# STOCK MARKET ANALYSIS USING DATA MINING TECHNIQUES

AYUSH SINGH
COMPUTER SCIENCE
SRMIST
CHENNAI, INDIA

## ABSTRACT

*Forecasting stock return has been an important financial subject that has found researchers' attention for many years. It involves an idea that basic information publicly available in the past has some tentative relationships to the future stock returns. This data helps the investor in the stock market to decide the better timing for buying or selling stocks based on the knowledge extracted from the existing prices of such stocks. The decision thus made will be based on decision tree classifier which is an important data mining techniques. To construct the ideal model, the CRISP-DM method is used over real pre-existent data of three major companies listed in Amman Stock Exchange (ASE).*

***Keywords***: *Data Mining, Data Mining, Data Classification, Decision Tree, Future stock return, data mining techniques, decision tree classifiers, CRISP-DM methodology, Amman Stock Exchange.*

## 1.INTRODUCTION

The stock market is a non-linear, non- parametric system that is very difficult to model with any authentic accuracy. Investors have been searching for a way to predict stock prices and to look for the right stocks at right time to buy or sell. To achieve such goals, and according to, some theories used the methods of basic analysis, where trading rules are introduced based on the factors such as macroeconomics, industry, and company. The authors of and said that fundamental analysis assumes that the price of a stock depends on its intrinsic value and expected return on investment. By making a detailed record about the company's operations and the market in which the company is operating can do this. Eventually, the stock price can be predicted using the required data. People mostly think that fundamental analysis is a fruitful method only on a long-term investment plan. However, for short- and medium- term speculations, fundamental analysis is generally not suitable.

Some other research utilized the methods of technical analysis, in which trading rules were based on the existential data of stock trading price and volume. Technical analysis as illustrated in and refers to the different numerous methods that aim to predict the future price displacement using past stock prices and volume information. It is dependent on the assumption that history repeats itself and that future market directions can be determined by examining existing price data. Thus, it is assumed that price trends and patterns exist that can be identified and utilized for profit. Most of the techniques used in technical analysis are highly subjective in nature and have been shown not to be statistically valid.

Currently, data mining techniques and artificial intelligence techniques like decision trees, rough set approach, and artificial neural networks have been applied to this area. Data mining stands for extracting or mining knowledge from large data stores or sets. A few of the functionalities are the discovery of concept or class descriptions, associations and correlations, classification, prediction, clustering, trend analysis, outlier and deviation analysis, and similarity analysis. Data segregation can be done in many different methods; one of those methods is the classification by using Decision Tree. It is a graphical representation of all possible outcomes and the paths by which they may be reached.

Decision trees and artificial neural networks can be implemented by using an appropriate learning algorithm. Following the assumption of technical analysis that co-relation exists in price data, it is possible in principle to implement data mining techniques to discover these patterns in an automated manner. Once these patterns have been found, future prices can easily be predicted.Today, the grand challenge of employing a database is to get useful rules from data during a database for users to form decisions, and these rules could also be hidden deeply

within the data of the database. Traditionally, the tactic of turning data into knowledge relies on manual analysis; this is often becoming impractical in many domains as data volumes grow exponentially. the matter with predicting stock prices is that the quantity of knowledge is just too large and large . This paper uses one among the info mining methods; which is that the classification approach on the historical data available to undertake to assist the investors to create their decision on whether to shop for or sell that stock so as to realize profit.

The main objective of this paper is to research the historical data available on stocks using decision tree technique together of the classification methods of knowledge mining so as to assist investors to understand when to shop for new stocks or to sell their stocks.

Analyzing stock price data over several years may involve a couple of hundreds or thousands of records, but these must be selected from millions. the info which will be utilized in this paper to create the choice tree are going to be the historical prices of three listed companies in Amman stock market over two years of your time .The remainder of this paper has been divided into four sections. Section 2 of the paper gives a literature review about the topic of using data processing techniques so as to undertake to predict the costs and therefore the trend of stocks, some related add that subject is shown during this section. Section 3 talks about the methodology utilized in building the classification model. Then section 4 shows the experiments that are done on the info collected using the model and evaluation of the results using one among the evaluation methods. Finally, a quick conclusion and therefore the future work about the subject is given in section 5.

