

Stock Price Prediction using Various Machine Learning Algorithm

Anushka Chaurasia, Mr. Rohit Kumar Tiwari

Computer Science and Engineering Madan Mohan Malaviya University of Technology Gorakhpur, Uttar Pradesh, India

Computer Science and Engineering Madan Mohan Malaviya University of Technology Gorakhpur, Uttar Pradesh, India

Submitted: 15-01-2021

Revised: 27-01-2021

Accepted: 31-01-2021

ABSTRACT—It is complex to evaluate the stock market where alliances between output and input are irregular. Forecasting the price of the share market is the most complicated activity of the financial time series. Prediction of the stock should be possible by utilizing the current and historical data available on the market. In this paper we studied SBI data to forecast the stock price we used distinct machine learning algorithms for predicting i.e. Ada random forest, Ada decision tree, Decision tree, random forest, SVM, AdaSVM, MLP classifier, Gradient boosting. Using Accuracy and Root mean square error the performance of all algorithms has been compared.

Index Terms—Ensemble Learning, Stock Prediction, Machine Learning, MLP.

I. INTRODUCTION

Stock market plays very crucial role in economic growth specially in large country like india. Stock market is a collection of markets and exchange where shares of public registers firms are traded either OTC(over-the-counter) or through centralized exchanges. It is also known as equity or share market. In other words we can say that firms will register their shares in the firms as small assets called stocks. A company registers its stock at a process called the IPO. This is an offer amount at which the firm trades the stock and increases capital after which these stocks are the equity of the proprietor and he may trade them at any cost to a customer at a trade like Bombay Stock Exchange or BSE. In any country people are afraid of investing in stock market because its return aren't guaranteed. The stock market forecast helps to cope up this problem as its act of trying to find out the future price of the firm stock or other financial instruments traded on an exchange. The successful forecast of a stock's future cost could produce a meaningful profit. The efficient-market hypothesis recommends that stock values reflect all presently

accessible information and any value changes that are not based on newly disclose information thus are constitutionally unpredictable. Others disagree and those with this point of view possess innumerable procedure and technologies which allegedly allow them to gain future cost news.

In the implemented work, we have used eight machine learning algorithm and their performance are compared in forecasting the stock price of SBI . These machine learning algorithms are: SVM, Decision tree, Ada-boost, MLP Classifier, Random Forest, Gradient Boosting and RBF. The performance of these classifiers has been compared using different matrices that include root mean square error Confusion matrix F1 Score, Precision and accuracy.

II. LITERATURE REVIEW

Stock price prediction is a very challenging and complex process because price movement just behaves like a random walk and time-varying. In the previous year many researchers have used a different algorithm to predict the stock price. Here we present a brief analysis of some significant research.

VK.Sai Reddy [1]. In his paper he uses SVM (Support Vector Machine) by RBF (Radial Bias Function) kernel for predicting the stock price. He uses four features: Price momentum, Price Volatility, Sector momentum, Sector Volatility for testing and training the data and $\log_2 c$ and $\log_2 g$ value to reduce the error. His model has generated the highest profit compared to the selected benchmark.

K.Hiba Sadia et.al [2]. In this paper SVM and Random Forest method is used to predict the stock price. Historical data is collected from the Kaggle website contain eleven attributes they are: HIGH, LOW, OPENP, CLOSEP, VCP, LTP, TRADE, VOLUME & VALUE. It contains 121608 records of different trading companies. The value inside the trading code is replaced by "GP", then the

time series graph is a plot between CLOSEP and DATE, candlestick plot was also generated using DATE, OPENP, HIGH, LOW, CLOSEP. They extracted a new feature in which if today CLOSEP is greater than yesterday CLOSEP then they assign it 1 otherwise -1. The accuracy of SVM is 0.0787 and Random Forest is 0.808.

M Ali. Ghazanfar et.al.[3]. This paper aims to predict the target volume values concerning upcoming n-days prices. For this they use the Saudi Stock Exchange (SSE) and Karachi Stock Exchange (KSE) data to perform stock prediction. The dataset contains the following attributes: Open, High, Low, Current, Change. They extracted a feature i.e. label for labeling the data, a threshold has been set on the input feature 'Change'. The attribute 'change' is range from "3 to -3", the value exactly matching to '0' is given class label "C", below '0' is class "B"

and above '0' is class "A". They use two cross-fold validation for training the data using 7 machine learning algorithm (SVM, RBF, KNN, Adaboost, Multi-layer perceptron, Naïve Bayes, Bayesian Network) from which only three algorithms gives the best accuracy, RMSA and MAE's they are Adaboost, Multi-Layer perceptron and Bayesian Network).

