TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 9141 - 9150

A Comparative Study of K-Nearest Neighbor, Naive Bayes and Random Forest techniques for Stock Market Prediction Model

Omar D. Madeeh, Hasanen S. Abdullah
University of Technology - Iraq

Article Info
Volume 83
Page Number: 9141 – 9150
Publication Issue:
May - June 2020

Article History
Article Received*: 19 November 2019*
Revised*: 27 January 2020*
Accepted*: 24 February 2020*
Publication*: 18 May 2020*

Abstract:
Nowadays, the stock market's prediction is a topic that attracted researchers in different countries. Stock market prediction is a process that requires a comprehensive understanding of the data and analysis it accurately. Therefore, it needs intelligent methods to deal with this task to ensure that the prediction be as correct as possible, which will return profitable benefits to investors. The large number of companies traded in the stock market basket for various sectors makes it difficult for investors to predict the shares of a particular company or sector. This study aim discusses using the techniques of data mining for selecting the best model for forecasting financial stocks for companies. The study proposes to study three powerful forecasting techniques, namely Naïve Bayes (NB), Random Forest (RF), and K-Nearest Neighbor (KNN) for NYSE stock market data. The results of the experiments showed that Random Forest technique is best for companies' stock prediction according to the error rate metrics, precision, recall, and F-measure.

Keywords: *Data mining, Stock market, K-Nearest Neighbor, Naïve Bayes, Random Forest.*

I. Introduction

The stock market's decision-making has beena task of highimportancebecause ofof 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

May-June 2020
ISSN: 0193-4120 Page No. 9141 - 9150

relations in large data-set. Such tools might involve machine learning approaches, mathematical algorithms, and statistical models [5].

The data mining is one of the major significant analysis steps related to the process of knowledge discovery in database (KDD). The major aim of data mining has been extracting significant information from the large raw data, then convert it to a form that is understandable to be used efficiently. Generally, the tasks of data mining might be divided into 2 categories: predictive and descriptive classification methods [10].

There are many techniques and methods that can be used to analyze and forecast stock market data [7]. This paper discusses the use of three powerful and more common techniques in data mining for build approach to predict financial stocks for companies. In this work are applied KNN, RF and NB techniques to the New York Stock Exchange (NYSE) as a comparative study to find the best predictive model among them.

II. 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).

May-June 2020
ISSN: 0193-4120 Page No. 9141 - 9150

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%.

III. 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 three Classifies namely: KNN, RF and NB. The Fig (1) shows the general structure of the suggested method.

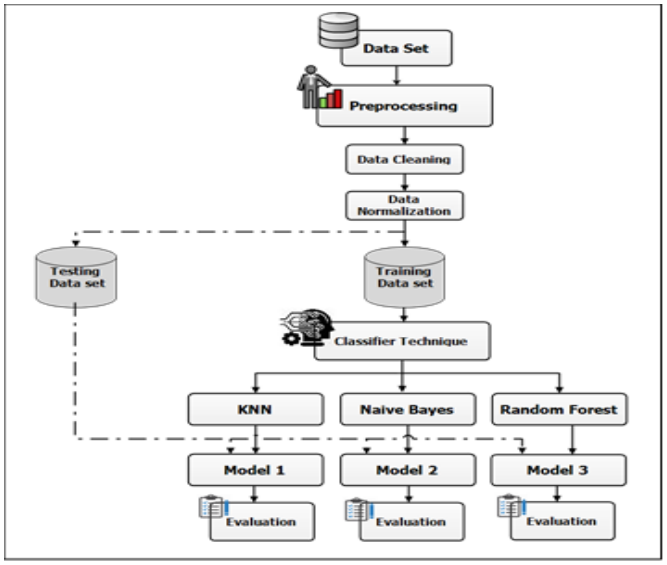

Fig 1. Structure design of the suggested approach

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 eighteen company and sector for a period of twelve months, with atotal of more than 4,500 trading days in stock market.