## 2. LITERATURE REVIEW

Over the past 20 years many important changes have taken place within the environment of monetary markets. the event of powerful communication and trading facilities has enlarged the scope of selection for investors. Forecasting stock return is a crucial financial subject that has attracted researchers' attention for several years. It involves an assumption that fundamental information publicly available within the past has some predictive relationships to the longer term stock returns. so as to be ready to extract such relationships from the available data, data processing techniques are new techniques which will be wont to extract the knowledge from this data.

For this reason, numerous researchers have focused on technical analysis and used advanced math and science. Extensive attention has been dedicated to the sector of AI and data processing techniques. A few methods are inducted and implemented using the given mentioned techniques, the authors have made an empirical study on building a stock buying/selling alert system using back propagation neural networks (BPNN), their NN was codenamed NN5. The system was trained and tested with past price data from Hong Kong and Shanghai Banking Corporation Holdings over the amount from January 2004 to December 2005. The empirical results showed that the implemented system was ready to predict short-term price movement directions with accuracy about 74%.

The research by used decision tree technique to create on the work of Lin where Lin tried to switch the filter rule that's to shop for when the stock price rises k% above its past local low and sell when it falls avoid its past local high. The proposed modification to the filter rule out was by combining three decision variables related to fundamental analysis. An empirical test, with the help of the stocks of electronic companies in Taiwan, showed that Lin's method outperformed the filter rule consistent with , in Lin's work, the standards for clustering trading points involved only the existing information; the longer term information wasn't considered in the least . The research by aimed to enhance the filter rule and Lin's study by considering both the past and therefore the future information in clustering the trading points. The researchers used the info of Taiwan stock exchange which of NASDAQ to hold out empirical tests. Test results showed that the proposed method outperformed both Lin's method and therefore the filter rule out the 2 stock markets.

The model of applied the concept of serial topology and designed a replacement decision system, namely the two- layer bias decision tree, for stock price prediction. The methodology developed by the authors differs from other studies in two respects; first, to scale back the classification error, the choice model was modified into a bias decision model. Second, a two-layer bias decision tree is employed to enhance purchasing accuracy. The empirical results have indicated that the present decision model produced an excellent purchasing accuracy, and it significantly over exceeds than random purchase.

The authors of presented an approach that used data processing methods and neural networks for forecasting stock exchange returns. an effort has been made during this study to research the predictive power of monetary and economic variables by adopting the variable relevance analysis technique in machine learning for data processing .

The authors around the globe have examined the usefulness of the neural network models used for level estimation and classification. The result shows that the trading methods implemented by the neural network classification models produced higher profits under an equivalent risk exposure than those suggested by other strategies.

The research by was basically a comparison between the work of Fama and French's model and therefore the artificial neural networks so as to undertake to predict the stock prices within the Chinese market. the aim of this study is to demonstrate the accuracy of ANN in predicting stock price movement for firms traded on the Shanghai stock market . so as to demonstrate the accuracy of ANN, the authors made a comparative analysis between Fama and French's model and therefore the predictive power of the univariate and multivariate neural network models. The results from this study identifies that artificial neural networks offer a significant opportunity for investors to enhance their prediction efficieny in selecting stocks, and more importantly, an easy univariate model appears to be more successful at predicting returns than a multivariate model.

Al-Haddad et al., presented a study that aimed to supply evidence of whether or not the company governance & performance indicators of the Jordanian industrial companies listed at Amman stock market (ASE) are suffering from variables that were proposed and to supply the important indicators of the connection of corporate governance & firms' performance which will be employed by the Jordanian industrial firms to unravel the agency problem. The study consists of random samples of different (44) Jordanian industrial firms. The study founds a positive direct relationship between corporate governance and company performance.