B.Narayananet.al[4]: This paper aims to develop an ensemble model namely Ada SVM and Ada Naïve Bayes and perform a Comparison with SVM an Naïve Bayes model regarding classification error and accuracy. For this the historical data is collected from www.datamarket.com which contains the following attributes: Date, quantity, end, min, max, start and cname. On this basis of this attributes the training and testing are performed and the results are:

Table 1 COMPARISION OF DIFFERENT ALGORITHM ON THE BASIS OF ACCURACY AND ERROR

Measures	Existi ng SVM	Ada SVM	Existing Naïve Bayes	AdaNai ve
Accuracy	93.86 %	94.33 %	88.32%	97.19%
Classificati on error	6.14%	5.67%	11.68%	2.81%

K.Pahwaet.al[5]: This paper introduces a linear regression method to forecast the future stock price for exchange using open source libraries. The outcome of this paper is absolutely based on numbers and consider a lot of axioms. They use google dataset which is downloaded from quandl. They two new attributes i.e. HL_PCT and PCT_CHANGE for prediction.

X. Ming Bai et.al[6] This paper proposed an Ada ANN model for forecasting which exploits Adaboost theory and ANN model for task prediction. The dataset is collected from the Chinese stock market and the international stock market. The single Ann forecasting model is used in this paper has the same architecture as the weak learner in the AdaANN.The experiment was performed in three groups and in each group AdaANN performs well in terms of statistical accuracy it increase the accuracy.

P.Rajesh et.al.[7] In this paper they used two algorithms for prediction ensemble learning and Heat map which was based on the percentage change in the stock price data that will classify the data into sell, hold or buy classes. The main purpose of this paper is to obtain the best performance with the minimum classification complexity of the stock trend. The heat map is created based on the correlation coefficient whereas the ensemble learning

model classifies the stock data into the mostvote based system. Random forest, SVM and KNN Classifier show the best results. The accuracy of the forecast model is more than 51 %

PP.SinghKholi et.al.[8] The main purpose of this paper is to forecast the attribute of BSE(Bombay Stock Exchange). They use attributes such as market history, FEX(Foreign Exchange rate) and commodity prices(Silver, Crude oil, Gold) that influence the stock trends. They used 9 months of data which is collected from <https://www.investing.com>. They used Adaboost, Gradient Boosting, SVM and Random Forest with different training sets (70%, 90%) and test set (30%, 10%) data, which gives different accuracy. It verifies that BSE has the highest dependency on the gold rate since the correlation factor is highest and lowest at the silver rate. Adaboost shows the highest accuracy in all the ML algorithms that are used in this paper of 76 %.

S.Lounnapha.et.al[9]: In this paper a Deep learning method CNN(Convolutional neural network)is used for stock price prediction. They use the Thai Stock market (BBL, CAPLL and PTT) for training and testing. For data preparation they use a sliding window approach. They aimto forecast the movement of the 3 listed in the SET 50 index,

manipulate some information from historical data. They use 1-D convolutional and max-pooling layers where, Subsamples=1, Pool length = 2 Number of filters =64, Activation Function=RELU and 1 & 2 hidden convolutional layers.

AI. Awan et.al.[10]: This paper, explored various components that influenced the behavior of the Stock exchange share process. The objective is to find and evaluate the effects of elements that are driven through various political and environmental situations in the country which drive the fall and rise of share prices. Based on the outcome of these analyses an algorithm is derived using Rule-Based System and Artificial Intelligence (AI) with Expert System, to forecast the behavior of stock values. This model uses the decision tree algorithm simultaneously with a novel Type, Effect and Weight (TEW) Algorithm, to logically forecast the pattern of stock prices. The goal of this research is to help users of the stock exchange and save capital for financiers.

III. METHOD USED

In this section, six ML algorithms are used and compared based on their test accuracy, confusion matrix and Root mean square error. These models are as follows.

