ANALYSIS OF STOCK PRICE PREDICTION USING DATA MINING APPROACH

by

Mukhariz bin Muhamad

12094

Final Dissertation submitted in partial fulfilment of
the requirements for the
Bachelor of Technology (Hons)
(Business Information System)

Supervisor: Dr. Yong Suet Peng

Co-Supervisor: Dr. Lai Fong Woon

May 2012

Universiti Teknologi PETRONAS
Bandar Seri Iskandar
31750 Tronoh
Perak Darul Ridzuan

CERTIFICATION OF APPROVAL

Stock Price Prediction Model using Data Mining Approach

by

Mukhariz Muhamad

A project dissertation submitted to the

Business Information System Programme

Universiti Teknologi PETRONAS

in partial fulfillment of the requirement for the

BACHELOR OF TECHNOLOGY (Hons)

(BUSINESS INFORMATION SYSTEM)

Approved by,

…….

(Dr Vivian Yong Suet Peng)

UNIVERSITI TEKNOLOGI PETRONAS

TRONOH, PERAK

May 2012

CERTIFICATION OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this project, that the original work is my own except as specified in the references and acknowledgements, and that the original work contained herein have not been undertaken or done by unspecified sources or persons.

_____

MUKHARIZ MUHAMAD

ABSTRACT

Financial forecasting is one of the most interesting subjects within the area of machine learning studies. Forecasting stock prices is challenging due to the nature of stock prices that are usually non-linear, complex and noisy. This paper would be discussing the most prominent forecasting method which is the time-series forecasting and its machine learning tools used to create the prediction. The aim of this project is to study the data mining approach on predicting stock price that offers accuracy and sustains its reliability in the system. Using Data Mining approach in training the algorithms that will produce the best results based on Public Listed Companies' stock price data that dates back until 1998. This system utilizes Artificial Neural Network and Support Vector Machine as its main inference engine with numerous methods to measure the accuracy of both. It is anticipated that this analysis would become a platform for producing a prediction application that is reliable for usage in the future.

ACKNOWLEDGEMENT

Greatest appreciation is given towards my supervisors, Dr. Vivian and Dr. Lai for guiding me from scratch to complete this project. Without their guidance and advices, I would not be able to grasp even the fundamental understanding of time-series forecasting alongside with the machine learning areas that are sometimes evaded by people due to its complexity in nature. Their guidance, knowledge, experience, supports and feedbacks gives me the driving force to complete this project.

I would like to express my gratitude towards my family members that have encouraged and supported my work from the beginning until the completion of the project. Family members do give the best comfort during times of pain.

Not to be forgotten friends and colleagues that assist me directly or indirectly, giving advices and sharing thoughts and passion for the project itself.

This project gives me such an amazing experience and I would to thank everyone involved in the remarkable journey.

Table of Content

List of Figures

List of Tables

ABBREVIATIONS AND NOMENCLATURES

PLC	Public Listed Company
ANN	Artificial Neural Network
SVM	Support Vector Machine
DIG	Ultra Oil & Gas Company
AMD	Advanced Micro Devices

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND OF STUDY

There are numerous stock price reports that usually determine the value of the stocks at a period of time done by prominent investment bankers or financial institutions. It either uses only the market forces to evaluate the prediction or it only focuses upon the company's condition, it exists in order to assist the investors to invest in the stock market. However, such reports would not come in handy.

This means that either the report can only be acquired when one is subscribed and paid the services of a stockbroking company or the report is too complex and technical that not many of investors can understand such report without putting a lot of efforts to understand the jargons and charts on the details. Therefore, such complication would not bring a good advantage towards the majority of shareholders especially that not all shareholders have a sound and solid financial background to easily understand such reports.

Furthermore, since stock price fluctuations mainly depend on the investors' confidence, the investors must rely on certain concrete information to get ahead of the crowd to gain more profits out of stock market. If the information gained is hard to be digested or too confusing, one can expect that it would further deter the crowd to join the stock market due to lack of ability to understand such information.

Besides that, there are numerous ways to predict stock prices but in the current terms, the most prominent method of doing it is mainly based on human-based methods that are mainly attributed towards experience, knowledge of the subject matter and expertise in the field of stock prices. However, there is no prominent system to effectively forecast stock price especially in a Malaysian context.

Therefore, this project would capitalize on the complexity of stock valuation reports and to simplify it. In order to do so, a sound and reliable system must be created in order to assist shareholders and also the public to invest in stock markets.

1.2 PROBLEM STATEMENT

1.2.1 PROBLEM IDENTIFICATION

There are numerous ways to predict stock prices being used by the investment managers, stockbroking firms and even banks. Even though the software or system can be bought, it is too expensive for an individual to acquire it and thus, making prediction system exclusive towards financial firms and financial experts only.

The general public, however, holds the majority numbers of investors in the stock market. The general public usually does not have a sound and firm financial background to effectively use the current software as the reports produced from it is very complex with a lot of financial jargons that only financial experts will find it easy to understand.

Hence, the problems from this situation are:

- Average people without a financial background cannot fully utilize the forecasting system which are mostly very complex and too technical
- Deters the usage for the general public to conduct financial forecasting

In essence, the general public must at least have a system that caters towards their need which is to gain as much information as possible before pledging their funds into the stock market. With the majority number of investors not knowing how to analyze the reports produced, it would further highlight their need to have a simple and yet efficient predicting system of the stock market.

1.2.2 SIGNIFICANCE OF THE PROJECT

This project would benefit the investors on a two-way basis. The first is that the system user can effectively predict the stock prices in the market and decide on which stocks to invest on based on the prediction and looking at the trend of the stocks. Such features would ease process of selecting an investment and would save cost as to not to hire external parties to help predict the stock prices for the investors.

Secondly, the system user can also detect anomalies in the stock prices, such as an unprecedented increase in the real stock price, which would also be the first

sign of a financial fraud. Hence, financial regulators can benefit from this system by detecting the early signs of fraudulent activities in the stock market and to make an efficient investigation into such cases.

It would also benefits the investors by giving signs of such dangerous PLCs and the investors can safeguard their investments by not investing the company or withdrawing their investments in said PLC. This is important to protect investors from greedy PLCs that would only seek to maximize their profits while jeopardizing the money from investors.

In all, such system would not only assist investors but also towards the regulators monitoring the market. Although the predictions might not be 100% correct, the system serves as an initial indicator towards the tendency that the stock price might rose or fall.

1.3 OBJECTIVES AND DELIVERABLES

Every new inventions and innovations are developed for a purpose. The main objective of this research is:

- To study the data mining approach on stock price prediction
- To create an analysis tools that is able to predict the stock price movements in short term trends

In essence, this project is to help investors to buy and sell stocks by providing them more options to gain information rather than the complex and technical reports.

This project deliverables would be:

- Data mining tools that is accurate and reliable for the usage of investors
- A system that would be able to generate a concise report on the predicted movement of stock price a selected business sector

1.3.1 SCOPE OF STUDY

This project aims at predicting public listed companies' stock price. This means that the scope of study revolves around PLC that already has several years of historical

data. The project would try to find several PLCs' data available in the internet and uses it to feed the data mining machine to give the desired results.

This project focuses on Malaysian stock market because the nature and systems used in each market is different from one region to another such as the financial terms, the Bursa Malaysia rules and regulations and the indexes used by the Bursa Malaysia is different with other stock markets.

The existing idea was to get the data from the Bursa Malaysia. However, due to certain difficulties, the data cannot be obtained without incurring a huge amount of cost within the region of thousands of Ringgit Malaysia. Therefore, the next alternative would be finding free source data from the internet.

The free source data must comprises of three years back-dated data for the purpose of the research. Since there are a lot of websites that provide the free data, the project would be scoped down into finding these websites and extract the given data for the system.

The three years data is to be fed into the inference engine to teach the machine to learn and recognize the patterns of the price movements and to predict the movements in the future. Besides than learning, the data would also be used as a test and validation data in order to ensure the accuracy and reliability of the predicted stock prices.

The nature of the stock data that would be used within this project would be the price fluctuations rather than the volume and other quotations that would usually associated with a stock data. This is because the stock price movement is already adequate to find the historical patterns and to predict the future movement of the stock prices.

The scope of this study would further assist in creating the accuracy and reliability that the system proposes. The training model, validation model and testing model each having one year of historical stock data to train, validate and test the patterns and predictions would be integral in order to train the system to be as accurate as possible.

1.4 RELEVANCY OF THE PROJECT

This system is relevant to the investors in the stock market whereby it will be useful to have an analysis tools that can guarantee accuracy and reliability of the predictions. Investors would want to have a secured and profitable return in which this system would provide such benefits to the investors.

Besides that, it would also serve the financial regulators a chance to get the first pointer of a financial fraud taking place in the market. Such advantages in the system would elevate the safety of investing in the stock market and would deter PLCs to commit financial crimes that would jeopardize the investors' money. It would mean that the Bursa Malaysia's activity can be spurred and become livelier due to the existence of this system to the most extensive impact possible.

The results would also indicate the efficiency of the market itself. The more the stock price can be predicted, the more the market is not efficient which means that the market can be dictated by several other factors rather than just the company's performance and economic outlook.

1.5 FEASIBILITY OF THE PROJECT WITHIN THE SCOPE AND TIME FRAME

1.5.1 TECHNICAL FEASIBILITY

In terms of technical feasibility, the software is feasible technically, although there are some risks taken into consideration. It concerns whether or not the system can be developed.

Technology Area (Medium Risk)

Generally, investors would have come in a lot of background and not necessarily those who are well versed in the financial areas only. Hence, it would be challenging to find the common qualities that would suit all people of various backgrounds. However, most investors do have a common knowledge in handling computers and are tech-savvy in which would help the developers to create a friendly user interface according to their preferences. The risk involved towards the users is medium

As for the developer point of view, the developer has some experiences in some programming languages such as Java. This knowledge will be very helpful in creating the system to ensure the system can be built. A part from that, open source development tools and software are available over the internet which will be used to develop this system such as the WEKA data mining software. Therefore, it can be concluded that the risk is medium for the system developer to transform the ideas into a working solution.

Familiarity of the Functional Area (Low to Medium Risk)

The system developer needs to explore on how to come up with the suggested and proposed system. Plus, the developer must have a comprehensive understanding of the financial indicators, the nature of the stock markets, the expected output that the stockholders would want to have and the primary advantage that would benefit the whole financial association in this nation. At the same time, a better understanding on how current systems work in which the methods currently in place albeit for different nation and region. However, with a background in financial management, the developer would already have acquired basic understanding on financial markets and its indicators. Hence, the risk is low to medium in this context area.

Project Size (Medium Risk)

Based on the suggested and proposed features that are made available in the system development, the project size of the proposed system is in medium scale. Due to the development of this system only focuses on the forecasting of stock price which relatively reduces the risk. Furthermore, the user involvement is required to be able to come up with this system.

In terms of time frame, a complete and thorough study of the subject matter especially in the algorithm approach requires a high time commitment. Since the research period is very short, it is limiting the extensive research outcomes and transforming ideas and solutions into a working system will be quite challenging. With all the constraints that may be encountered throughout the development phase, the risk on the project size is medium.

High compatibility of the proposed system with existing technology

The compatibility of the system with the technology is great. The most common software of data mining being used is the WEKA data mining system which is compatible with certain programming language such as R. The user-interface integrated with WEKA would be created using R programming platform with the 'rWeka' library package to assist the usage of algorithms. Plus, R basically uses the same logical functions like Java and even C++ that would not give a lot of problem to the developer.

1.5.2 OPERATIONAL FEASIBILITY

Operational feasibility refers to the acceptance of users on the system development, how they feel about the solution provided, and it is a measurement whether a system can work and will work to solve the problem addressed (Castro & Mylopoulos, 2002).

The proposed system helps in providing a platform to predict the trend of the stock prices for the usage of investors and future investors. The system also suits the tolerance of risk and the need of investors to find the desired trend of stocks that the investors would want to spend in. this would immediately and significantly help investors to get the desired return on investment.

In order to cater the different background and levels of technological acceptance by the users, the system will issue internal training and user guide manual to the users. It also would provide internal troubleshooting advices towards the users of the system. This would greatly assist the investors in order to efficiently use the system.

It is believed that such suggestions being proposed into the system would effectively addresses the problem that investors currently have and to ease the work of financial regulators in the nation. Hence, it is feasible in terms of operation to create the system.

CHAPTER 2: LITERATURE REVIEW

2.1 STOCK PRICE PREDICTING MODELS

There are several methods currently being used by financial experts to analyze and evaluate stocks. Such methods would eventually lead towards predicting the future value of the stocks either in short or long term period of time. According to Gitman, Joehnk, & Smart (2011), there are two main stock evaluation methods which are fundamental analysis and technical analysis. Both have different attributes and focuses on two distinct aspects of the stock and the company.

Fundamental analysis focuses at the financial condition and operating results of a specific company. It means that the analysis is on the actual financial performance of the company and the relation of that with the current stock price. It would be the indicator whether a stock price is undervalued or overvalued in which the market would soon adjust itself towards the current fundamental valuation of the company. In essence, fundamental analysis assist investors to formulate expectations on the future performance of the company and its stocks (Gitman, Joehnk, & Smart, 2011).

Therefore, we can conclude that fundamental analysis is the evaluation of the company's intrinsic value in which the market would adjust the stock prices according to the intrinsic value. Thus, giving a solid ground to predict the stock's future prices. The situation of fundamental analysis is as follows: if the stock price is higher than the company's intrinsic value (overvalued), then the market would eventually strive to lower down the stock price to create an equilibrium of the stock price and its intrinsic value and vice versa. With fundamental analysis, stockbrokers can predict the future price according to its intrinsic value.

Technical analysis on the other hand is to analyze the market behavior rather than the financial condition of the company. It involves studying the various market statistics such as volume of trading, amount of short selling or buy/sell patterns of the stocks. It would try to evaluate the position of its current price to predict the future movement of the stock price. Based on historical data, it would try to recognize the pattern and trend as a basis to predict the future stock price (Gitman, Joehnk, & Smart, 2011).

Base on the excerpt, it can be concluded that the trend of stock price is that it would climb until it reaches it peaks and begin to fall. It would then fall until it reaches the trough where it would begin climbing up again. Hence, the technical analysis would try to evaluate the current position of the stock price. In order to do so, it must use the historical data to plot its current trend and base on that evaluation, try to predict its future movement. Technical analysis is a feasible yet effective way to predict the future stock price movement.

Within the scope, feasibility, size and aim of this project, technical analysis is preferred than fundamental analysis. This is because the fundamental analysis requires a lot more information and processing which might not be feasible for the size and scope of the project initiated such as financial statements, past performance of the company and the pro-forma cash flow which must be analyzed before the intrinsic value of the company can be evaluated. Therefore, technical analysis suites this project better than fundamental analysis.

2.1.1 TECHNICAL ANALYSIS

It is mentioned that technical analysis would be the basis of the project. Technical analysis, as described by Credit Suisse (2010), is the study of market action through examining the price changes that occur on a daily basis or weekly basis or over any other constant time period displayed ingraphic form, called charts. Hence the name chart analysis. Technical analyst believes that the relevant information of the market has been reflected in the price of the stocks except for natural disasters or other news within the same wavelength in which this reflection occurs very quickly.

This means that the price of a particular stock would eventually adjust itself towards any news announced by the company or the current economical condition in such swift transition. This would mean that the current market is highly responsive towards such news that the price of a stock alone can be a mirror image of the condition of the company or the economy of a nation. Therefore, watching the price movements alone is already adequate and precise enough. The price, on the other hand, is being reflected in charts and graphs.

The financial market alone can give trends, momentum and patterns that tend to repeat over a period of time albeit not the same way all the time. It would also reveals the mood of the investors and not the fundamental factors of the company (Credit Suisse, 2010).

As being explained above, the technical analysis is the best tool to find patterns and trends based on historical data of the stocks. Therefore, since patterns and trends can also be deduced using machine learning tools, it would mean that a combination of both technical analysis and machine learning would be sufficient and efficient method to predict stock prices.

Technical analysis is more superior than fundamental analysis, in the context of finding trends and patterns, that would suite the needs and capabilities of common machine learning tools such as the artificial neural networks, support vector machines and also fuzzy grey prediction.

2.2 TIME-SERIES FORECASTING: DEFINITIONS AND CONCEPTS

Since it has already being discussed that technical analysis would be used within the scope of the project, therefore the next best thing to be discussed is about its implementation of it within the machine learning scope. In this case, the time-series forecasting is the machine equivalent of a technical analysis and would be further explained in this section.

Time series approach means that historical events do have a pattern that would be applicable to predict the future events. Causal approach would usually find a reason behind the historical patterns and uses it to generate the forecast. Judgmental approach is the gathering the knowledge and opinions from experts to predict the future events and lastly, experimental approach would mean experiments are conducted to become the data to rely upon for forecasting (IPredict, 2012).

However, time-series forecasting is the most widely used due to its model that only take into consideration that historical events are bound to happen again given the same factors of the events, (IPredict, 2012). Hence, it is considered to be the most popular method of forecasting within the financial world especially whereby the

events of financial world do have the tendency to repeat it although its degree of effects is unknown.

Time-series is a sequence of data points taken at uniform interval point of time such as the stock market index. Time-series are very frequently plotted via line charts, (wikipedia.org, 2011). A time-series chart would usually shows some patterns that depict the movement of data points along the timeline. The time-series chart is as shown in Figure 1.

