An Efficient Prediction Model based on Machine Learning Techniques for Prediction of the Stock Market

An Efficient Prediction Model based on Machine Learning Techniques for Prediction of the Stock Market

Omar D. Madeeh [1*]; Hasanen S. Abdullah [2]

[1*]Department of Computer Sciences, University of Technology, Baghdad, Iraq.

[2]Department of Computer Sciences, University of Technology, Baghdad, Iraq.

omar.dahfer@gmail.com

Abstract. Nowadays, the stock market's prediction is a topic that attracted researchers in the world. Stock market prediction is a process that requires a comprehensive understanding of the data stock movement and analysis it accurately. Therefore, it needs intelligent methods to deal with this task to ensure that the prediction is as correct as possible, which will return profitable benefits to investors. The main goal of this article is the employment of effective machine learning techniques to build a strong model for stock market prediction. The work involved three stages, the first stage involved preprocessing for the stock market data set, then the second stage which involved employing two from supervised machine learning techniques namely K-Nearest Neighbor (K-NN) and Random Forest (RF), and finally, the evaluation stage of accuracy and efficiency of the prediction for the two proposed models. The results experiments showed that the two proposed models achieved a high accuracy ratio and the RF model was the best of prediction accuracy, where it reached 93.23%, 93.12% and 93.17% respectively according to evaluation measures precision, recall, and F-measure.

1. Introduction

The stock market's decision-making has been a task highly difficult because of the complicated behavior, also unstable nature, that is related to market. There has been a required for exploring huge amounts of significant data created via stock market for various companies or sectors [1]. Stock market prediction which is also referred to as the attempt to specify share's future values regarding certain company or financial instrument that is traded on exchange, in which efficient prediction related to the future value of stock might be yielding considerable profits to customers [2]. One of the main components of the stock market is the customer basket, which represents the traded shares group for a number of companies and sectors.

Prediction techniques encompass different statistical analytics approaches specified via machine learning, predictive modelling, and data mining, which is analyzing historical and current facts for making predictions related to future or unrecognized events [3] [6]. With regard to business, the predictive models exploiting the patterns indicated in the historical as well as transactional data for determining future value as much as possible to avoid risks and achieve greater opportunities in profit [4].

Data mining includes using complicated tools of data analysis for discovering formerly undefined, significant patterns, also the relations in large data-set. Such tools might involve machine learning approaches, mathematical algorithms, and statistical models [5].

The process of right decision-making in data mining tasks depends on employing the techniques and methods of machine learning whether supervised or unsupervised. Machine learning can be defined is the study of computer techniques that can improve automatically. These techniques create a mathematical model for the purpose of making decisions or predictions for a certain task. There are many techniques and methods that can be used to analyze and forecast stock market data [7].

This article contributes to employment effective machine learning techniques for the prediction of sectors and companies for the stock market, which helps investors achieve better profit and avoid loss as possible.

2. Related Works

In last few years the stock market is interesting field of research. Many more works are proposed by researcher to predict the stock market. The presented sections involved some studies which are specified as related works for checking the stock market prediction system with the use of the data mining approaches and they are summarized as follows.

Khalid, A., Hassan, N., Ismail, H., Mohammed, K. and Ali, S. [8], in this work, they presented applied KNN as well as non-linear regression method, which are two of machine learning algorithms for the purpose of predicting the stock prices with regard to a sample from 6 companies that are listed on Jordanian stock exchange for assisting investors, users, decision-makers, and management to make the informed and right investments decisions. Based on the results, KNN has high robustness with simple error ratio; therefore, the results have been good to very good. Also, on the basis of data for actual stock prices; prediction results have been approximately parallel to the actual stock prices.

Maryam Farshchian and Majid Vafaei Jahan [9], they developed adequate model to increasing the precision related to forecasting the behavior of Stock market Exchange in Tehran with the use of Hidden Markov Model. The model was trained for three specific industries. The dataset has been collected from Tehran Stock market Exchange from 26-03-2011 to 09-12-2014 for three different industries, Shargh Cement Company, Shiraz Petrochemical Company, Jaber Ebne Hayyan Pharmaceutical Company. The dataset was divided into 70% for the purpose of training and the remaining 30 % for testing. The implementation results showing that the maximum precision, F1 measure as well as accuracy have been for Jaber Ebne Hayyan Pharmaceutical Company with 78.57% , 79.37% and 82.37% with the use of Hidden Markov Model, while, the rest of the industries gave forecast accuracy rang 69% to 82% only.