The Table (1) showing 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 formatyear-month-day
Open	price related to stock at market open (all data in in USD).
High	Maximum daily prices
Low	Minimumdaily 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)

2. Preprocessing

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,

TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 9141 - 9150

data-set's normalization and cleaning are achieved prior to conducting prediction approach.

2.1 Data Cleaning

This is considered as the initial step in the stage of data pre-processing, which is used tofind smooth noise data, missing values, recognize outliers as well as correct inconsistent. These unwanted data will effect 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.

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 equation (1) below:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

Where

X: refer to original value in dataset features.

Xmin: refer to minimum value in dataset features.

Xmax: refer maximum value in dataset features.

3. Prediction Techniques suggested in the Approach

There are many prediction techniques, this section explains the techniques that have been proposed for study in building our model.

A. 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 equation (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.

B. Naïve Bayes (NB)

This is considered as simple and important train classifier which is determining the possibility of outcome providing set of conditions with the use of Bayes' theorem [19].NB classifier can be considered as on of the supervised learning algorithm, also uncomplicated probabilistic classifier which estimates the set of probabilities through counting the frequency as well as the collections of values in certain data-set. NB classifier indicates the appearance of attributes in category isn't relying on the appears related to other ones[17]. NB classifier approach is on the basis of Bayesian Theorem, also it is utilized in the case when input's dimensionality has been high. Also, theBayesian classification depends on Bayes Theorem, which will be indicated as follows:

AssumingAhas been data sample for stock with Class B, specifying the name of company, assuming H is

May-June 2020
ISSN: 0193-4120 Page No. 9141 - 9150

certain hypothesis, in a way that data sample Amight be belonging to certain class B. The Bayes theorem has been applied to calculate posterior probability P(B|A), from P(B), P(A), and P(A|B) [18]. The Bayes theorem might be specified from equation (3) as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (3)$$

Where:

P(A):can be defined as prior probability related to class which is represent Company name in stock market.

P(B|A):can be defined as the likelihood that has been probability of given class's predictor, which is represent likelihood certain stock for certain company.

P(B):specified as the prior probability related to class's predictor.

There are a number of advantages of NB technique,the most important of which are [18]:

- It is requiring not much training's computational time.
- Improving the performance of classification via eliminating unrelated features.
- Good performance.

C. 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 beentypically 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.

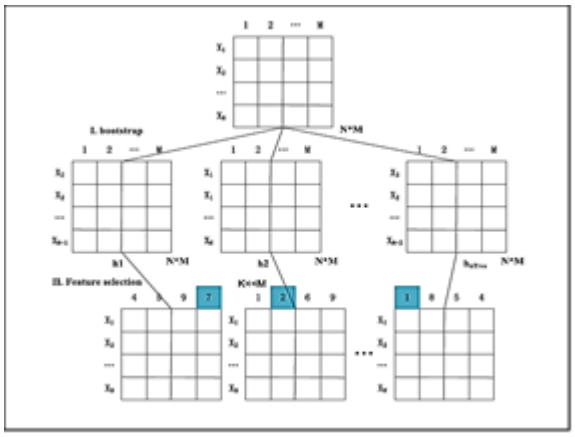

Fig 2. Random Forest procedures in training sets

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 Fig (3) shows how random forest trees predict the end result.

May-June 2020
ISSN: 0193-4120 Page No. 9141 - 9150

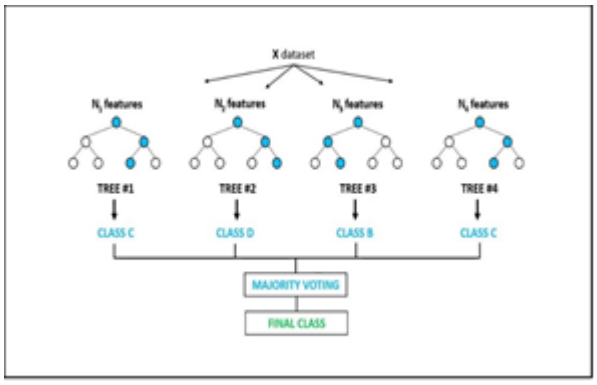

Fig 3. prediction method in Random Forest trees of the end result

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.

