National College of Ireland

Predicting The Outcome Of The Horse Race Using Data Mining Technique

MSc Research Project
Data Analytics

Priyanka Selvaraj
x16133218

School of Computing
National College of Ireland

Supervisor: Thibaut Lust

National College of Ireland
Project Submission Sheet – 2017/2018
School of Computing

Student Name:	Priyanka Selvaraj
Student ID:	x16133218
Programme:	Data Analytics
Year:	2017
Module:	MSc Research Project
Lecturer:	Thibaut Lust
Submission Due Date:	11/12/2017
Project Title:	Predicting The Outcome Of The Horse Race Using Data Mining Technique
Word Count:	5374

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature:	
Date:	10th December 2017

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. Please do not bind projects or place in covers unless specifically requested.
3. Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting The Outcome Of The Horse Race Using Data Mining Technique

Priyanka Selvaraj
x16133218
MSc Research Project in Data Analytics

10th December 2017

Abstract

The sole purpose of the research is to ascertain the outcome of the horse race with higher degree of accuracy using advanced data mining techniques. The difference between the existing predictive models and this model will be choosing the relevant attributes by knowing their weights and how much do they contribute in finding their label variable. Hence ensuring that the model takes into consideration only the most relevant attributes and neglects the less relevant attributes thereby reducing the noise in the system. The most efficient data mining techniques are implemented on the most relevant data for the prediction of the outcome of the horse race. Therefore, a sensible model is created, and efficiency of the system is checked by calculating its accuracy. The research is carried across two datasets, one is web scrapped from the live website and the other dataset from the data repository. The difference between their efficiency is showcased to explain the importance of choosing the attributes or features contributing to build an efficient model.

1 Introduction

Sports has been an amazing and interesting field ever since it was known. One of the oldest and worldwide largest spectator sport is horse racing. In 1868, prompted the beginning of organized horse racing in united states. It just started as a normal sport and now became the largest public entertainment business. Horse racing is a sport which involves running of thoroughbred horses and the gamblers bet money on a horse, predicting it to be the winner of the race. Hence, horse racing is an industry that relies on gambling as its main source of income, since the gate receipts make much less income, unlike other sports. so, prediction plays a very important role to help the gamblers to end up profitable. Though the predictions made by horse racing experts seems to work it is not always as efficient as machine learning techniques and manual prediction is always not possible with huge datasets. It is always wise to choose the less time consuming and reliable source of output which is produced by data mining techniques. The main objective of this research is to predict the outcome of a horse race by considering the most relevant features and implementing the most effective data mining techniques to it and thereby building an efficient model. Two datasets were chosen, the first dataset is being web scrapped from a live sporting website sporting life.com and the second dataset was taken from Kaggle. Same techniques are performed on both the dataset and their differences are showcased to highlight the importance of features to be considered or knowledge of attributes needed to be known to evaluate a highly efficient model. This research will focus on educating the future research people to have an in-depth knowledge of the attributes they are considering in order to carry out an effective research.

The paper below summarizes four sections, section 2 explains the related work done in this field . section 3 details the methodology whereas section 4 outlines the implementation. Section 5 details the model performance and evaluation the final section 6 summarizes the conclusion and future work.

2 Related Work

The previous research work done in sports prediction seems to be very vast but still gives place for research. A lot of research has been done in prediction in almost all sports but gambling sports like horse racing and greyhound racing pay their way for more research for more efficient models since it involves a lot of money and time by the investors (Philpott and Teirney; 2004; Burns and Garrick; 2006).

Below are the some of the literature work done on this topic. Each paper and source were considered carefully and their drawbacks, limitations, advantages were analyzed and effectively implemented in my model.

Takahashi (2015) in his paper The effect of age on the racing speed of thoroughbred horses has explained how age influences the performance of thoroughbred racehorses. The purpose of the study was to reveal the average racing speed of the horses and observe changes in their average speed in accordance with their age. A set of racing data was taken, and the racing speed of each horse was calculated. A common characteristic found was average speed increase until the first half of the age 4 and after the latter half of the age 4, the speed remained constant only with little variation. Overall the running performance of the horses reported being a peak at the age of 4 and 5. Only age attribute was considered here to measure the performance of the racehorses which is not an effective way of measuring performance because age is not the only attribute that contributes for the performance hence other attributes should also be given equal importance.

Similarly, Allinson and Merritt (1991) have explained the prediction of horse racing using neural networks which makes use of multilayer perceptron. They have used data from 200 horses from sporting life. In order to increase the chances of success only two-year-old horses were considered which is seeming to be a restricted way of prediction and also if there is data for several hundred horses, jockey, and trainer this method will not be significant since it is designed only for a particular dataset.

Silverman and Suchard (2013) in their paper" Predicting horse race winners through regularized conditional logistic regression with frailty" have added two new contributions to the existing conditional logistic regression which is parameter regularization and frailty parameter in order to increase the efficiency of the system. Historical data from 3681 races were taken and this method was performed which turned out to be better than the existing model, but this model is more into mathematical calculation instead machine learning technique will be more flexible and will produce better results.

Schumaker and Johnson (2008), In An investigation of SVM regression to predict longshot greyhound races paper they have performed SVM regression algorithm on 1953 races across 31 different race tracks and have explored the role of the simple betting engine on a wide range of wager types and they have evaluated three performance measures such as accuracy, pay out and betting efficiency. They got the betting efficiency of about 87.4% for high accuracy and low pay out or vice-versa. This method seems to be quite efficient since the performance measures they have considered, and the system seems to be reliable (Lazar; 2004).

According to Davoodi and Khanteymoori (2010), the author has analyzed the performance of five supervised Neural network algorithms such as backpropagation, backpropagation with momentum, Quasi-newton, levenberg-Marquardt and conjugate gradient descent learning algorithm for the prediction of horse racing. 100 races data were taken and analyzed and the results imply that all the five methods have given an accuracy of about of 77% on average proving that neural networks are appropriate methods for racing prediction. Among these five, the BP algorithm slightly performs better than other algorithms but need a longer training time.

According to Sameerchand Pudaruth and Dookhun (2013), the author has predicted the winning horse in the horse race using statistical methods which seems to give 58% accuracy and also finds 4.7 out of 8 horses will be winners on an average but the professional experts predicted that only 3.6 out of 8 horses will win the race which shows less efficiency than the statistical software. Though the statistical software outperformed the expert prediction, statistical methods seem to be less efficient in comparison with other machine learning techniques so the application of any machine learning algorithm like a decision tree, ANN, SVM can be performed to increase the performance.

Johansson and Sonstrod (2003), has used the artificial neural network to demonstrate its power when applied to hard data mining task such as for the prediction of the winner in greyhound racing, a similar sport like horse racing and again neural networks stand out with the results. Several betting formats such as win, quinella and exacta are evaluated. Neural networks found to produce better results with harder formats even. The proposed method seems to form a base for more refined prediction tool.

