AN ANALYSIS OF DECISION TREE ALGORITHMS USING STOCK MARKET PREDICTION

G.Haripriya,
M.Phil Research Scholar,
Department of Computer Science,
Thiruvalluvar Government Arts College,
Rasipuram,Tamilnadu,India.

Dr S. Sathiyabama,
Assistant professor,
Department of Computer Science,
Thiruvalluvar Government Arts College,
Rasipuram,Tamilnadu,India.

Abstract: Data mining is a process, which finds useful patterns from large amount of data. Data mining is looking for hidden, valid, and potentially useful patterns in huge data sets [1]. Data mining is classified into predictive and Descriptive models. In the present work the techniques of data mining used classification, were classification is a predictive method. The predictive method makes prediction about values of data. Classification techniques are using various decision tree algorithms. In this Paper, four decision tree algorithms (ID3, C4.5, Random forest and CART) have been applied on the historical data for predicting the Stock Market performance in Price. The efficiency of various decision tree algorithms can be analyzed based on their accuracy and time taken to derive the tree. The predictions obtained from the system have helped the Investor to identify and improve their performance. Particularly, this work is carried out to compare the four decision tree algorithms in the prediction of the performance accuracy in Stock Market data. All the algorithms are applied for Stock Market data to classify the data set for classification and prediction. Among these four methods, this work concludes the best algorithm for the chosen input data on decision tree supervised learning algorithms to predict the best classifier. This learn about tries to help the buyers in the inventory market to decide the better timing for buying or promoting shares based totally on the information extracted from the historical prices of such stocks. The decision taken will be based on the best decision tree algorithms.

Keywords: Data mining Techniques, Knowledge discovery process, Data mining, Classification, and Decision tree Algorithms.

I. INTRODUCTION

Data mining is concerned with extracting meaningful rules or interesting patterns from the large amount of data collected through different sources. Every data mining techniques can be used to perform the job efficiently. It is to be noted that a single technique cannot be used for all types of data because depending on different type of data, appropriate technique is available for extraction of information[1]. The purpose of the data mining technique is to mine information from a large data set and a reasonable form for supplementary purpose. Data mining is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts cost, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angels, categories it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases [2].

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD) an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspect and pre-processing, model an inference considerations, interestingness metrics, complexity consideration, post-preprocessing of discovered structures, visualization, and online updating. Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses [3].

II. LITERATUREREVIEW

Abhishek Gupta, et al [4] Proposed a methodology implemented for the prediction of stock markets consists of two stages. First is to apply clustering algorithm such as K-means and then the clustered values are partitioned into number of parties and apply horizontal partition based decision tree algorithm. The algorithms are tested on two datasets, one is monthly and other is yearly dataset.

TamalDattaChaudhuri, et al [5] Demonstrate that volatility prediction in stock markets has to be preceded by a study of the number of predictors and the number of clusters. The data on historic volatility may not be homogenous and the presence of many clusters would validate that. If there are too many clusters then it implies that volatility is random and would be difficult to predict. Further, the choice of the predictors has to be mapped with the number of clusters.

BiniB.Sa et al, [6] made an analysis which helps the people to identify the more profitable companies using data mining approaches is proposed. The clustering and regression are the two techniques of data mining used here, Validation index is used for analysing the performance of different clustering methods such as partitioning technique, hierarchical technique, model based technique and density based technique. Among the different clustering techniques experimented, partitioning technique and model based technique give high performance i.e. K-means and EM clustering algorithm respectively. For prediction of future stock price multiple regression technique is used which helps the buyers and sellers to choose their companies from stock.

Ayman E. Khedr, et al, [7]proposed a model is to predict the stock market behavior, whether it is falling or rising. The proposed architecture combines the analysis of the stock market news and the historical prices together, in order to boost the classification accuracy of the stock market behavior. In this investment the text analysis on stock market news to determine the polarity of the news articles. Moreover, stock market historical prices opening, high, low and closing prices (OHLC) are analyzed to predict the future trends.

SarikaShrivastava, et al, [8] provides the analytical study of various data mining based predictive models to predict the stock in financial domain and find the most consistent prediction model among them. In Stock Market, Data mining play a very important role for the analysis of data. Data mining techniques can be applied on past and present financial data to generate pattern and decision making. The predictive techniques like Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Random Tree (RT) and Multilayer Perceptron (MLP) for analysis and perdition of stock market. We have used Bombay Stock Exchange (BSE) data set to analysis for the stock market prediction.