Radu Iacomin [11], he presented method for prediction of a future trend for the values of the stocks in stock market. He used one of the algorithms related to machine learning for make a safe prediction of stock values. Support Vector Machines algorithm (SVM) was used with the help of feature selection Principal Component Analysis (PCA) for the purpose of making a correct predictive decision. The dataset has been collected from Bloomberg, which is a platform for trading stocks where he chosen sixteen forex stocks in Swiss securities market. The results indicated that the proposed model (PCASVM) given an accuracy rate of 68% according to performance standard Rate of Recognition (ROR).

Ayman E., Salama S. and Nagwa Y. (2017) [1], they developed an approach to predict market securities in future patterns with a little blunder proportion and improve the precision of expectation. This forecast model relies upon a notion investigation regarding securities and financial news for exchanging prices, such model provides optimum accuracy results over every single past investigation by considering many kinds of news identified with the market, also with the authentic stock costs. A data-set involves stock costs from 3 organizations which are utilized. The underlying advance is to examine news sentiment to get the substance limit utilizing the credulous Bayes calculation. This process is achieved to estimate exact results running from 73% to 86%. The approach can give final results with a precision of 89%.

3. Proposed Approach

In this research work, it is proposed to build a model for forecasting the financial stocks of companies and sectors for the NYSE stock market, based on two types of machine learning Classifiers namely: KNN, RF. The Figure 1 shows the general structure of the suggested method.

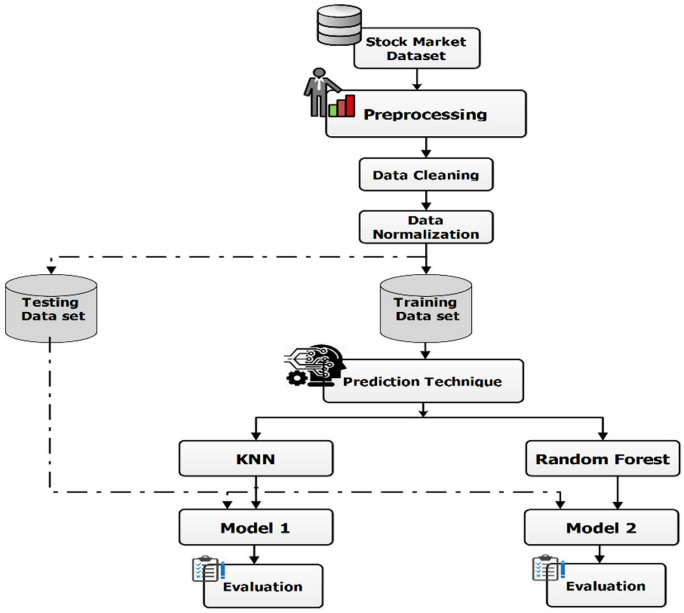

Figure 1. Structure design of the suggested approach

3.1. Data Set

Data-set has been a very basic module and the initial step towards the approach build. The dataset has been collected from New York Stock Exchange (NYSE) from Kaggle repository for Dow Jones and is an industrial index of the largest US manufacturing companies on NYSE.

The data-set includes (Name, Close, High, Date, Low, Open, Close, and Volume) attributes. It included more fifteen company and sector for a period of twelve months, with a total of more than 4,500 trading days in stock market. Table 1 shows description related to features in data set.

Table 1. Description of data-set Features

Feature	Description
Date	The date specified for each trading day in a format year-month-day
Open	price related to stock at market open (all data in in USD).
High	Maximum daily prices
Low	Minimum daily prices
Close	The stock price when the stock market is closed for a specific trading day.
Volume	Number of traded shares
Name	Ticker name of socks (company or sector)

3.2. Preprocessing

The preprocessing represents the first stage in the prediction model build. Recently, the majority of data in real- world have been incomplete involving aggregate, noisy as well as missing values. Due to the

fact that the quality decision is on the basis of quality mining that depends on quality data, preprocessing is a task of high importance to be done prior to achieving mining processes.

The main tasks in the preprocessing of data have been integration, reduction, cleaning, and transformation of data [24]. With regard to such phase, data-set's passed in two steps, normalization and cleaning.