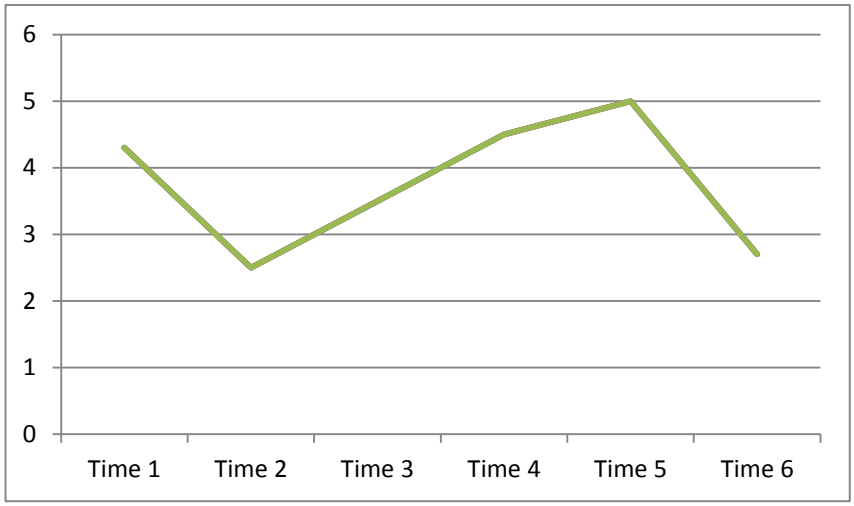

Figure 1: Time-series Chart

According to Hadavandi, Shavandi, & Ghanbari (2010), time-series forecasting basically uses the past data in order to predict the future performance according to the relevant forces that determine the past data. In the other hand, time-series forecasting is a process that has uses past data and the factors that influence the past performance and relate such information to predict the future performance.

The concept behind time-series forecasting depicts that it would create a pattern from the past data in order to plot the future data according to the similar patterns from the past. It is simply based on facts and figures which create a precedence that such events would occur again in the future.

2.3 PREDICTION METHODS USING DATA MINING IN FINANCE

In the financial world, the ability to make the right decision at the right time is vital towards the stability and the ability of one party to generate much wealth base on the current financial market status. To make the right decision at the right time requires

information and data that would eventually assist the decision-maker to undertake such informed choice.

The same goes towards the decision in buying and selling stock prices. Since stock prices are very hard to determine and involves a lot of factors, in which mentioned by Esfahanipour & Aghamiri (2010) that stock price movements are an active complicated domain knowledge and also highly non-linear in which making it very difficult to press a button on which factors that truly determine the stock prices. This would mean that predicting the right stock to buy at the right time is very difficult feature.

However, financial time-series forecasting is not an easy feat as being mentioned by numerous research papers (Hadavandi, Shavandi, & Ghanbari, 2010), (Tay & Cao, 2001), (Kim, 2003) & (Chang & Liu, 2008). This is because time-series data for stock prices are usually noisy and non-stationary as mentioned by Tay and Cao, 2001. The noise comes from the incomplete past financial data to relate the past price with the future price in order to come up with a dependable prediction and non-stationary in terms of the time-series data that keeps on changing through time.

This would mean that the exact data must be extracted and used in order to clear the noise as much as possible. The exact data must be perfect in order to still allow forecasting to take place while not jeopardizing the accuracy of the forecast. Chang & Liu (2008) give the idea there are two common analytical approaches towards predicting stock prices in the terms of finance. The first is the fundamental analysis and the second is technical analysis. Combination on these two analyses, the prediction of stock prices is possible and much more accurate.

Based on this idea, it can be derived that fundamental analysis is significant in long-term prediction while technical analysis for short-term prediction. As being mentioned above, technical analysis would have more emphasize on short-term stock predicting. Hence, it would further support the notion that technical analysis would be preferred in this project paper.

Furthermore, to integrate both technical analysis with data mining approach, a few machine learning tools has been used in such field such as artificial neural networks,

support vector machines and fuzzy grey predictions (Kim & Han, 2000), (Roh, 2007), (Chang, Wang, & Zhou, 2012) & (Kim, 2003). This project would try to use both methods in order to accurately predict stock price movements over short-term period of time.

2.3.1 ARTIFICIAL NEURAL NETWORKS (ANN)

Artificial neural networks (ANN) are programs that are based on the geometry of the human brain. It is useful to detect complex patterns in which it would detect the patterns by receiving large number of inputs and expected outputs. ANN has the ability to learn and recognize new patterns and use it to predict the output (Landt, 1997)

ANN can also be used for classification purpose, noise reduction and prediction of patterned outputs. It does so by recognizing patterns and extrapolates on historical data to predict the desired output. The most significant ability of ANN is that it has the capability to learn those patterns which is gathered from sample inputs being fed into the machine (Cheung & Canons, 2002).

This means that ANN is one of the tools to predict the future stock price using historical data of the stock price alone which is the essence of technical analysis. It can show the movement of future trend of the stock price only and it is based on the historical data being provided into it. In other words, it can show the movement of the stock prices and not to the extent of the percentage of movement and the length of time the movement would stay.

ANN do have a lot of limitations such as it is unreliable and unpredictable depending on the input data being fed to it in which stock market data is often incomplete and complex (Kim, 2003), the developer believes that it is still an integral part of time-series forecasting especially for stock market price. This is because an artificial neural network also becomes a platform for other tools used in the system such as the support vector machines.

ANN in essence would be complementing the technical analysis as the historical stock price data is already sufficient enough to find the future movement trend of the stock price itself. Its limitations, in this project's output, is that it can only predict the movement of stock prices and not to have the percentage of movement and the length

of time of that particular movement. It will be complemented with other data mining techniques that will be discussed later in this project paper.

2.3.2 SUPPORT VECTOR MACHINE (SVM)

According to Tay & Cao (2001), support vector machine is part of neural network algorithm which implements the empirical risk minimization principle which seeks to minimize the factors of the generalization error rather than minimize the training error. In other words, dependency of generalization error on the training error has been reduced which eventually results in better generalization performance.

The basic idea of SVM is to find a hyperplane which separates the data into two classes regardless whether or not the data is linearly separable. This would lead to the solution of some regression tasks where the system can be trained to output a numerical value rather than only "yes/no" classification. In other words, SVM can produce more than one dimension of output (Boswell, 2002).

SVM also have been widely used in financial areas for prediction in numerous investment properties and are a good alternative for financial time-series forecasting as compared to back-propagation neural networks and case-based reasoning. SVM is known as the algorithm that finds a special kind of linear model, the maximum margin hyperplane which gives the maximum separation between the decision classes SVM implements the structural risk minimization principle and this leads to better generalization than conventional techniques (Kim, 2003).

As a conclusion, SVM with its capabilities would sufficiently overcome the problems in the artificial neural networks and by using it side-by-side, it would generally make the prediction system better. This means that with SVM, the output of the project would include the percentage of change as well. Combining SVM with ANN would create a good multi-dimensional output in terms of having to predict the movement of stock prices and also its proportion of movement. This would be further enhanced and researched as the project is in its development phase.

2.3.3 FUZZY GREY PREDICTION

According to Hellman, fuzzy system gives another dimension towards approaching classification problem. It would focus on solving the problem rather than trying to model the system. It is not rigid as a mathematical equation as it also takes into

consideration the “grey” area which makes fuzzy logic to require sufficient expert knowledge for the formulation of the rule base, the combination of the sets and the defuzzification. The employment of fuzzy logic can be helpful for complex processes such as nonlinear processes in which this case is about stock price prediction.

Fuzzy grey prediction is really helpful to nonlinear processes in order to classify such processes. Within the terms of this project, the nonlinear processes would be the stock price data, the classification would be the prediction of the patterns that is gained from the historical data. Hence, fuzzification techniques are really helpful in terms of predicting stock prices.

Fuzzy grey prediction is widely used for prediction methods to predict stock price at any given hour. Since the prices of stock are distinctive for each hour, the capabilities of fuzzy grey technique to undergo such prediction is a paramount success. Fuzzification techniques are employed to predict the stock prices promptly and effectively (Wang, 2002).

The excerpt from Wang (2002) has confirmed that fuzzy grey prediction are useful in terms of predicting stock prices. Another aspect that can be taken from that journal is that fuzzy grey prediction can be used to predict stock prices with relative of a time frame. The time frame output would also be useful in terms of predicting the movement of the stocks for a period of time that can be quantified using fuzzy grey prediction. Therefore, fuzzy grey prediction is a useful machine learning tools to incorporate it with SVM and ANN in order to achieve the three levels of desired output of this project.

2.4 IMPROVEMENT AND IDEAS FROM EXISTING SYSTEMS

Saad, Prokhorov, & Wunsch. II, (1998) has used data mining techniques to predict short term stock trends using historical daily closing prices with time delay, recurrence or probabilistic neural networks. Within that published paper, it can be confirmed that such short term prediction would makes it more accurate and feasible to implement in a particular system. short term prediction would mean using short-term data to produce a short term output would be sufficient.

This is inline with project’s goal which is to make a short-term prediction of stock prices. Therefore, short-term is within the range of 22 working days towards 100

days. Within the development of the project, the time period would be further confirmed with more accuracy training being implemented on the system itself.

Another approach is by Roh (2007), whereby that published article has confirmed that a hybrid model can be used. Within the experiment, Roh has used a combination of ANN and also financial time series model which is the econometrics and also the market and financial theories. Since financial time series model is not reflective of the market variables, ANN has been implemented towards the financial time series model. The result is that the hybrid model, can enhance the predictive power of the deviation and direction accuracy.

This means that a hybrid prototype between financial forecasting models and also machine learning tools are already there. The idea now is that to improve on the system, which is to integrate the more machine learning tools within the financial model to further increase the accuracy, reliability and its effectiveness. Even if the project would fails, it would also confirms that the machine learning tools cannot be combined together and the project would try to satisfy the objective of this study which is to find the most suitable tools to predict stock prices.

The ideas garnered from these project papers are abundant. It gives a view that machine learning tools are effective in terms of predicting stock prices on different levels and it also indicates that a hybrid model would also enhance the current predictive powers of the financial models. Therefore, this project would try to enhance these ideas by implementing the ANN, SVM and fuzzy grey machine with technical analysis that would gives a better prediction tools for stock prices.

CHAPTER 3: METHODOLOGY

3.1 RESEARCH METHODOLOGY

Research is conducted to gather more information about a specific topic. It is a process of finding facts and data that would be useful for the project besides than improving the current system that has been implemented by other researchers. This project research area would be gathering information on historical stock price data and basic concepts regarding time-series forecasting and its tools and also the current financial forecasting model.

Basic concepts for time-series forecasting and financial forecasting model are acquired through discussion and brainstorming sessions between the experts on the subject matter. The idea extracted would then be materialized within this project paper. The concepts and theories would then be backed up by numerous other reading materials that are available for the research area.

Other than brainstorming and discussion sessions, reading materials are also useful in giving a firm understanding on the concepts and theories behind time-series forecasting, its tools and also the combining factors between time-series forecasting and financial forecasting methods. Since the area of research is thoroughly researched by a lot of experts, there are a lot of reading materials to be referred to for this project paper.

Since most of the data used for the development of the system are secondary data, which is existing information that is readily available for usage, the data gathering would mainly consist of gathering information from reliable resources such as Bursa Malaysia. The data can be requested by the party involved or generated manually.

In essence, the literature review and also the findings from the brainstorming session would be extensive and would hugely aid in combining both machine learning tools and the financial forecasting model of the system. The combination of both aspects of forecasting would be done via developing and testing the system to discover the best combination of tools and models that produces the best results.

The methodology used in developing this system is Rapid Application Development (RAD). RAD means that the system would have prototypes being developed iteratively to add functions and features progressively until the system is fully operational, after numerous testings along the way. Therefore, this methodology is parallel and suitable for the system development where the final output will be a working prototype and will be delivered within a shorter time scale.

Figure 2 depicts the general picture of the RAD.

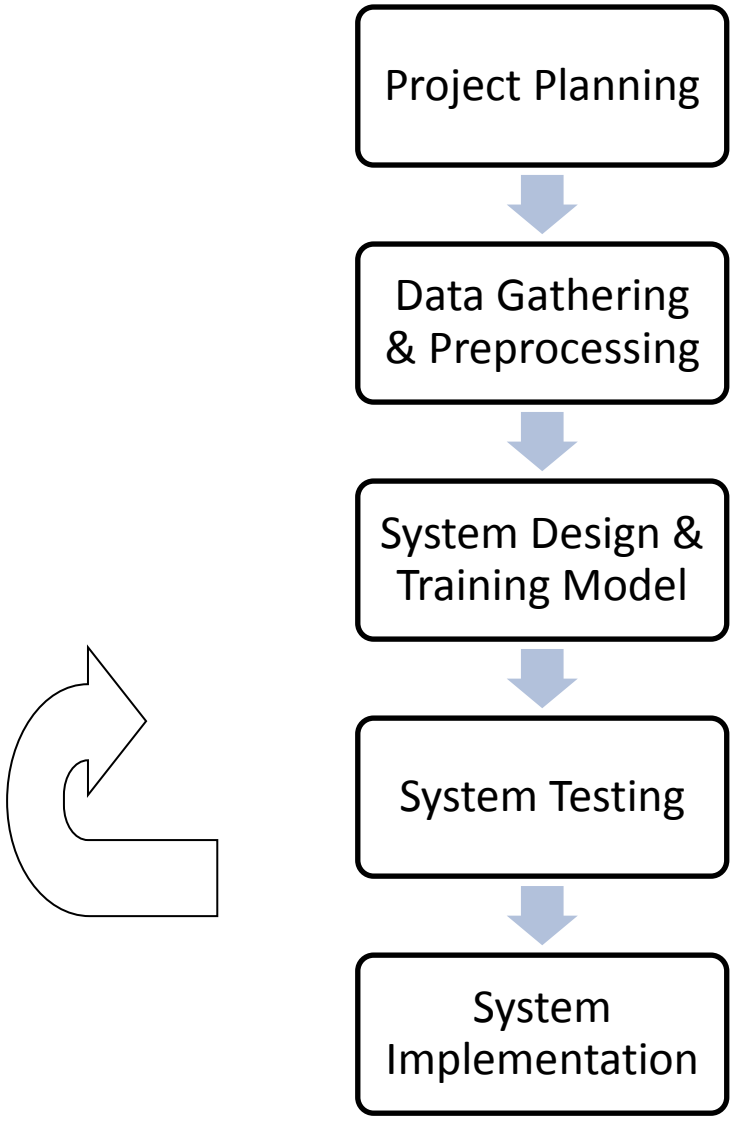

Figure 2: The RAD Phases

3.2 PROJECT ACTIVITIES

3.2.1 PROJECT PLANNING

Reading materials were extensively used in order to gain understanding on the conceptual foundation on the research area which are the time-series forecasting, machine learning tools and financial forecasting models. Brainstorming session with supervisors to find the direction needed for the project is also conducted. The tools used for the system is chosen which is R programming language and WEKA as its data mining machine.

In the end, the output of the project planning is apparent in the earlier parts of the reports which contain a lot of project plan information such as scope, feasibility analysis and literature review.

3.2.2 DATA GATHERING AND PREPROCESSING

Secondary data are the main focus of this project. The data gathering would mean a reliable source of prediction compared to generated data, regardless of the origin country of the PLC. A real stock price data would mean that the system is ready to be used for the market and most of the accuracy and reliability testing would be validated with a real data.

Activities to be done to gather data would be conducted mostly via the internet. Since internet has a lot of websites that hosted data to be purchased by their consumers, there are also certain websites that gives free promotional data to everyone that is interested to use their services. Even though the data is free, it is reliable as the websites would want to attract future customers by giving a free sample first.

The said website would be the www.kibot.com in which the main purpose of the company is to sell historical stock price data. However, it also has several promotional data as said earlier. Based on that, there are two prominent data that are extracted and used to feed the inference engine of the system which are the Advanced Micro Devices (AMD) and Ultra Oil & Gas (DIG) from the US stock market.

The format of the data is that it contains the date, time, open price, high price, low price, close price and volume for each day starting from 1998 and 2007 for AMD

and DIG respectively. It is given in text file format which means that it is easy to process the data to be used by the data mining tools and the system. The data received are intraday stock price data which means that the data also gives specific stock price according to the time of the same date. In the end, the data gathered contains around 90 000 lines of data that must be processed to get the one closing price for each day for one year. An example of the raw stock price data can be seen in *Appendix A.*

The data would then being preprocessed to get one daily closing price only which is called the Train Data. The preprocessing tools mainly revolved around WEKA which apparently have a preprocessing function to ease the process of cutting down 90 000 lines of data into an average of 250 lines of data. This is because for one year, an average of around 250 days the stock market would be open and the preprocessed data must depict the same exact amount of days that the market is operational during the year. Examples of the Train Data are available in *Appendix B.*

After the data has been preprocessed, the next step is to create the Test Data. The Test Data contains the same variables with the Train Data but there are some omissions for the closing price of the Train Data during the last few days of one year of stock price data. The omission would be at intervals of 30 days, 20 days, 10 days, 7 days and 4 days to assess the capability of the system itself.

In all, there two sets of data after preprocessing procedure has been implemented. Test Dataset and Train Dataset. All of the 21 years of historical data are altered into the two datasets mentioned. All of the datasets are then saved in the arff format to ease the loading into the algorithm that can read arff format data.

3.2.3 SYSTEM DESIGN AND TRAINING MODEL

During the design phase, the system framework and architecture will be established to identify and describe the fundamental software system abstraction and their relationship. There are basically two major parts of the system which are the inference engine, which contains the algorithms, and the R GUI itself, to produce the output. The system would be designed by using Java-based R programming language as its main system language and the inference engine is based on the WEKA data mining tool. The R language is able to integrate the WEKA algorithms into its model

and the main advantage of the R language is that it supports the generation of graphical output.

1) System Framework and Architecture

Figure 3 describes the system framework:

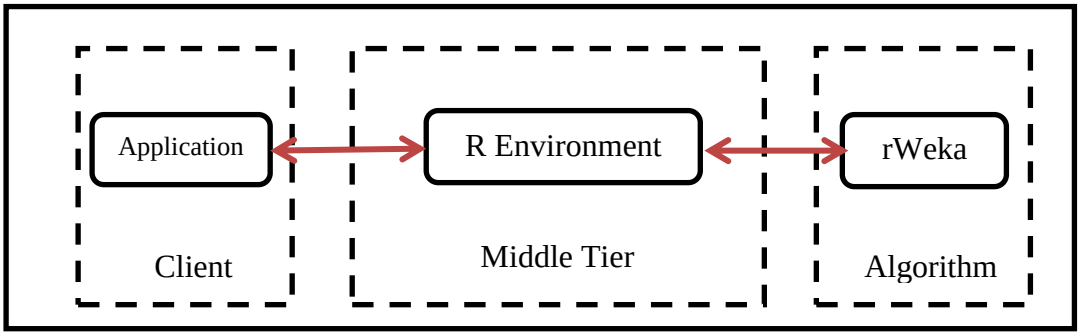

Figure 3: System Framework

The user/client will make the transaction on the application program which it will send the request to commit the algorithm through the R Environment (Middle Tier). The user can choose the particular algorithm that it wanted to implement towards a certain PLC and the 'rWeka' will implement the algorithm and send it back to the user in a user-friendly report depicted as a graph. The implementation of the algorithm can be further elaborated in the Data Flow Diagram in Figure 7.

