		Justin Lin	Star Trek	122	Sofia Bo	Melissa	3626451	53607	Lydia WR	432	English	USA	1850000	2016	ant'		100	
		David Ya	The Log.	110	Christop	Alexande	3626451	42372	Casper	239	English	USA	1800000	2016	*	SYSTEM		
erators		Bryan Si.	X-Men A.	144	Jennifer -	Robin M	3302490	148379	Tye Sher	822	English	USA	1780900	2016	-	English (United States)		
erators		Jon Favr.	The Jun.	123	Scarlett J	Bill Murra		106072	Garry Sh.	398	English	UK	1750000	2016	advanced parameters	English (United States)		
		Roland	Indepen	120	VIVICE A.	Sela Ward		58137	Judd Hir	520	English	USA	1650000	2018	tos compatibility (7.4.000)			
Data Access (45)		Hideaki	Gadalite	120	Mark Chi.	Shin ya .	1435234	374	Atsuko M	13	Japanese	Japan	1650000	2016				
Elementing (27) Cleansing (26)		James B.	Alice Thr.	113	Johrany	Alan Ric.	1304686	21352	Anne Hat	131	English	USA	1700000	2018	100			
Modeling (129)		Duncan	Warcraft	123	Dominic	Callum -	1260888	111609	Ruth Nee	781	English	US4	1600000	2016	ead Excel			
Scoring (9)		C ettors					-					Ignore e	ontors Sho	w only prors	Ludhimar Studio Coni			
Validation (29)		Row, Colur			Error			Ong	inal value			Anssage			ad Insport East, Dave Stan 8	II. Car Mirrisoft Spinarsteina Dinas	122	
Usiley (86)																		
Edensions (17)															Wator reads an ExampleSet I Tutorial Process	from the specified Excel file.		
										- Bres	tous -	·	Einten	Cancel	- Conditional Process			
												-		Desc	iption			
														This o	perator can be used to load da	ata from Microsoft Excel spreadsheets		
		Leverage the												to del	operator is able to read data from Excel 95, 97, 2000, XP, and 2003. The user has to define which of the spreadsheets in the workbook should be used as data			
Get more operators from the Marketplace			Activate Wisdom of Crowds table. 1												such that each row is an example and			

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.

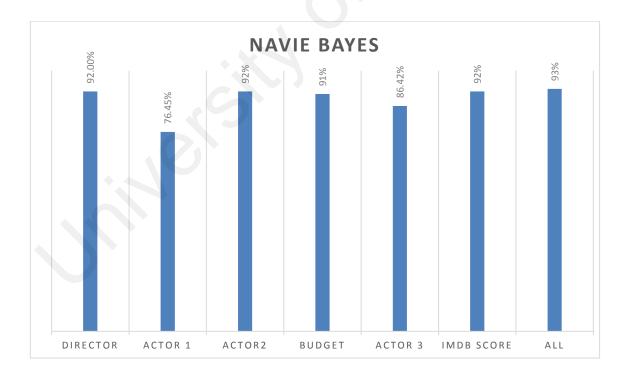


Figure 4.9: Naïve Bayes Result based on 100 movies

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.

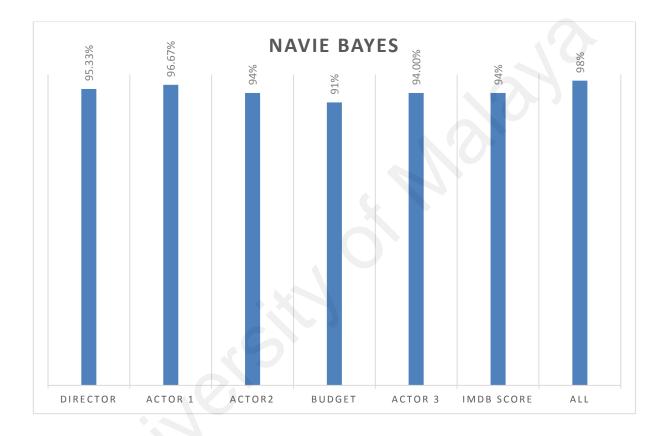


Figure 4. 10: Naïve Bay Result based on 150 movies

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.