According to (Bigus; 1996), again the author uses neural networks, a part of artificial intelligence for the prediction of the outcome of sports like rugby and football. Multilayer perceptrons were used in predicting the result. Using multilayer perceptrons give an additional advantage to neural networks. This seems to be efficient method since the multilayer perceptron adapts very quickly and performs well in prediction. Similarly, this method can be employed to other gambling sports like horse racing.

According to (Williams and Li; 2008) its a case study of performing artificial neural networks on horse race prediction in Jamaica, this again specifies the high performance of neural networks in horse race prediction. This method is being performed from a horse race dataset from races happened in Jamaica.

According to (Gabriel Rushin and Beling; 2017) the author has detected credit card fraud using logistic regression, deep learning, and gradient boosted tree. In this deep learning, algorithm seem to perform better than other two algorithms in finding the credit card fraud, logistic regression is the poorest performer and gradient boosted tree performs better than logistic regression but less than deep learning. All the three methods are evaluated for five different datasets and the results are concluded based on the outcomings. Proper reasoning is not given on the performance of three algorithms so its hard to know the drawbacks.

Geraghty (2014) in his research paper, Analysis of horse jump racing data and using predictive analytics to predict how many horses will fall in a race has predicted how many horses will fall in the horse jump racing data using machine learning techniques.

Bunker and Thabtah (2017) in his paper Machine learning and New Zealand horse racing prediction has explored the importance of machine learning techniques and used neural networks for the prediction of horse racing prediction.

According to Asha Gowda Karegowda and Jayaram (2010), in a Comparative study of attribute selection using gain ratio and correlation-based feature selection have stated the importance of feature selection and stated the importance between correlation-based feature selection and radial basis function. They have explored that the accuracy of the system after feature selection seems to be more hence making the system more efficient than normal.

Information gain ratio as term weight: the case of summarizing of our results, in this paper the author has stated the importance of weight by information gain ratio and examined the results to be more efficient with performing this method. Corporate distress analysis: comparisons using linear discriminant analysis and neural networks, this paper is a comparison between two machine learning methods linear discriminant analysis and neural networks. LDA seems to be an efficient method of neural networks having a good efficiency and it works well with the system where the independent variable is not normally distributed (Martin G.S.; 1996; Oki H.; 1994; Setterbo J.J.; 2009).

Silverman (2012) in his paper A hierarchical Bayesian analysis of horse racing has predicted the running speed of the horse using Bayesian approach. Data of about 36,006 observations from 2973 distinct horse races from Hong Kong race tracks were taken and different Bayesian models were evaluated. Running speed of all the horses were calculated but simply betting on the horse with the highest predicted speed is not enough to predict the winner of the race and this system does not seem to be efficient when compared to machine learning techniques.

Schumaker and Johnson (2017) in his paper Expert prediction, symbolic learning, and neural networks an experiment on greyhound racing has applied various machine learning techniques like a decision tree, game playing scenarios, neural networks for the prediction of greyhound racing. It has considered around 50 performance variables for eight competing dogs in a race. Again, neural network has been an effective method to perform but considering all 50 attributes without any feature selection seem to be a drawback in this system.

All the available papers were critically analyzed and their drawbacks were stated. When it comes to horse racing it can be interpreted as more than other algorithms, ANN algorithm seems to be more efficient and has been showing higher performance in most cases. Other methods have not been effectively done. So, my methodology will interpret a new process which will check for better accuracy and overcome the flaws of the previous works.

3 Methodology

The methodology used in this research is CRISP-DM. Cross-industry standard process for data mining method is commonly known as CRISP-DM. It is a leading data mining process model commonly used by data mining experts to solve many data mining problems. It provides a structured approach for planning a data mining project. This model is an idealized sequence of events and breaks down into six major phases such as

- Business understanding
- Data understanding
- Data preparation
- Modelling
- Evaluation

- Deployment

Each of them is most important and dependent on each other hence differentiating it to be the most efficient process model.

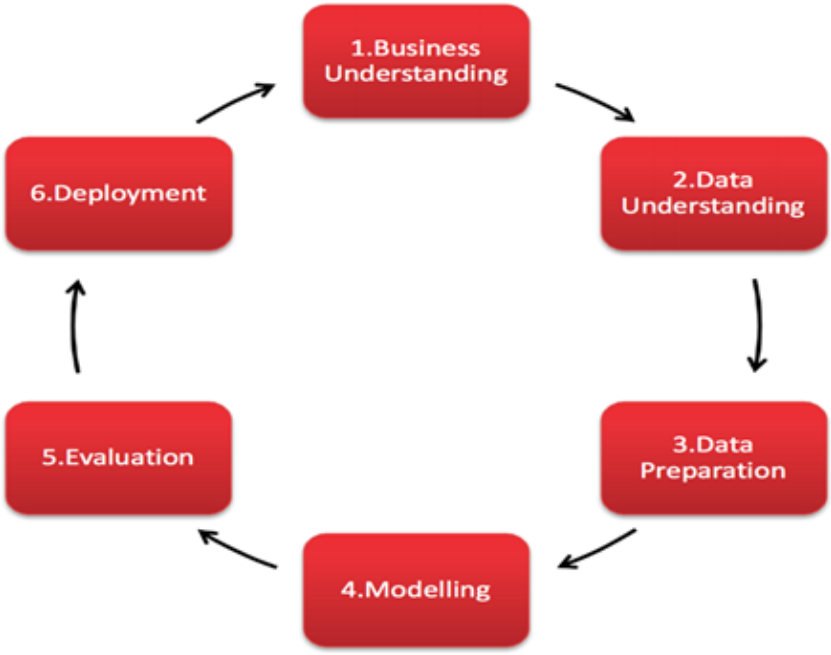

Figure 1: CRISP-DM Methodology

3.1 Business Understanding

This is the initial phase of the process which involves the proper understanding of project objective then converting it into the data mining problem and sketching a plan towards the objectives of the solution. A proper domain knowledge and also awareness about issues with existing model coupled with knowledge of previous research work in this domain is needed to have a complete understanding of the project. This method can be breakdown into two subcategories such as

3.1.1 Setting Objectives

This involves describing your primary objective from a business perspective and can be attaining a certain accuracy in your data mining model. This should be clearly understood to be successful in this project. The major objective of my research is to help the punter to invest on a correct horse which is likely to win the race and reject the horse that is likely to show less performance thereby saving the time and money of the punter.

3.1.2 Produce Project Plan

In this stage, we describe the plan for achieving data mining and business goals. This stage also describes the steps to be followed in the rest of the project including the selection of tools and techniques. My research will focus on implementing advanced data mining techniques such as K Nearest neighbour, Logistic regression and Linear discriminant analysis to determine the output of horse race.