2) System Architecture

The Figure 4 illustrates the System Architecture. The preprocessed data is divided into two which are the Training Data and the Test Data. The Test Data will be loaded into the algorithm when the user/client request for the specific set of test data. The Training Data is automatically loaded into the algorithm whenever the application starts. The algorithm, which has been trained earlier, will run the Test Data to produce the prediction based on the patterns and trends recognized from both the Test Data and also the Training Data.

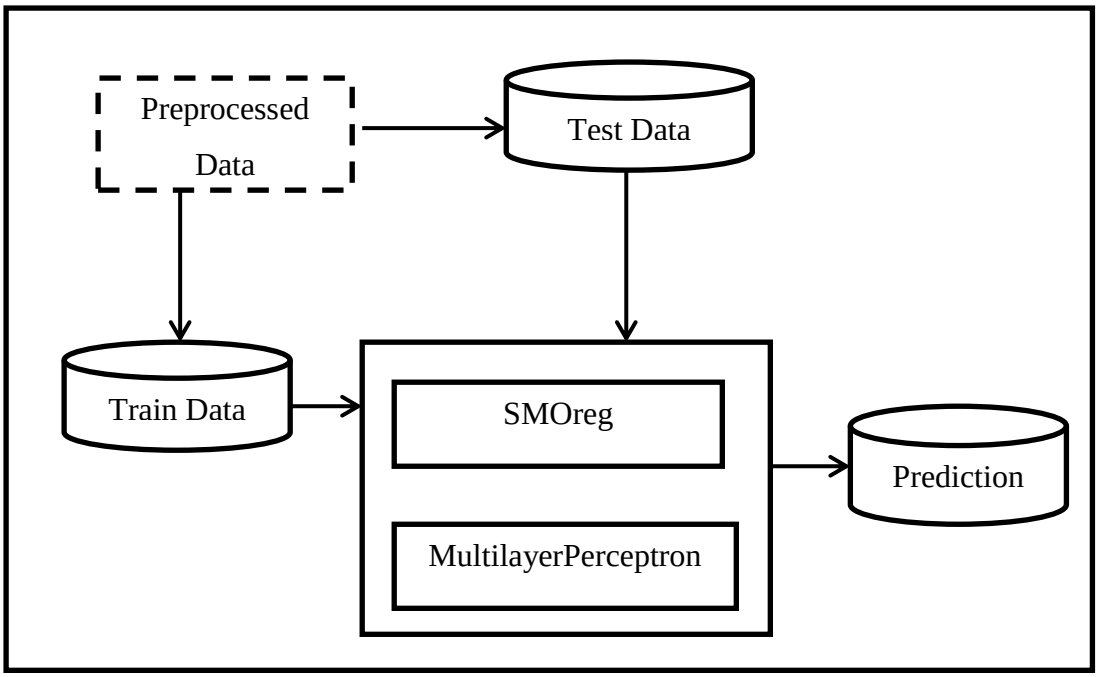

Figure 4: System Architecture

3) Training Model

Since there are 21 years of historical data overall, the training of the inference engine would use all the processed data to allow the algorithm to learn. The training model requires the usage of all available data within the system to train the system. The algorithm would run all the historical data available to serve this purpose. WEKA is used for this purpose of building and training of the algorithm.

Building the algorithm would mean that the search to find the best parameters to be parsed into the algorithm and the search for the best available algorithm to implement the parameters. For the ANN machine, the selected algorithm is the 'MultilayerPerceptron' function and for the SVM machine is the 'SMOreg' function. The selections are based on reading of materials from the WEKA guidelines and also the try-and-error procedure to find the least error percentage. The parameters are selected based on the accuracy level of the results. The 'MultilayerPerceptron' parameters are depicted in Figure 5.

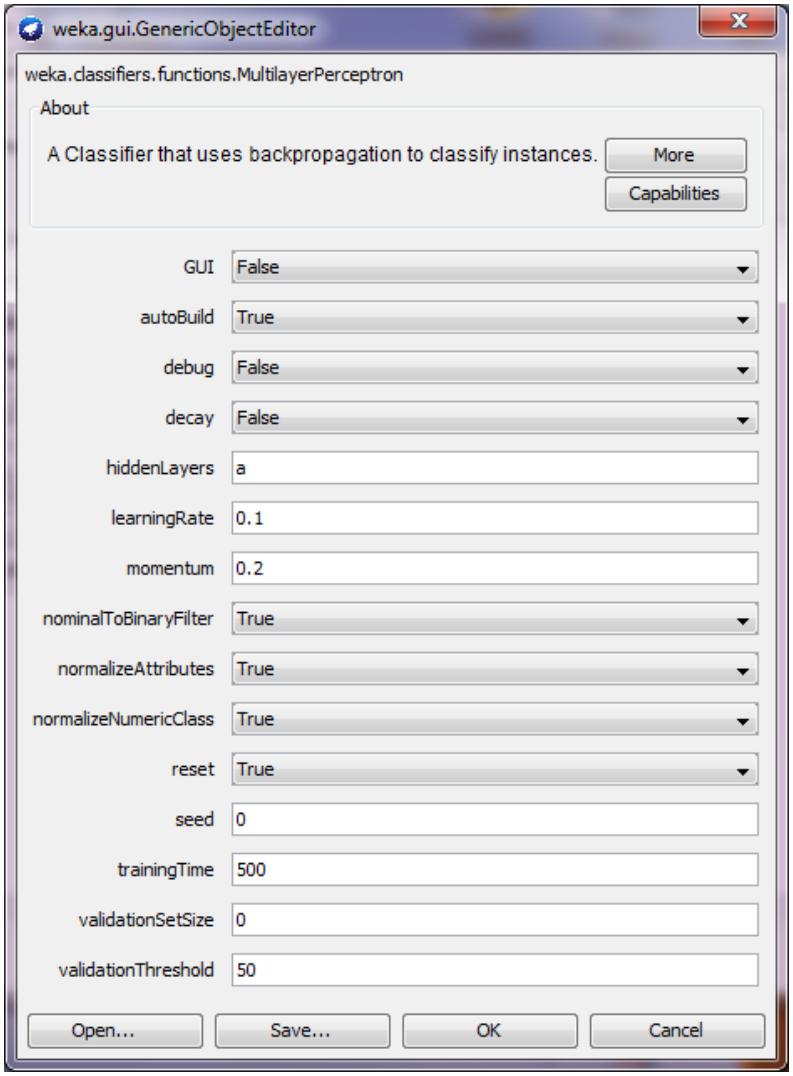

Figure 5: 'MultilayerPerceptron' Parameters

Based on Figure 5, the 'autoBuild' parameter is set as TRUE to indicate that the algorithm will add and connect the hidden layers from the data. The 'LearningRate' would change the amount of time the weights are updated throughout the course of the training process. The 'Momentum' parameter is to set the amount of momentum given to the weights before it updates. The 'validationThreshold' parameter is to stop the algorithm if the amount of error reaches a certain values. The significant parameters and its uses are described in order to show the changes of the parameters for this training model.

The ‘SMOreg’ parameters are depicted as follows:

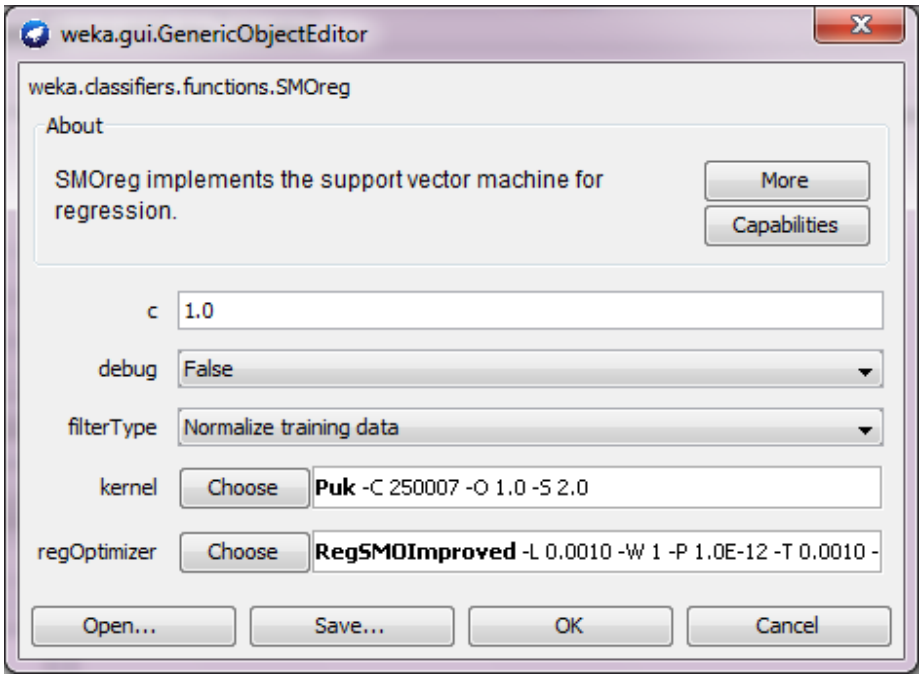

Figure 6: ‘SMOreg’ Parameters

For Figure 6, the parameter that has been changed is the ‘kernel’ and ‘regOptimizer’ parameters. The ‘Puk’ kernel is the best kernel to implement the algorithm on this set of data with Sigma value at 2.0 and Omega value at 1.0. The ‘regOptimizer’ function is to choose the learning algorithm of the SVM algorithm. For this system, the learning algorithm chosen is the ‘RegSMOImproved’ algorithm with its ‘epsilonParameter’ at 0.001 and the rest is set at default. This is the best parameters of the algorithm.

The training model is run mostly on WEKA first to validate the best results and then parsed into R using the ‘rWeka’ package. The ‘rWeka’ package contains most of the algorithms in WEKA to be implemented in R using simple coding lines. The training model is then used in the back-end of the system in which the user cannot see the process.

3.2.3 SYSTEM IMPLEMENTATION AND TESTING

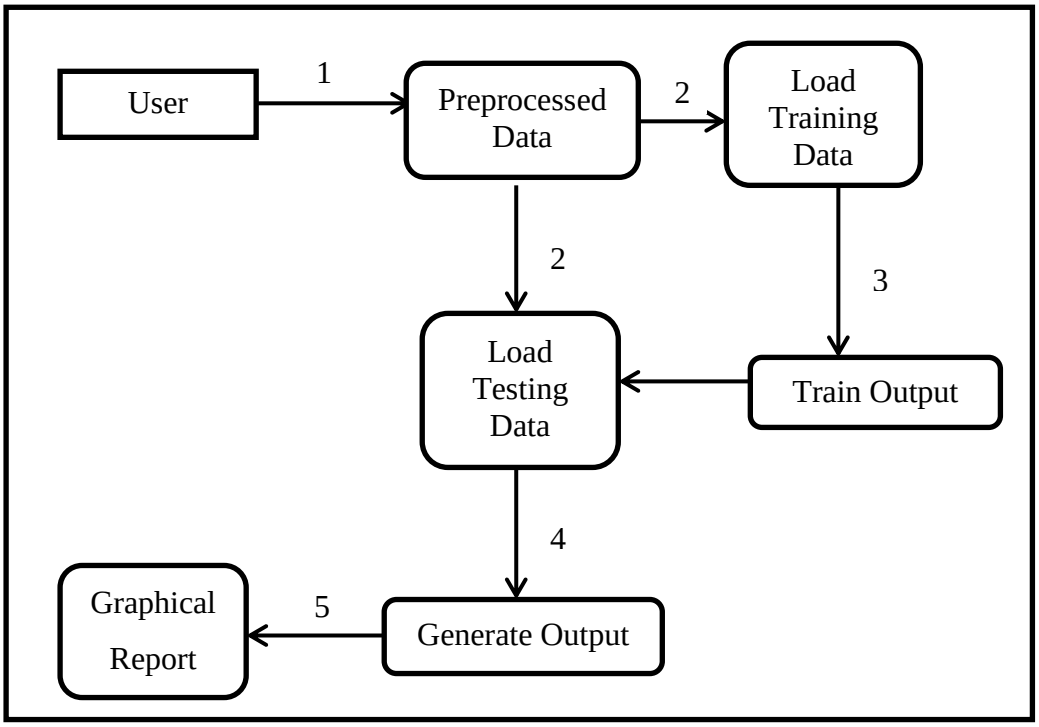

Figure 7: Data Flow Diagram

The Data Flow Diagram in Figure 7 depicts on the implementation flow of the system. The user would initiate the application by choosing one of the two companies available in the system. From that moment on, the program would load both Training and Testing Data to be used by the algorithm to generate the desired output. Before the output is generated, the algorithm is trained first by the Training Data before both of the trained output and the Testing Data is loaded into the algorithm that will be used to generate the desired output according to the user's preferences. Then, the application will generate a graphical report to depict the output in a more user-friendly way rather than numbers and formulas.

The testing of the application will be using the Testing Data whereby the data would omit the Close Price attributes at several intervals such as 30 days, 20 days, 10 days, 7 days and 4 days at the end of the year of the data. Example of the Testing Data and how it differs from the Train Data are illustrated in *Appendix C*.

3.3 GANTT CHART

See appendix D (Gantt chart)

3.4 TOOLS

Table 1: Tools used to develop the system

Elements	**Software/ Platform**
Software	WEKA and R
Platform	Windows XP/ Windows 7
Programming Language	R
Data Mining Machine	WEKA

3.4.1 HARDWARE REQUIREMENT

The minimum hardware requirement to develop the system is shown on **Table 2**.

Table 2: Minimum Hardware Requirement

Hardware	**Descriptions**
CPU	Intel Centrino 1.6 Ghz Processor or higher or others
Memory	At least 512 MB Recommended: 1 GB or higher
Hard Disk Space	At least 10 GB
Others	Other required standard computer peripherals (Mouse, Keyboard, etc)

CHAPTER 4: RESULT AND DISCUSSIONS

There are two sets of data used which are the Training Set and also the Testing Set. The Training Set is basically the actual stock price data of each company (AMD and DIG) in which the AMD has 14 years of stock price data (1998 until 2012) meanwhile DIG has 6 years of stock price data (2007 until 2012) which gives a total of 20 years of stock price data. The Training Set will be used to train the algorithm to recognize the patterns and also to be used for the Test Set later on.

The Test Set has 10 years of stock price data in general. For the early purpose of building the Inference Engine, there are 5 groups of Test Set which indicates the days of stock price removed in order to predict stock price of the removed stock prices. The first group is the 30 days stock price removed, second group is the 20 days stock price removed, third group will be the 10 days stock price removed, and forth group is the 7 days removed while the last group is 4 days removed.

Each group have two stock price data such as the first group has two AMDs stock price data removed of its last 30 days. The results of each Test Set group will then be calculated and summarized to get the best prediction result of the ANN and SVM algorithms according to the groups allocated that stemmed on the accuracy of each prediction. The best result would be the indicator of the Test Set that will be used in the application system.

Therefore, there are two main parts of the development which are the inference engine and also the R GUI system. This is to ensure that the accuracy of the stock prices is solidified first before using it within the system.

4.1 INFERENCE ENGINE

The software used for the inference engine is the WEKA data mining software that contains the algorithms of ANN and SVM.

4.1.1 FLOWCHART

The required tasks that need to be completed first are the building of the basic inference engine of the system. Since there are two algorithms that will be used in the inference engine, Figure 8 illustrates the work flow of it from start to the end.

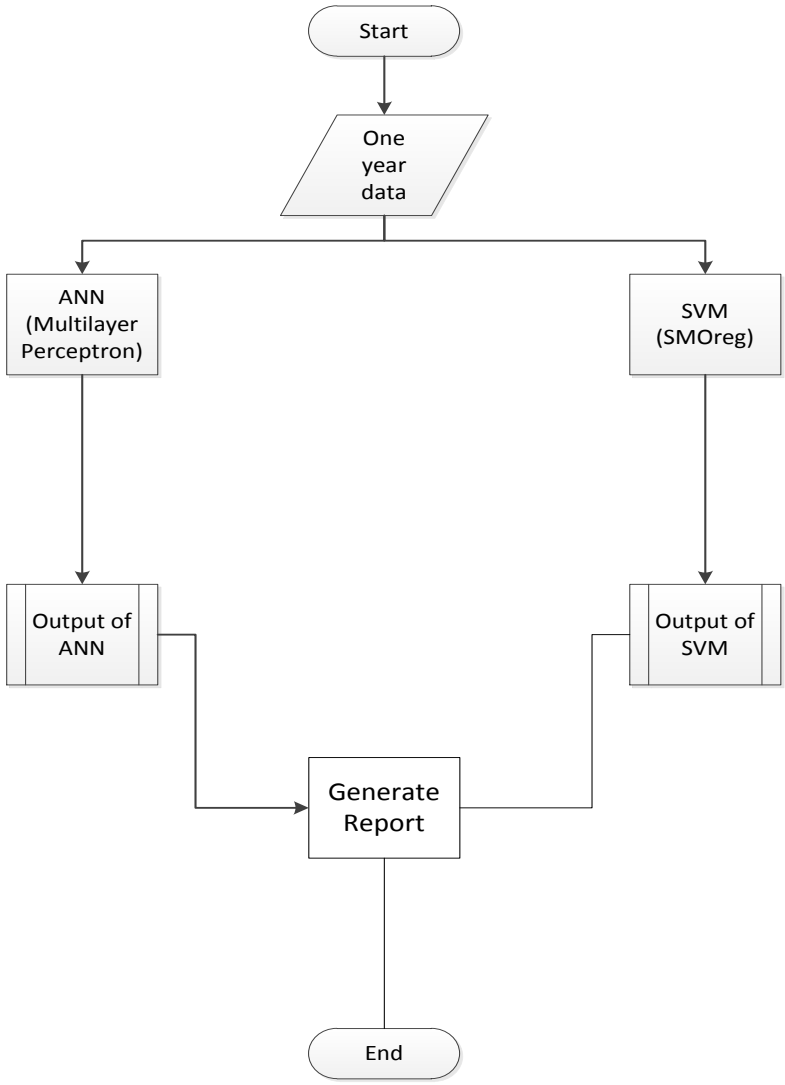

Figure 8: Flowchart of the Inference Engine

The inference engine flowchart depicts the back-end process of the system. It is suitable for both Training Data and also Testing Data.

