International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 06 | Jun 2022 www.irjet.net p-ISSN: 2395-0072

Performance Comparisons among Machine Learning Algorithms based on the Stock Market Data

Nusrat Mehzabin[1], Mithun Kumar[1], Mirza A.F.M. Rashidul Hasan[2]

[1]Department of Computer Science and Engineering, Bangladesh Army University of Engineering & Technology (BAUET), Natore-6431, Bangladesh
[2]Department of Information and Communication Engineering, University of Rajshahi, Rajshahi-6205, Bangladesh

---

Abstract - *Stock trading is the vital activities which is nonlinear in the real world and analysis on the stock market the crucial aspects of the financial world. Prediction of the financial values of the stock market based on the stock market data is an act that tries to evaluate the future financial value. The stock market is one of the large business platforms where people invest based on some prediction. To avoid investment risk people search for the best algorithms and tools which will increase their profits. The traditional basic methods and technical research may not confirm the effectiveness of the prophecy. This can be done by machine learning algorithms. Therefore, the paper explains the prediction of a stock market using machine learning approaches and shows a comparison among the approaches. In this paper, we identify an efficient approach for predicting future stock market performances. The successful prediction of the stock market will have a very positive consequence on the stock market institutions and the investors also. The paper focuses on applying machine learning algorithms like Linear Regression, Random Forest, Decision Tree, K-Nearest Neighbor (KNN), Logistic Regression, Linear Discriminant Analysis, XG Boost Classifier, Gaussian Naive Bayes on three types of datasets including Combine news, Reddit News, 8 years value of the stock market. We evaluate the algorithms by finding performance metrics like accuracy, recall, precision and fscore. The results suggest that the performance of Linear Discriminant Analysis (LDA) can be predicted better than the other machine learning techniques.*

Key Words: K-Nearest Neighbour; Linear Regression; Linear Discriminant Analysis; Machine Learning; Stock Market; Random Forest; XG Boost Classifier.

1. INTRODUCTION

The stock market has a vital importance in the rapid growing economic country. The stock market includes various customers and dealers of inventory. The country growth are highly related with the stock market hence, there is a linear relation between them [1]. The fundamental approach analyzes stocks that investors perform before investing in a stock where the investors look at the intrinsic value of stocks and performance of the industry, economy, etc. to decide whether to invest or not. Rather than the technical analysis evaluate stocks by studying the statistics generated by market activity like past prices. Stock market prediction method figuring out the destiny scope of marketplace. A system is critical to be built which could work with the most accuracy and it should take into account all crucial elements that would affect the result. Sometimes the marketplace does nicely even if the economic system is failing due to the fact there are numerous motives for the income or lack of a share [2]. Predicting the overall performance of an inventory marketplace is difficult because it takes into consideration numerous elements. The aim is to discover the feelings of investors. It is usually hard as there ought to be a rigorous evaluation of countrywide and global events. It may be very crucial for an investor to recognize the modern-day rate and get a near estimation of the destiny rate. Therefore need more committed output of the prediction algorithms can change the mindset of the people for the business. Currently, the better analysis of machine learning in business fields has inspired many traders to implement machine learning based for highly predicted output. Various researches have already been finished to predict the stock market. There are some mechanisms for stock price prediction that comes under technical analysis such as Statistical method, Pattern Recognition, Machine learning, Sentiment analysis. In this research, we use machine learning algorithms which is the subfield of AI that's extensively described the functionality of a machine to emulate intelligent human behaviour. Machine learning algorithms are either supervised or unsupervised. In Supervised Machine learning, labelled input data is trained and the algorithm is applied. Classification and regression are forms of supervised machine learning. Unsupervised Machine learning has unlabelled data that has a lower managed environment that analyses pattern, correlation, or cluster.

The dataset is an important part of machine learning methods. In this research, various machine learning approaches are employed on a dataset obtained from Kaggle. The paper aims to implement various machine learning algorithms on the stock market data and findings the best approach for the dataset. The rest of the paper consists of the following: First discusses the related works. Then discusses all the prediction methods and finally discusses the experimental results with a conclusion where the last section involves the references.