Hajizadeh et al. provided an summary of application of knowledge mining techniques like decision tree, neural network, association rules, and correlational analysis and available markets.

Prediction stock price or financial markets has been one among the most important challenges to the AI community. Various technical, fundamental, and statistical indicators are proposed and used with varying results. Soni surveyed some recent literature within the domain of machine learning techniques and AI wont to predict stock exchange movements. Artificial Neural Networks (ANNs) are identified to be the dominant machine learning technique available market prediction area.

El-Baky et al., proposed a replacement approach for fast forecasting of stock exchange prices. The proposed method uses a new high speed time delay neural networks (HSTDNNs). The authors used the MATLAB tool to simulate results to verify the theoretical computations of the approach.

## 3.THE METHODOLOGY OF THE STUDY

Data mining methodology is meant to make sure that the info mining effort results in a stable model that successfully addresses the matter it's designed to unravel . Various data processing methodologies are proposed to function blueprints for a way to arrange the method of gathering data, analyzing data, disseminating results, implementing results, and monitoring improvements. to create the model that analyses the stock trends using the choice tree technique, the CRISP-DM (Cross- Industry Standard Process for data mining) is employed . this system was proposed within the mid-1990s by an eu consortium of companies to function a non-proprietary standard process model for data processing . This model consists of the subsequent six steps:

- Understanding the rationale and objective of mining the stock prices.
- Understanding the collected data and the way it's structured.
- Preparing the info that's utilized in the classification model.
- Selecting the technique to create the model.
- Evaluating the model by using one among the documented evaluation methods.
- Deploying the model within the stock exchange to predict the simplest action to be taken, either selling or buying the stocks.
- Understanding the rationale and objective of building the model

The main reason and objective of building the model is to undertake to assist the investors within the stock exchange to make a decision the simplest timing for purchasing or selling stocks

supported the knowledge extracted from the historical prices of such stocks. The decision taken are going to be supported one among the info mining techniques; the choice tree classifiers.

***Understanding the collected data***

The Oracle database of Amman stock market (ASE) contains the historical prices of the 230 companies listed within the exchange from the year 2000. because the amount of such data is extremely large and sophisticated , the choice was taken to settle on three companies listed within the exchange. the choice of those companies was supported the subsequent five criteria which represent the companies' size and liquidity: market capitalisation , days traded, turnover ratio, value traded and therefore the number of shares traded, also the world representation was considered during the choice of those companies. These companies are "Arab Bank", its' code within the stock exchange "ARBK" and it is a part of the banking sector, "United Arab Investors Company", its' code is "UAIC" which belongs to the services sector, and "Middle East Complex for Engineering, Electronics and Heavy Industries", whose code is "MECE" which is a part of the economic sector. the amount that was selected is from April 2005 to May 2007, which presented the present and actual status of the market at that period of your time .

At the start , the info collected contained 9 attributes; this number was reduced manually to six attributes because the other attributes were found not important and not having an immediate effect on the study. Table1 shows the 6 attributes selected with the respective descriptions and the possible outcomes. the category attribute is that the investor action whether to shop for or sell that stock and it's named, "Action". the info of this attribute was taken also from ASE database, which is that the net position of 1 of the most important brokers handling the above mentioned stocks a day . internet position might be either buying or selling that stock for that day.

***Preparing the data***

At the start, when the info was collected, all the values of the attributes selected were continuous numeric values. Data transformation is utilized by generalizing data to a higher-level concept so as all the values become discrete. The criterion that was made to rework the numeric values of every attribute to discrete values trusted the previous day price of the stock. If the values of the attributes open, min, max, last were greater than the worth of attribute previous for an equivalent trading day, the numeric values of the attributes were replaced by the worth Positive. If the worth s of the attributes mentioned above were but the value of the attribute previous, the numeric values of the attributes were replaced by Negative. If the worth s of these attributes were adequate to the value of the attribute previous, the worth s were replaced by the value Equal. Table 2 shows a sample of the continual numeric values of the info before selecting the 6 attributes manually and before generalizing them to discrete values, while table3 shows an equivalent sample after selecting the 6 attributes and after transforming them to discrete values.