Shubhangi S. Umbarkar, et al, [9] presented a method for prediction of stock market and considers only closing prices of the shares instead of textual information about the stocks. This reduces the efforts that are required for the extraction of news information. The trading strategies consider technical trading indicators, which are used to generate the superior returns. The technical trading indicator give decision that is more appropriate. Data mining technique such as association rule miming and Naïve Bayes algorithm generates significant signals within the polynomial time. It also increased the accuracy of the prediction system by accepting the accurate closing prices of the stock.

III.PROPOSEDMETHODOLOGY

3.1 PERFORMANCE ANALYSIS OF DECISION TREE ALGORITHMS

The proposed model consists of the following five steps:

1. Preprocessing the stock market data set.
2. Building the model for various decision tree algorithms such as ID3, C4.5, CART and Random Forest using training data set.
3. Evaluation of classifiers using test data.
4. Comparing classifiers for accuracy and time.
5. Best classifier has been chosen for stock market prediction.

The main reason and objective of building the model is to try to help the investors in the stock market to decide the best timing for buying or selling stocks based on the knowledge extracted from the historical prices of such stocks. The decision taken will be based on one of the data mining techniques; the decision tree classifiers.

Types of Decision

Trees Decision trees used in data mining are mainly of two types:

1. Classification tree in which analysis is when the predicted outcome is the class to which the data belongs. For example outcome of loan application as safe or risky [13].
2. Regression tree in which analysis is when the predicted outcome can be considered a real number. For example population of a state. Both the classification and regression trees have similarities as well as differences, such as procedure used to determine where to split. There are various decision trees algorithms namely ID3(Iterative Dichotomiser 3), C4.5, CART(Classification and Regression Tree), CHAID(CHi- squared Automatic Interaction Detector). This research work analysis the popular algorithms ID3, C4.5, CART and Random Forest.

In this research work, C4.5, ID3, CART and Random forest algorithms are one of the decision tree generator using top-down approach. They use the information to increase ratio as the splitting criterion for each internal node. Like other similar top-down approaches, C4.5, ID3, CART and Random forest uses a greedy searching strategy with looking one step ahead to find the way of splitting instance space, so it often suffers from being poor at local optimum and therefore performs poorly in dealing with some hard classification tasks, in which training data are described by high dimensional attribute vectors and the concept to be learned is complex [15].

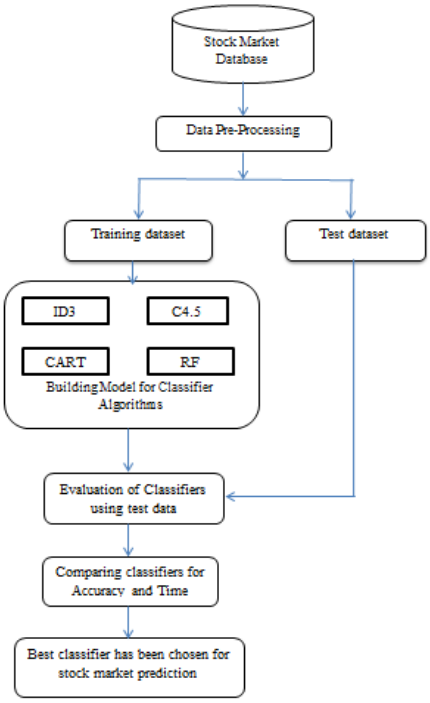

Fig. 3.1 Stock Prediction framework

There have been proposed several non-greedy search approaches, which choose a splitting model automatically by observing the structure of data, so the splitting models are different nodes to nodes. Construction of optimal or near-optimal decision trees using a two stage approach has been attempted by many authors. In the first stage, a sufficient partitioning is induced using any reasonable greedy method. In the second stage, the tree is refined to be as close to

optimal as possible. A different temptation of generating non-greedy decision tree is to use genetic method. In the proposed approach C4.5, ID3, CART and Random forest generates a decision tree using the standard TDIDT (Top down Induction of Decision Trees) approach, recursively partitioning the instance space into smaller subspaces, based on the value of a selected attribute. It begins with a set of instances, called training instances, already separated into classes[15]. Each instance is described in terms of a set of attributes, which can be numerical or symbolical. The overall approach uses a greedy search strategy to choose the attribute that divides the instances best into their classes. This process is functional recursively to each partitioned subset, with the procedure ending when all instances in the current subset belong to the same class; C4.5, ID3, CART and Random forest uses an information gain ratio for measuring how well an attribute can divide the instances into their classes.