2. PREDECTION METHODS

The stock market prediction seems a complex problem because many factors have yet to be addressed and it doesn't seem statistical at first. But by proper use of machine learning algorithms, one can relate previous data to the current data and train the machine to learn from it and make appropriate assumptions [6]. The dataset being utilized for analysis was picked up from Kaggle. It consisted of various sections namely "Date, Open, High, Low, Close, Volume and Adj close". All the data was available in a file of CSV format which was first read and transformed into a data frame using the Pandas library in Python. Data preprocessing involves data collecting and removal of noisy and irrelevant something from data is the approach of data cleaning. In this paper, we are applying machine learning techniques to the data to measure overall accuracy, sensitivity and false-positive rate. Although machine learning as such has many models this paper focuses on linear regression, linear discriminant analysis, KNN, support vector machine, random forest and XG Boost for simulation and analysis. All these approaches have been described in this section.

2.1 Support Vector Machine (SVM)

The Support Vector Machine is a discriminative classifier that separate the hyperplane. The SVM is a very famous supervised machine learning technique having a predefined goal variable that may be used as a classifier in addition to a predictor. The outputs of the algorithm is optimal hyperplane for the labeled training data. In the two-dimensional space, this hyperplane is a line dividing a plane into two parts wherein each class lay on either side Support Vector Machine is considered to be one of the most suitable algorithms available for the time series prediction. Both the regression and classification approach uses the supervised algorithm. The SVM involves the plotting of data as a point in the space of n dimensions. These dimensions are the attributes that are plotted on particular coordinates [4].

Many hyperplanes could classify the information. One affordable preference is because the fine hyperplane is the only one that represents the most important separation, or margin, among the 2 classes. So we select the hyperplane so that the gap from it to the closest information factor on every facet is maximized. If this sort of hyperplane exists, its miles are referred to as the maximum margin hyperplane, and the linear classifier it defines is referred to as a maximum-margin classifier or equivalently, the perceptron of most suitable stability.

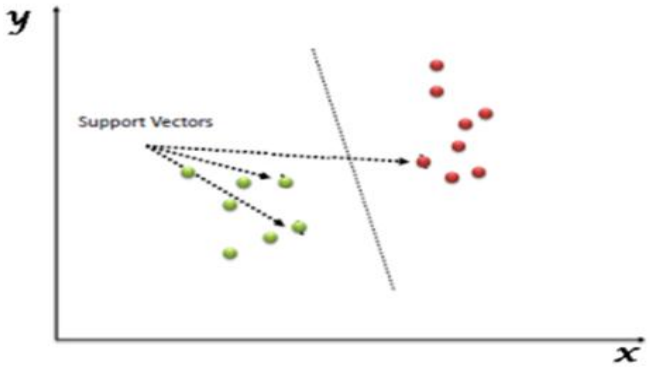

Fig -1 Support Vector Machine.

More formally, a support vector system constructs a hyperplane or set of hyperplanes in a high or infinite-dimensional space, which may be used for classification and regression. Support Vector Machine is one of the maximum famous supervised learning algorithms that's used for classification in addition to regression troubles. The purpose of the SVM set of rules is to create an excellent line or selection boundary that may segregate n-dimensional area.

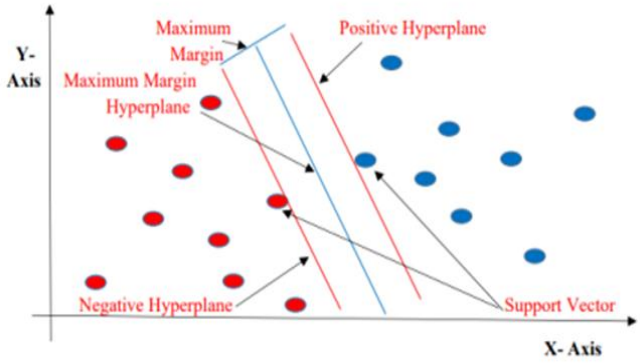

Fig -2 Support Vector Machine (Margin, Hyperplane, Support Vector).