A. Ada Boost

Ada Boosting is also known as Adaptive Boosting. It is one of the successful machine learning algorithms used for regression and classification processes. It can boost the performance of any machine learning algorithm, it includes an ensemble of multiple weak classifiers to assemble a single strong one. Generally AdaBoost is used for decision tree having 1 depth known as stumps. The algorithm process is as follows:

Step-1: Initialize equal weight to each training instance that is being used.

$$\text{Initial weight } (w_i) = \frac{1}{N} \quad (1)$$

Where i = number of instances and N = total no of instances.

Step-2: Calculate the weight error (e) of the stumps, it is a wrong prediction out of total instance.

$$e = (\text{correct} - N)/N \quad (2)$$

Step-3: It appoints a higher weight to incorrect one so that in the next iteration these observations will get the high probability for classification.

Step-4: Also, It assigns the weight to the trained classifier in each iteration according to the accuracy of the classifier. The more accurate classifier will get high weight.

Step-5: This process iterates until the complete training data fits without any error or until it reached to the specified maximum number of estimators.

Step-6: To classify, perform a "vote" across all of the learning algorithms you built.

B. Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning (ML) algorithm that uses for both regression and classification problems. I use the classification problem for two grouped classification. In SVM we plot a single data item in an N-Dimensional space (N- number of attributes) with the value of each attribute being the value of a particular coordinate. It can be done by finding the hyperplane that discriminates the two classes. It is used to find a plane that has the max distance between data points of two classes, i.e. max margin. SVM uses three types of the kernel they are Linear, Polynomial, and RBF(Radial Bias Function).In this paper we use the RBF kernel which is used when the boundaries are hypothesized to be curved in shape.

C. Neural Network

The neural net is an artificial portrayal of the human brain. That tries to simulate its learning process,It consists of three basic units: The input layer receives the information from the outside world. The output layer is responsible for computation and transferring the information to the outside world from the network. and Hidden Layer It does intermediate computation between input and output layers. There can be multiple hidden layers.this paper multilayer perceptron classifier is used which is a combination of multiple perceptron and also known as Feed-Forward neural network. The process by which MLP learns is known as Backpropagation neural network. In backward propagation and weight updation, we calculate the total error at the output node

$$e = \frac{1}{2} \sum_i^n (ac^i - nt^i)^2 \quad (3)$$

Where ac is actual output and nt is network output and propagate these errors back through the network using backpropagation to compute gradients. Then we use a gradient descent optimization method to adjust all weight in the network for minimizing the error at the output layer.

D. Random Forest

Random forest is a supervised machine learning(ML) algorithm that uses an ensemble model for regression and classification problems. It takes the help of Bagging (Bootstrap Aggregation) and Decision Trees. Thus, it reduces the issue of overfitting in decision trees. Bagging reduces the variance problem of high variance methods like decision trees (CART). An RF algorithm is a group of unpruned regression and classification trees that

are obtained from the subsamples of the training dataset.

E. Decision Tree

The decision tree is a supervised machine learning (ML) algorithm used for classification, regression, and prediction. It is a flowchart like a tree structure, where each internal node denotes a test on an attribute, each branch represents a result of the test and each leaf (terminal node) denotes an outcome (continuous or categorical value). Then are multiple algorithms that are used with a decision tree:

CART, MARS, ID3, CHAID, C4.5.

F. Gradient Boosting

Gradient boosting is a powerful machine learning algorithm. It can do ranking, classification and regression. It can train many models sequentially. The loss function is deliberately decreased by each new prototype. In Gradient Boosting, we suppose a uniform distribution say

$$A_1 = 1/n(4)$$

for all n observations. Then the process of algorithm is as follow:

Step-1: Suppose an $\alpha_{(y)}$.

Step-2: Calculate a weak classifier $h(y)$.

Step-3: Update the population distribution for the next step.

$$A_{y+1}(i) = \frac{A_y(i) \exp(-\alpha_y z_i h_y(x_i))}{D_y} \quad (5)$$

$$D_y = \sum_{i=1}^m A_y(i) \exp(-\alpha_y z_i h_y(x_i)) \quad (6)$$

Step-4: Now, used the new population distribution to search the next learner.

Step-5: Repeat Step 1–Step 4 until no hypothesis is found which can further boost the accuracy.

Step-6: Take the weighted average of the boundary using all the learners used till now. Weight is the alpha values calculated as:

$$\alpha_y = \frac{1}{2} \ln \left(\frac{1 - \epsilon_y}{\epsilon_y} \right) \quad (7)$$

IV. EXPERIMENT SETUP

We have conducted an experiment, in which we focused on predicting stock price using various machine learning algorithms. We proposed this model as “Stock Price Predictor” in which we took the past 20 years of data for training and testing. In this system we used NumPy to clean and manipulate the raw data, matplotlib to visualize the data and sci-kit-learn, which was used for real analysis and prediction.