3.2.1. Data Cleaning

This is considered as the initial step in the stage of data pre-processing, which is used to find smooth noise data, missing values, recognize outliers as well as correct inconsistent. These unwanted data will affect on mining procedure and led to unreliable and poor results [12].

In this step the missing values in data set was address by using the attribute mean strategy for filling the missing value where the approach works by replacing the missing value for particular attribute by the average value for that attribute.

3.2.2. Data Normalization

Data Normalization is method works by an adjusting the data values into a specific range such as between 0-1 or -l-l. This method is useful for mining techniques. Normalization is used to scale the data attributes and can be used to speed the learning stage [13]. The normalization is calculate using Eq. (1) below:

$$\chi_{scaled} = \frac{\chi - \chi_{min}}{\chi_{max} - \chi_{min}} \quad (1)$$

Where

X: refer to original value in dataset features.

Xmin: refer to minimum value in dataset features.

Xmax: refer maximum value in dataset features.

4. Prediction Techniques

This section describes second stage after preprocessing stage which represents the techniques that have employed in building the proposed model. Section (I.) describes the K-NN technique, while section (II.) describes the Random Forest (RF) technique.

4.1. K-Nearest Neighbor (KNN)

This is considered as statistical approach and considered simplest machine learning. It attempts on classifying the unrecognized samples on the basis of recognized classifications regarding its neighbors [14]. The major aim of this algorithm is memorizing training set, after that predicting the label related to all new instance based on the labels related to the closest neighbor in training set [15]. KNN was majorly applied in classification problems. It has been on the basis of distance function measuring the similarity or difference between 2 instances. Standard Euclidean distance d (x, y) between 2 instance x and y has been utilized as distance function [16]. Distance function has been specified as Eq. (2):

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (a_i(x) - a_i(y))^2} \quad (2)$$

There are 2 major benefits related to KNN, which are flexibility and efficiency. It is majorly known for its scalability, simplicity speed, and effectiveness.

4.2. Random Forest (RF)

The Random Forests algorithm is a classification technique developed by Breiman [20]. It can be defined as one of the supervised classification algorithms, it is creating multiple decision trees on the basis of

random sub-samples from data, each with the ability to produce results in the case when provided with prediction values. High number of trees, will lead to high accuracy and minimum overfitting risks in comparison to the other models. RF model creates 'n' number of trees as weak classifiers as well as merging all trees in forest. In the case when RF model has been applied for regression the mean related to the resulting values from all decision trees has bene the resulting prediction value, in the case when utilized for classification, resulting class has been the mode of resulting classes from decision tree [21]. There have been 2 random processes in RF.

The first one is that the training sets have been created with the use of boot-strap method randomly with replacement. The other procedure is that the random features have been chosen with the non-replacement from total features in the case when the trees' nodes have been split. Size κ related to feature sub-set has been typically less in comparison to the size related to total features, M. The initial step is randomly selecting κ features, calculating information gain related to κ split and selecting optimum. Therefore, the size of candidate features will be M− κ. After that, continue [22]. Figure 2 showing RF procedures in the training sets.

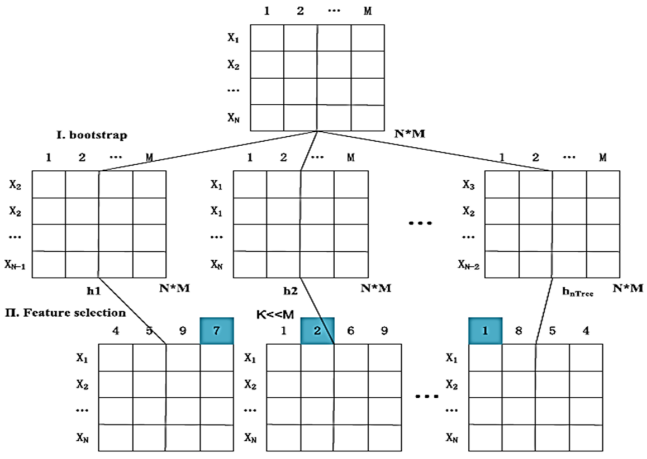

Figure 2. Random Forest procedures in training sets

A result, there will be many trees trained in a weaker way and each of them will produce a different prediction. The ways to interpret these results based on a majority vote (the most voted class will be considered correct) or averaging results, which yields very accurate predictions. The Figure 3 shows how random forest trees predict the end result.