3.2 DATA PREPROCESSING

Initially, all the records with missing values were removed from the dataset in order to improve theaccuracy of the prediction. Then the data was further partitioned into two parts:

Training dataset (d.t): It is the data with which the machine is trained. Various classificationalgorithms are trained on this data. 60% of the data is taken as training data.

Validation dataset (d.v): It is the data which is used for the purpose of cross-validation. It is usedto find the accuracy rate of each algorithm. The remaining 40% of the data is taken as validationdata.

Divide it into two subsets:

Training set is a subset of the dataset used to build predictive models. Test set or unseen examples is a subset of the dataset to assess the likely future performance of a model[19]. The stock prediction framework proposed in this research work is portrayed in Fig. 3.1 stock prediction framework. Data mining methodology is designed to ensure that the data mining effort leads to a stable model that successfully addresses the problem it is designed to solve. Various data mining methodologies have been proposed to serve as blueprints for how to organize the process of gathering data, analyzing data, disseminating results, implementing results, and monitoring improvements. To build the model that analyses the stock trends using the decision tree technique, the CRISP-DM (Cross- Industry Standard Process for data mining) [17] is used.

3.2.1 Understanding the collected data

Table 3.1 shows the attributes selected with their descriptions and their possible values. The class attribute is the investor action whether to buy or sell that stock and it is named, "Action". The net position could be either buying or selling that stock for that day.

Table 3.1 Attribute Description

Attribute	Description	Possible Values
Open	Current day open price of the stock	Positive, Negative, Equal
High	Current day maximum price of the stock	Positive, Negative, Equal
Low	Current day minimum price of the stock	Positive, Negative, Equal
Close	Previous day close 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

Classification divides the text data into two sentiments, the positive and negative. The result of the classification is influenced by the learning model created from training data. The better models are made, the better the classification which done by machines [18]. The parameter false positive, true positive, false negative, true negative sensitivity, deviation, precision, and accuracy are used to compute the classifiers performance.

- True Positive (TP) represents the examples that are properly predicted as normal.
- True Negative (TN) shows the instances which are correctly predicted as an attack.
- False Positive (FP) recognizes the instances there are predicted as attack while they are not.
- False Negative (FN) signifies the cases which are prefigured as normal while they are attack in reality.
- The sensitivity is described as the ratio of true positives to the sum of false negatives and true positives.
- The specificity is described as the ratio of true negatives to the sum of false positives and true negatives.

Accuracy: These refer to the capability of the classifiers to correctly measure the intrusions from training dataset and defined as ratio of the correctly classified data to total classified data.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Experimentation result shows that the proposed model is more accurate as compare to other data mining technique. The accuracy of a proposed model is near about 100 percentage.

Precision Ratio: It is also called as Positive Predictive Value. It measures the relevant instance that is retrieved after classification. A classifier that has high precision means that classifiers or algorithm returns more relevant results.

$$PositivePredictiveRatio = \frac{TP}{TP + FP}$$

Recall: It is also known as sensitivity. This is also used to measure the relevant instance that is selected. The higher value of recall more the relevant data is selected for classification. It is defined as:

$$Recall = \frac{TP}{TP + FN}$$

F-Measure: It is mostly used to measure the effectiveness of

classifiers. This is harmonic mean of the precision and recall. It is also called as traditional F-measure otherwise balanced F-score. It is defined as:

$$F - Measure = 2\frac{Precison * Recall}{Precison + Recall}$$

Table 3.2 Shows the Description of processing training and testing set. The performance is calculated by using help of confusion matrix which is generated when algorithm is implemented on dataset. Confusion matrix for CART, C4.5, Random Forest, and ID3 is taken from weka by opening the Training Features Set and Testing Features Set there directly

Table 3.2 Description of processing training and testing set

Features	Values
Size of Training Set	251
Size of Testing Set	150
Tweet Classes	Positive, Negative, Equal
Methods	CART, C4.5, Random Forest, ID3

.