4.1.2 MECHANISM OF SVM & ANN

The usage of ANN and SVM algorithms hugely impeded upon the nature of such algorithms to begin with. Detailed explanations of both the nature of ANN and SVM would further explain the algorithms.

According to Tan, Steinbach, & Kumar (2006) the ANN algorithms attempted to simulate the biological neural systems which have neurons which are linked to each other via strands of fibre called axons basically. Neurologists have discovered that the human brain learns by changing the strength of the synaptic connection between neurons upon repeated stimulation by the same impulse. Therefore, the ANN tries to imitate the same functions of the human neurological features by having input layer, learning algorithms and its output layer.

The inference engine consists of ANN is because of the general characteristics of the ANN which are:

1) Multilayer neural networks (used in the inference engine) can be used to approximate any target functions which means that it can basically predict the output given its parameters are set accordingly.
2) ANN can handle redundant features because the weights are automatically learned during the training step. The weights for redundant features tend to be very small.

This shows that ANN is able to handle stock price data effectively but perhaps not very accurate.

SVM on the other hand is almost similar to ANN but with more upgrades. It adds dimensionality in its decision making by choosing the best hyperplane that separates the data set into two distinct classes as good as possible. This decision boundary method is also known as the support vectors. The Figure 9 would depict the workings of support vectors.

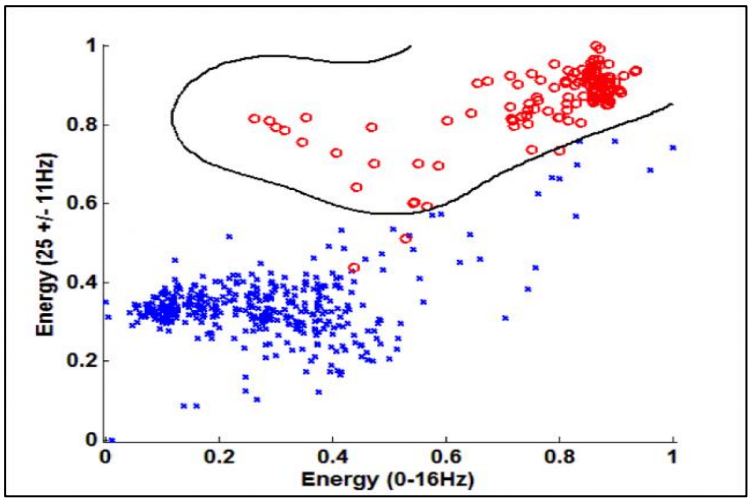

Figure 9: Decision Boundaries (MIT Computer Science and Arficial Intelligence Laboratory)

As depicted, the SVM can separate two non-linear classes and since stock price is a non-linear data, it can classify the stock prices to predict the output of the non-existent data. SVM can also show empirical results that would add another dimension towards the output which is the value of the stock price at the exact day required.

4.1.3 RESULTS AND DISCUSSIONS

There will be two different inference engines which are the ANN inference engine and also the SVM inference running the same Test Set data containing each group of data. The examples shown are the instances of the Test Set data, although it might not contain all of the results but a summarized table is shown in Table 9.

4.1.3.1 ANN INFERENCE ENGINE

Using the Testing Data, the result of the inference engine is illustrated by figures subsequently in next pages.

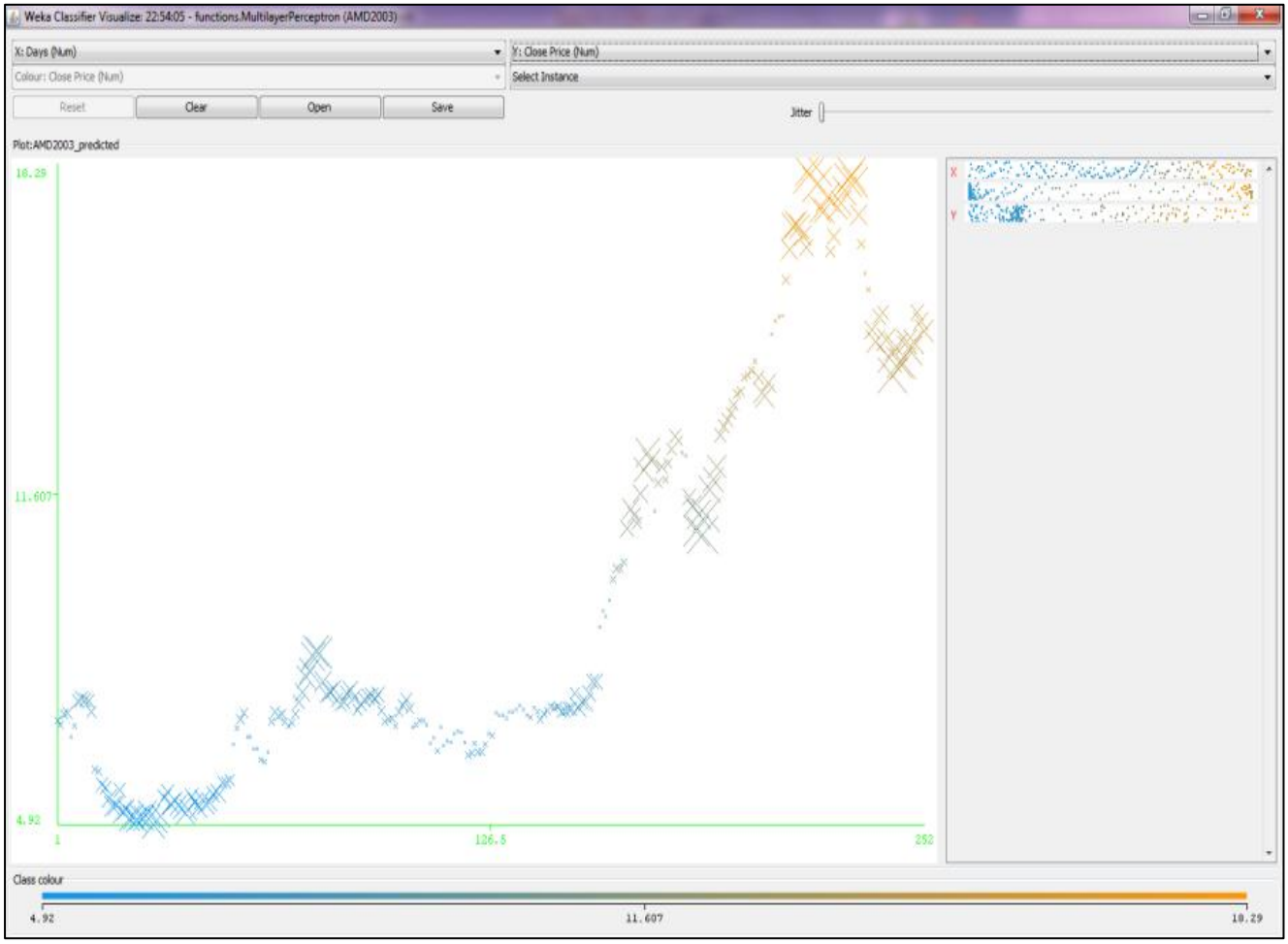

Figure 10: AMD 2003 Stock Price

Figure 10 depicts the AMD 2003 original stock price data. The x-axis represents the closing price for each attributes while the y-axis represents the number of transaction days in the year 2003. In this instance, the number of days is 252 whereby the minimum closing price is 4.92 and maximum is at 18.29. The colours depict the location of the points according to the closing price. Blue would mean closer to the minimum while orange would mean closer to maximum.

The first algorithms that are implemented on this stock data are the ANN algorithm. With the Testing Data used, the predicted result is as follows. The AMD 2003 Test Set has been removed the last 30 days of the stock price data.

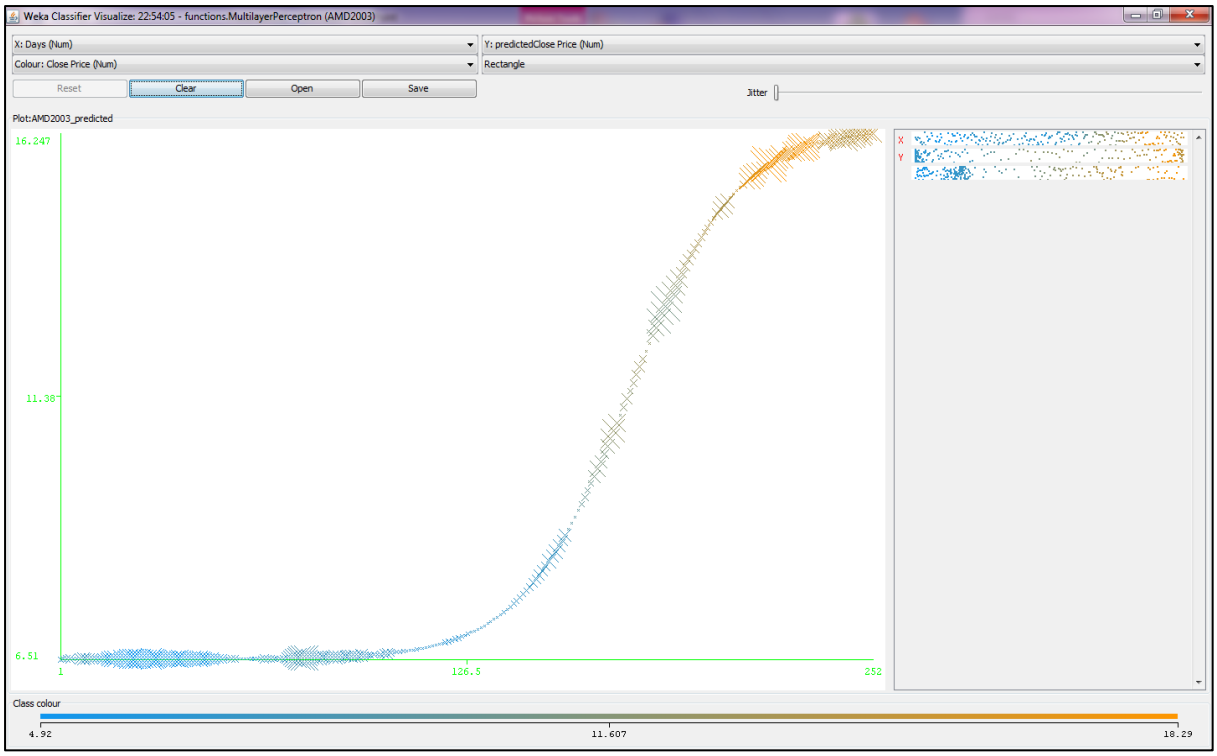

Figure 11: AMD 2003 Training Set using ANN

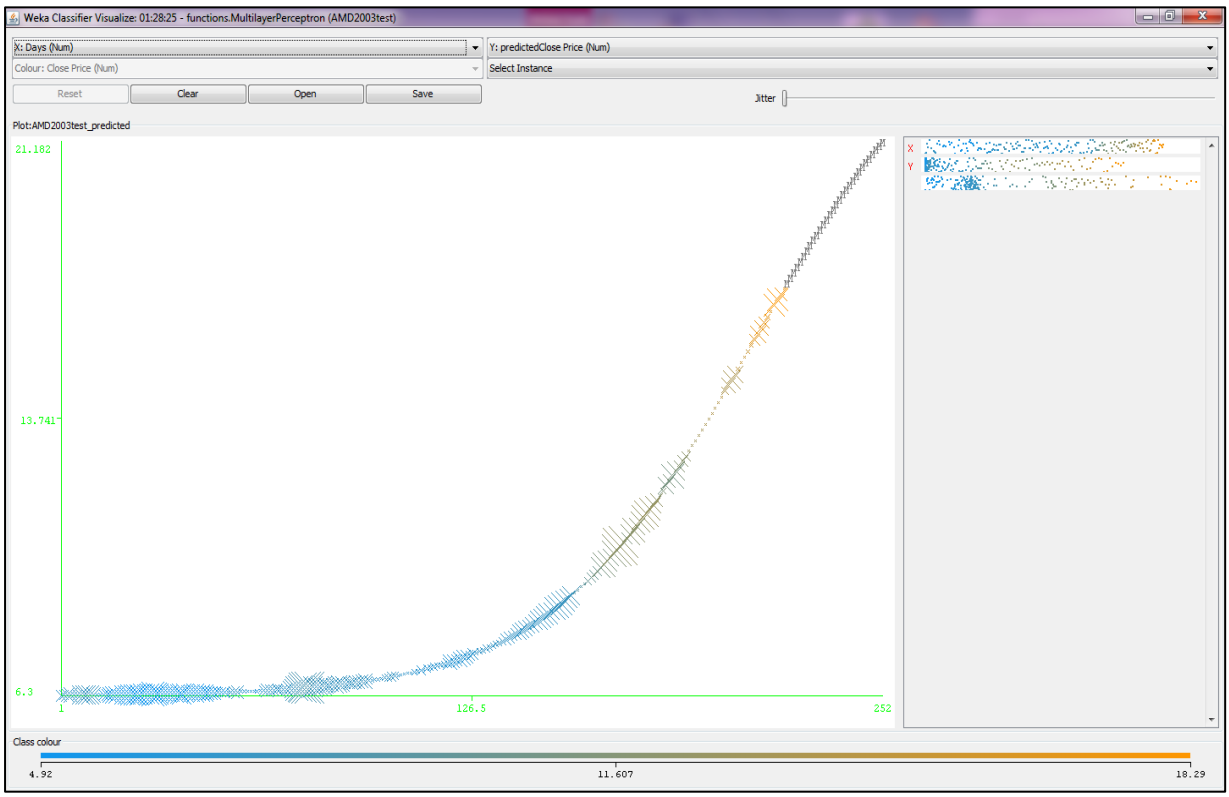

Figure 12: AMD 2003 Testing Set using ANN

Based on Figure 12 and Figure 13, assessing the patterns by observation would show that the patterns are almost similar to each other. However, to get a more conclusive result, the mean squared error of both predictions is conducted.

Table 3: Mean Squared Error on AMD 2003 using ANN

Test Set	Train Set	Squared Error	Mean Squared Error
15.78162	17.3717	2.528348046	
15.81311	17.52181	2.919662525	
15.84294	17.67108	3.342081234	
15.87121	17.81944	3.795607926	
15.89798	17.96683	4.28014446	
15.92333	18.11318	4.795469301	
15.94732	18.25843	5.341243299	
15.97002	18.40251	5.917031925	
15.99149	18.54537	6.522297947	
16.01179	18.68694	7.156406121	
16.03099	18.82717	7.818633777	
16.04914	18.96601	8.508177267	
16.06628	19.10341	9.224158637	
16.08249	19.23933	9.965632472	
16.09779	19.3737	10.73158633	
16.11225	19.5065	11.52095343	
16.1259	19.63768	12.33263389	
16.13878	19.76721	13.16546798	
16.15095	19.89504	14.01825486	
16.16243	20.02116	14.8897895	
16.17326	20.14552	15.77880979	
16.18349	20.2681	16.68409604	
16.19313	20.38888	17.60433485	
16.20223	20.50784	18.53827747	
16.21081	20.62495	19.48463194	
16.21891	20.74021	20.4421356	
16.22654	20.85358	21.40950842	
16.23374	20.96508	22.38552142	
16.24053	21.07467	23.36891921	
16.24693	21.18236	24.35847916	**11.96094316**

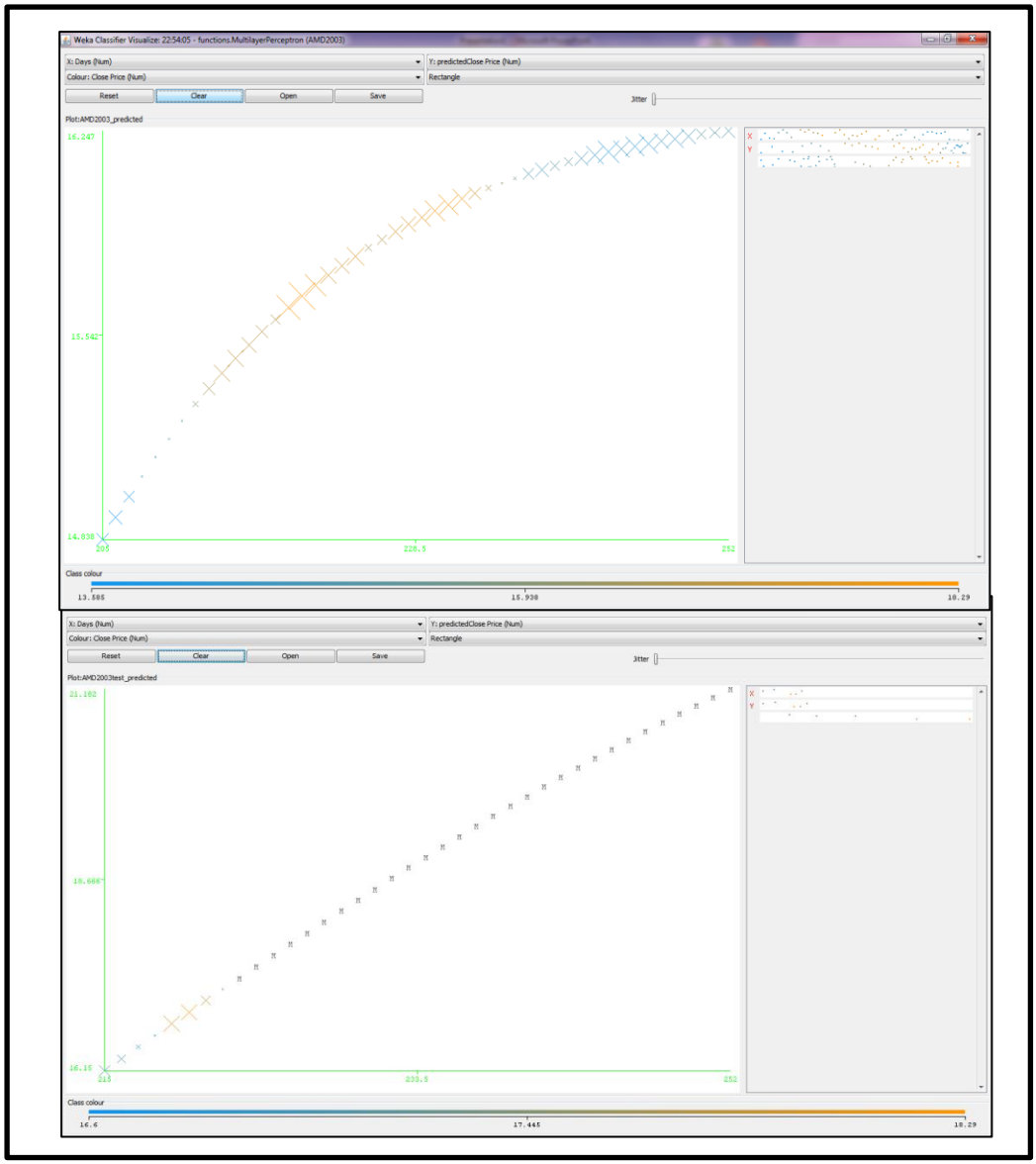

Figure 13: AMD 2003 using ANN (Train Set and Test Set)

Figure 13 describes the close up view on the variances of the Test Set with the Train Set. The mean squared error calculation is to find the exact numbers of the differences.