The SVM chooses the acute points/vectors that assist in developing the hyperplane. These intense instances are referred to as assist vectors and therefore set of rules is named a Support Vector Machine.

2.2 Decision Tree (DT) Classifier Algorithm

The decision tree algorithm is one of the family of supervised learning algorithms which can be used for solving regression and classification problems also. Therefore the intention of using a decision tree approach is to generate a training model that can use to calculate the values of target variables by learning decision rules. Compared with other classifiers it is easy to understand that solve the problems based on the tree representation. The every internal node of the tree assemble to an attribute where each leaf node corresponds to a class label. For predicting a class label for record travers from the root of the tree. Then compare the values of the

root attribute with the record's attribute of the tree and based on the comparison we succeed the branch corresponding to that value and then jump to the next node. Hence, for predicting the class values the process continue comparing our record's attribute values with other internal nodes of the tree.

2.3 K-Nearest Neighbor (KNN)

K-Nearest Neighbor (KNN) characterization is a standout amongst the most basic and straightforward arrangement strategies. K Nearest Neighbor is also known as a lazy learning classifier. Classification typically involves partitioning samples into training and testing categories.

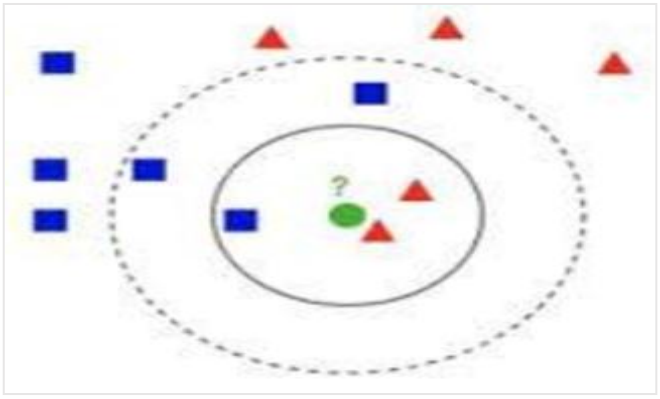

Fig-3 K-Nearest Neighbor (KNN).

During the training process, we use only the true class of each training sample to train the classifier, while during testing we predict the class of each test sample [8]. KNN is a "supervised" classification method in that it uses the class labels of the training data. Unsupervised classification methods, or "clustering" methods, on the other hand, do not employ the class labels of the training data.

2.4 Random Forest

Random Forest is a supervised algorithm and a sort of ensemble learning process. It is a flexible algorithm that can appear in both regression and classification. It is constructed on multiple decision trees. It's mainly building multiple decision trees and merges them for processing results [7]. In this supervised algorithm, a subset of features is taken into consideration. The working procedure is:

1. Randomly select m features.

2. For a node, find the best split.

3. Split the node using best split.

4. Repeat the first 3 steps.5. Build the forest by repeating these 4 steps.

2.5 Logistic Regression

Like many other machine learning techniques, it is borrowed from the field of statistics and despite its name, it is not an algorithm for regression problems, where to predict a continuous outcome. Instead, Logistic Regression is the go-to method for binary classification. Logistic Regression is specially fit for those dependent variables for binomial or multinomial classification. It gives a discrete binary outcome between 0 and 1. Logistic regression measures the relationship between the dependent variable (our label, what we want to predict) and the one or more independent variables (or features) by estimating probabilities using its underlying logistic function. These probabilities must then be transformed into binary values to make a prediction. This is the task of the logistic function, also called the sigmoid function. The Sigmoid-Function is an S-shaped curve that can take any real-valued number and map it into a value between the range of 0 and 1, but never exactly at those limits. These values between 0 and 1 will then be transformed into either 0 or 1 using a threshold classifier. The picture below illustrates the steps that logistic regression goes through to give the desired output.