Table 3.3 Confusion Matrix

	Correct Labels	
	Positive	**Negative**
Positive	True Positive	False Positive
Negative	True Negative	False Negative

Table 3.3 Shows the Confusion matrix. It is calculated by the analysis results and with the help of confusion matrix precision, recall and f-measure is calculated. An analysis result shows True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN) for each sentiment class.

A confusion matrix shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data [50].NxN,where N is the number of target values (classes). Performance of such models is commonly evaluated using the data in the matrix. Table 3.4 displays a 2x2 confusion matrix for two classes (Positive and Negative).

- Accuracy: The proportion of the total number of predictions that were correct.
- Positive Predictive Value or Precision: the proportion of positive cases that were correctly.
- Negative Predictive Value: the proportion of negative cases that were correctly identified.
- Sensitivity or Recall: the proportion of actual positive cases which are correctly identified.
- Specificity: The proportion of actual negative cases which are correctly identified.

Table 3.4 Confusion Matrix

		Actual	
		Positive	Negative
Prediction	Positive	Positive True	Positive False
	Negative	Negative True	Negative False
	Accuracy (ACC) = (Σ True positive +Σ True negative)/Σ Total population	True positive rate (TPR) = Σ True positive / Σ Condition positive	False positive rate (FPR) = Σ False positive / Σ Condition negative

3.2.2 Preparing the Data

At the beginning, when the data was collected, all the values of the attributes selected were continuous numeric values. Data transformation was applied by generalizing data to a higher-level concept so as all the values became discrete. The criterion that was made to transform the numeric values of each attribute to discrete values depended on the previous day closing price of the stock [18]. If the values of the attributes open, High, Low, last were greater than the value of attribute previous for the same trading day, the numeric values of the attributes were replaced by the value Positive. If the values of the attributes mentioned above were less than the value of theattribute previous, the numeric values of the attributes were replaced by Negative. If the values of those attributes were equal to the value of the attribute previous, the values were replaced by the value Equal.

Table 3.5 shows a sample of the continuous numeric values of thedata before selecting the 6 attributes manually and beforegeneralizing them to discrete values, while Table3.6 shows a same sample after selecting the 6 attributes and aftertransforming them to discrete values.

Table 3.5 Sample of historical data before selecting relevant attributes and after generalization

Open	High	Low	Close	Mean	Action
21.61	21.94	21.61	21.68	21.73	**Sell**
21.39	21.61	21.53	21.47	21.53	**Sell**
21.21	21.64	21.30	21.49	21.47	**Sell**
21.42	21.49	20.40	21.45	21.11	**Buy**
21.09	21.74	21.64	21.56	21.64	**Sell**
20.97	21.53	20.50	21.50	21.17	**Buy**
21.50	21.32	20.40	21.32	21.01	**Buy**

Table 3.6 Sample of historical data before selecting relevant attributes and after generalization

Open	High	Low	Close	Mean	Action
Positive	Positive	Positive	Negative	Positive	**Sell**
Negative	Positive	Positive	Negative	Positive	**Sell**

Negative	Negative	Equal	Negative	Positive	**Sell**
Negative	Negative	Equal	Negative	Negative	**Buy**
Negative	Positive	Equal	Negative	Positive	**Sell**
Positive	Equal	Positive	Equal	Negative	**Buy**
Positive	Equal	Positive	Positive	Negative	**Buy**

Investors are familiar with the saying, "buy low, sell high" but this does not provide enough context to make proper investment decisions. Before an investor invests in any stock, he
needs to be aware how the stock market behaves [20]. Investing in a good stock but at a bad time can have disastrous results, while investment in a stock at the right time can bear profits.

3.4.3Building the Model

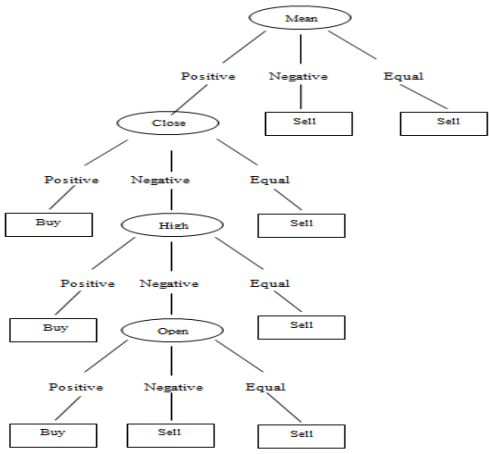

Fig. 3.2 Decision Tree

Fig. 3.2 shows as the process of building the tree goes on, all the remaining attributes were used to continue with this process. After building the complete decision tree, the set of classification rules were generated by all the paths of the tree. The maximum number of attributes that were used in some of the classification rules generated were 4 attributes, while some classification rules used only 1 attribute.