My project would research on how important it is to choose most relevant attributes in the model to build an efficient system and how does it affect the accuracy if the relevant attributes are not available in the model. So, the ultimate motive of the research is to improve the process of decision-making process through extensive data analysis combined with mining techniques.

3.2 Data Understanding

This stage in the CRISP-DM process involves the collection of required data from the sources, understanding the data, ensuring its quality and making it ready for the analysis. This step creates an initial data collection report which summarises the list of data sources being used, the technology used to get the data, data format, size of data and the description of data for better understanding. Verifying the data quality is also a part of data understanding. Verifying the data quality means checking whether the data is complete, is the data correct or it contains errors and creates data quality reports stating the quality issues that need to be sorted in the data. Data understanding is important since a better understanding of data is needed to produce an efficient model. I have used two datasets for my research.

3.2.1 DataSet 1

The first dataset for my research was web scrapped from a live website sportinglife.com. sporting life is a multi-sports website which provides data and race cards for various sports. The required data was scrapped using a web scrapping tool mozenda web scrapper and exported as a excel file for nearly 700 rows. The attributes are Rank, horse name, dam, sire, sex, age, odds, trainer, jockey, owner, no. of matches, type. There are around four numerical attributes and seven-string variables.

ATTRIBUTES	**DESCRIPTION**
RANK	Rank of the horses in the race e.g. 1, 2, 3.
DAM	Dam is the male parent of the horse.
SIRE	Sire is the female parent of the horse.
SEX	Sex is the gender of the horse.
AGE	Age of the horse.
ODD	Odd is the amount of money bet on the horse.
TRAINER	Trainer is the person one who trains the horse.
JOCKEY	Jockey is the person who rides the horse in the race.
OWNER	Owner is the person who owns the horse.
NO. OF MATCHES	This is the count of the matches that the horses have been through.
TYPE	Type is the grounds type where the horses race.

Table 1: Description of Attributes of Dataset 1

3.2.2 Dataset 2

The second dataset for my research was taken from famous data repository Kaggle. Kaggle is a platform that has datasets uploaded by companies and users. It is highly used by data mining experts to perform best predictive models. Data was available in csv format and it was about 10,000 rows. There are around 34 numerical attributes and 3 string attributes in this dataset.

ATTRIBUTES	DESCRIPTION
Race_id	Unique identifier of each race.
Horse_no	The unique number assigned to this horse in each race.
Horse_id	Unique identifier for the horse.
Result	Finishing position of each horse in this race.
Won	Whether the horse won (1) or lost(0). It is the label variable.
Lenghts_behind	From the finishing position how, many lengths does the horse is behind.
Horse_age	Age of the horse at the time of the race.
Horse_country	Country from which the horse belongs to.
Horse_type:	Sex of the horse.
Horse_rating	Rating number assigned by HKJC for the horse.
Horse_gear	It is the string representing the gear carried by the horse in the race
Declared_weight	Declared weight of horse and jockey in lbs.
Actual_weight	Actual weight carried by the horse in the race.
Draw	post position number of the horse in the race.
Position_sec1	Position of the horse at the end of section 1 of the race.
Position_sec2	Position of the horse at the end of section 2 of the race
Position_sec3	Position of the horse at the end of section 3 of the race.
Position_sec4	Position of the horse at the end of section 4 of the race.
Position_sec5	Position of the horse at the end of section 5 of the race.
Position_sec6	Position of the horse at the end of section 6 of the race
Behind_sec1	Lengths behind the leader at the end of section 1 of the race.
Behind_sec2	Lengths behind the leader at the end of section 2 of the race.
Behind_sec3	Lengths behind the leader at the end of section 3 of the race.
Behind_sec4	Lengths behind the leader at the end of section 4 of the race.
Behind_sec5	Lengths behind the leader at the end of section 5 of the race.
Behind_sec6	Lengths behind the leader at the end of section 6 of the race.
Time1	Time taken by the horse to cross the section 1 of the race(sec).
Time2	Time taken by the horse to cross the section 2 of the race(sec).
Time3	Time taken by the horse to cross the section 3 of the race(sec).
Time4	Time taken by the horse to cross the section 4 of the race(sec).
Time5	Time taken by the horse to cross the section 5 of the race(sec).
Time6	Time taken by the horse to cross the section 6 of the race(sec).
Finish_time	Finishing time of the horse in seconds.
Win_odds	Win odds of the race at the start of the race.
Place_odds	Placing odds for the horse at the start of the race.
Trainer_id	Unique identifier for the trainer of the horse at the race.
Jockey_id	Unique identifier for the jockey of the horse at the race.

Table 2: Description of Attributes of Dataset 2

So, the required data is acquired and checked for data integrity. Both the datasets were checked for data quality ensuring a reliable data is used. Now, the data is ready for next step of the process data preparation.

3.3 Data Preparation

In this stage of the project, we decide which part of the data or which attributes to use for analysis. Selecting required attributes for the analysis is a major step of the project. This selection is made based on the relevance of data based on data mining goals, quality of data, technical constraints and business requirements. Major requirement of data preparation is cleaning the data.

3.3.1 Data Cleaning

Data cleaning is the process of raising the data quality to the level that is required to carry out the analysis, which includes replacing missing values, changing the data to required formats and modifying the data according to needs. This can be done manually or using any other data cleaning technology.

- **DATASET 1:** Since the first dataset was web scrapped which was changed from semi-structured

to a structured had lot of inconsistencies like missing value, format mismatch etc. I used rapid miner to clean the data by replacing the missing values with average, changing the formats and ensure the data quality making it ready for analysis.

- **DATASET 2:** The second dataset was available in the structured format and it also contained data inconsistencies. There were many missing values which was replaced by their average value using rapid miner. Data formats were checked and thus made the data ready for analysis.

3.3.2 Constructing Required Data

This means constructing required data from existing data, making modifications or slight changes in the data so that the new data created will work more efficiently than the existing data.

- **DATASET 1:** The attribute rank was modified and categorized into three categories such as winner class, moderate class, loser class making it simple for the user to understand. Then the attribute odds were changed into whole numbers from fractions making it simple to use.
- **DATASET 2:** This dataset dint wanted many modifications since it was already available in a simpler way. The data was checked against consistency and made ready for analysis.

Thus this stage of the process ensures the cleaning of data, constructing required data from the existing data for analysis and ensures data integrity to create an efficient model.

3.4 Data Modelling

This is the most important part of the project since it involves a lot of research and understanding to discover a proper method that takes all the considerations from previous work, all the risks and consequences involved and promising to produce an efficient system. Data modeling is a part of implementation process since it explores the research methodology and their implementation to their datasets. On consideration of various aspects, once the quality data is ready for analysis implementation process takes place.