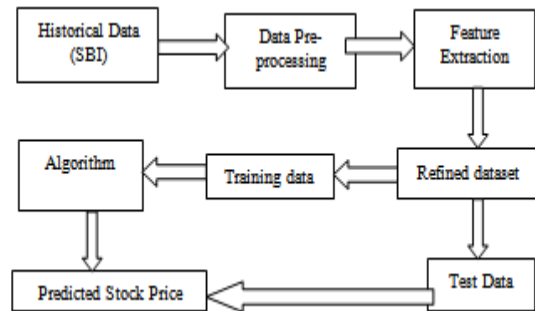


Fig 1: System Architecture

G. Dataset

In this module we use historical price data of SBI which is extracted from yahoo finance. It is a 20-year data set (2000-2020). The following feature of dataset: ‘Date’, ‘Open’, ‘High’, ‘Low’, ‘Close’, ‘Adj Close’ and ‘Volume’.

H. Data Pre-Processing

Data pre-processing is a part of data mining in which the raw data transformed into a more systematic format. Raw data usually inconsistent and contain many errors. In this section we remove the attribute that is not required they are Adj. Close and Low columns. Now we need to fill the missing values to make our dataset consistent. For this we used mean function. As we are using python we don’t need to write the whole equation we can use just mean().

Table 2 RAW DATASET OF SBI

Date	Open	High	Low	Close	Adj Close	Volume
2000-01-24	22.32 8400	22.6 7750	21.4 2740	21.54 0600	1.02 2121	1.966 714e+07
2000-01-25	21.32 3601	21.5 0280	21.0 0280	21.21 5099	1.00 6676	2.359 792e+07
2000-01-26	163.2 08134	165. 5336	160. 6070	162.9 52856	114. 8204	2.185 570e+07

I. Feature Extraction

The feature extraction purpose is to truncate the number of attributes in a dataset by creating new attributes from the actual one. These new truncated set of attributes should then be able to encapsulate most of the information contained in the original set of attributes. In this way, an encapsulated version of the original attributes can be built from a combination of the original set. In, this section we used six features: Open, High, Close, Volume, O_C and I_D. where, O_C is a new feature that is extracted from the open and close column.

$$O_C = \sum_{i=1}^n Open_i - Close_i \quad (8)$$

I_D is also a new feature which contains 0 and 1

$$I_D = \begin{cases} 1, & O_C > 0 \\ 0, & O_C < 0 \end{cases} \quad (9)$$

J. Training Data

In this 80% dataset is used for training and a 20 % dataset is used for test and predict the values.

K. Evaluation Metrics

1) Confusion Matrix: It is a chart that is often used to define the efficiency of a classification model (or "Classifier") on a set of test data for which the true values are known. It is a summary of prediction results on a classification problem.

- (TP) True Positive: Dataset is positive and is forecasted to be positive.
- (TN) True Negative: Dataset is negative and is forecasted to be negative.
- (FP) False Positive: Dataset is negative but is forecasted positive
- (FN) False Negative: Dataset is positive but is forecasted negative

2) Measurement Factor.

a) Accuracy: estimate the closeness of determined value to the acknowledge or standard value.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

b) Recall: is the ratio of the total number of correctly classified positive values divided by the total number of positive values. It has also known as sensitivity.

$$R = \frac{TP}{TP+FN} \quad (11)$$

c) Precision: estimate the closeness of two or more determined data to each other.

$$Precision = \frac{TP}{TP+FP} \quad (12)$$

$$F_{measure} = \frac{2 * R * Precision}{R + Precision}$$

(13)

d) RMSE :Root Mean Squared Error

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (Predicted_i - Actual_i)^2}{N}} \quad (14)$$

Here N represents the total number of samples that are forecasted.

V. RESULTS

In this research we have used different classification algorithms for forecasting i.e. Ada Decision, Ada SVM, CNN, Gradient Boosting, Decision tree, Ada random forest and SVM with RBF kernel. Here, in fig2 the OHLCV graph is plotted which shows the increasing and decreasing momentum of stocks it is a bar graph that shows high, low, open, and closing prices for each period.