IV EXPERIMENTAL RESULTS AND DISCUSSION

4.1 PERFORMANCE ANALYSIS OF DECISION TREE ALGORITHMS

Each of the classification algorithms were first trained using the training data which contained 60% of the dataset [21]. Then the remaining 40% was used for the purpose of cross-validation. From the predictionproduced by each of the algorithms, the confusion matrix was obtained.

4.1.1 Performance of ID3 Algorithm

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Accuracy= 92+32/92+16+10+32*100=82.6%

Table 4.1 shows the performance of ID3, which shows the result of various parameters such as Time taken to build model, Absolute error and Accuracy. And also table 4.1 shows the precision, Recall, F-Measure.

Table 4.1 Performance of ID3 Algorithm

Parameters	Performance of ID3
Build time (Seconds)	0.53
Absolute error	0.543
Accuracy (%)	82.6
Precision Ratio	83.82
Recall	81.07
F-Measure	82.93

Table 4.1 is used to identify the performance of ID3. It takes build time as 0.53 seconds and absolute error is 0.5 and accuracy for ID3 is 82.6%. It can detect different types of attacks with best accuracy. As a result, the proper option of the classifier is critical to yield high-quality classifications in hidden instances.

4.1.2 Performance of C4.5 Algorithm

Accuracy= 102+38/102+10+0+38*100=93.3%

Table 4.2 shows the performance of C4.5, which shows the result of various parameters such as Time taken to build model, Absolute error and Accuracy. And also table 4.2 shows the precision, Recall, F-Measure.

Table 4.2 Performance of C4.5 Algorithm

Parameters	Performance C4.5
Build time (Seconds)	0.08
Absolute error	0.348
Accuracy (%)	93.3
Precision Ratio	89.59
Recall	92.87
F-Measure	91.48

Table 4.2 is used to identify the performance of C4.5. It takes build time as 0.08 seconds and absolute error is 0.3482 and accuracy for C4.5 is 93.3%. It can detect different types of attacks with best accuracy. As a result, the proper option of the classifier is critical to yield high-quality classifications in hidden instances.

4.1.3 Performance of CART Algorithm

Accuracy= 81+32/81+16+21+32*100=82.6%

Table 4.3 shows the performance of CART, which shows the result of various parameters such as Time taken to build model, Absolute error and Accuracy. And also table 4.3 shows the precision, Recall, F-Measure.

Table 4.3 Performance of CART Algorithm

Parameters	Performance of CART
Build time (Seconds)	1.32
Absolute error	0.659
Accuracy (%)	75.3
Precision Ratio	72.24
Recall	77.71
F-Measure	76.76

Table 4.3 is used to identify the performance of CART. It takes build time as 1.32 seconds and absolute error is 0.659 and accuracy for CART is 75.3%. It can detect different types of attacks with best accuracy. As a result, the proper option of the classifier is critical to yield high-quality classifications in hidden instances.

4.1.4 Performance of Random Forest Algorithm

Accuracy= 103+30/103+10+7+30*100=88.6%

Table 4.4 Performance of Random Forest Algorithm

Parameters	Performance of Random Forest
Build time (Seconds)	0.14
Absolute error	0.473
Accuracy (%)	88.6
Precision Ratio	86.22
Recall	88.56
F-Measure	87.41

Table 4.4 shows the performance of ID3, which shows the result of various parameters such as Time taken to build model, Absolute error and Accuracy. And also table 4.4 shows the precision, Recall, F-Measure. Table 4.4 is used to identify the performance of Random Forest. It takes build time as 0.14 seconds and absolute error is 0.437 and accuracy for Random Forest is 88.6%. It can detect different types of attacks with best accuracy. As a result, the proper option of the classifier is critical to yield high-quality classifications in hidden instances.