3.4.1 Modelling Techniques

After a careful review of various processes, I have a chosen the following techniques:

- **Weight by information gain ratio:** Weight by information gain ratio is a feature selection technique used to calculate the weight of the attributes with respect to the label attributes by using the information gain. Information gain is used to decide which are the most relevant attributes to determine the label attribute based on their weight. Higher the weight more the attribute is relevant By this we get to the weights of the attributes that contribute the most to find the label attribute so only most relevant attributes are selected for the analysis.
- **K nearest neighbor:** KNN is one of the simple classification technique which is very easy to implement and uses very less processing time. It is a lazy learning algorithm but efficient when it comes to processing time and easy implementation.
- **Logistic regression:** Logistic regression is a simplified version of the generalized linear model which has binomial label variable and numerical or nominal feature variables. This is an effective regression technique.
- **Linear discriminative analysis:** The linear discriminative analysis is a machine learning technique that is used to find the linear combination of features which best separates two or more classes of example (rapid miner). This has a predictive objective hence used for my methodology.

3.4.2 Tools Used

The tool used Rapid Miner 8.0. Rapid miner is a data science software platform which provides an integrated environment of data preparation, machine learning, deep learning and predictive analytics. It is extensively used for education, research, rapid prototyping, and application development since it supports all the steps of machine learning process right from data preparation to model validation.

According to a research rapid miner provides 99% of an advanced analytical solution through template-based frameworks that guarantee speedy delivery and reduce errors by nearly eliminating the need to write the code. Rapid miner is a flexible and an easy to implement technology since it provides GUI to design and execute analytical workflows. Rapid miner also promises to give better accuracy levels than other technologies I have used. In considering all these merits rapid miner seems to be a promising technology to use.

4 Implementation

Before conducting the experiments it is necessary perform exploratory data analysis. In the beginning of the process, to understand my data in depth, few visualizations were created to check the outliers and their distribution criteria. By doing this, data can be easily understood. If it requires any changes it can be done and according to the dataset, the methodologies can be executed. So, this gives a better understanding of the data which is very important to create an efficient model. The second dataset was checked for its outliers and it seems to have no outliers making the data perfectly distributed along the boundaries.

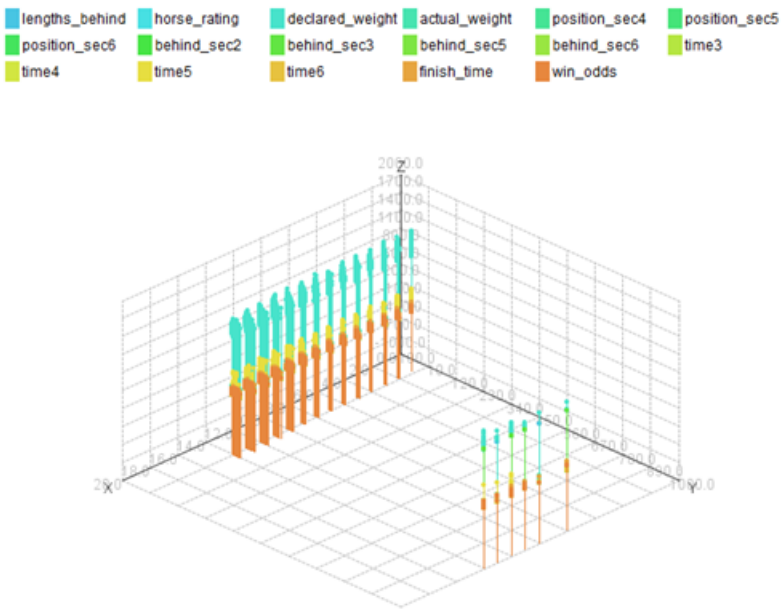

Figure 2: Exploratory Data Analysis

Each attribute values were plotted, and their distribution range was seen. The horse finish time graph is shown below and seems there is a high variation in the values. Similarly, charts for other attributes were created and analyzed.

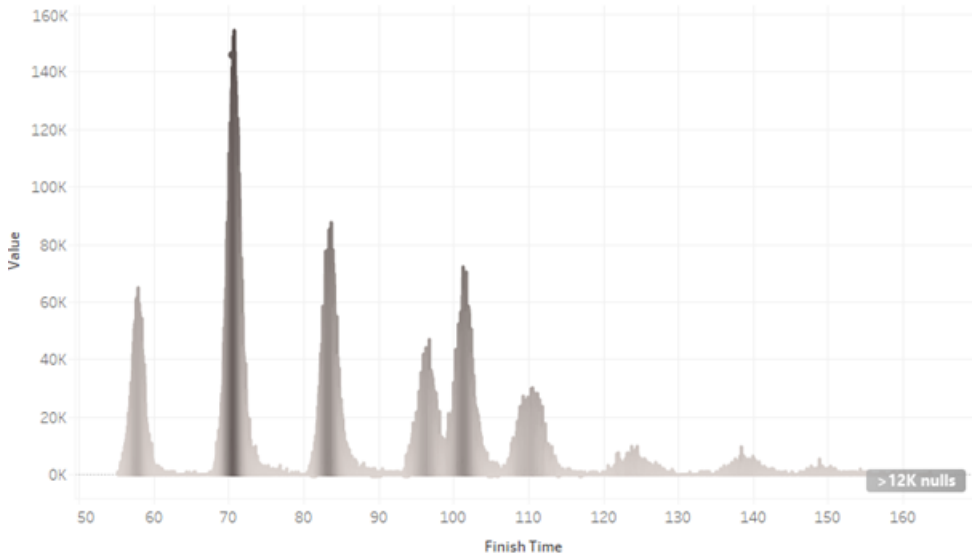

Figure 3: Attribute Analysis

4.1 Dataset 1

The first dataset undergoes a feature selection method known as weight by information gain ratio initially and then used to implement three different data mining techniques.

4.1.1 Weight By Information Gain Ratio:

Weight by information gain ratio is been applied to the first dataset. There is a separate operator available in rapid miner to perform this function. It takes all the attributes in consideration and performs weight to information gain ratio and produce the weights of each attribute. From this, we get to know the most relevant attributes and less relevant attributes from which we choose top-tier attributes which are contributing the most to finding the label variable. The attributes and their weights are shown in a bar graph below. I have considered attributes such as jockey, sire, trainer, odds, number of matches, owner, and dam as top-tier attributes and eliminated age, type, sex since they show very fewer weights meaning they contribute very less to find the label variable. Rank is the label variable here. It has three categories 1,2 and 3.