The next Test Set used is the AMD 2004 Test Set with the same 30-days-stock-price-removed data as the AMD 2003

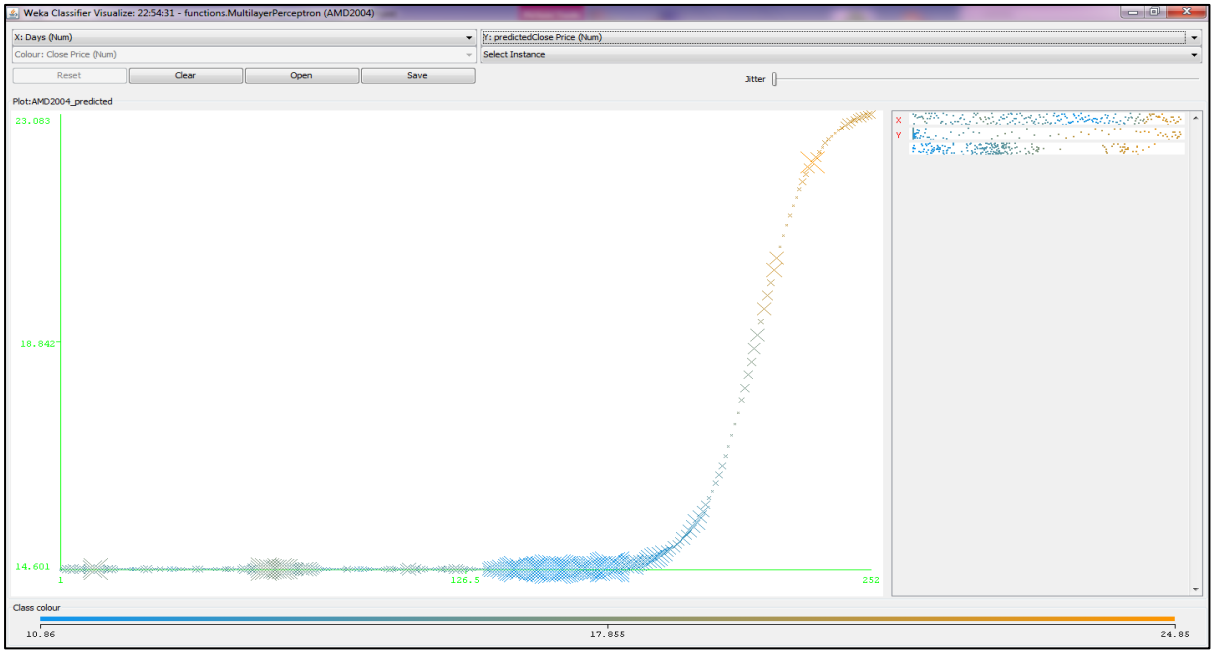

Figure 14: AMD 2004 Train Set using ANN

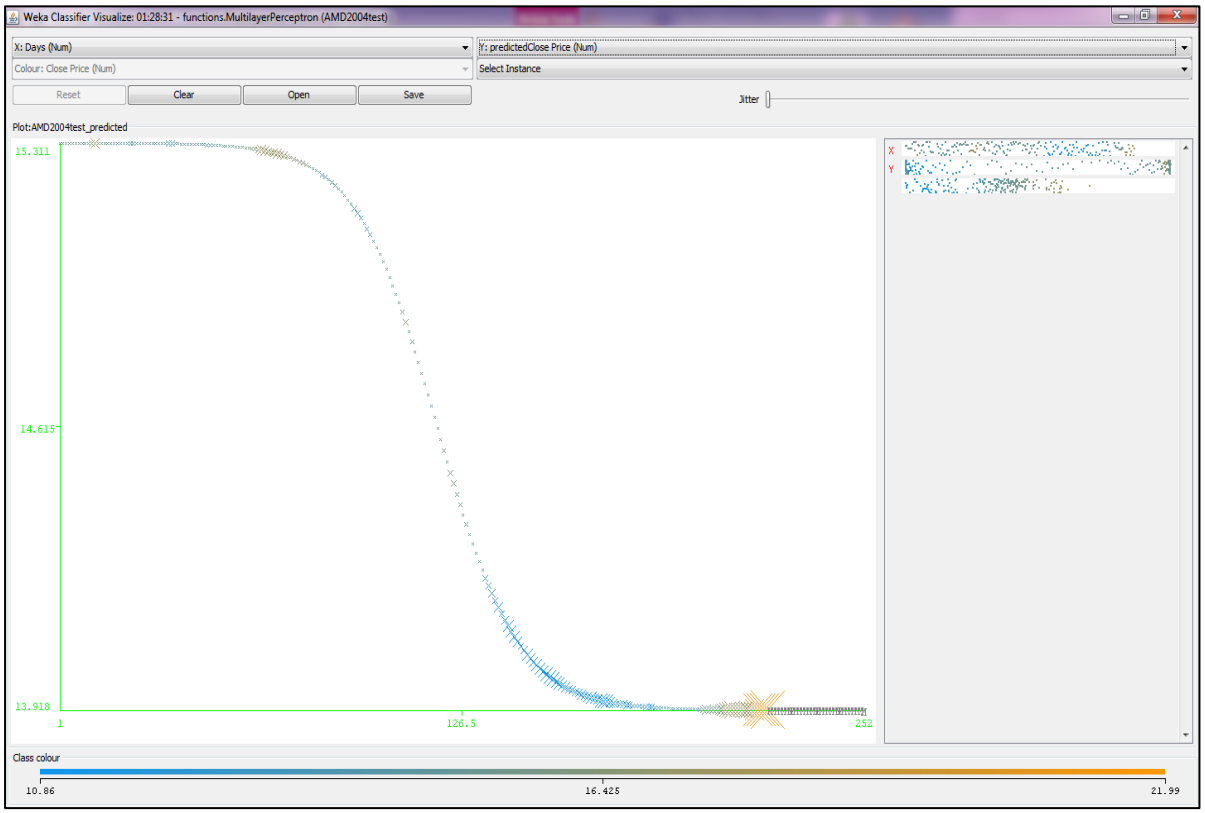

Figure 15: AMD 2004 Test Set using ANN

Graphs on Figure 14 and Figure 15 illustrate bigger differences in the predicted output. The mean squared error for both graphs is:

Table 4: Mean Squared Error on AMD 2004 using ANN

Test Set	Train Set	Squared Error	Mean Squared Error
13.91866	20.39611	41.9573585	
13.91863	20.61123	44.79086799	
13.9186	20.81681	47.58527361	
13.91858	21.0122	50.31951564	
13.91856	21.19694	52.97493188	
13.91854	21.37075	55.53552331	
13.91852	21.53351	57.98817931	
13.9185	21.68527	60.32273177	
13.91848	21.82619	62.53189326	
13.91847	21.95656	64.61095515	
13.91845	22.07675	66.55776099	
13.91844	22.18719	68.37221003	
13.91843	22.28838	70.05604626	
13.91842	22.38084	71.61253533	
13.91841	22.46511	73.04616636	
13.9184	22.54175	74.36225146	
13.91839	22.61131	75.56685813	
13.91838	22.67432	76.66641524	
13.91837	22.7313	77.66764706	
13.91837	22.78275	78.57726824	
13.91836	22.82914	79.40207149	
13.91835	22.87093	80.14856332	
13.91835	22.90851	80.82306673	
13.91834	22.94229	81.43165555	
13.91834	22.97262	81.98004064	
13.91833	22.99983	82.4735696	
13.91833	23.02422	82.91728733	
13.91833	23.04608	83.31582006	
13.91832	23.06564	83.67348148	
13.91832	23.08315	83.99418225	**69.7087376**

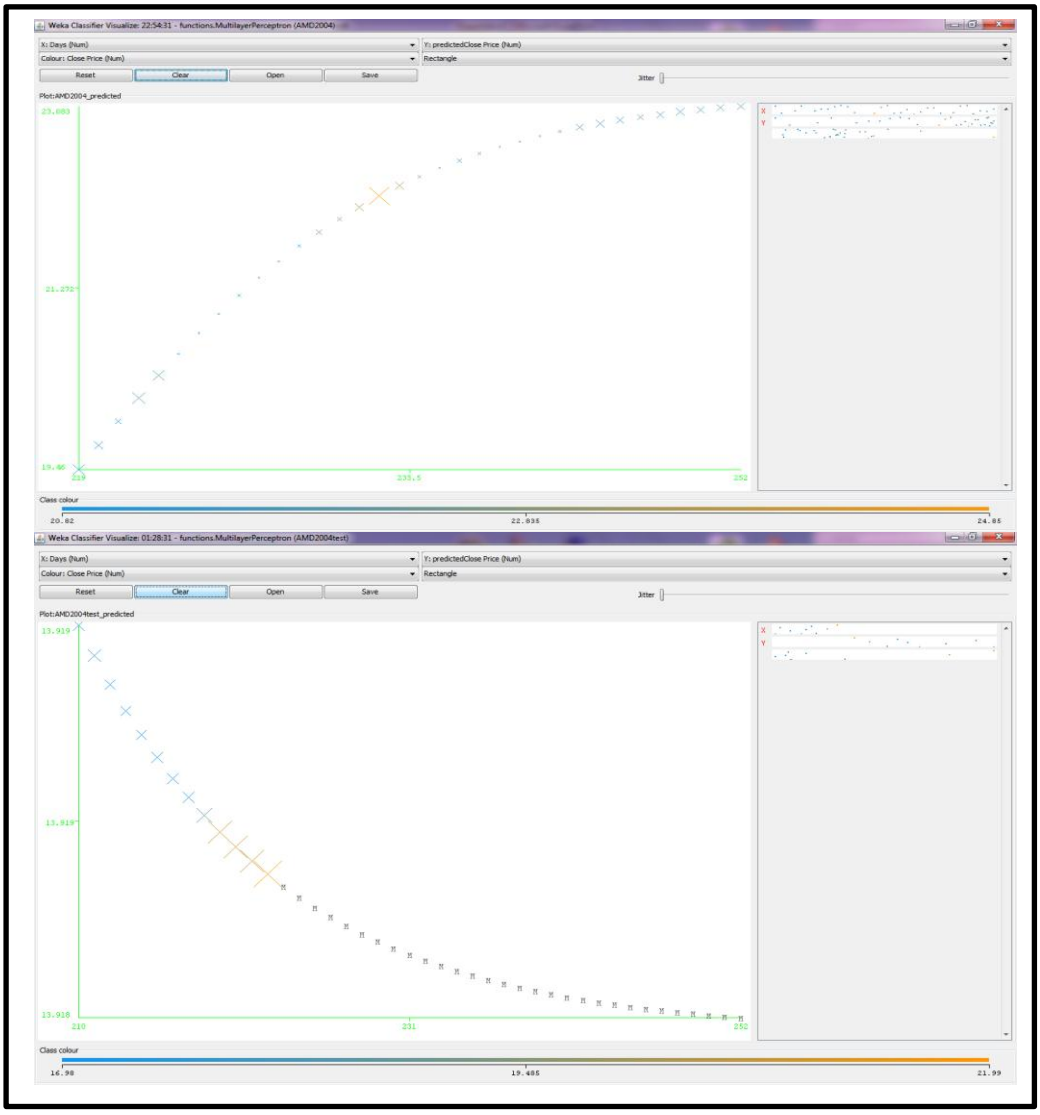

Figure 16: AMD 2004 using ANN (Train Set and Test Set)

Figure 16 depicts the close up view on the location where the stock price are removed and the ANN prediction. The big amount of mean squared error indicates that the prediction is not accurate for the AMD 2004 Test Set running on the 'MultilayerPerceptron' algorithm. However, that does not mean that the ANN is not reliable.

The next Testing Data will be using the 20 days missing data evaluation. The AMD 2005 data is used for the evaluation.

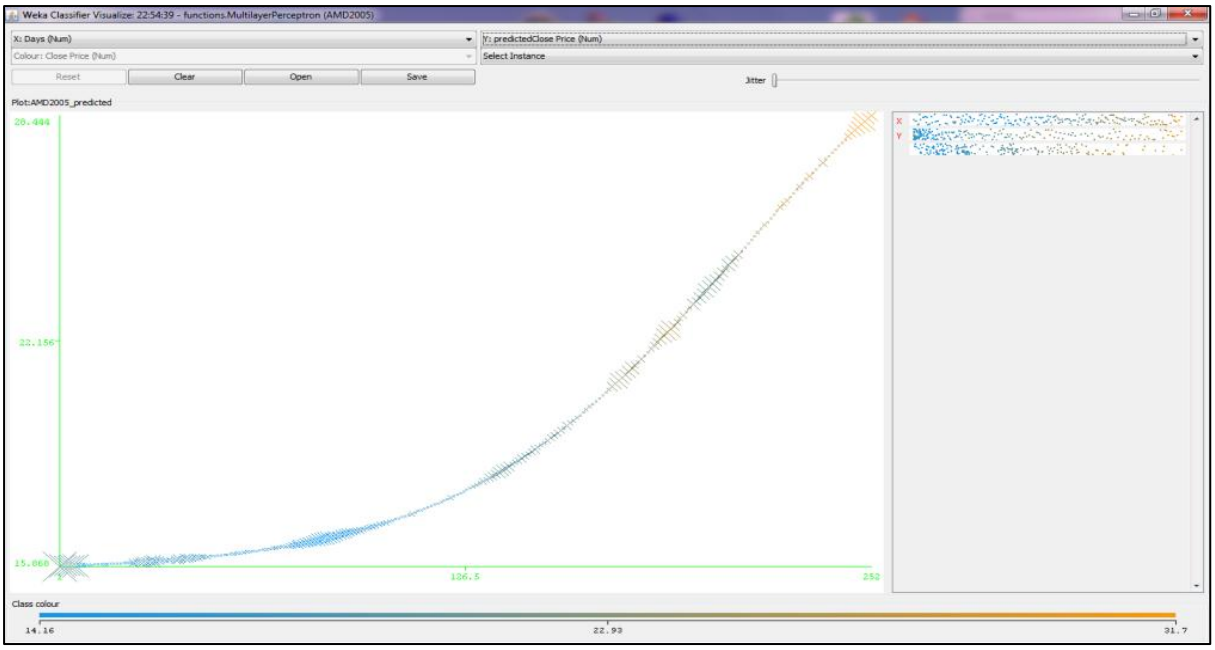

Figure 17: AMD 2005 Train Set using ANN

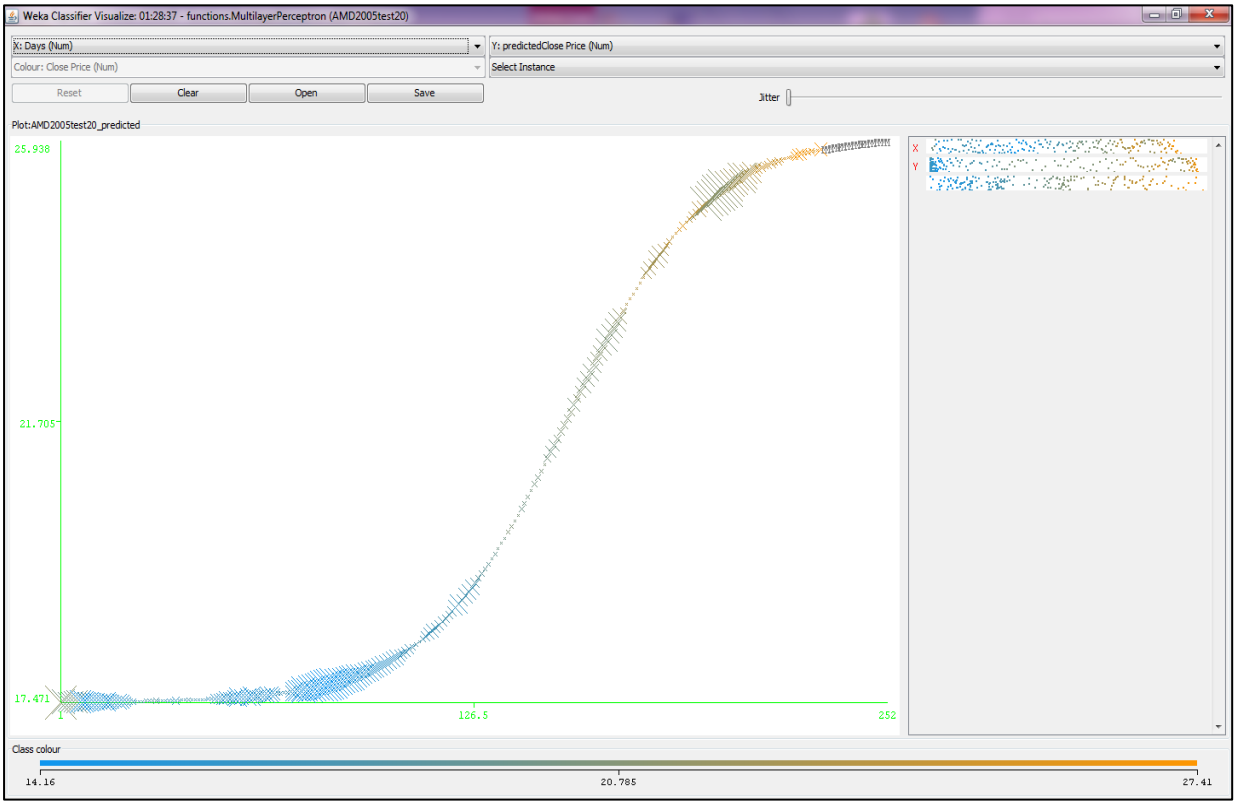

Figure 18: AMD 2005 Test Set using ANN

Using the ’20-days-data-removed’ Test Set; the results show a significant amount of alterations towards the Test Set prediction. The mean squared error calculations are as in Table 5.