4.2 COMPARISONS AND DISCUSSIONS

Table 4.5 Comparative table for CART, C4.5, Random Forest, ID3 Accuracy

Dataset	Classifier	F-Measure	Recall	Precision	Accuracy
Stock Market Prediction	ID3	82.93	81.07	83.82	82.6
	CART	76.76	77.71	72.24	75.33
	Random Forest	87.41	88.56	86.22	88.67
	C4.5	91.48	92.87	89.59	93.38

Table 4.5 shows that scores of the above methods are relatively stable in experiments. In above four methods, the classification effects of CART are poor, and score is only around 79%. When using ID3 Algorithm and Radom Forest suitable for sequence modeling, RF can reach around 89%. The method proposed in this Research, which can capture the information of the context more effectively, so the analysis works best. C4.5 is significantly improved, reaching around 93%.

Figure 4.1 shows the accuracy parameters in graph form for visualization and comparison. Clearly ID3 and CART and random forest showed higher accuracy as compared to ID3, random tree Algorithms.

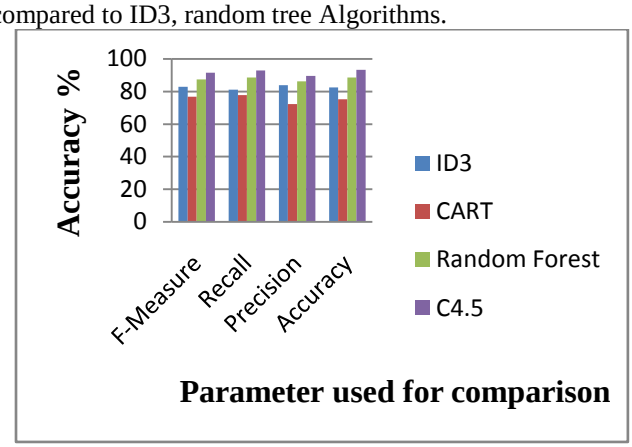

Fig. 4.1Visualization of accuracy parameters

Table 4.6 shows Comparative table for CART, RF, ID3, C4.5 performance. Finally, it is concluded that the C4.5 can detect different types of attacks with best accuracy compared to other systems. As a result, the proper option of the classifier is critical to yield high-quality classifications in hidden instances.

Table 4.6 Comparative table for CART, C4.5, Random Forest, ID3 performance

	ID3	CART	Random forest	C4.5
Build time	0.53	1.32	0.14	0.08
Absolute error	0.54	0.65	0.47	0.34
Accuracy (%)	0.82	0.75	0.88	0.93

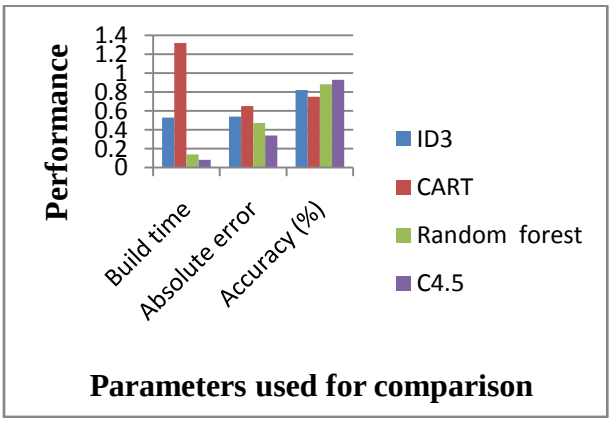

Fig. 4.2Graphical Representation for Performance comparison

Figure 4.2 gives the bar graph of various classifier and its

parameters for better visualization of output results. Though most classifier gave consistent results, but ID3 and random forest gave more accurate results comparatively. It infers about the accuracy of classifier and reveals that C4.5 and random forest have proved to be most accurate classifier with less error and FP rate.

Finally, it is concluded that the C4.5 can detect different types of attacks with best accuracy compared to other systems. As a result, the proper option of the classifier is critical to yield high-quality classifications in hidden instances. Four decision tree algorithms are compared using various parameters such as Accuracy, Time taken to build, Absolute Error rate, precision, recall, and F-Measures. C4.5 Algorithms provides better results with highest Accuracy, least error rate and less time taken to build compared with other decision tree algorithms. And also C4.5 gives the best results for precision, recall, and F-Measures.