***Building the model***

After the info has been prepared and transformed, subsequent step was to create the classification model using the choice tree technique. the choice tree technique was selected because the development of decision tree classifiers doesn't require any domain knowledge, thus it's appropriate for exploratory knowledge discovery. Also, it can handle high dimensional data. Another usefulness is that the steps of decision tree induction are quite simple and fast. Generally, decision tree accuracy is taken into account good. the choice tree method depends on using the knowledge gain metric that determines the foremost useful attribute. the knowledge gain depends on the entropy measure.

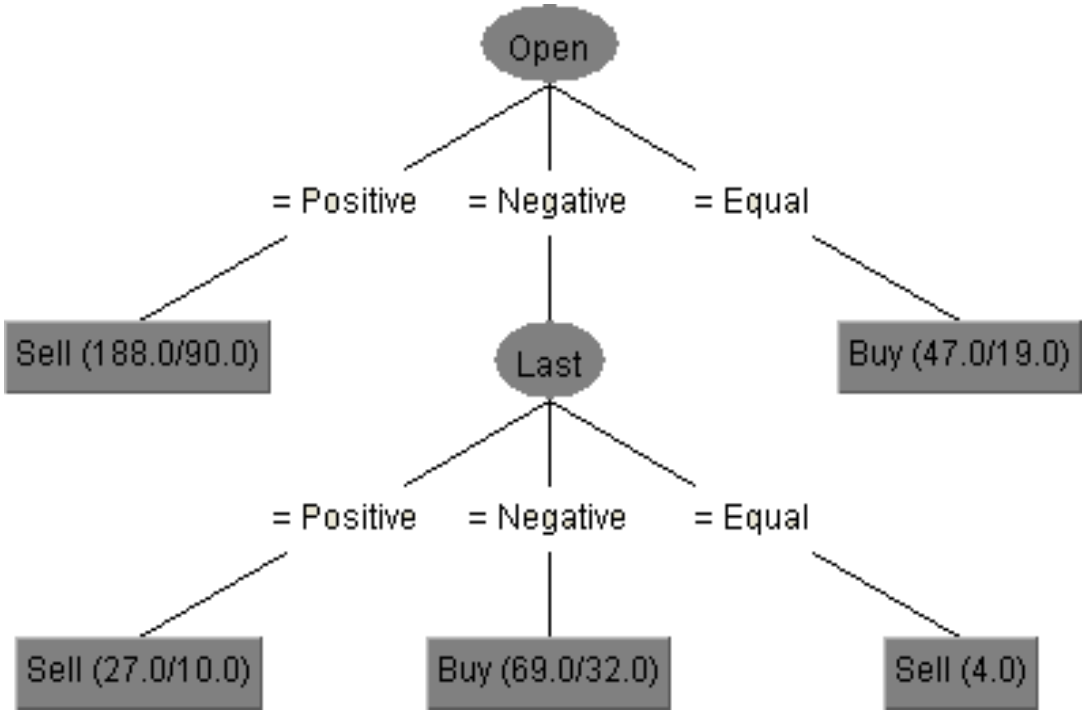
| Previous | Open | Max | Min | Last | Action |
|---|---|---|---|---|---|
| Positive | Positive | Positive | Negative | Negative | **Sell** |
| Negative | Positive | Positive | Negative | Negative | **Buy** |
| Negative | Negative | equal | Negative | Negative | **Buy** |
| Negative | Negative | equal | Negative | Negative | **Sell** |
| Negative | equal | Positive | Negative | Positive | **Buy** |
| Positive | Negative | Positive | Negative | Positive | **Buy** |
| Positive | Positive | Positive | Positive | Positive | **Buy** |
| Positive | equal | Positive | Negative | Negative | **Buy** |
| Negative | Positive | Positive | Negative | Negative | **Sell** |