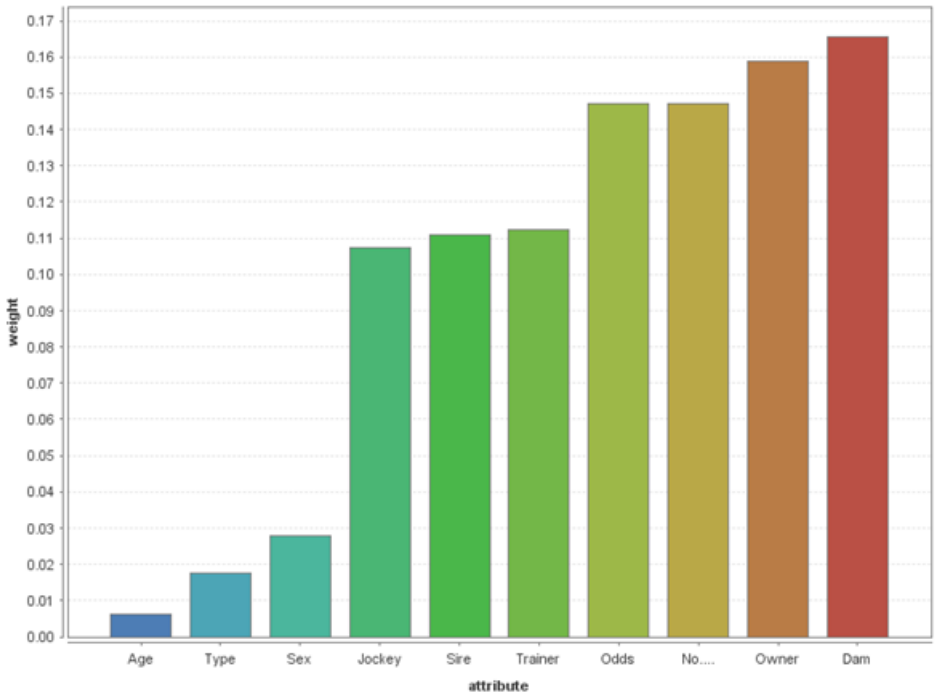

Figure 4: Weight By Information Gain Ratio on Dataset 1

Thus taking the most relevant attributes I perform the efficient modeling techniques which are as follows

4.1.2 K Nearest neighbor algorithm:

Using the top tier attributes, KNN algorithm is performed. KNN is performed using rapid miner. Initially, the KNN workflow has read excel operator to load the data into the model. Then set role operator is used to differentiate the label variable from another variable which is rank here, this helps the model to understand the dataset. Then, split data operator is used where the train and test data is split into 70 and 30 in order to train the model. After splitting the data, 70% of the data gets trained using KNN operator and 30% of data goes to apply model operator. When the 70% of the data is trained then it is applied to apply model operator and the system is tested. Then performance attribute is used to measure the accuracy thereby knowing how efficient the model works. KNN seems to produce 33.97% accuracy seeming to show that the system is not as efficient as we wanted it to be. Hence cross-validation method was used to improve the accuracy of the model but still, the accuracy remains 35.14% which is again not much efficient. The diagram below shows the workflow and confusion matrix of the KNN algorithm.

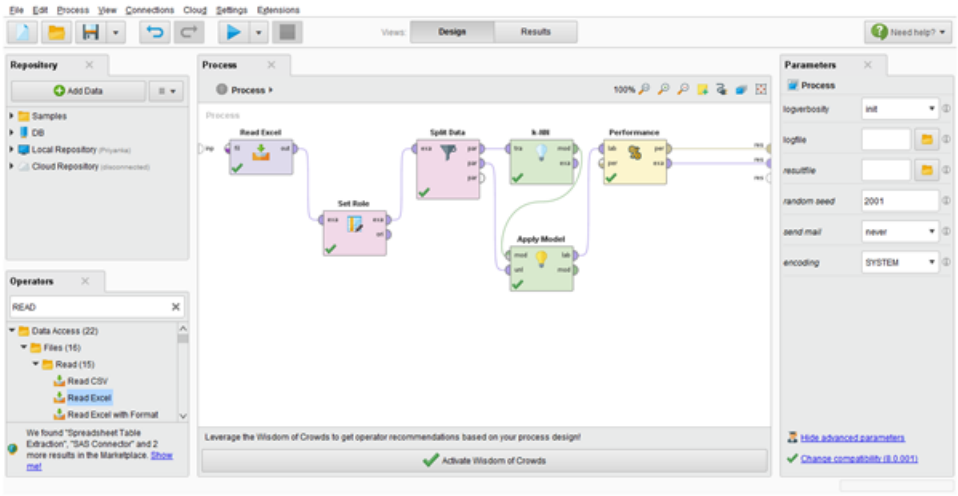

Figure 5: Workflow of KNN on Dataset 1

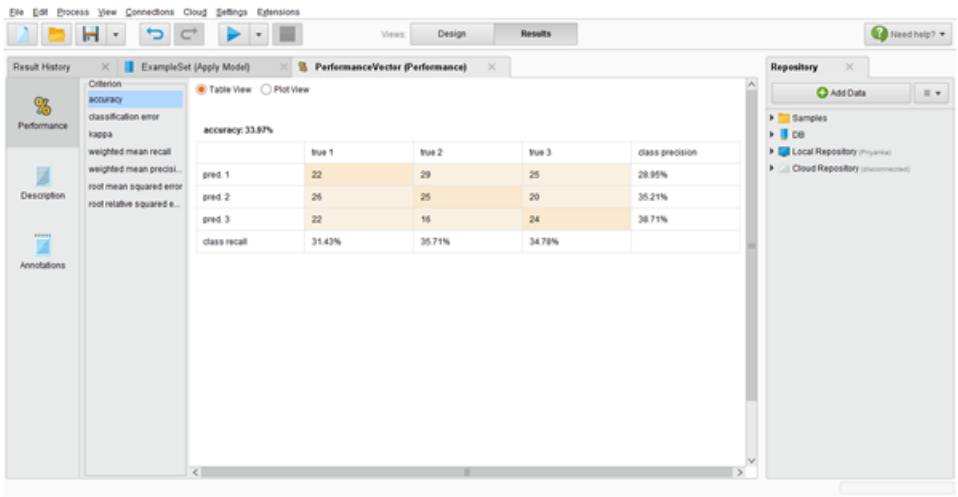

Figure 6: Confusion Matrix of KNN on Dataset 1

4.1.3 Linear Discriminant analysis:

Again, using top tier attributes LDA is performed. LDA workflow is same as KNN workflow but the KNN operator is alone replaced with LDA operator to perform LDA method and also after the data is being fetched into the model replace missing value operator is used since this method strictly not allows missing values to take place. The data is split into 70 and 30 and it is trained , tested and gives an accuracy of about 33.33%. This accuracy is also not much efficient hence cross-validation was performed which again fetched less accuracy. The screen shot below shows the confusion matrix of LDA method.