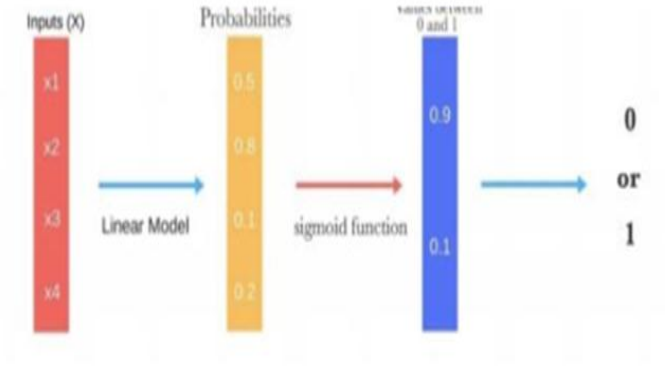

Fig-4 Logistic Function.

2.6 Linear Discriminant Analysis (LDA)

The Linear Discriminant Analysis (LDA) reduces the dimensionality which is commonly used for supervised classification problems. The LDA separates two or more classes based on the modeling differences within groups that is used to project the features in higher dimension space into a lower dimension space [9]. Suppose two sets of data points belonging to two different classes that we want to classify. As shown in the given 2D graph (Fig. 5.), when the data points are plotted on the 2D plane, there is no straight line that can separate the two classes of the data points. Hence, in this case, LDA is used to reduce the 2D graph into a 1D graph to maximize the separability between the two classes.

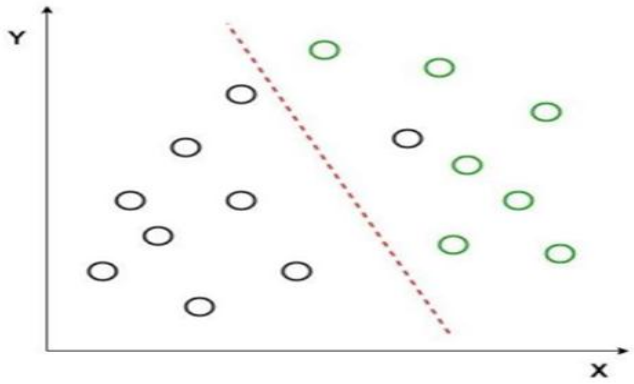

Fig-5 Linear Discriminant Analysis (LDA).

From Fig-5, it can be seen that a new axis is generated and plotted in the 2D graph such that it maximizes the distance between the means of the two classes and minimizes the variation within each class. In simple terms, this newly generated axis increases the separation between the data points of the two classes. After generating this new axis all the data points of the classes are plotted on this new axis. But LDA fails when the mean of the distributions are shared, as it becomes impossible for LDA to find a new axis that makes both the classes linearly separable [10]. In such cases, we use a non-linear discriminant analysis.

2.7 XG Boost Classifier

XG Boost is a decision-tree-based outfit machine learning algorithm that employments a gradient boosting system. In prediction issues including unstructured data (images, content, etc.) artificial neural networks tend to beat all other algorithms or systems. In any case, when it comes to small-to-medium structured/tabular information, choice tree-based algorithms are considered best-in-class right presently. XG Boost algorithm was developed as a research project at the University of Washington. The algorithm differentiates itself in the following ways:

1. A wide range of applications: Can be used to solve regression, classification, ranking and user-defined prediction problems.

2. Portability: Runs smoothly on Windows, Linux, and OS X.

3. Languages: Supports all major programming languages including C++, Python, R, Java, Scala, and Julia.

4. Cloud Integration: Supports AWS, Azure and Yarn clusters and works well with Flink, Spark etc.

XG Boost and Gradient Boosting Machines (GBMs) are both outfit tree strategies that apply the rule of boosting weak learners utilizing the gradient descent architecture. XG Boost approach is attractive for the following purposes.

1. Tree Pruning: The stopping criterion for tree splitting within GBM framework is greedy in nature and depends on the negative loss criterion at the point of split. XG Boost uses max_depth parameter for specified the values than criterion first and computing pruning trees reversely which improves the computational performance significantly of the algorithm.

2. Hardware Optimization: This algorithm has been designed to make efficient use of hardware resources. This is accomplished by cache awareness by allocating internal buffers in each thread to store gradient statistics.