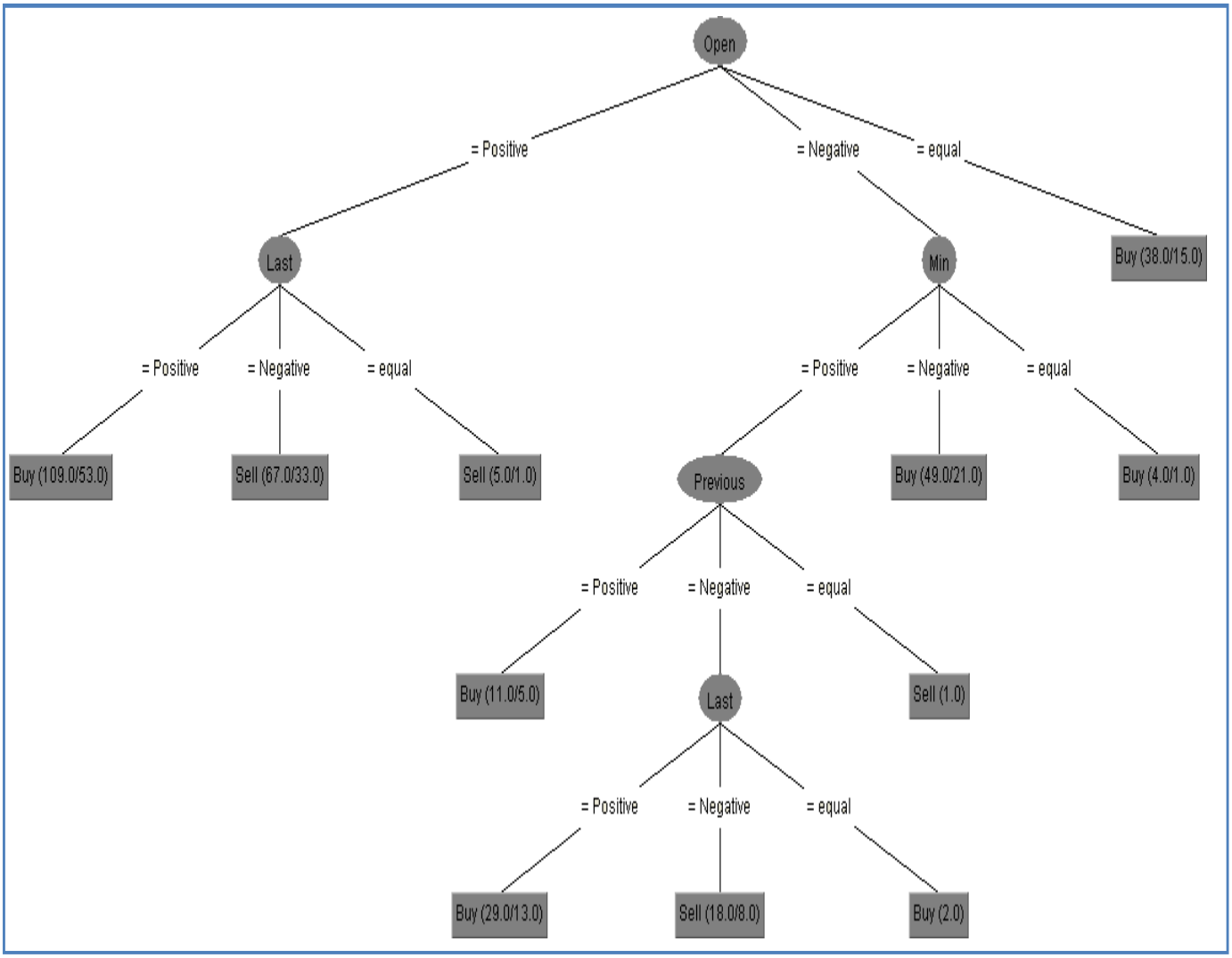
| Previous | Open | Max | Min | Last | Action |
|---|---|---|---|---|---|
| 25.82 | 25.99 | 26 | 25.41 | 25.67 | Sell |
| 25.67 | 25.68 | 25.68 | 25.2 | 25.3 | Buy |
| 25.3 | 24.8 | 25.3 | 24.41 | 24.9 | Buy |
| 24.9 | 24.8 | 24.9 | 24.3 | 24.87 | Sell |
| 24.87 | 24.87 | 25.55 | 24.85 | 25.3 | Buy |
| 25.3 | 25.25 | 26 | 25.25 | 25.82 | Buy |
| 25.82 | 25.99 | 26.4 | 25.99 | 26.3 | Buy |
| 26.3 | 26.3 | 26.7 | 26 | 26.02 | Buy |
| 26.02 | 26.09 | 26.25 | 25.55 | 25.63 | Sell |