Table 5: Mean Squared Error on AMD 2005 using ANN

Test Set	Train Set	Squared Error	Mean Squared Error
25.81506	26.7436	0.862182817	
25.8245	26.83875	1.028715234	
25.83352	26.93338	1.20968542	
25.84215	27.02745	1.404945572	
25.85039	27.12095	1.614325255	
25.85827	27.21387	1.837632382	
25.8658	27.30617	2.074654214	
25.873	27.39785	2.325161423	
25.87987	27.48888	2.588906744	
25.88644	27.57926	2.86562601	
25.89272	27.66896	3.155039195	
25.89871	27.75797	3.456847748	
25.90444	27.84628	3.770738702	
25.9099	27.93386	4.096401938	
25.91513	28.02072	4.433500826	
25.92012	28.10682	4.781691877	
25.92488	28.19217	5.140631152	
25.92943	28.27676	5.50994874	
25.93377	28.36056	5.889285436	
25.93792	28.44356	6.278266889	**3.055048457**

Although the graphical differences might look quite significant, the mean squared error indicates that the error is acceptable. The usage of the mean squared error is to gauge numerically the differences between the Test Set predictions with the Train Set predictions.

The next Train Set will be in the category of the ’10-days-data-removed’ in which it uses the AMD 2010 stock price data to conduct the testing.

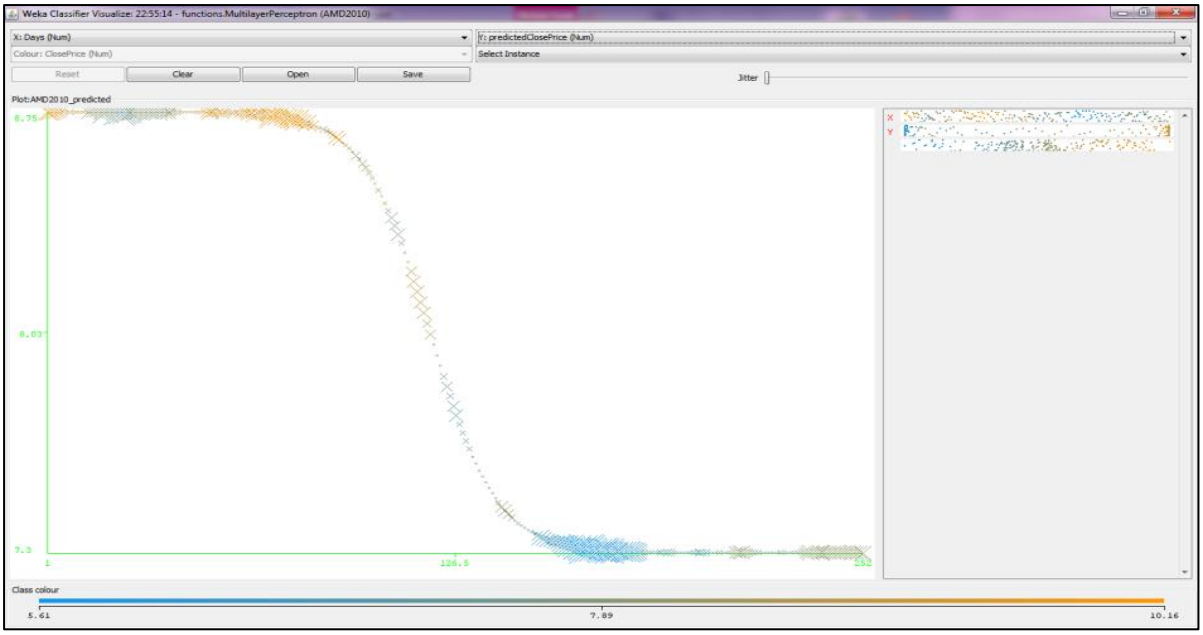

Figure 19: AMD 2010 Train Set using ANN

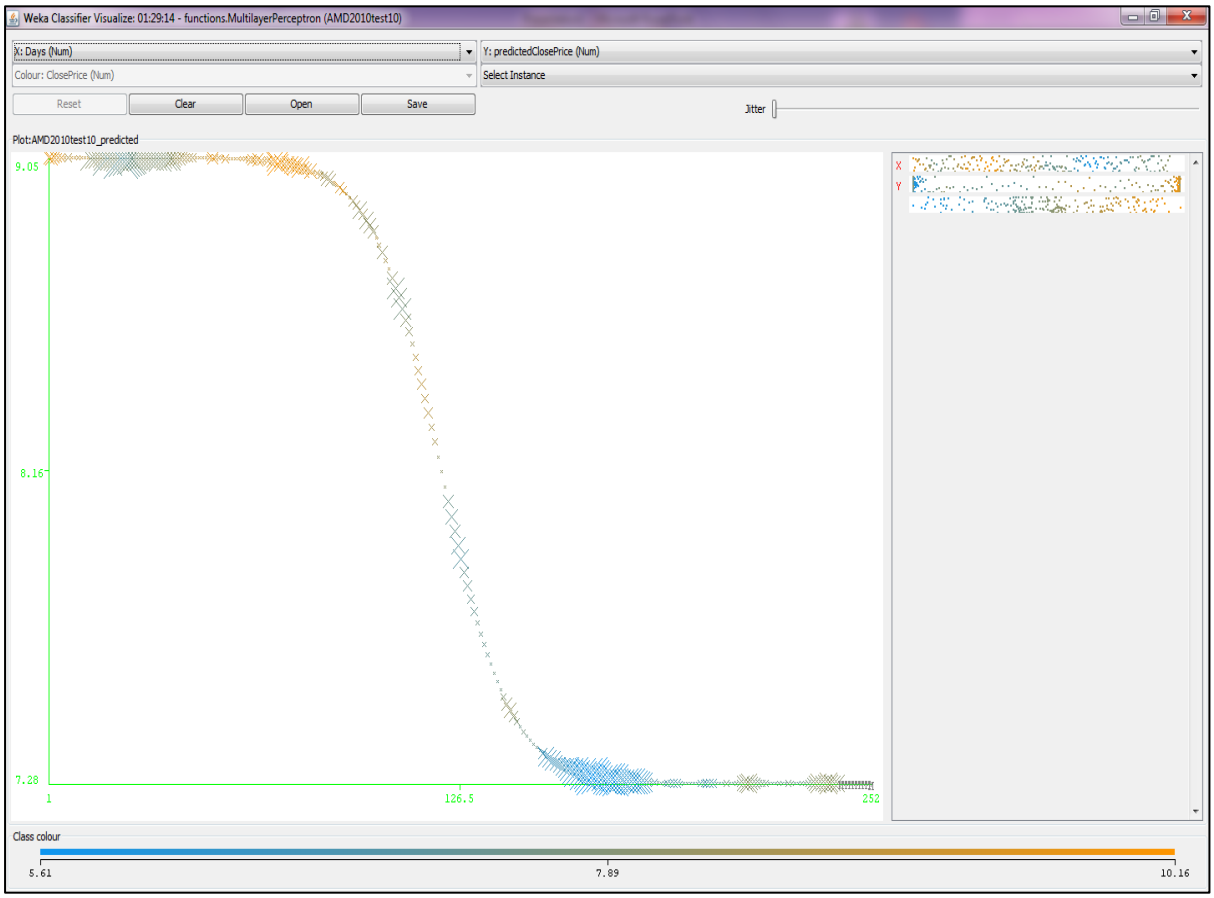

Figure 20: AMD 2010 Test Set using ANN

Figure 21: AMD 2010 using ANN (Train Set and Test Set)

Based on Figure 19, Figure 20 and Figure 21, there are not much of a difference between the two predicted graphs. The mean squared error is also low.

Table 6: Mean Squared Error on AMD 2010 using ANN

Test Set	Train Set	Squared Error	Mean Squared Error
7.279755	7.302595	0.000521666	
7.279753	7.302594	0.000521711	
7.279752	7.302593	0.000521711	
7.279751	7.302593	0.000521757	
7.27975	7.302592	0.000521757	
7.279749	7.302592	0.000521803	
7.279748	7.302591	0.000521803	
7.279747	7.302591	0.000521848	
7.279747	7.30259	0.000521803	
7.279746	7.30259	0.000521848	**0.000521771**

Next Test Set has been removed 7 days of data at the end of the year. The DIG 2007 stock price data is used for this purpose.

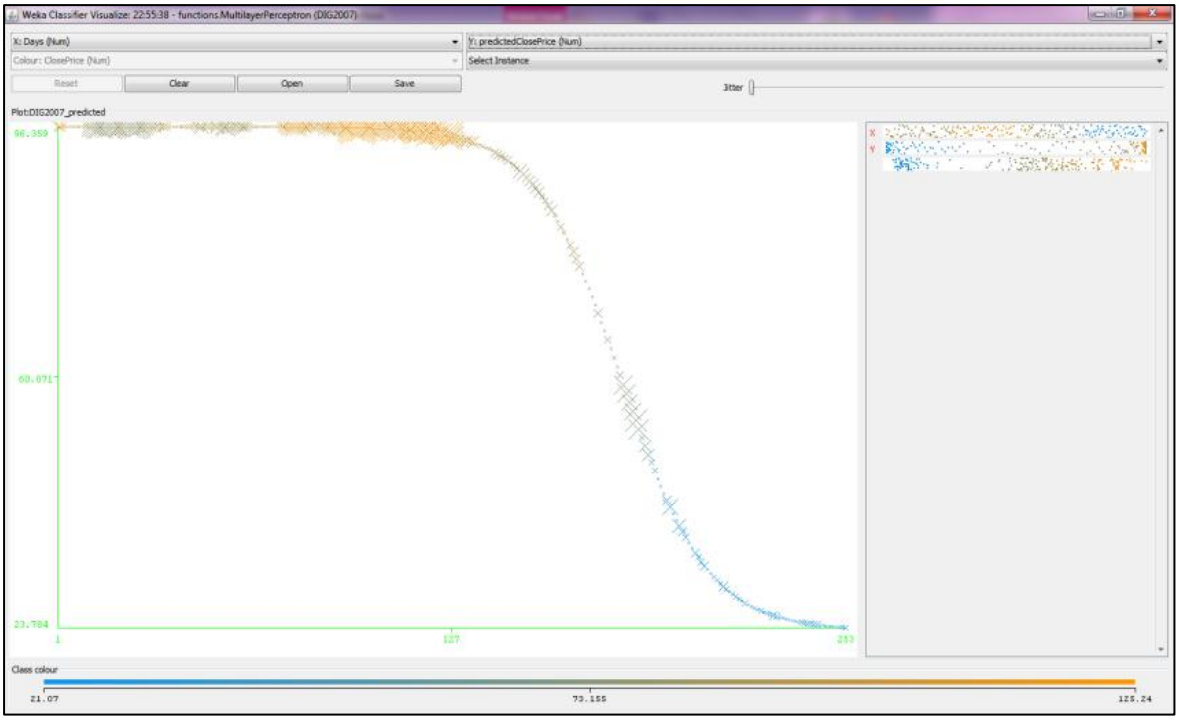

Figure 22: DIG 2007 Train Set using ANN

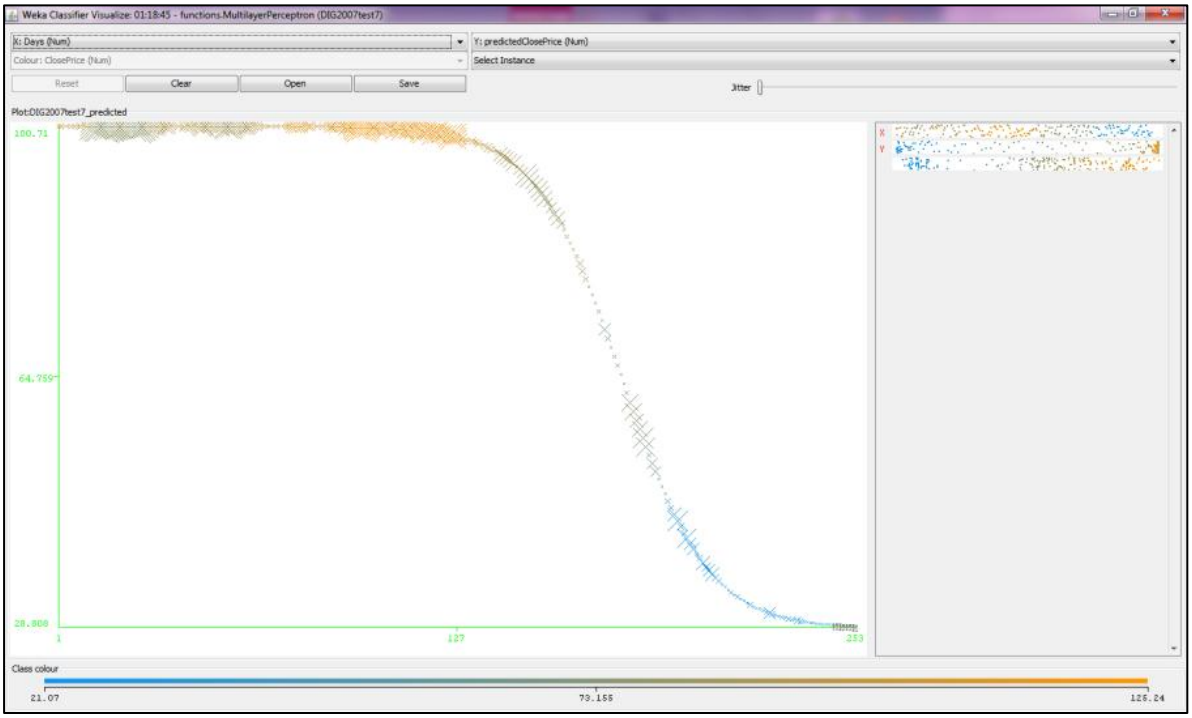

Figure 23: DIG 2007 Test Set using ANN

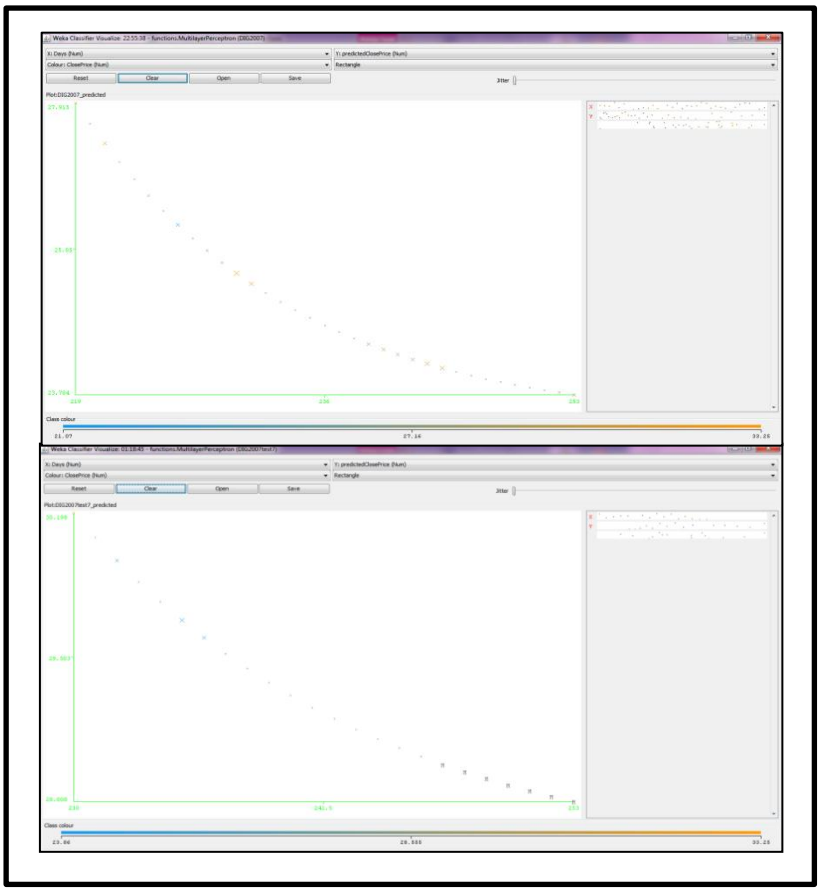

Figure 24: DIG 2007 using ANN (Train Set and Test Set)

Within the illustrations of Figure 22, Figure 23 and Figure 24, the graph looks accurate. Table 7 would give a concrete result to approve or disprove the initial observation.

Table 7: Mean Squared Error on DIG 2007 using ANN

Test Set	Train Set	Squared Error	Mean Squared Error
28.98554	24.00785	24.77739774	
28.95033	23.96365	24.86702729	
28.91758	23.92246	24.95122381	
28.88712	23.88409	25.03027916	
28.85878	23.84835	25.10449897	
28.83243	23.81505	25.17416227	
28.80793	23.78403	25.23953102	**25.02058861**

Despite the graph looks similarly close to each other, the mean squared error shows otherwise. The high mean squared error indicates that the predictions are not quite accurate.