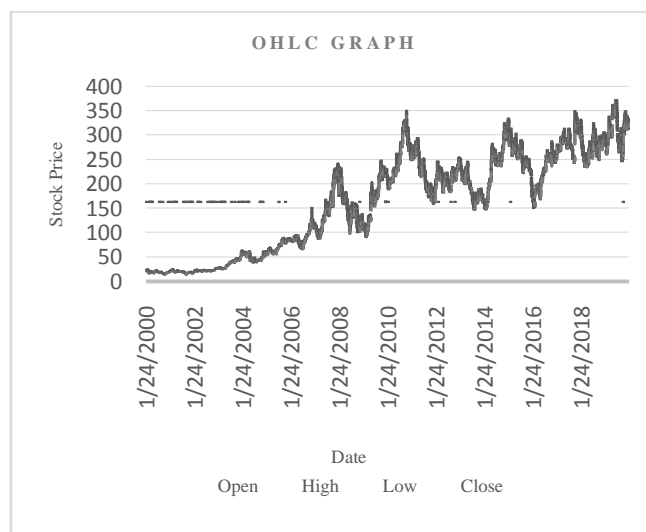


Fig 2 OHLC graph.

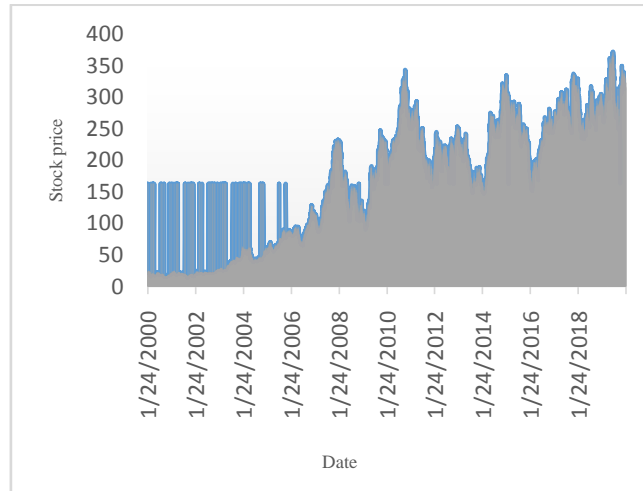


Fig 3 Graph between Close price and date

Fig 3 shows the relation between date and close price it indicates that on which date the closing price is low and on which date the closing price is high.

Table III represent heat map of Ada decision tree which is an ensemble learning method

in which ada boost algorithm is combined with decision tree in which base classifier is taken as a Decision tree with criterion is entropy and max depth is 4 .it gives 88%accuracy and0.35 of rmse error with a learning rate of 0.2

TABLE 3 ADA DECISION TREE CONFUSION MATRIX

		Predicted	
		0	1
Actual	0	501	58
	1	65	378

Table IV is a heat map of the Decision tree where the criterion is entropy and n_estimator is 100 .it gives 81% accuracy and 0.663 of rmse error with a learning rate of 0.2. Here 405 is truly classified value for 0.

Table 4 DECISION TREE CONFUSION MATRIX

		Predicted	
		0	1
Actual	0	405	154
	1	287	156

Table V is a heat map of Gradient boosting which is an ensemble learning method .In which n_estimator is 2000 and the learning rate is 0.2. It gives 88% accuracy and 0.34 of root mean square error.

TABLE 5 GRADIENT BOOSTING CONFUSION MATRIX

		Predicted	
		0	1
Actual	0	507	52
	1	69	374

Ada random forest is an ensemble learning method in which the ada boost algorithm is combined with random forest where the base classifier is taken as a random forest with criterion is entropy and n_estimator is 100. In ada boost the

learning rate is 0.2 and n_estimator is 50 .it gives accuracy 81 % and 0.4388 of root mean square error.Table VI shows the confusion matrix of ada random forest..

TABLE 6 ADA RANDOM FOREST CONFUSION MATRIX

		Predicted	
		0	1
Actual	0	465	94
	1	99	344

Random Forest is used to training and tests the datasets where the criterion used is entropy and n_estimator is 1000. It gives 80% accuracy and 0.446

of rmse error. The confusion matrix of the random forest is shown in Table VII.

TABLE 7 RANDOM FOREST CONFUSION MATRIX

		Predicted	
		0	1
Actual	0	464	95
	1	105	338

SVM is used for training and testing the data here RBF kernel is used for prediction which gives 56% accuracy and 0.665 of root mean squared

error. The confusion matrix of the support vector classifier is shown in Table VIII.