The gain ratio is employed to rank attributes and to create the choice tree where each attribute is found consistent with its gain ratio. When the choice tree model was applied on the info of the three companies using the WEKA software version 3.5, the basis attribute for both ARBK and UAIC company was the Open, while the attribute Last was the basis for the choice tree of the MECE company. because the process of building the tree goes on, all the remaining attributes were wont to continue with this process. After building the entire decision tree, the set of classification rules were generated by following all the paths of the tree. the utmost number of attributes that were utilized in a number of the classification rules generated were 4 attributes, while some classification rules used just one attribute. Both the ID3 and C4.5 algorithms were utilized in building the choice trees and therefore the pruning technique was utilized in the C4.5 algorithm so as to scale back the dimensions of the produced decision trees. Table 4 gives a summary about the numbers of the classification rules that resulted after building the choice trees for every company using the C4.5 algorithm.

| Company | Number of classification rules without pruning | Number of classification rules with pruning |
|---|---|---|
| **ARBK** | 21 | 11 |
| **UAIC** | 31 | 5 |
| **MECE** | 21 | 9 |

| Attribute | Description | Possible Values |
|---|---|---|
| Previous | Previous day close price of the stock | Positive, Negative, Equal |
| Open | Current day open price of the stock | Positive, Negative, Equal |
| Min | Current day minimum price of the stock | Positive, Negative, Equal |
| Max | Current day maximum price of the stock | Positive, Negative, Equal |
| Last | Current day close price of the stock | Positive, Negative, Equal |
| **Action** | The action taken by the investor on this stock | Buy, Sell |