Next result is the DIG 2009 stock price data which contains a test set of 4 days data removed.

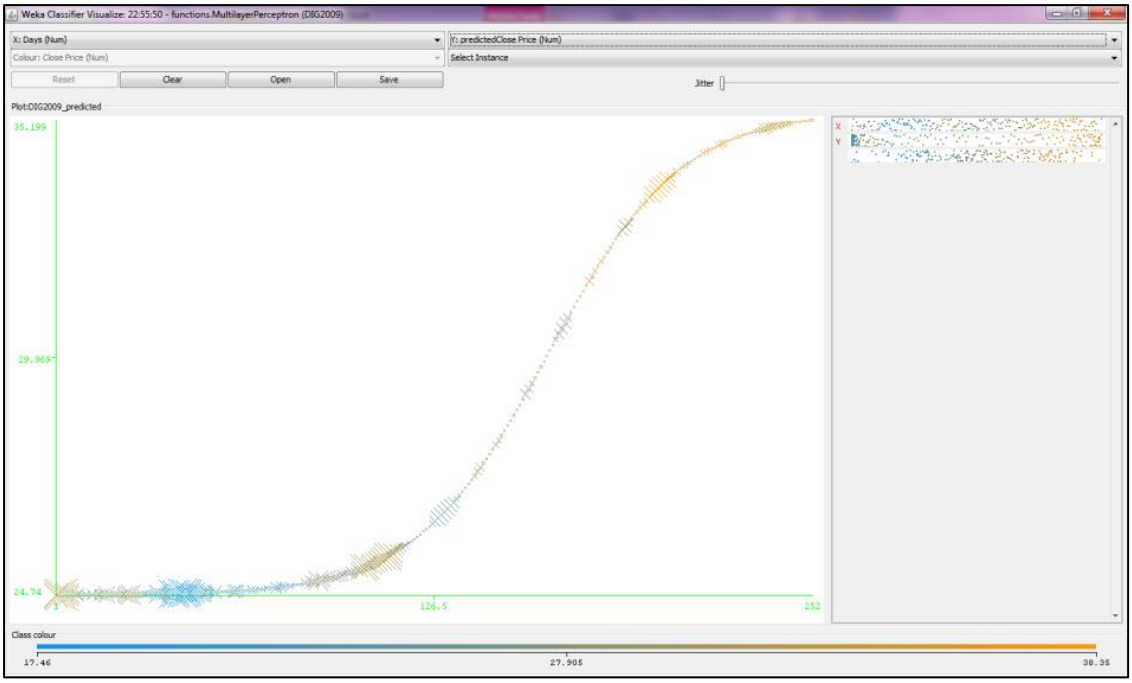

Figure 25: DIG 2009 Train Set using ANN

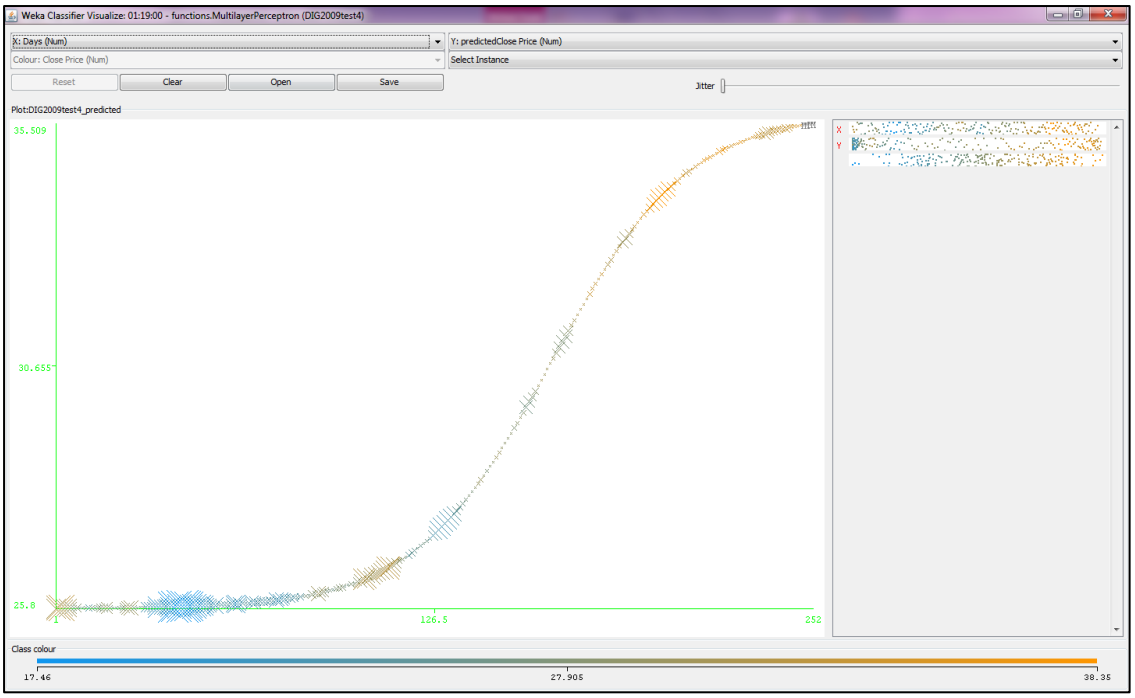

Figure 26: DIG 2009 Test Set using ANN

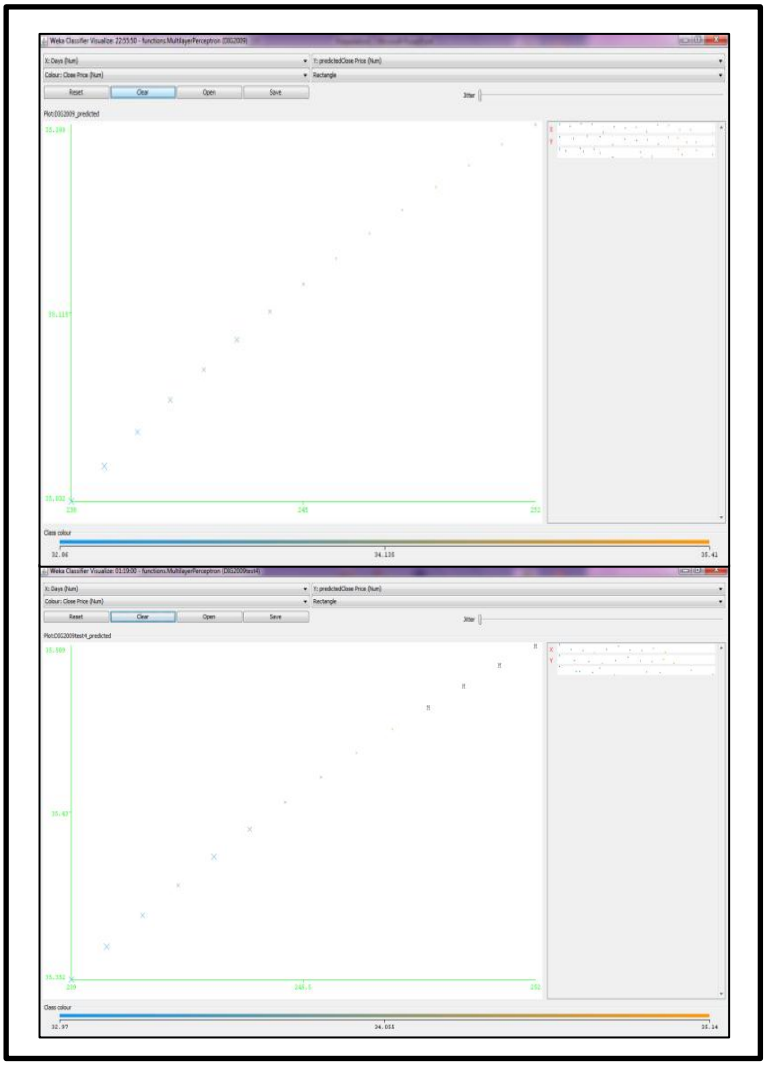

Figure 27: DIG 2009 using ANN (Train Set and Test Set)

Table 8: Mean Squared Error on DIG 2009 using ANN

Test Set	Train Set	Squared Error	Mean Squared Error
35.48014	35.17113	0.09548409	
35.49027	35.18083	0.095754351	
35.49998	35.19011	0.096022516	
35.50929	35.19899	0.096287331	**0.095887072**

For DIG 2009, the figures above show direct correlation with the mean squared error which is accuracy can be obtained using the ANN algorithm with the parameters set earlier.

The mean squared error for all the 10 Test Sets have been calculated and summarized in Table 9.

Table 9: Mean Squared Error on Test Sets using ANN

Data Name (Days-Data-Removed)	Mean Squared Error on Train Set with Actual Data	Mean Squared Error on Test Set with Actual Data	Mean Squared Error on Test Set with Train Set
AMD 2003 (30)	2.1861	18.0897	11.9609
AMD 2004 (30)	0.8755	67.9059	69.7087
AMD 2005 (20)	2.4233	9.6920	3.0550
AMD 2006 (20)	4.7669	1.3013	1.3101
AMD 2010 (10)	0.6633	0.7009	0.0005
AMD 2011 (10)	0.04476	0.0377	0.0021
DIG 2007 (7)	7.3759	9.3679	25.0201
DIG 2008 (7)	69.0051	184.4813	29.4730
DIG 2009 (4)	0.2976	0.6031	0.0959
DIG 2010 (4)	0.1149	0.0733	0.0230
AVERAGE	**8.775336**	**29.22531**	**14.06493**

In essence, the ANN inference engine is acceptably accurate with its test set having a 30 mean squared error. The lower the mean squared error, demonstrates that the more accurate the model for that particular Test Set. Within Table 3, the most accurate Test Set predictions are for the Test Sets with the 4 days data removed. This means that the GUI will only calculate 4 days of ANN prediction to give the best accurate result to the user to utilize. This has proved that the best ANN prediction is to for a short term period of 4 working days.

5.1.3.2 SVM INFERENCE ENGINE

The graphical interface is similar between the SVM and the ANN engine. The most important aspect of the SVM prediction results would be its accuracy which is measure by the mean squared error of the test set with the actual closing price. The same indicator for the test sets are used for the SVM inference engine. The summary of the mean squared error is shown in Table 10.

Table 10: Mean Squared Error on Test Sets using SVM

Data Name (Days-Data-Removed)	Mean Squared Error on Train Set with Actual Data	Mean Squared Error on Test Set with Actual Data	Mean Squared Error on Test Set with Train Set
AMD 2003 (30)	3.6553	15.7062	6.9386
AMD 2004 (30)	4.5466	10.3112	2.3588
AMD 2005 (20)	2.6333	8.5920	2.4412
AMD 2006 (20)	0.4313	0.4545	0.1293
AMD 2010 (10)	0.01229	0.019556	0.018386
AMD 2011 (10)	0.415634	0.047148	0.000831
DIG 2007 (7)	9.4979	35.9026	9.7576
DIG 2008 (7)	45.7077	93.6079	9.2542
DIG 2009 (4)	0.1775	0.1787	0.000457
DIG 2010 (4)	0.035911	0.109847	0.043817
AVERAGE	**6.711344**	**16.49297**	**3.0194319**

Comparing Table 9 and Table 10, the SVM algorithm has produced more accurate predictions rather than the ANN algorithm. The algorithms reacted similarly on the parameters for the Test Sets, with the 10-days-data-removed being the best category followed by the 4-days-data-removed. Furthermore, taking away 7 days of data produced the worst accuracy result for both of the algorithms.

As a conclusion, this accuracy test has proven that the inference engine is accurate and reliable to predict the stock price for a short period of time. This would be a stepping stone in developing the R GUI so that it will focus on the Test Sets that has been proven to be accurate for the algorithms. The most important aspect is to have huge loads of Training Data that will assist the system to learn the pattern of the data before implementing it to the Testing Data.

4.2 R GUI APPLICATION

4.2.1 FLOWCHART

The GUI implementation of the software uses the R programming language that is integrated with WEKA capabilities to run data mining algorithms and also capable of producing reports that are user-friendly. However, all systems must have a flowchart of its own and since this project requires two focal operations which are the inference engine and also the R GUI system, it must also have two flowcharts.

The inference engine within the R GUI application takes the same parameters during the testing of the algorithms within WEKA. The R GUI would take the exact same algorithms and implement the algorithms on the Training Data which are loaded into the system for the learning of the algorithms. The Test Data are loaded when the user gives input to the system that indicates the desired Test Data to be loaded into the algorithms.

The output would be in the means of graphical medium that would be easily understood by the users to utilize the prediction functions of the system. Figure 28 illustrates the flow of actions that the system will undergo.

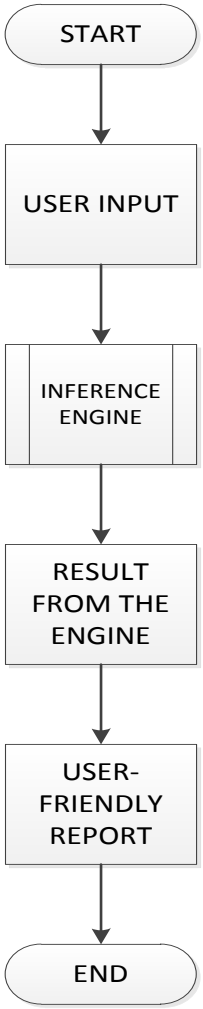

Figure 28: Flowchart of the GUI System for the Stock Prediction Application

4.2.2 PROTOTYPE SCREENSHOT

Basically, this is a simple GUI system to facilitate the user and to produce a desired output for their knowledge. Based on the accuracy test implemented prior to designing the GUI, the system has scoped down into two sets of Testing Data which are the 10 days and 4 days removed data sets. The inference engine will learn the algorithm automatically on start and the user input comes in terms of choosing the company to be used by the system. The system would then produce the output in terms of a graphical interface that would summarize the prediction system and also compare it with the actual stock price.

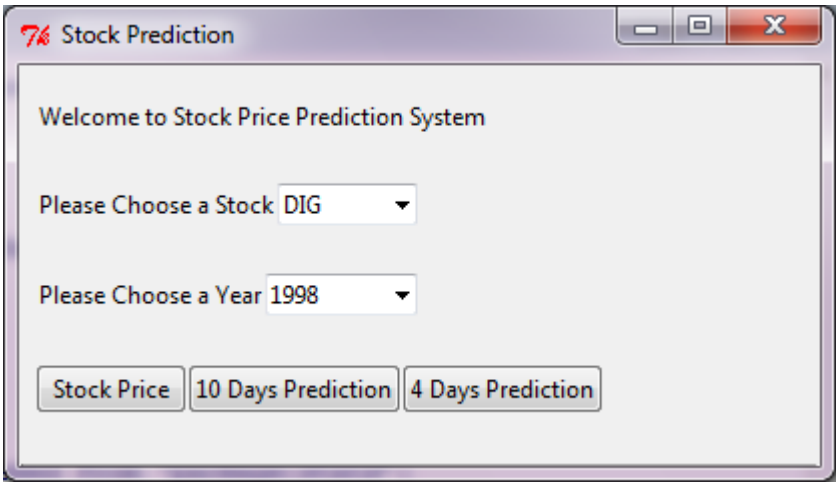

Figure 29: Main Window

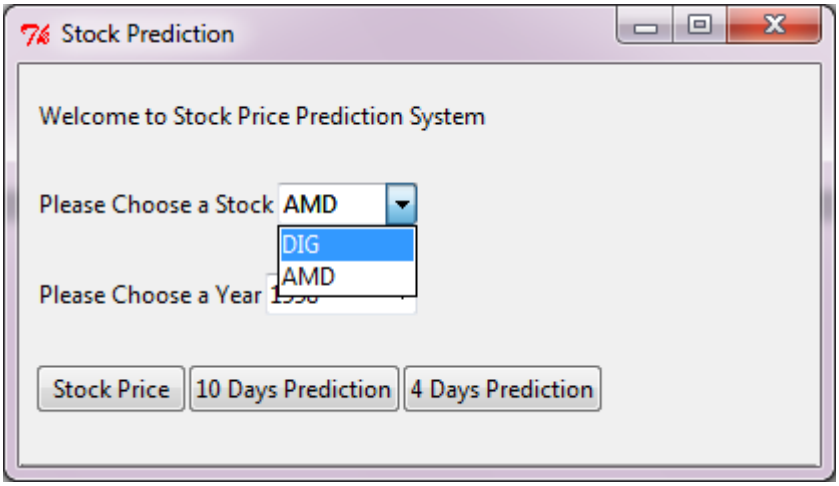

Figure 30: Combo box Functionality

There are three options the user can choose from that the system will work for them. The three options are ‘Stock Price’ button that allows the user to view the actual stock price, ’10 Days Prediction’ button which allow the user to view the prediction based on 10 days data removed and ‘4 Days Prediction’ button that implements the same function as the previous button.

The algorithms chosen for the prototype purpose is the ANN for the ‘4 Days Prediction’ button while SVM for the ’10 Days Prediction’ button. This is to indicate the ability of the system to run on both SVM and ANN at the same time by just a click of a button.