3. EXPERIMENTAL RESULTS

The data was collected and developed so that it can be converted into the form that can be used in the model as inputs. The feature selection methods have been developed in Python programming language with Anaconda, version 1.9.7. The combined dataset consists of top 25 newspapers data in date perspective and the dataset about stock market consists of feature Open, Close, High, Low and Volume. Here we marge these datasets to create a new class label that will have binary values (either 0 or 1). Now we trained datasets using a model and then the test data is run through the trained model. We obtain a confusion matrix that represents the values of "True positive, False negative, False positive and True negative".

True positive is the number of correct predictions that a value belongs to the same class. True negative is the number of correct predictions that a value does belong to the same class. False-positive is the number of incorrect predictions that a value belongs to a class when it belongs to some other class. False-negative is the number of incorrect predictions that a value belongs to some other class when it belongs to the same class. For measuring the performance of the classifiers we applied the measurements of precision, f-score, re-call, support, macro average, weighted average, false-positive rate, and overall accuracy. Here, TP, TN, FP and FN correspond to true positive, true negative, false positive and false negative respectively. The ROC curve analysis was also performed in our study.

Sensitivity is described as the probability of accurately recognizing some conditions. Sensitivity is calculated with the following formula:

$$\text{Sensitivity} = TP / (TP+FN) \quad (1)$$

Precision points to how familiar estimations from separate samples are to each other. The standard error is an example of precision. When the standard error is little, estimations from different samples will be alike in value and vice versa. Precision is measured as follows:

Precision= TP / (TP+FP) (2)

The F score is measured concurring to the accuracy and review of a test practiced under consideration. F- Score is estimated with the help of the following formula:

F-Score= 2TP / (2TP+FP+FN) (3)

In statistics, when conducting various comparisons, a false positive ratio is the probability of incorrectly discarding the null hypothesis for a distinct test. The false-positive rate is determined as the ratio between the numbers of negative results incorrectly classified as positive (false positives) and the total amount of actual negative results.

False Positive Rate= FP / (FP+TN) (4)

Overall accuracy is the possibility that a sample will be accurately matched by a test that is the total of the true positives and true negatives divided by the total number of individuals examined that is the sum of true positive, true negative, false positive and false negative. However, overall accuracy doesn't show the actual performance as sensitivity and specificity may differ despite having higher accuracy. Overall accuracy can be estimated as follows:

Overall Accuracy= (TP+TN) / (TP+TN+FP+FN) (5)

Table 1. Overall accuracy value for different algorithms.

Algorithms	Sensitivity	Precision	F1 Score	Support	Accuracy
LR	1.00	.53	.69	317	0.530988
LDA	.91	.97	.95	317	0.943048
KNN	.45	.43	.44	317	0.458961
CART	.72	.57	.64	317	0.566164
NB	.99	.57	.69	317	0.532663
SVM	.99	.57	.67	317	0.532663
RF	.67	.57	.55	317	0.559463
XG Boost	.60	.79	.67	317	0.591289

Table 1 shows the acquired values of accuracy, sensitivity, precision and f-score for the algorithms that are implemented on the dataset.

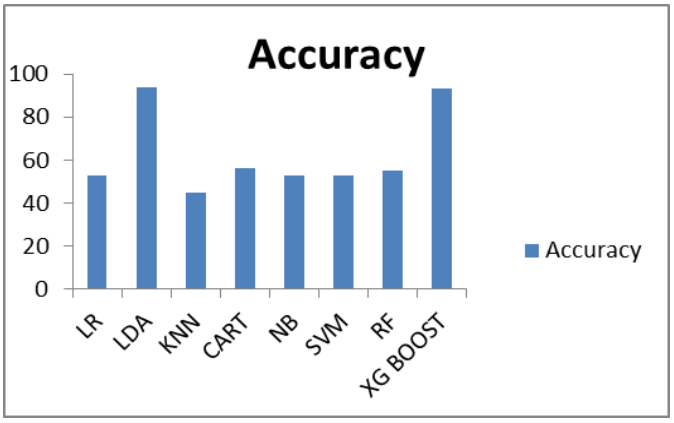

Fig -6 Comparisons among the algorithms based on the accuracy.