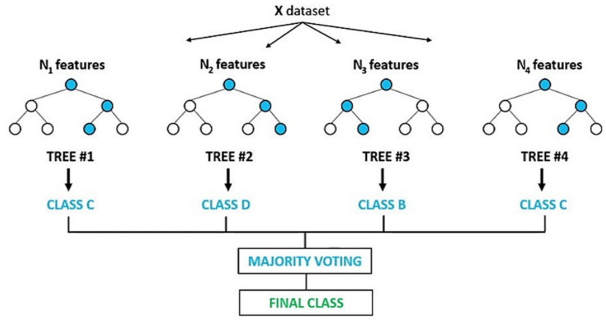

Figure 3. prediction method in Random Forest trees of the end

Random forest is very useful for data interpretation and prediction goals. One of its most important applications is the stock market prediction Because of the high volatility in the stock market, the task of predicting will become quite challenging.

RF provides many significant characteristics, for instance, the ability for dealing with high dimensional data, difficult correlations, and interactions. Furthermore, it is one of the major powerful algorithms, as extremely robust and accurate approach in prediction that it is not suffering from overfitting problem, since it is taking average for all predictions, that conceal biases.

5. Performance Evaluation Measures

The final stage of our work is to check the prediction model, for this purpose, we used common measures for evaluating the efficiency and accuracy of the prediction model, this work provides 3 measures, which are precision, recall, accuracy and F1-measure was applied. Estimating such measures is on the basis of assessing the confusion matrix that it is matrix in which the test values have been distributed through creating 2 classes as can be seen in Table 2 [23].

Table 2. Confusion Matrix Classes

	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Where:

- True positive (TP): specifying the number of positive instances which have been adequately classified
- False Negative (FN): specifying the number of positive instances which have been inadequately classified.
- False Positive (FP): specifying the number of negative instances which have been inadequately classified.
- True negative (TN): specifying the number of negative instances which have been adequately classified.

5.1. Precision Measure

This is a proportion related to the predicted shares for certain class which have been classified correctly.

$$Precision = \frac{No.of\ true\ positives}{No.of\ true\ positives + false\ positives} \quad (3)$$

5.2. Recall Measure

This is defined as the proportion related to all shares for certain class which have been classified correctly.

$$Recall = \frac{No\ of\ true\ positives}{No\ of\ true\ positives + false\ positives} \quad (4)$$

5.3. F1-Measure

Might be utilized for estimating the performance related to shares classifiers through mixed recall and precision.

$$F1 = 2 * \frac{precision * recall}{precision + recall} \quad (5)$$

Error metrics is also has been used, which is specified as the percentage related to data-set classified inadequately, MAE and RMSE has been utilized on each prediction model for the two techniques. the Eq. (7) defined as MAE, and Eq. (8) defined as RMSE.

$$MAE = \frac{1}{N}\sum_{k=1}^{N}|y_k - \hat{y}_k| \quad (6)$$

$$MSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2} \quad (7)$$

Where:
Yk : refer to actual value.
Y^k: refer to predicted value.
N: refer to number of data samples.

6. The Results and Discussion

This study was implemented by using python as programming language, under Microsoft windows 10, 64-bit OS, 4GB RAM, and with CPU 2.7GHz core i7. The proposed approach is trained and tested over the dataset taken from NYSE stock market. The data set was divided into two parts, 70% for the purpose of training the model and the remaining 30% for testing and evaluating the model. The Tables 3 and 4 show the results related to the implementation of two prediction models, KNN and RF.

The structure of Tables 3 and 4 consists of four columns, the first column represents The company or sector class in the NYSE stock market whose stock values have been predicted and the remaining three columns represent the percentage of prediction accuracy for each company based on the company's stock data.

Table 3. prediction accuracy results for KNN technique

Company Class	Precision	Recall	F-Measure
MMM	97.5	93.9	95.7
AXP	92.2	86.5	89.2
AAPL	98.6	100	99.3
KO	92.1	93.3	92.7
XOM	88.8	83.5	86.1
GE	97.5	100	98.7
GS	89.2	95.7	92.3
INTC	94.9	90.2	92.5
JNJ	92.9	97.5	95.2
MRK	95.9	93.3	94.6
MSFT	83.3	91.5	87.2
PG	83.6	88.4	85.9
TRV	92.9	87.8	90.3
UTX	86.5	84.9	85.7
GOOGL	74.7	84.4	79.3
AMZN	78	67.6	72.4

It is observed of the results of Table (3) that the K-NN model achieved a high prediction accuracy for companies and sectors for the (NYSE) market. and the average accuracy ratio were 90.21%, 90.13% and 90.15% respectively according to accuracy measure precision, recall, and F-measure.