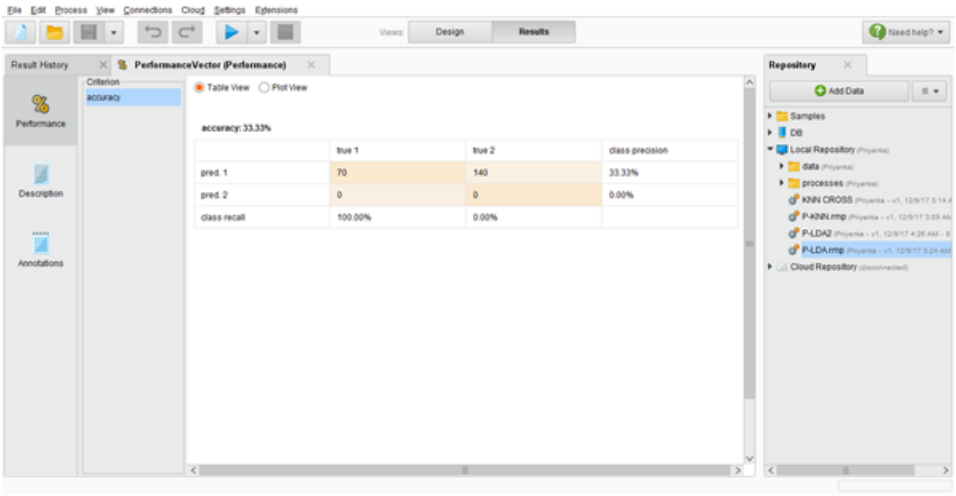

Figure 7: Linear Discriminant analysis on Dataset 1

4.1.4 Logistic Regression

Logistic regression is applied when the label variable is binominal but here the label variable is polynomial so performing logistic regression in this dataset was not possible. So, I tried performing polynomial regression which allows polynomial label variable. Rapid miner dint fetches me confusion matrix instead I got root mean square value which was 0.9 which is really high since RMSE value should be between 0-1 and should be as low as possible for the model should be efficient.

4.2 Dataset 2

The second dataset is structured dataset got from Kaggle. Since the performance of the first dataset was not convincing the second dataset was chosen from Kaggle and data mining techniques were applied to analyze the drawbacks of the methodology and dataset so attributes with more depth information were chosen and methods were implemented, and their performance was analyzed.

4.2.1 Weight By Information Gain Ratio:

The new dataset was first implemented with feature selection method weight by information gain ratio using rapid miner and their weights were noted. All the attributes in the data were considered and corresponding weights were generated by performing weight by information gain ratio. Attributes with higher weights were selected as top-tier attributes since they are the most relevant attributes. Bar graph of the attributes based on their weights are shown below. From this dataset I have chosen result, won, lengths behind, horse rating, declared weight, actual weight, position_sec6, position_sec5, position_sec4, behind_sec3, behind_sec5, behind_sec6, time3, time4, time5, time6, finish_time, win_odds , place odds and eliminated other attributes since they exhibited less weights . Won is the label attribute here which has values 0 and 1. 0 means losing the race and 1 means winning the race.

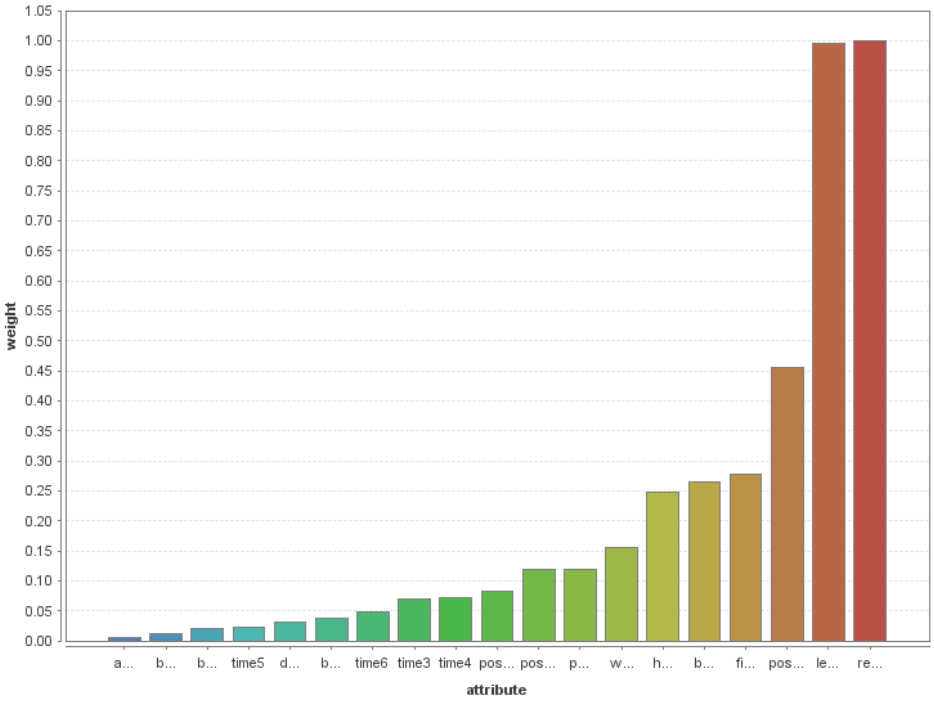

Figure 8: Weight By Information Gain Ratio on Dataset 2

Then the top tier attributes are chosen and data mining techniques are applied to them.

4.2.2 K Nearest neighbor algorithm:

Using the top tier attributes KNN method is been performed using rapid miner. The procedure is same as the procedure followed for the first dataset. The data is read using read csv operator and then sent through set label operator where the label attribute is set which is won here. Then the data is sent to split data attribute where the data is split into train and test as 70 and 30. Then the training data is applied to KNN operator to train the data and the test data is sent to apply model data where the model is trained and tested. Finally, confusion matrix of the model is measured using performance operator. KNN with this dataset seems to be more efficient since it produces an accuracy of 91.40% which is shown below. This creates an efficient model and it is seen that it is purely based on the performance of the attributes.

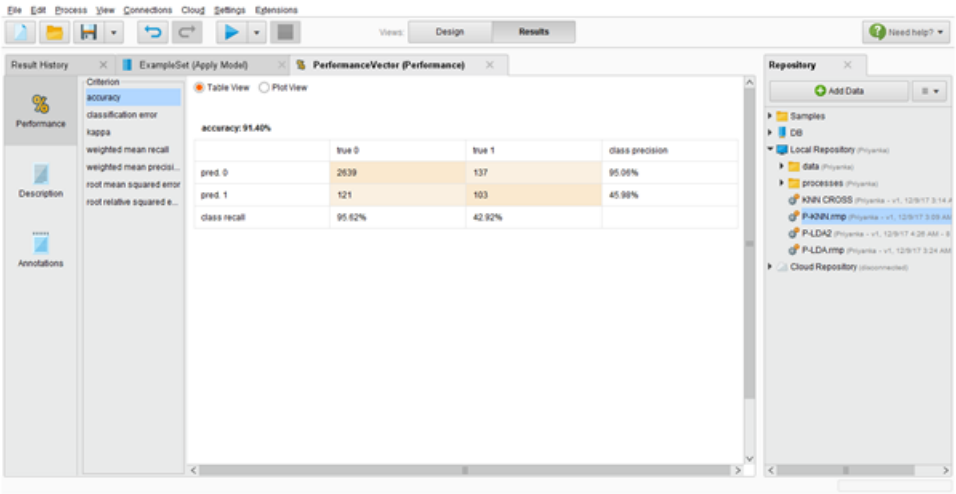

Figure 9: Confusion Matrix of KNN on Dataset 2

4.2.3 Linear Discriminant analysis:

LDA is performed using top tier attributes in the rapid miner. The method follows the same steps as performed for the first dataset. Same as the implementation of KNN process but it uses LDA operator to perform LDA algorithm and an additional operator known as replacing missing values operator is used to replacing the missing values in the data. After performing LDA its performance is measured in terms of accuracy using the confusion matrix which is 92%. The model seems to be efficient which again purely based on attributes performance. A screenshot of the confusion matrix of LDA method is shown below.