Fig -6 shows the comparisons among the algorithms based on the accuracy. From Fig -6 we understand that Linear Discriminant Analysis (LDA) shows the better performance than other algorithms.

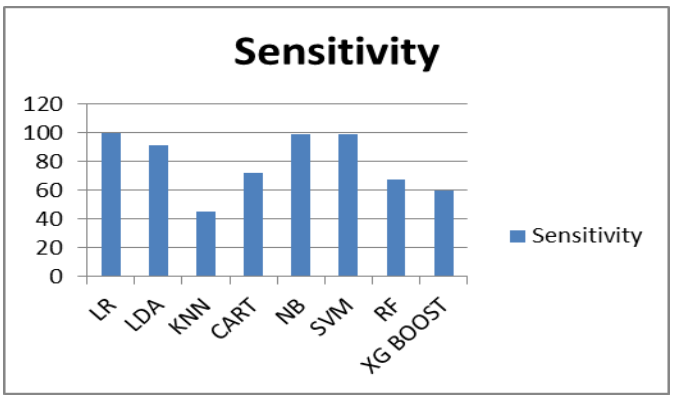

Fig -7 Comparisons among the algorithms based on the sensitivity.

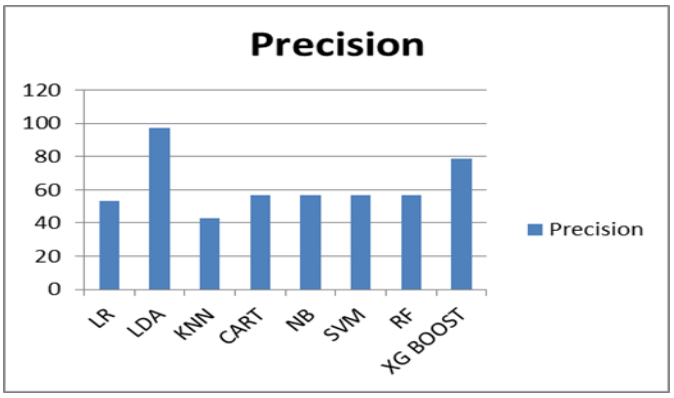

Fig -8 Comparisons among the algorithms based on the precision.

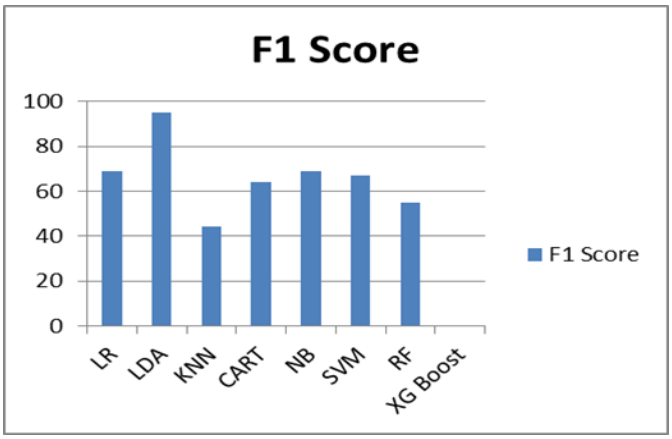

Fig -9 Comparisons among the algorithms based on the F1 Score.

Fig -8 and 9 show the comparisons among the algorithms based on the precision and F1 Score where LDA performs better than other algorithms. The LDA value goes to around 95 to 97 for precision and F1 score which is the better measurement for stock market analysis based on the type of dataset that used in this research. From Table 1 and Fig -6, it can be assured that the linear discriminant analysis algorithms outperform other methods. However, for further evidence, the ROC curve analysis was performed as well.