Table 4. prediction accuracy results for Random Forest technique

Company Class	Precision	Recall	F-Measure

MMM	98.7	95.1	96.9
AXP	91.4	88.5	89.9
AAPL	98.6	100	99.3
KO	94.7	96	95.4
XOM	90.1	85.9	88
GE	100	100	100
GS	94.4	98.6	96.5
INTC	96.3	95.1	95.7
JNJ	92.8	95.1	93.9
MRK	100	94.7	97.3
MSFT	89.5	95.8	92.5
PG	85.3	92.8	88.9
TRV	88	89.2	88.6
UTX	85.7	79.2	82.4
GOOGL	93.2	89.6	91.4
AMZN	88.7	92.6	90.6

It is observed of the results of Table (4) that the RF prediction model achieved better prediction accuracy for companies and sectors for the (NYSE) market from the K-NN model. where the average accuracy ratio was 93.23%, 93.12% and 93.17% respectively according to accuracy measure precision, recall, and F-measure.

Figure 4 shows the results related to prediction accuracy ratios for 2 prediction models (K-NN) and (RF), and indicating that RF is the best in its accuracy ratio according to precision, recall, and F-measure.

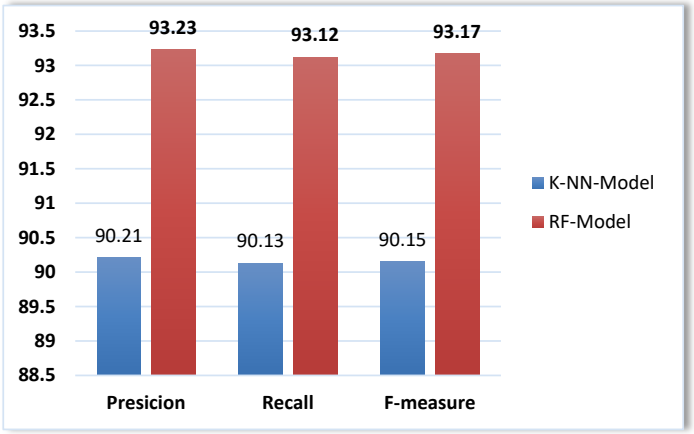

Figure 4. Accuracy of prediction for proposed

Table 5 shows summary of the results of accuracy of two prediction model according to evolution measures has used and errors rate for each technique.

Table 5. Results summary for prediction models

Prediction technique	Performance Measures			Error Rate	
	Precision	Recall	F-Measure	MAE	RMSE
K-NN	90.21%	90.13%	90.15%	0.0156	0.0967
Random Forest	93.23%	93.12%	93.17%	0.0127	0.0809

7. Conclusion

Stock market prediction is a very difficult task because of the volatile nature of the movement of financial share data for sectors and companies in the stock market. The ideal solution for achieving efficient and accurate forecasting is employing artificial intelligence in applying machine learning techniques. Two powerful techniques have been proposed to build an Efficient prediction model for forecasting companies and sectors for the New York Stock Exchange (NYSE).

The results of the experiments demonstrated that the RF model achieved the highest prediction accuracy in precision, recall, and F-measure with a ratio of 93.23%, 93.12%, and 93.17% respectively, comparing with the K-NN model that gave the prediction accuracy with a rate of 90.1%.

The Random Forest technique is proved Efficient in terms of efficiency and accuracy because of its strategy in building the predictive model. Where it works is to create a large number of predictive trees and each tree gives a specific result. Therefore, the final result of the prediction relies on voting or averaging the results that make it more accurate compared to the other prediction models. Therefore, employing the Random Forest (RF) prediction model in the stock market It can provide more profits for investors and reduce the risk of loss as much as possible.

References

[1] Khedr, Ayman E., and Nagwa Yaseen. "Predicting stock market behavior using data mining technique and news sentiment analysis." International Journal of Intelligent Systems and Applications 9.7, pp. 22. (2017).