***Deploying the model***

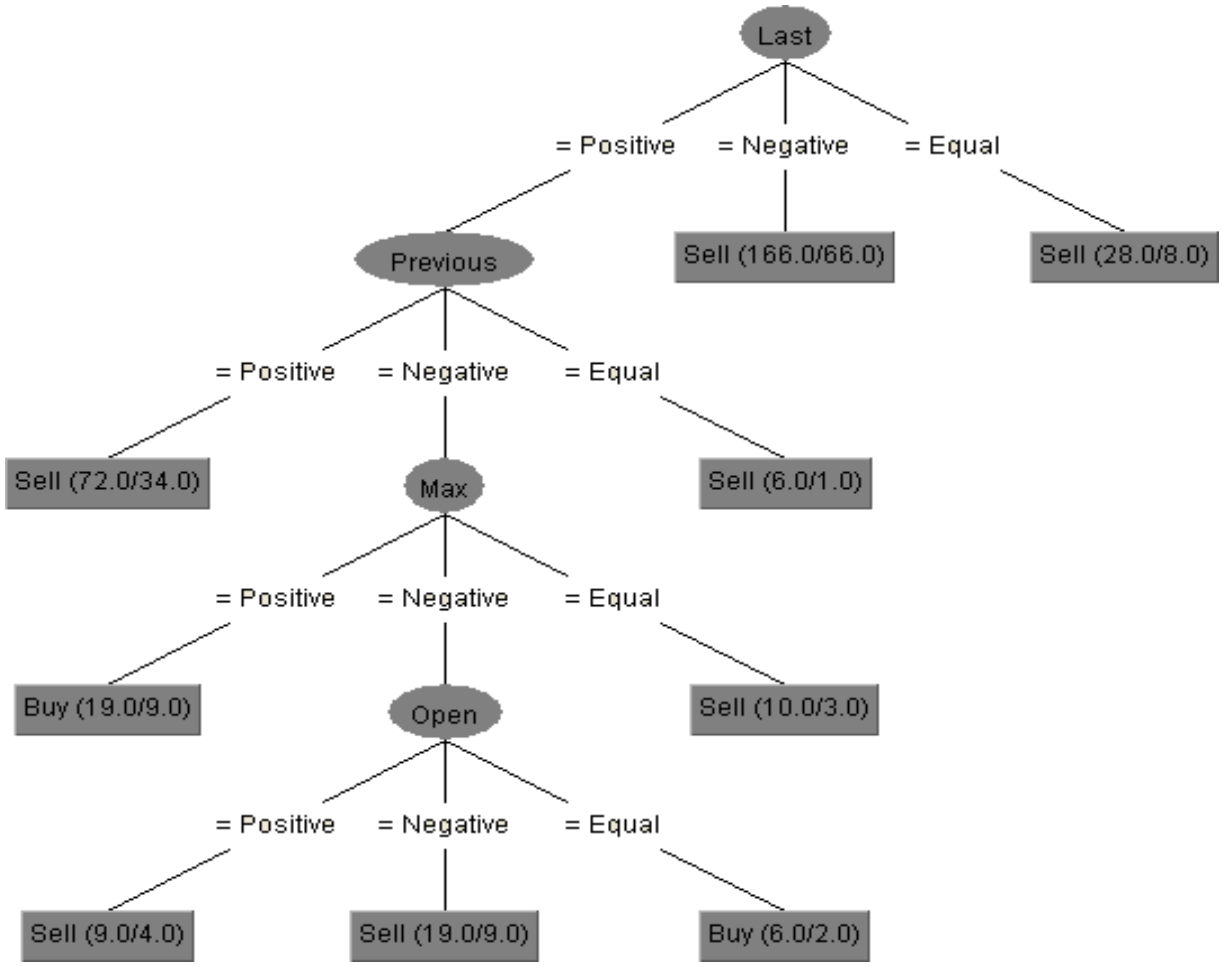
The classification rules that were generated from the choice tree model are often used and integrated during a system that predict the simplest action and timing for the investors, either to shop for or sell the stocks thereon day.

# 4. RESULTS AND DISCUSSION

This section presents step 5 of the CRISP methodology which was wont to build the model. it's simply about evaluating the model by using one or more of the documented evaluation methods. so as to guage the model, the WEKA software was wont to calculate the accuracy of the classification model. Two evaluation methods were used, the K-Fold Cross Validation (K-CV) where K=10 folds and therefore the refore the percentage split method where 66% of the info was used for training and the remainder for testing. Both of the evaluation methods were used on the ideology of ID3 and C4.5 decision tree classification methods. Table 5 shows us the accuracy of all the classifiers created using both of the classification methods and both evaluation methods.

As we will see from the table, the resultant classification accuracy from the choice tree model isn't very high for the training data used and it varies from one company to a different . the rationale for such a coffee accuracy is that the corporate 's performance within the stock exchange is suffering from internal financial factors such as; news about the company, financial reports, and therefore the overall performance of the market. Also, external

factors can affect the performance of the corporate within the market such as; political events and political decisions. Thus, it are often difficult to possess a model that provides a high accuracy classification for all the businesses at an equivalent time because the performance of those companies differs.