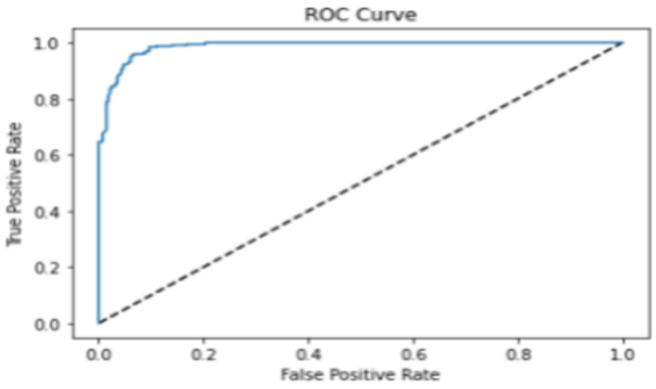

Fig -10 ROC curve.

The observations made from the performance of the algorithms are:

1. Linear Discriminant Analysis (LDA) gives the highest accuracy rate for prediction.

2. Logistic Regression (LR) reaches highest sensitivity.

3. Linear Discriminant Analysis (LDA) reaches highest precision and f-score.

4. KNN is the worst algorithm among these algorithms for prediction in terms of accuracy.

Therefore, from all these experimental results the liner discriminant analysis (LDA) model outperformed than all other studies of stock market analysis for included datasets.

4. CONCLUSIONS

Our research study aims to analyze the stock market data by implementing machine learning algorithms on the datasets. In the stock market business, prediction plays a vital role which is very difficult and challenging process due to the variable nature of the stock market. We applied eight algorithms: Logistic Regression, Linear Discriminant Analysis, Random Forest, SVM, KNN, CART, Random Forest and XG BOOST on the dataset. This paper was an attempt to determine the analysis of the stocks of a company with greater accuracy and reliability using machine learning techniques. We conclude that Linear Discriminant Analysis (LDA) is the best algorithm out of the implemented algorithms with an accuracy rate of 94.3%. In the future, this paper would be adding more parameters that will predict better estimation.

REFERENCES

[1] A. Sharma, D. Bhuriya and U. Singh, "Survey of stock market prediction using machine learning approach," In the Proceedings of International conference of Electronics, Communication and Aerospace Technology (ICECA), pp. 506-509, 2017.

[2] M. Usmani, S. H. Adil, K. Raza and S. S. A. Ali, "Stock Market Prediction Using Machine Learning Techniques", In the Proceedings of 3rd International Conference On Computer And Information Sciences (ICCOINS), pp. 322-327, 2016.

[3] M. P. Naeini, H. Taremian and H. B. Hashemi, "Stock Market Value Prediction Using Neural Networks", In the Proceedings of International Conference on Computer Information Systems and Industrial Management Applications (CISIM), pp. 132-136, 2010.

[4] K. Pahwa, N. Agarwal, "Stock Market Analysis using Supervised Machine Learning", In the Proceedings of International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), pp. 197-202, 2019.

[5] Z. Hu, J. Zhu and K. Tse, " Stocks Market Prediction Using Support Vector Machine", In the Proceedings of 6th International Conference on Information Management, Innovation Management and Industrial Engineering, pp. 115-120, 2013.

[6] M. Ballings, D. V. D. Poel, N. Hespeels and R. Gryp, "Evaluating multiple classifiers for stock price direction prediction", Journal of Expert System Application, 2015, Vol. 42, pp. 7046-7056.

[7] S. Jain and M. Kain, "Prediction for Stock Marketing Using Machine Learning", An International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 6(4), pp. 131-135.

[8] M. S. Babu, N. Geethanjali and B. Satyanarayana, "Clustering Approach to Stock Market Prediction", An International Journal of Advanced Networking and Applications, 2012, vol. 03(04), pp.1281-1291.

[9] T. Tantisripreecha and N. Soonthornphisaj, "Stock Market Movement Prediction using LDA-Online Learning Model", In the Proceedings of 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), pp. 135-139, 2018.

[10] I. Parmar, N. Agarwal, S. Saxena, R. Arora, S. Gupta, H. Dhiman and Lokesh Chouhan, "Stock Market Prediction Using Machine Learning", In the Proceedings of First International Conference on Secure Cyber Computing and Communication(ICSCCC), pp. 574-576, 2018.