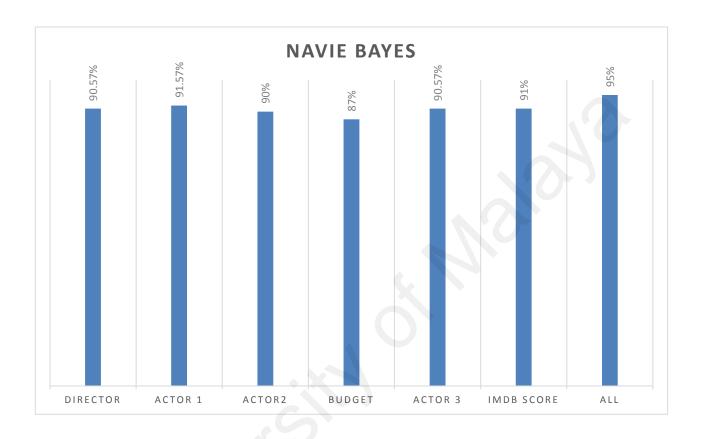


Figure 4. 11 : Naïve Bay Result based on 200 movies

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.

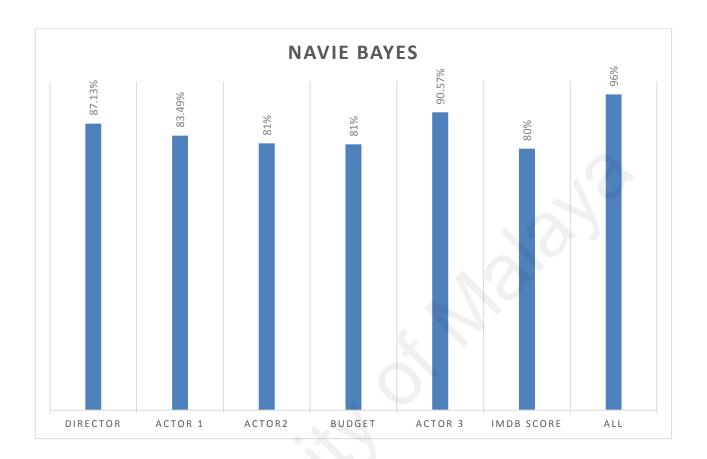


Figure 4. 12 : Naïve Bayes Result based on 300 movies

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.

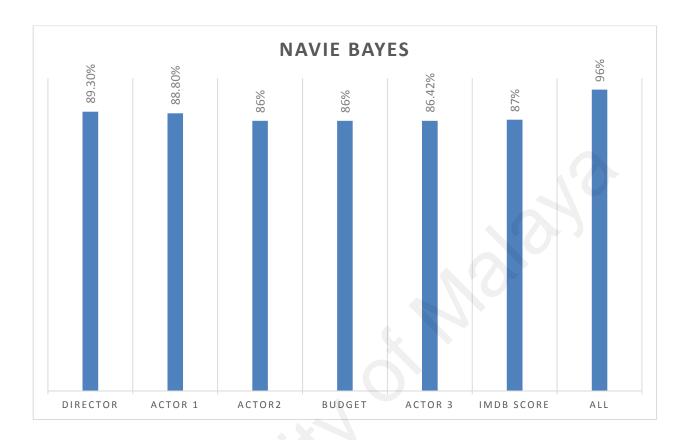


Figure 4. 13 : Naïve Bay Result based on 400 movies

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.

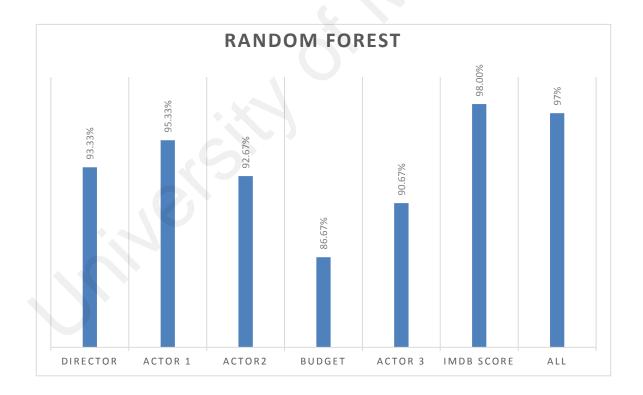


Figure 4. 14 : Random Forest Result based on 100 movies

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.