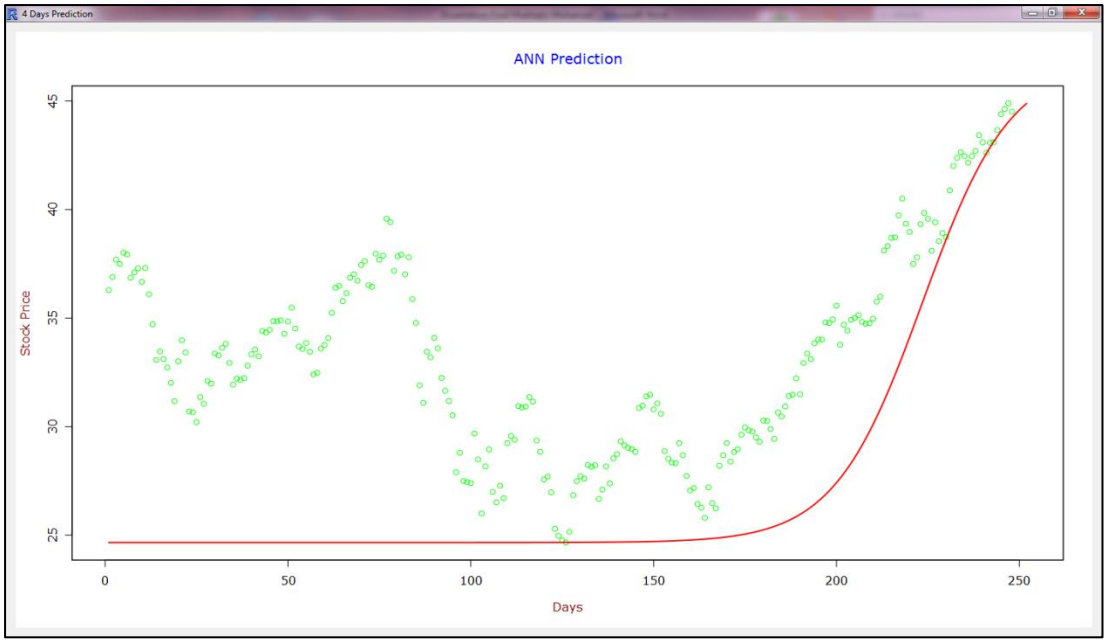

Figure 31: Results of ANN Algorithms

Figure 31 would occur when the '4 Days Prediction' button is clicked by the user. The company chosen is DIG and the year is 2010. The red line depicts the prediction while the green line depicts the Test Set data.

To compare the result with the stock price data, the 'Stock Price' button is clicked to get view the actual stock price. Figure 32 depicts the graphical interface of the 'Stock Price' button.

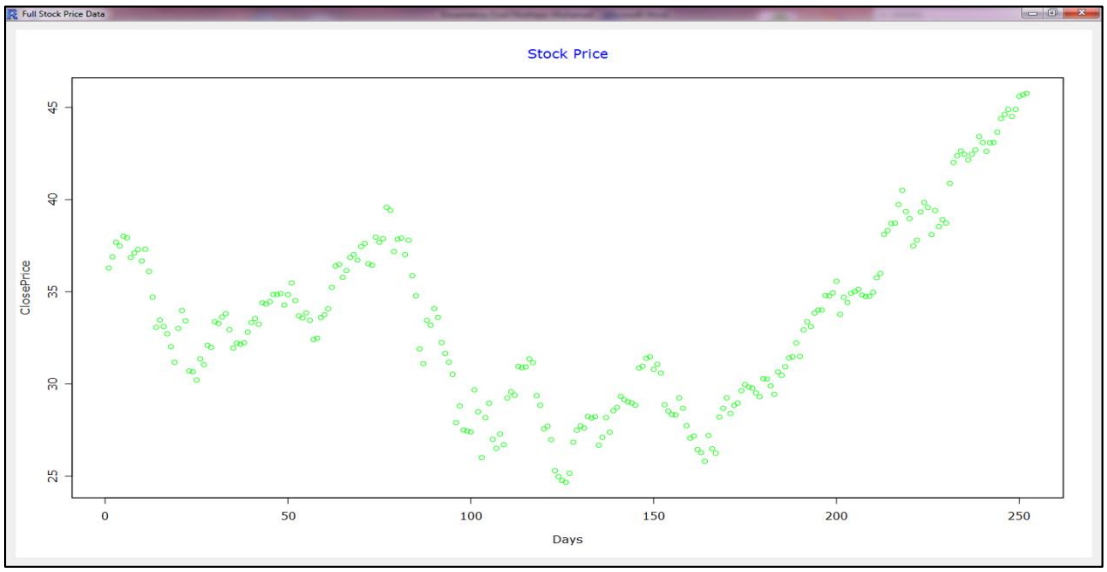

Figure 32: Original Stock Price Graph of DIG 2010

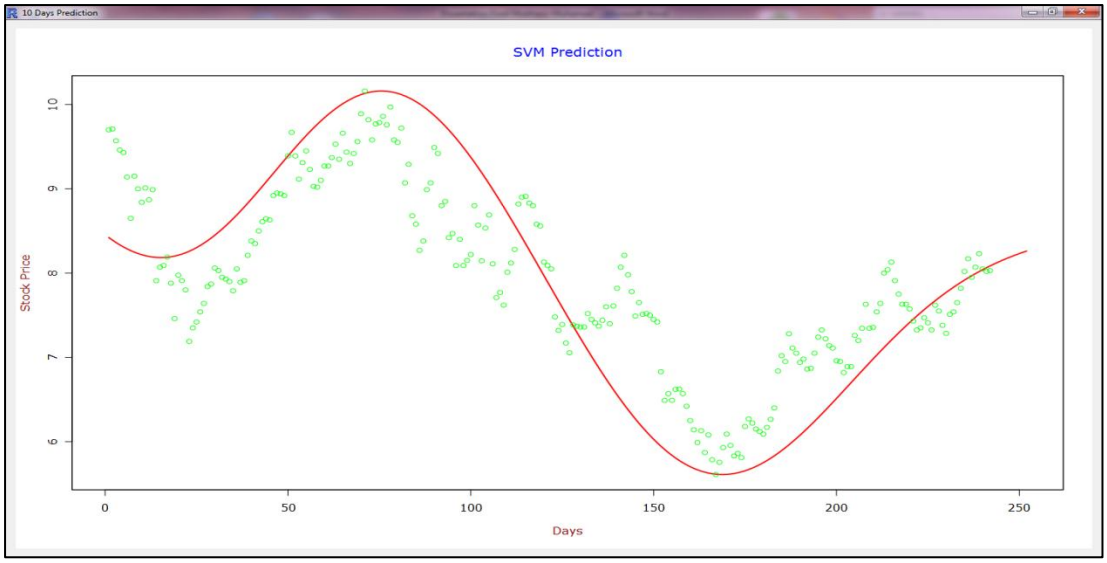

Figure 33: Results of SVM Algorithm

Figure 33 is the output from the button ’10 Days Prediction’ that runs the SVM algorithm and implements it in the 10-days-removed-data test sets available within the system. The company chosen is AMD 2010 and the red line depicts the prediction while the green dots depict the Test Set data.

Again, the user can compare the last 10 days prediction with the actual data in which depicted in Figure 34.

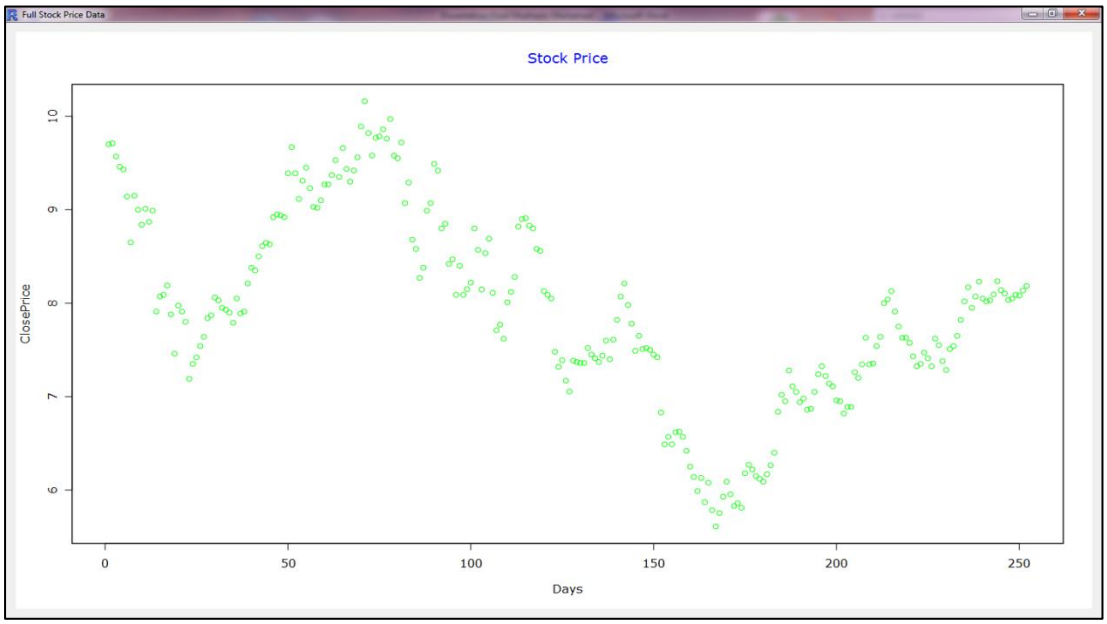

Figure 34: Original Stock Price Graph of AMD 2010

CHAPTER 5: CONCLUSION AND RECOMMENDATION

5.1 CONCLUSION

Data mining has emerged to be very important research area that provides solutions in a lot of aspect of organizations or even individuals. In the past, it might have been absolutely hard to predict the future stock price due to the lack of knowledge that most investors do not possess except for the analysts that always do the stockbroking reports. Data mining has given new powers to the investors to not just rely on the analysts for the reports alone.

This project has proposed a simple application that enables an analysis of the data mining tools of to predict the future stock price. Using the data mining approaches, the model is built with adherence to accuracy and also reliability so that the results can further expanded into a full-pledged system. Using ANN and SVM algorithms to complete the task, it seems like the prediction powers of the algorithms have been tested and proven reliable albeit on a short term prediction.

In conclusion, the data mining approach to predict stock prices accurately, precisely and to maintain its reliability has achieve its main objectives. The prediction power of the inference engine has been proven to be reliable and able to convince its users to use the software for stock price prediction.

5.2 RECOMMENDATION

Future works related would be to further develop the GUI to make it more interactive to the end users. The addition of functions that would allow users to add in personal stock price data would further create a more user-friendly application. The basis of the inference engine and basic connection between the inference engine and the front-end has been developed therefore, future works should focus on the usability of the application.

REFERENCES

Boswell, D. (2002, August 6). Introduction to Support Vector Machine.

Castro, J., & Mylopoulos, J. (2002). Information System Analysis and Design. *Journal of Feasibility 3*.

Chang, P.-C., & Liu, C.-H. (2008). A TSK type fuzzy rule based system for stock price prediction. *Expert Systems with Application 34*, 135-144.

Chang, P.-C., Wang, D.-d., & Zhou, C.-l. (2012). A novel model by evolving partially connected neural network for stock price trend forecasting. *Expert Systems with Applications 39*, 611-620.

Cheung, V., & Canons, K. (2002, May 27). An Introduction to Neural Networks. *The Notes on Neural Networks*. Winnipeg, Manitoba, Canada: University of Manitoba.

Credit Suisse. (2010). *Global Technical Research & Behavioral Finance: Credit Suisse.* Retrieved April 1, 2012, from Credit Suisse Web Site: https://www.credit-suisse.com/legal/pb_research/technical_tutorial_en.pdf

Esfahanipour, A., & Aghamiri, W. (2010). Adapted Neuro-Fuzzy Inference System on indirect approach TSK fuzzy rule base for stock market analysis. *Expert Systems with Applications 37*, 4742-4748.

Gitman, L., Joehnk, M., & Smart, S. (2011). *Fundamentals of Investing.* Boston: Pearson Education Inc.

Hadavandi, E., Shavandi, H., & Ghanbari, A. (2010). Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. *Knowledge-Based Systems 23*, 800-808.

Hellman, M. (n.d.). *Fuzzy Logic Introduction.* Retrieved March 2012, from School of Electronic Engineering, Xidian University Webpage: see.xidian.edu.cn/faculty/xbgao/FuzzySystem/Introduction/fuzzy.pdf

IPredict. (2012, January). *Resources: iPredict*. Retrieved February 2012, from iPredict: http://www.ipredict.it/TimeSeriesForecasting.aspx

Kim, K.-j. (2003). Financial time series forecasting using support vector machines. *Neurocomputing 55*, 307-319.

Kim, K.-j., & Han, I. (2000). Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index. *Expert Systems with Applications 19*, 125-132.

Landt, F. O. (1997, August 4). Stock price prediction using neural networks. *Master Thesis*. Leiden University.

MIT Computer Science and Arficial Intelligence Laboratory. (n.d.). *Non-linear SVM separation*. Retrieved July 27, 2012, from DDMG: Data Driven Medicine Group: http://groups.csail.mit.edu/ddmg/drupal/content/non-linear-svm-separation

Roh, T. H. (2007). Forecasting the volatility of stock price index. *Expert Systems with Applications 33*, 916-922.

Saad, E. W., Prokhorov, D. V., & Wunsch. II, D. C. (1998). Comparative Study of Stock Trend Prediction Using Time Delay, Recurrent and Probabilistic Neural Networks. *IEEE TRANSACTIONS ON NEURAL NETWORKS 9*.

Tan, P.-n., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining.* Boston: Pearson Education Inc.

Tay, F. E., & Cao, L. (2001). Application of support vector machines in financial time series forecasting. *Omega 29*, 309-317.

Wang, Y.-F. (2002). Predicting stock price using fuzzy grey prediction system. *Expert Systems with Applications 22*, 33-39.

wikipedia.org. (2011, July 12). *Time Series*. Retrieved February 2012, from Wikipedia: The Free Encyclopedia: http://en.wikipedia.org/wiki/Time_series

APPENDICES

Appendix A

Date	Time	Open	High	Low	Close	Volume
01-02-98	9:32	9.12	9.12	9.12	9.12	51630
01-02-98	9:33	9.12	9.12	9.12	9.12	19210
01-02-98	9:34	9.12	9.12	9.12	9.12	11210
01-02-98	9:35	9.12	9.12	9.12	9.12	9400
01-02-98	9:36	9.15	9.15	9.15	9.15	11610
01-02-98	9:37	9.15	9.185	9.15	9.15	15810
01-02-98	9:38	9.185	9.215	9.15	9.185	61030
01-02-98	9:39	9.215	9.215	9.185	9.185	1200
01-02-98	9:40	9.215	9.215	9.185	9.185	10610
01-02-98	9:41	9.215	9.215	9.185	9.215	8400
01-02-98	9:42	9.215	9.215	9.215	9.215	8200
01-02-98	9:43	9.215	9.245	9.215	9.245	3400
01-02-98	9:44	9.215	9.215	9.215	9.215	12810
01-02-98	9:45	9.215	9.215	9.215	9.215	11210
01-02-98	9:46	9.245	9.245	9.245	9.245	6000
01-02-98	9:47	9.245	9.245	9.245	9.245	20210
01-02-98	15:51	9.65	9.65	9.62	9.62	9970
01-02-98	15:52	9.62	9.65	9.62	9.65	3390
01-02-98	15:53	9.65	9.65	9.65	9.65	6180
01-02-98	15:54	9.65	9.65	9.65	9.65	1400
01-02-98	15:55	9.62	9.62	9.59	9.59	26540
01-02-98	15:56	9.62	9.62	9.59	9.59	2390
01-02-98	15:57	9.59	9.62	9.59	9.59	5980
01-02-98	15:58	9.62	9.62	9.59	9.62	6980
01-02-98	15:59	9.59	9.65	9.59	9.65	18150
01-05-98	9:33	9.68	9.68	9.645	9.68	311520
01-05-98	9:34	9.647	9.682	9.647	9.647	8810
01-05-98	9:35	9.683	9.713	9.648	9.683	41630
01-05-98	9:36	9.745	9.745	9.685	9.685	57230
01-05-98	9:37	9.745	9.745	9.65	9.65	30620
01-05-98	9:38	9.684	9.684	9.649	9.649	38420
01-05-98	9:39	9.684	9.744	9.619	9.649	26420
01-05-98	9:40	9.618	9.648	9.559	9.559	46430
01-05-98	9:41	9.588	9.683	9.558	9.683	33220
01-05-98	9:42	9.558	9.618	9.558	9.558	26620
01-05-98	9:43	9.587	9.587	9.557	9.557	6810
01-05-98	9:44	9.557	9.557	9.522	9.522	13410
01-05-98	9:45	9.522	9.522	9.522	9.522	1000

Appendix B

Days	Close Price
1	9.65
2	9.65
3	9.921
4	9.5
5	9.47
6	9.12
7	8.88
8	10.018
9	9.22
10	9.03
123	8.38
124	8.38
125	8.235
126	9
127	9.06
128	7.94
129	7.807
130	7.72
131	8.161
161	8.5
162	7.625
163	7.59
164	6.5668
165	7.204
166	7.355
167	7
168	7.09
169	7.585
240	13.935
241	14.03
242	14.44
243	14.38
244	13.85
245	14
246	13.65
247	13.75
248	14.545
249	14.562

Appendix C

Train Set Data	
Days	ClosePrice
1	9.7
2	9.71
3	9.57
4	9.46
5	9.43
6	9.14
7	8.65
8	9.15
240	8.05
241	8.02
242	8.03
243	8.095
244	8.235
245	8.14
246	8.105
247	8.035
248	8.05
249	8.09
250	8.08
251	8.135
252	8.185

Test Set Data 10 Days	
Days	ClosePrice
1	9.7
2	9.71
3	9.57
4	9.46
5	9.43
6	9.14
7	8.65
8	9.15
240	8.05
241	8.02
242	8.03
243	?
244	?
245	?
246	?
247	?
248	?
249	?
250	?
251	?
252	?

Appendix D

No.	Detail/Week	1	2	3	4	5	6	7		8	9	10	11	12	13	14
1	Selection of Project Topic - Selection of supervisor								M I D							
2	Proposal Submission															
3	Approval of Proposal Submission								S							
4	Reseach Class								E							
5	Preliminary Research Work - Data Gathering - Requirement Gathering								M							
6	Submission of Extended Proposal								B							
7	Proposal Defense								R							
8	Project Work Continues								E							
9	Submission of Interim Draft Report								A							
10	Submission of Interim Report								K							

No.	Detail/ Week	1	2	3	4	5	6	7		8	9	10	11	12	13	14	15
1	Project Work Continues								Mid								
2	Submission of Progress Report								Sem								
3	Project Work Continues								B								
4	Pre-EDX								R								
5	Submission of Draft Report								E								
6	Submission of Dissertation (soft bound)								A								
7	Submission of Technical Paper								K								
8	Oral Presentation																
9	Submission of Project Dissertation (Hard Bound)																

Figure 35: Gantt chart for FYP