4. Performance Evaluation Measures

Evaluation Measures use for evaluating the efficiency of the prediction's method, 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.

A. Precision Measure: this is a proportion related to the predicted shares for certain class which have been classified correctly.

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

B. Recall Measure: this is defined as the proportion related to all shares for certain class which have been classified correctly.

$$\textbf{Recall} = \frac{No\ of\ true\ positives}{No\ of\ true\ positives + false\ negatives} \quad (5)$$

C. F1-Measure:might be utilized for estimating the performance related to shares classifiers through mixed recall and precision.

$$\textbf{F1} = 2 * \frac{precision * recall}{precision + recall} \quad (6)$$

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

May-June 2020
ISSN: 0193-4120 Page No. 9141 - 9150

each prediction model for the three techniques.the equation (7) defined as MAE, and equation (8) defined as RMSE.

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

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(y_k - \hat{y}_k)^2} \quad (8)$$

Where:

$\mathbf{Y_k}$: refer to actual value.

$\mathbf{Y\hat{}_k}$: refer to predicted value.

N: refer to number of data samples.

IV. EXPERIMENTAL RESULTS

This study is implemented by using python as programming language, under Microsoft windows 10, 64-bit OS, 4GB RAM, and with CPU 2.67GHz 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), (4) and (5) show the results related to implementation of three approaches, KNN, NB and RF.

Table (3) : predictionaccuracy results for KNN technique

Class	Precision	Recall	F-Measure
MMM	90	92.1	91.6
AXP	90	91.9	91.4
AAPL	99.6	99.4	99.4
KO	90	85	84.2
XOM	86.5	86.8	86
GE	98.8	97.3	97.1
GS	94.4	90.6	90.1
INTC	86.5	87.9	87.2
JNJ	97.6	94.8	94.5
MRK	84.9	84.4	83.4
MSFT	92	88.8	88.2
NKE	77.3	80.2	79.1
PG	89.6	89.5	88.8
TRV	89.2	89.2	88.6
UTX	82.5	85.9	85.2
VZ	68.1	72.3	70.9
GOOGL	77.7	75.7	74.3
AMZN	72.5	74.3	72.8
	87.1	**87.1**	**87**
	Weighted Average		

Table (4) : predictionaccuracy results for Naïve Bayes technique

Class	Precision	Recall	F-Measure
MMM	95.5	76.5	85
AXP	75.6	75.3	75.4
AAPL	97.7	84.1	90.4
KO	68.1	98	80.4
XOM	81	81.3	81.1
GE	100	100	100
GS	81.6	97.2	88.7
INTC	99.5	80.9	89.2
JNJ	88.1	82.5	85.2
MRK	70.5	80.1	75
MSFT	94.6	56.2	70.5
NKE	74.4	93.6	82.9
PG	82	92.4	86.9
TRV	87.7	84.9	86.2
UTX	71.8	89.2	79.6
VZ	93	58.6	71.9
GOOGL	93.3	88.4	90.8
AMZN	88.7	93.6	91.1
	85.7	**84**	**83.9**
	Weighted Average		

Table (5) : predictionaccuracy results for Random Forest technique

Class	Precision	Recall	F-Measure
MMM	96.3	92.8	94.5
AXP	94.7	92	93.3
AAPL	99.6	99.6	99.6
KO	81.6	86.5	83.9
XOM	88.1	88.4	88.3
GE	100	100	100
GS	93.5	96.8	95.1
INTC	92.9	93.2	93
JNJ	92.3	96	94.1
MRK	87.8	88.8	88.3
MSFT	88.6	92.8	90.7
NKE	85.8	84.1	84.9
PG	91	92.8	91.9
TRV	91.8	89.2	90.5
UTX	87.5	86.5	87

TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 9141 - 9150

VZ	80.1	72.1	75.9
GOOGL	94.1	89.2	91.6
AMZN	89.8	94.4	92
	90.9	**90.9**	**90.8**
	Weighted Average		

The class field in the results tables (3), (4) and (5) represents the symbol of the company or sector in the NYSE stock market whose stock values have been predicted. The remaining three fields represent the percentage of prediction accuracy for each company based on the company's stock data.

The Fig (4) showing the results related to prediction accuracy ratios for 3 approaches suggested in the work, and indicating that RF is the best in its accuracy ratio according to precision, recall and F-measure.

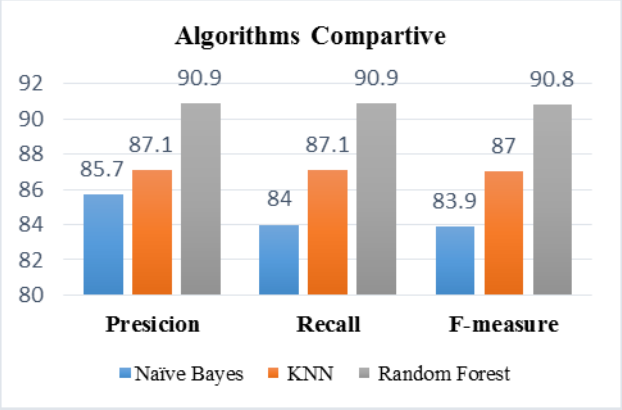

Fig 4. Accuracy of prediction for proposed techniques

Table (6) shows summary of the results of accuracy of classifiers according to evolution measures has used and errors rate for each classifier.

Table (6) :Results summary for all techniques

Technique	Performance Measures			Error Rate	
	Precision	Recall	F-Measure	MAE	RMSE
KNN	87.1%	87.1%	87%	0.0204	0.1024
Naïve Bayes	85.7%	84%	83.9%	0.0191	0.123
Random Forest	90.9%	90.9%	90.8%	0.0144	0.0882

V. Conclusion

Stock market prediction is a very difficult task because of the random nature of the company's financial share data. The ideal solution for achieving robust and accurate forecasting is employing artificial intelligence in applying machine learning techniques.Three powerful techniques have been proposed for the implementation separately to build an approach to forecasting companies and sectors for the New York Stock Exchange (NYSE).

The results of the experiments demonstrated that the highest precision, recall, and F1 measure were 90.9,90.9 and 90.8 respectively, in implement for Random Forest technique, while the rest of the other two techniques gave prediction accuracy rang 84 % to 87 % only.

We conclude, after studying the results of the three techniques that Random Forest technique is the best for stock market prediction 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 two techniques.

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 (2017): 22.
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, 2011: 186-195.

May-June 2020
ISSN: 0193-4120 Page No. 9141 - 9150

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) June.
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 (2019): 26.
8. Alkhatib, Khalid, et al. "Stock price prediction using k-nearest neighbor (kNN) algorithm." International Journal of Business, Humanities and Technology 3.3 (2013): 32-44.
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), (2013), 2319-7064.
13. Alasadi, Suad A., and Wesam S. Bhaya. "Review of data preprocessing techniques in data mining." Journal of Engineering and Applied Sciences 12.16 (2017): 4102-4107.
14. R.C. Neath, M.S. Johnson. "Discrimination and Classification". International Encyclopedia of Education (Third Edition), 2010, PP 135-141, Elsevier.
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 (2015): 258.
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 (2016): 1842-1845.
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

May-June 2020
ISSN: 0193-4120 Page No. 9141 - 9150

random forests." BMC bioinformatics 18.1 (2017): 169.

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 (2012): 322.
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. 2012: 14-16