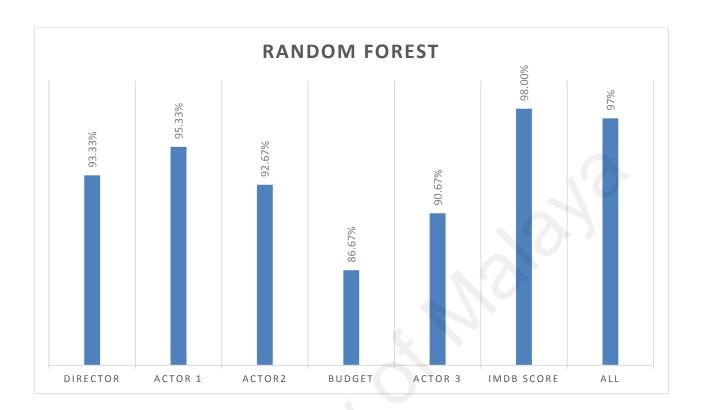


Figure 4. 15 : Random Forest Result based on 150 movies

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.

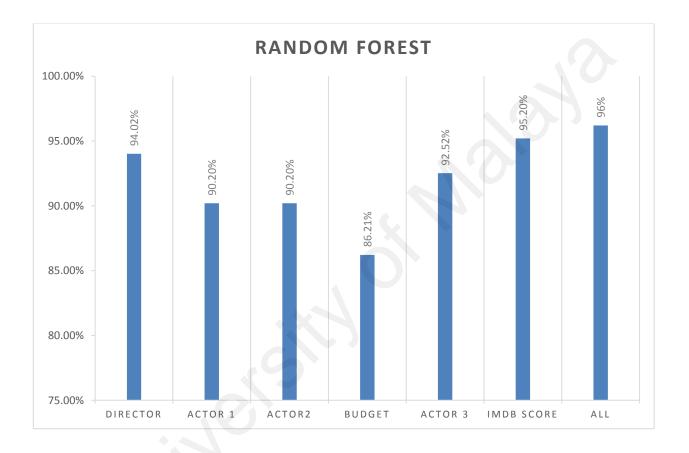


Figure 4. 16 : Random Forest Result based on 200 movies

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.

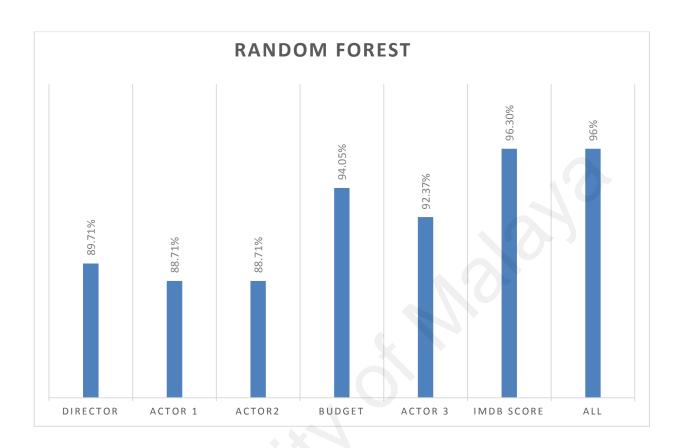


Figure 4. 17: Random Forest Result based on 300 movies

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

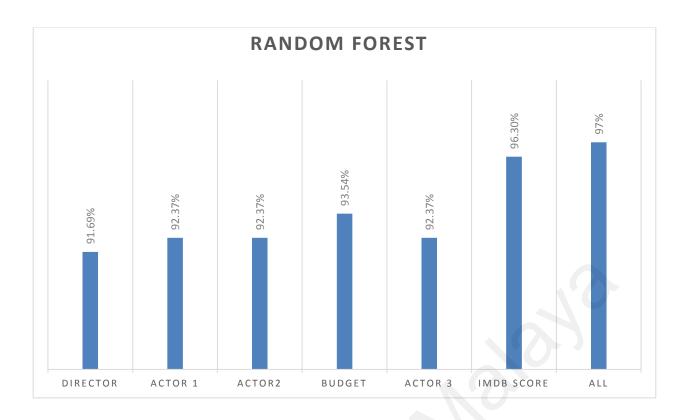


Figure 4. 18: Random Forest Result based on 400 movies

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.