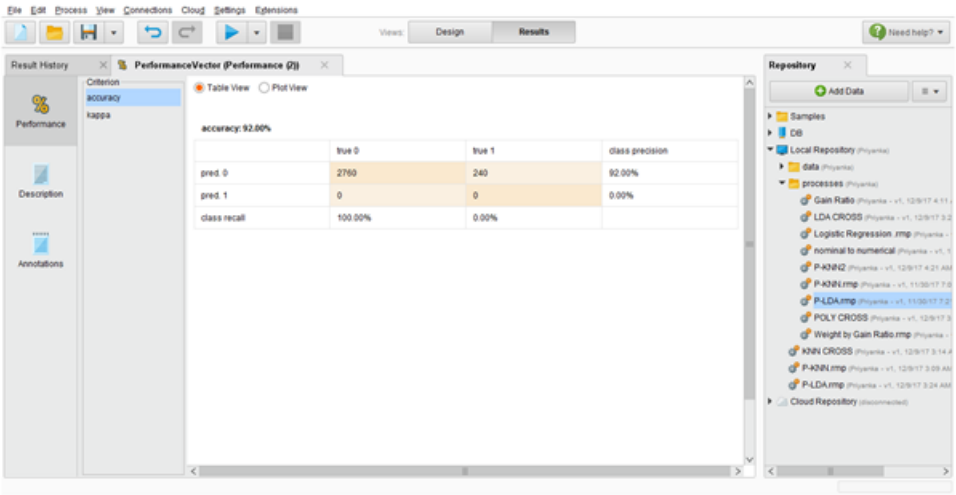

Figure 10: Linear Discriminant analysis on Dataset 2

4.2.4 Logistic Regression

Logistic regression is used here since the label variable is binomial which has only two values 0 and 1. Using top tier attributes logistic regression is performed in the rapid miner. Logistic regression workflow is shown below which is same as KNN workflow but replacing the KNN operator with Logistic regression operator. Split data operator is used, and the data is split into test and train data. The model is trained and tested, and the performance is measured using confusion matrix which seems to have an accuracy of 89.58%. This model also seems to be efficient since the accuracy is a good count. This method has

also performed efficiently purely based on the contribution of the attributes. The design process and confusion matrix of the logistic regression is shown below.

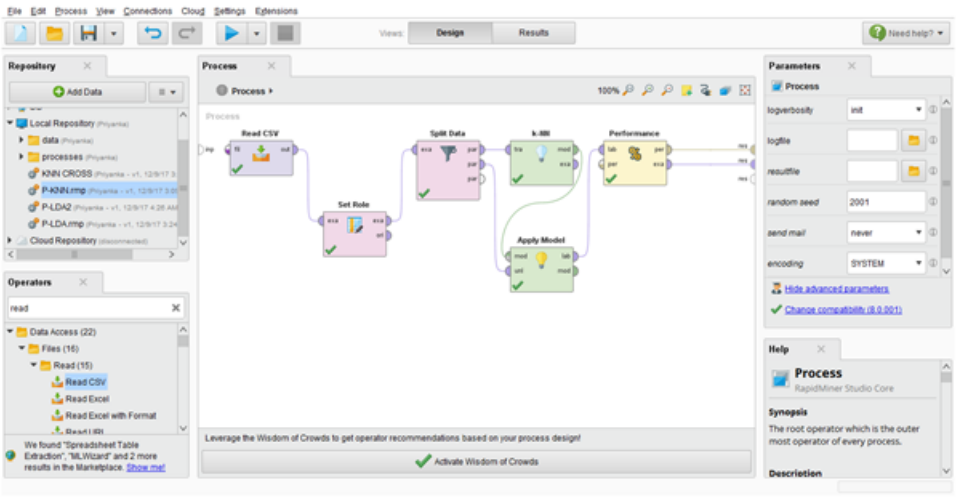

Figure 11: Workflow of Logistic Regression on Dataset 2

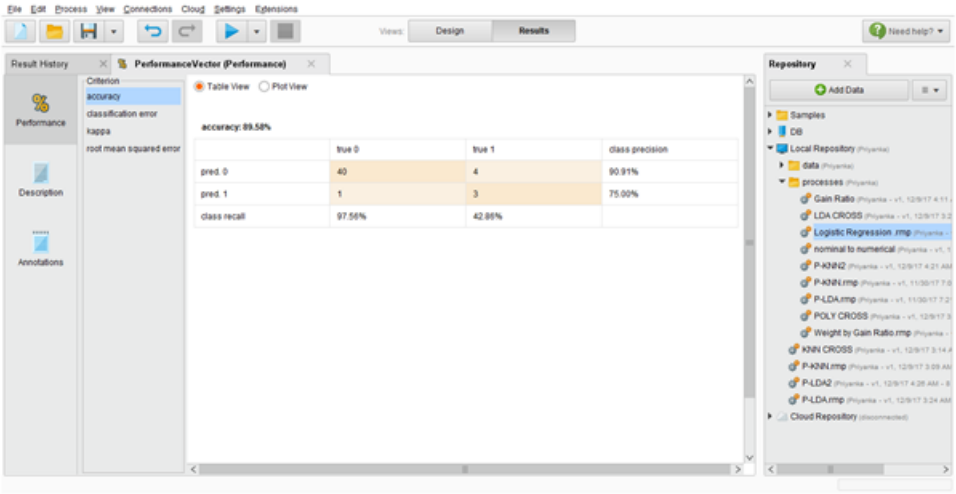

Figure 12: Confusion Matrix of Logistic Regression on Dataset 2

5 Evaluation

An important criterion to be considered while creating an efficient model is choosing the correct parameters or features from the data for the analysis. When this criterion is being satisfied along with the implementation of efficient data mining techniques the model will be efficient else if there is any comprise regarding the feature or attribute selection the model efficiency will be affected. The methods performed on both the datasets were same, but the efficiency of the models differ drastically which can be seen in the accuracy table below

	KNN	LDA	LOGISTIC REGRESSION
Dataset 1	33.97%	33.33%	N/A
Dataset 2	91.40%	92%	89.58%

Table 3: Performance Measure of the Models

The efficiency obtained for each model is purely based on the performance of the attributes. Even after applying cross-validation to the first data its accuracy was not much varied. This is the good example to know that proper selection of features can lead to an efficient model.