V CONCLUSION

In this Research work, four decision tree algorithms (ID3, C4.5, Random forest and CART) have been applied on the historical data for predicting the Stock Market performance in Price. The efficiency of various decision tree algorithms can be analyzed based on their accuracy and time taken to derive the tree. The predictions obtained from the system have helped the Investor to identify and improve their performance. C4.5 is the best algorithm for small datasets among all the four because it provides better accuracy and efficiency than the other algorithms. The proposed technique implemented here for the prediction of stock market provides efficient results as compared to other existing technique. The proposed methodology provides close prediction of actual value; hence the results will be more accurate and efficient. Final result gives various classifier and its parameters for better visualization of output results. It infers about the accuracy of classifier C4.5 and random forest have proved to be most accurate classifier with less error and less Build time. Finally, it is concluded that the C4.5 can detect different types of attacks with best accuracy compared to other systems. As a result, the proper option of the classifier is critical to yield high-quality classifications in hidden instances. Four decision tree algorithms are compared using various parameters such as Accuracy, Time taken to build, Absolute Error rate, precision, recall, and F-Measures. C4.5 Algorithms provides better results with highest Accuracy, least error rate and less time taken to build compared with other decision tree algorithms. And also C4.5 gives the best results for precision, recall, and F-Measures.

VI REFERENCES

[1]. P. Adriaans and D. Zantinge."Data Mining". Addison-Wesley: Harlow, England, 1996.

[2]. S. Chaudhuri and U. Dayal,"An overview of data warehousing and OLAP technology"ACMSIGMODRecord, 26:65{74, 1997.

[3]. Jiawei Han and MichelineKamber, "Data Minsing Concepts and Techniques", 3rd edition, published by Morgan Kauffman, 2006.

[4]. Abhishek Gupta, Dr. Samidha and D. Sharma, "Clustering-Classification Based Prediction of Stock Market Future Prediction", Abhishek Gupta et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2806-2809,August 2015.

[5]. TamalDattaChaudhuri and IndranilGhosh, "Using Clustering Method to Understand Indian Stock Market Volatility", Communications on Applied Electronics (CAE) – ISSN : 2394-4714 Foundation of Computer Science FCS, New York, USA Volume 2 – No.6, August 2015.

[6]. BiniB.Sa and TessyMathewb," Clustering and Regression Techniques for Stock Prediction", International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST - 2015), Procedia Technology 24 -1248 – 1255,(2016).

[7]. Ayman E. Khedr, S.E.Salama and NagwaYaseen, "Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis", I.J. Intelligent Systems and Applications, 2017, 7, 22-30 Published Online July 2017 in MEC.

[8]. SarikaShrivastava and A. K. Shrivas, "Performance Analysis of Data Mining Techniques for Stock Price Forecasting, International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 6 Issue 05, May – 2017.

[9]. Shubhangi S. Umbarkar and Prof. S. S. Nandgaonkar, "Using Association Rule Mining: Stock Market Events Prediction from Financial News", International Journal of Science and Research (IJSR), ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | (2013).

[10]. A.Subashini and Dr. M. Karthikeyan," Forecasting on Stock Market Time Series Data Using Data Mining Techniques", International Journal of Engineering Science Invention (IJESI) ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726 www.ijesi.org || PP. 06-13.2014.

[11]. RohanTaneja and Vaibhav, "Stock Market Prediction Using Regression", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056, Volume: 05 Issue: 05 | May-2018.

[12]. PujanaPaliyawan, "Stock Market Direction Prediction Using Data Mining Classification", ARPN Journal of Engineering and Applied Sciences, ISSN 1819-6608, VOL. 10, NO. 3, February 2015.

[13]. Senanu R. Okuboyejo and Daniel O. Adeyoju, "Design and Implementation of a Stock Market Monitoring System": ISSN: 2321-3 Available at: http://ijmcr.com, Accepted 5 December 2013, Available online 01 January 2014, Vol.2 (Jan/Feb 2014 issue).

[14]. ArtiBuche, and Dr. M. B. Chandak, "Stock Market Prediction using Text Opinion Mining , International Journal of Advanced Research" in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 6, Issue 6, June 2016.

[15]. R. Duda and P. Hart, " PatternClassification and Scene Analysis Wiley" New York, 1973.