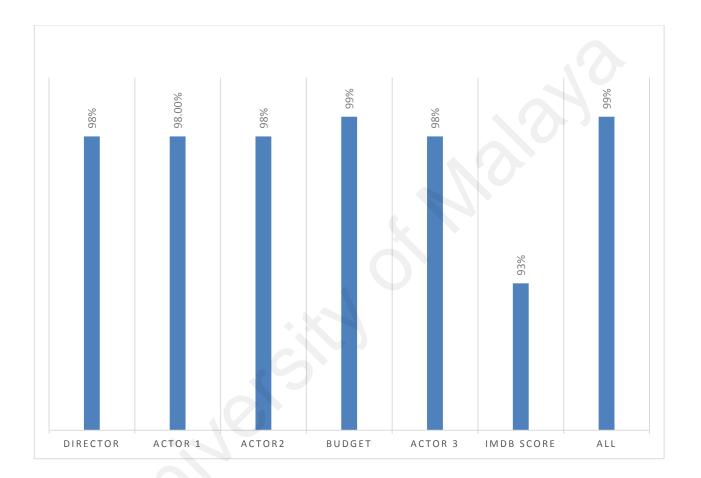


Figure 4. 20 : K-NN Result based on 150 movies

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.

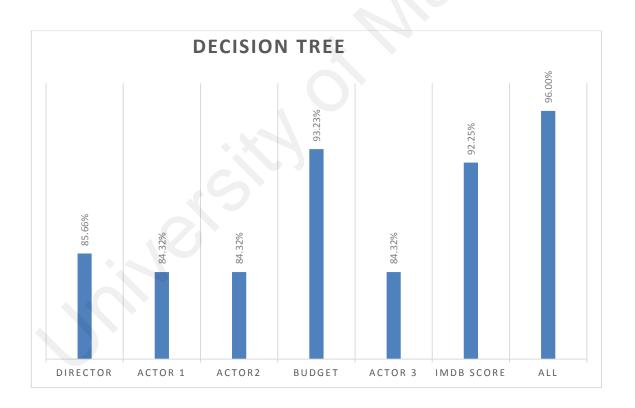


Figure 4. 24 : Decision Tree Result based on 100 movies

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.

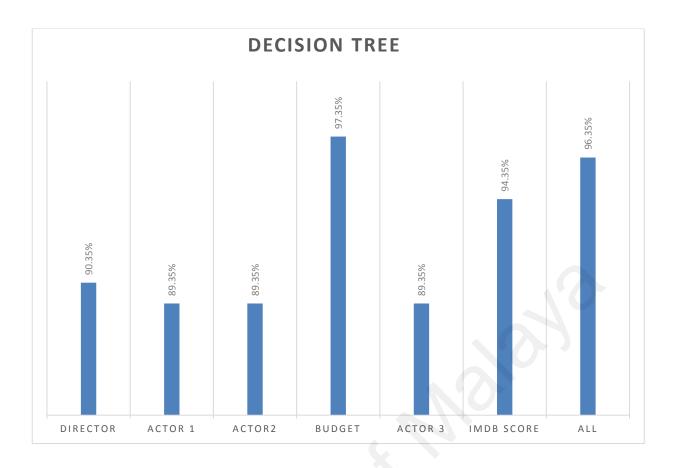


Figure 4. 26 : Decision Tree Result based on 200 movies

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.

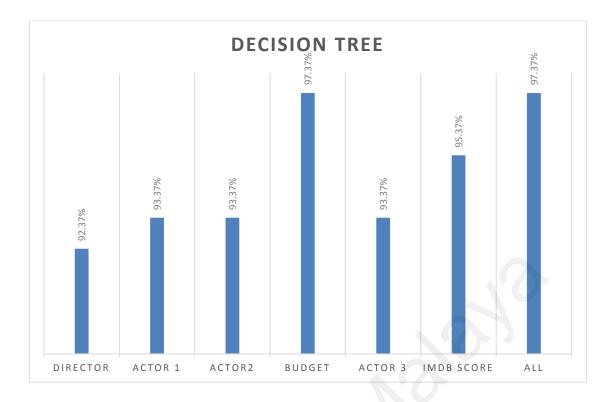


Figure 4. 27: Decision Tree Result based on 200 movies

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 methods and 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.

university

university