TABLE 8 SVM CONFUSION MATRIX

		Predicted	
		0	1
Actual	0	493	66
	1	378	65

Table IX is a heat map of Ada SVM which is an ensemble learning method,in which the SVM is used as a base classifier in ada boost to boost the

performance of SVM, the n_estimator is 100.and the kernel is RBF.it gives 56% accuracy and 0.664 of rmse error with a learning rate of 0.2.

TABLE 9 ADA SVM CONFUSION MATRIX

		Predicted	
		0	1
Actual	0	486	73
	1	375	68

Fig X is a heat map of neural network in which a multilayer perceptron Classifier algorithm is used for training and testing the data it gives 89% accuracy with 0.329 Root mean square error.

TABLE 10 NEURAL NETWORK CONFUSION MATRIX

		Predicted	
		0	1
Actual	0	453	106
	1	3	440

VI. CONCLUSION

The conclusion of this experiment concludes that a machine learning algorithm can be used to forecast stock market performance. All the

algorithm is applied on the same training and testing dataset The result indicates that the RMSE and accuracy of the Decision tree, Ada decision tree, Ada random Forest, Support Vector Machine, Ada

SVM, Gradient Boosting and MLP classifier is given in table.

TABLE 11 COMPARISON BETWEEN DIFFERENT ML ALGORITHM

Measures	RMSE	Accuracy
Decision tree	0.663 = 66.3%	81%
Ada decision tree	0.350 = 35.0%	88%
Random Forest	0.446 = 44.6%	80%
Ada random forest	0.438 = 43.8%	81%
SVM	0.665 = 66.5%	56%
Ada SVM	0.664 = 66.4%	56%
Gradient Boosting	0.347 = 34.7%	88%
MLP classifier	0.329 = 32.9%	89%

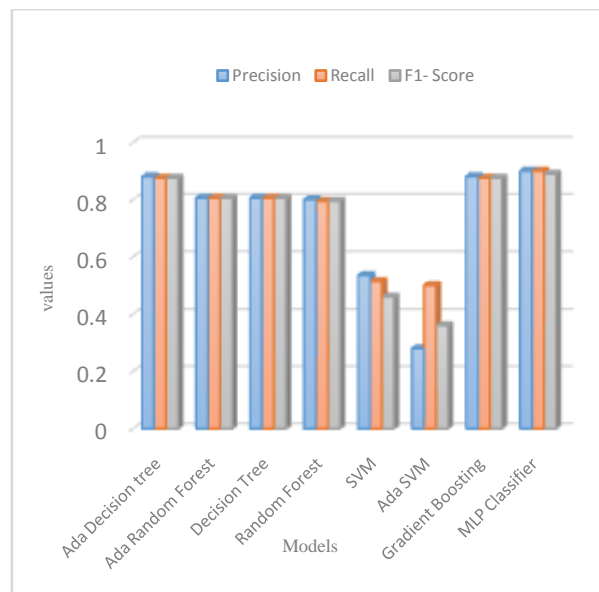


Fig 4 Table shows Comparison of Precision, Recall and F1 score of the different machine learning algorithm