| | | 10-CV | | | Holdout 66% | | |
|---|---|---|---|---|---|---|---|
| **Company** | **Classification Method** | **Total Instances** | **Correctly classified** | **Accuracy %** | **Total Instances** | **Correctly classified** | **Accuracy %** |
| **ARBK** | **ID3** | 499 | 233 | 44.689 | 170 | 73 | 42.941 |
| | **C4.5** | | 237 | 47.495 | | 83 | 48.824 |
| **MECE** | **ID3** | 502 | 255 | 50.797 | 171 | 84 | 49.123 |
| | **C4.5** | | 265 | 52.789 | | 91 | 53.216 |
| **UAIC** | **ID3** | 502 | 269 | 53.586 | 171 | 88 | 51.462 |
| | **C4.5** | | 264 | 52.590 | | 94 | 54.971 |

# 5.CONCLUSIONS

This study presents a proposal to use the choice tree classifier on the historical prices of the stocks to make decision rules that give buy or sell recommendations within the stock exchange. Such proposed model are often a helpful tool for the investors to require the proper decision regarding their stocks supported the analysis of the historical prices of stocks so as to extract any predictive information from that historical data. The results for the proposed model weren't perfect because many factors including but not limited to political events, general economic conditions, and investors' expectations influence stock exchange.

As for the longer term work, there's still big room for testing and improving the proposed model by evaluating the model over the entire companies listed within the stock exchange. Also, the evaluation of a bigger collection of learning techniques like neural networks, genetic algorithms, and association rules can represent an upscale area for future investigation. Finally, reconsidering the factors affecting the behavior of the stock markets, like trading volume, news and financial reports which could impact stock price are often another rich field for future studying.

## REFERENCES

[1] Wang, Y.F., (2003) "Mining stock price using fuzzy rough set system", Expert Systems with Applications, 24, pp. 13-23.

[2] Wu, M.C., Lin, S.Y., and Lin, C.H., (2006) "An effective application of decision tree to stock trading", Expert Systems with Applications, 31, pp. 270-274.

[3] Al-Debie, M., Walker, M. (1999). "Fundamental information analysis: An extension and UK evidence", Journal of Accounting Research, 31(3), pp. 261–280.

[4] Lev, B., Thiagarajan, R. (1993). "Fundamental information analysis", Journal of Accounting Research, 31(2), 190–215.

[5] Tsang, P.M., Kwok, P., Choy, S.O., Kwan, R., Ng, S.C., Mak, J., Tsang, J., Koong, K., and Wong, T.L. (2007) "Design and implementation of NN5 for Hong Kong stock price forecasting", Engineering Applications of Artificial Intelligence, 20, pp. 453-461.

[6] Ritchie, J.C., (1996) Fundamental Analysis: a Back- to-the-Basics Investment Guide to Selecting Quality Stocks. Irwin Professional Publishing.

[7] Murphy, J.J., (1999) Technical Analysis of the Financial Markets: a Comprehensive Guide to Trading Methods and Applications. New York Institute of Finance.

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

[9] Han, J., Kamber, M., Jian P. (2011). "Data Mining Concepts and Techniques". San Francisco, CA: Morgan Kaufmann Publishers.

[10] Enke, D., Thawornwong, S. (2005) "The use of data mining and neural networks for forecasting stock market returns", Expert Systems with Applications, 29, pp. 927- 940.

[11] Wang, J.L., Chan, S.H. (2006) "Stock market trading rule discovery using two-layer bias decision tree", Expert Systems with Applications, 30(4), pp. 605-611.

[12] Lin, C. H. (2004) Profitability of a filter trading rule on the Taiwan stock exchange market. Master thesis, Department of Industrial Engineering and Management, National Chiao Tung University.

[13] Cao, Q., Leggio, K.B., and Schniederjans, M.J., (2005) "A comparison between Fama and French's model and artificial neural networks in predicting the Chinese stock market", Computers & Operations Research, 32, pp. 2499-2512.

[14] Fama, E.F., French, K.R., (1993) "Common risk factors in the returns on stocks and bonds", The Journal of Finance, 33, pp. 3-56