[2] Chittineni, Suresh, et al. "A Comparative Study of CSO and PSO Trained Artificial Neural Network for Stock Market Prediction." International Conference on Computational Science, Engineering and Information Technology. Springer, Berlin, Heidelberg,pp. 186-195,2011.

[3] M. Kumari and S. Godara, “Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction”, IJCST ISSN: 2229- 4333, vol. 2, no.2, (2011).

[4] Khalid, Balar, and Naji Abdelwahab. "Big Data and Predictive Analytics: Application in Public Health Field.”, International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol6, No.5, 2016.

[5] S.Archana and Dr. K.Elangovan, “Survey of Classification Techniques in Data Mining”, International Journal of Computer Science and Mobile Applications, Vol. 2 Issue. 2, February 2014.

[6] Nyce, Charles. "Predictive Analytics White Paper, sl: American Institute for Chartered Property Casualty Underwriters." Insurance Institute of America, p.1, (2007).

[7] Shah, Dev, Haruna Isah, and Farhana Zulkernine. "Stock Market Analysis: A Review and Taxonomy of Prediction Techniques." International Journal of Financial Studies, 7.2, pp. 26 (2019).

[8] Alkhatib, Khalid, et al. "Stock price prediction using k-nearest neighbor (kNN) algorithm." International Journal of Business, Humanities and Technology 3.3, pp. 32-44, (2013).

[9] Farshchian, Maryam, and Majid Vafaei Jahan. "Stock market prediction with hidden markov model." 2015 International Congress on Technology, Communication and Knowledge (ICTCK). IEEE, 2015.

[10] BhaveshPatankar and Dr. Vijay Chavda, “A Comparative Study of Decision Tree, Naive Bayesian and k-nn Classifiers in Data Mining”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 4, Issue 12, December 2014.

[11] Iacomin, Radu. "Stock market prediction." 2015 19th International Conference on System Theory, Control and Computing (ICSTCC). IEEE, 2015.

[12] Maingi, Mathew Ngwae. "Survey on Data Preprocessing Concept Applicable in Data Mining." International Journal of Science and Research (IJSR), pp. 2319-7064, (2013).

[13] Alasadi, Suad A., and Wesam S. Bhaya. "Review of data preprocessing techniques in data mining." Journal of Engineering and Applied Sciences, 12.16, pp. 4102-4107, (2017).

[14] R.C. Neath, M.S. Johnson. “Discrimination and Classification”. International Encyclopedia of Education (Third Edition), , Elsevier , PP 135-141, 2010.

[15] Shalev-Shwartz, Shai, and Shai Ben-David. “Understanding machine learning: From theory to algorithms”. Cambridge University Press (CUP), 2014.
[16] Jiang, Liangxiao, et al. "Survey of improving k-nearest-neighbor for classification." Fourth international conference on fuzzy systems and knowledge discovery (FSKD 2007). Vol. 1. IEEE, 2007.
[17] Yildirim, Pinar. "Filter based feature selection methods for prediction of risks in hepatitis disease." International Journal of Machine Learning and Computing, 5.4, pp. 258, (2015).
[18] Jadhav, Sayali D., and H. P. Channe. "Comparative study of K-NN, naive Bayes and decision tree classification techniques." International Journal of Science and Research (IJSR) 5.1, pp. 1842-1845, (2016).
[19] Bonaccorso, Giuseppe. Machine learning algorithms,Reference guide for popular algorithms for data science and machine learning,1st Edition.Packt Publishing Ltd, 2017.
[20] Abdulsalam, Hanady, David B. Skillicorn, and Patrick Martin. "Streaming random forests." 11th International Database Engineering and Applications Symposium (IDEAS 2007). IEEE, 2007.
[21] Maini, Sahaj Singh, and K. Govinda. "Stock market prediction using data mining techniques." 2017 International Conference on Intelligent Sustainable Systems (ICISS). IEEE, 2017.
[22] Ma, Li, and Suohai Fan. "CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests." BMC bioinformatics, 18.1, pp. 169, (2017).
[23] Santra, A. K., and C. Josephine Christy. "Genetic algorithm and confusion matrix for document clustering." International Journal of Computer Science Issues (IJCSI) 9.1, pp.322. (2012).
[24] Kumar, R. Kishore, G. Poonkuzhali, and P. Sudhakar. "Comparative study on email spam classifier using data mining techniques." Proceedings of the International Multi Conference of Engineers and Computer Scientists. Vol. 1., pp. 14-16, 2012.