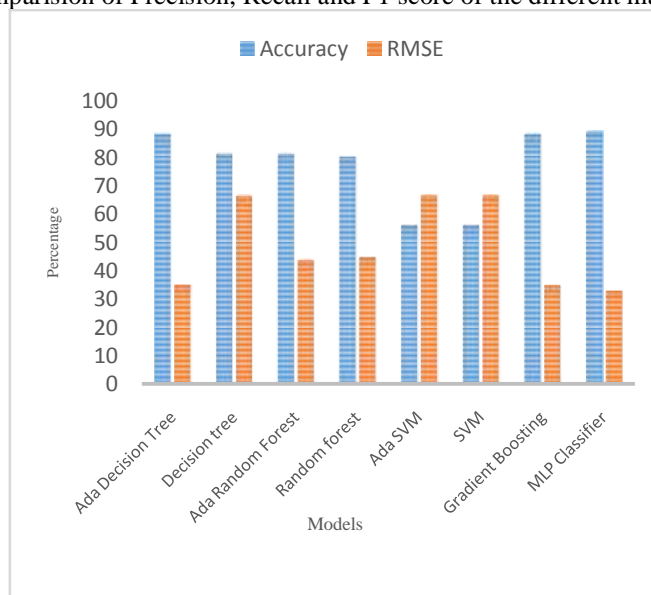


Fig 5 Comparison of Accuracy and root mean square error in percentage

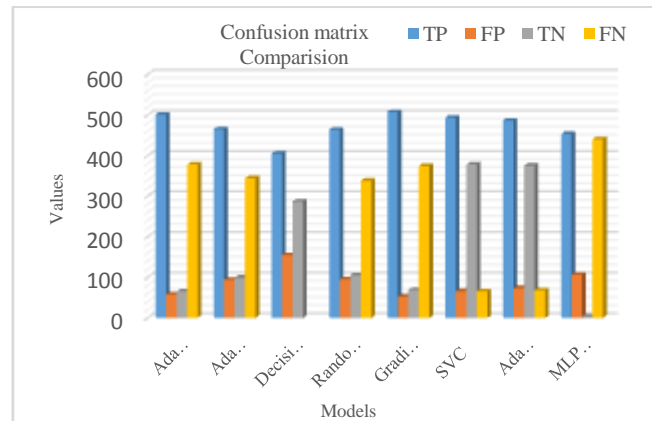


Fig 6 Confusion matrix comparison of the different ML algorithm

ACKNOWLEDGMENT

This paper is written based upon a project of the author during his degree programme in the year 2020 after the invaluable guidance by Mr. Rohit Kumar Tiwari (Asst. Prof.) MMMUT, Gorakhpur. The author of this paper, does not claim rights to any of the algorithm, code, data, formula used, definitions, problem solving approach, as her property. She has only used her brain in compiling it all together and made efforts in obtaining results and putting it together in the format of an IEEE paper.

REFERENCES

- [1] V Kranthi Sai Reddy, "Stock Market Prediction Using Machine Learning," International Research Journal of Engineering and Technology (IRJET), vol:05, Issue:10| Oct 2018.
- [2] I. S. Jacobs and C. P. Bean, "Fine particles, thin films, and exchange anisotropy," in Magnetism, vol. III, G. T. Rado, and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
- [3] M.Ali Ghazanfar, S.Ali Alahmari, Y. Fahad Aldhafiri, A. Mustaqeem, M. Maqsood and M.A. Azam, "Using Machine Learning Classifiers to Predict Stock Exchange Index," International Journal of Machine Learning and Computing, Vol. 7, No. 2, April 2017.
- [4] B.Narayanan and M.Govindarajan, "Prediction of Stock Market using Ensemble Method", International Journal of Computer Applications (0975 – 8887), Volume 128 – No.1, October 2015.
- [5] K.Pahwa and Neha Agarwal, "Stock Market Analysis using Supervised Machine Learning," International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16th Feb 2019.
- [6] Xiao-Ming BAI, Cheng-Zhang WANG, "AdaBoost Artificial Neural Network for Stock Market Predicting", 2016 Joint International Conference on Artificial Intelligence and Computer Engineering (AICE 2016) and International Conference on Network and Communication Security (NCS 2016).
- [7] P.Rajesh, N.Srinivas, K.Vamshikrishna Reddy, Vamsi Priya, Vakula Dwija, M.D.Himaja, "Stock trend prediction using Ensemble Learning technique in python", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-5 March 2019.
- [8] P.P. Singh Kohli, S.Zargar, S.Arora and P.Gupta, "Stock Prediction Using Machine Learning Algorithms", Applications of Artificial Intelligence Techniques in Engineering, Advances in Intelligent Systems and Computing 698 Springer Nature Singapore Pte Ltd. 2019.
- [9] S.Lounnapha, Wu Zhongdong, and C.Sookasame, "Research on Stock Prediction Method Based on Convolutional Neural Network", 2019 International Conference on Virtual Reality and Intelligent Systems (ICVRIS).
- [10] Aizaz Imtiaz Awan, Zeeshan Bhatti, "Predicting Stock Exchange Behaviour using Decision tree and Type Effect Weight (TEW) algorithm", University of Sindh Journal of Information and Communication Technology (USJICT), Volume 2, Issue 3, July 2018.



**International Journal of Advances in
Engineering and Management**

ISSN: 2395-5252



IJAEM

Volume: 03

Issue: 01

DOI: 10.35629/5252

www.ijaem.net

Email id: ijaem.paper@gmail.com