The ultimate aim of my research was to showcase the importance of feature selection and to know how accurate the data mining techniques will predict the outcome of the horse race. I used two datasets from which one is web scrapped from a live website with available attributes and the other data from Kaggle with in-depth attributes. Initially the feature selection method was applied to both the datasets to get the most relevant attributes and they were considered to be the top tier attributes.

DATASET 1	**DATASET 2**
Jockey	Result
Sire	Lengths Behind
Trainer	Horse Rating
Odds	Declared Weight
No of Matches	Actual Weight
Owner	Position_sec6
Dam	Position_sec5
	Position_sec4
	Behind_sec3
	Behind_sec5
	Behind_sec6
	Time3
	Time4
	Time5
	Time6
	Finish_Time
	Win_odds
	Place_Odds

Table 4: Performance Measure of the Models

The above are the top tier attributes for each dataset. The first dataset has only basic attributes and it was more about the horse rather than race information whereas the second dataset has more in-depth attributes regarding the race such as odds information the first dataset has only odds value whereas the second dataset has winning odds value and placed odds value which is in more detail. Similarly, even for the other attributes, the first dataset has generalized horse information like sire, dam, owner, trainer but lags race information. Whereas the second dataset has in-depth attributes more about the race like finish time of the horse, time at the end of section 3 of the race, the position of the horse at the end of section 6 of the race. The features selected for the analysis should be in-depth and it should include every single information available in order to make the model more efficient and effective. So, it is clearly understood that in-depth attributes about the horse and race are required to predict the horse racing with higher accuracy. On using the second data set all the models were efficiently done reflecting if required features are applied, all the methods approximately produce 90% of accuracy.

After the successful evaluation of the model, its merits and drawbacks were analyzed and it will be successfully deployed in real time prediction of horse racing in the domain of horse racing or any similar gambling sport.

6 Conclusion and Future Work

The existing work on prediction of horse racing lacks machine learned feature selection, it exists in few cases, but it is manual feature selection. While the results of the predicting algorithms coupled with machine learning feature selection gives more than adequate efficiency than the existing systems. Machine-learned feature selection combined with domain knowledge can predict any system with greater efficiency. The failure to adequately select proper features in the first data set was emphasized and the answer to it was to be found by analyzing the second dataset with appropriate features and applying data mining techniques. The ultimate aim of the research is to check how accurate were the data mining techniques in predicting the horse race. The accuracy of all the models were approximately 90% thus giving an efficient model. Although feature selection plays an important role in this method, it alone

cannot account for a model complete success or failure. Machine learned featured selection combined with data mining technique optimally account for the success of the model. Due to lack of time, research work had been limited and therefore, I would like to extend my research in future. However, a possible future research on this paper will focus on extracting more attributes from bet fair or time form website where it contains live information about the horse and the race and applying feature selection, mining techniques on them and produce a live prediction of horse racing with higher accuracy than the existing systems and my research will be extended to apply all the possible data mining techniques and finding the highest efficiency producing model. My future research will also extend on applying other possible technologies and analyse their efficiencies. The neural network proves to be the highest efficiency producing model in the prediction of horse racing.

References

Allinson, N. and Merritt, D. (1991). Successful prediction of horse racing results using a neural network, *IEEE Colloquium In Neural Networks: Design Techniques and Tools* .

Asha Gowda Karegowda, A. M. and Jayaram, M. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection, *International Journal of Information Technology and Knowledge Management* .

Bigus, J. P. (1996). Data mining with neural networks: solving business problems from application development to decision support, *McGraw-Hill, Inc.* .

Bunker, R. P. and Thabtah, F. (2017). A machine learning framework for sport result prediction, *Applied Computing and Informatics* .
URL: *http://www.sciencedirect.com/science/article/pii/S2210832717301485*

Burns, E., R. E. and Garrick, D. (2006). The effect of simulated censored data on estimates of heritability of longevity in the thoroughbred racing industry., *Genetic Molecular Research* .

Davoodi, E. and Khanteymoori, A. R. (2010). Horse racing prediction using artificial neural networks, *Recent Advances in Neural Networks, Fuzzy Systems Evolutionary Computing* .

Gabriel Rushin, Cody Stancil, M. S. S. A. and Beling, P. (2017). Horse race analysis in credit card fraud-deep learning, logistic regression, and gradient boosted tree, *In Systems and Information Engineering Design Symposium* .

Geraghty, M. (2014). Analysis of horse jump racing data and using predictive analytics to predict how many horses will fall in a race, *PhD thesis, Dublin, National College of Ireland* .

Johansson, U. and Sonstrod, C. (2003). Neural networks for gold at the greyhound racetrack., *Proceedings of the International Joint Conference In Neural Networks* .

Lazar, A. (2004). Income prediction via support vector machine., *International Conference on Machine Learning and Applications* .

Martin G.S., Strand E., K. M. (1996). Use of statistical models to evaluate racing performance in thoroughbred, *J. Am. Vet. Med. Assoc* .

Oki H., Sasaki Y., W. R. (1994). Genetics of racing performance in the japanese thoroughbred horse: Ii. environmental variation of racing time on turf and dirt tracks and the influence of sex, age, and weight carried on racing time., *J. Anim. Breed. Genet* .

Philpott, A., S. H. and Teirney, D. (2004). A simulation model for predicting yacht match race outcomes, *Operations Research* .

Sameerchand Pudaruth, N. M. and Dookhun, Z. B. (2013). Horse racing prediction at the champ de mars using a weighted probabilistic approach, *International Journal of Computer Applications* .

Schumaker, R. P. and Johnson, J. W. (2008). An investigation of svm regression to predict longshot greyhound races, *Communications of the IIMA* .

Schumaker, R. P. and Johnson, J. W. (2017). Expert prediction, symbolic learning, and neural networks: An experiment on greyhound racing, *IEEE* .

Setterbo J.J., Garcia T.C., C. I. R. J. M. J. K. S. H. M. S. S. (2009). Hoof accelerations and ground reaction forces of thoroughbred racehorses measured on dirt, synthetic, and turf track surfaces, *Am. J. Vet. Res* .

Silverman, N. (2012). A hierarchical bayesian analysis of horse racing, *Journal of Prediction Markets* .

Silverman, N. and Suchard, M. A. (2013). Predicting horse race winners through regularized conditional logistic regression with frailty, *Journal of Prediction Markets* .

Takahashi, T. (2015). The effect of age on the racing speed of thorough bred race horses, *Journal of equine science* .

Williams, J. and Li, Y. (2008). A case study using neural networks algorithms: horse racing predictions in jamaica, *In Proceedings of the International Conference on Artificial Intelligence* .