Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 03, March-2022

Performance Evaluation of Binary and Multi-Class Dataset using Ensemble Classifiers

Anupama Jha[1], Dr. Meenu Dave[2], Dr. Supriya Madan[3]
[1]Research Scholar, Jagannath University, Jaipur, India
[2]Professor, Jagannath University, Jaipur, India
[3]Professor, VIPS, GGSIPU, New Delhi, India

***Abstract*: Now a day's digital data is dominating the entire globe. Every day it is increasing as it is generated from various domains such as social media, healthcare, education, banking etc. as well as through smart devices, IoT Devices etc., which we call these days Big Data. Due to the availability of Big Data, ensemble machine learning methods represent an attractive approach that can be used to deal with mining large and complex datasets. Ensemble models are now standard in Big Data Mining due to the fact that combining multiple classifiers together on a large dataset can often produce a much powerful classifier. The main principle behind this approach is that when weak classifiers are correctly combined, we can get better results.**

With this research paper, we have made a noble attempt to compare the performance of proposed methodology with existing research study using the various data mining classifiers. This study also compares the performance of basic data mining classifiers with the ensemble classifiers to solve the two major classes of classification problems: Binary-Class and Multi-Class in terms of accuracy. Ensemble approaches are implemented here to improve the performance of simple models and reduce overfitting of more complex models.

The experimental results show that for the Multi-class classification task, Bagging performs well in comparison of Binary-class. But for Binary Class dataset, it is found that in most of the models, basic classifiers perform better than the ensemble classifiers. Moreover, it is observed that the Bagging performs well for all types of training-testing splits of the datasets.

Keywords:- Big Data Mining, Classification, Ensemble method, Bagging, Booting

1. INTRODUCTION

The main objective of Big Data mining is to uncover hidden insight from the large volume dataset that can be useful for many organizations to make better decision [1]. In recent year, it has attracted more and more attention due to the fact that it has been successfully applied to many domains such as Data science, Big Data Analytics, Business Intelligence, WWW, Sentiment Analysis [2] etc. In data mining, classification approach is considered to be the most important data mining approach as it becoming a fascinating topic to the researchers that precisely and effectively describes data.

In this research work, we have made an extension to improve the performance by using the ensemble learning methods. In view of that, the performance of 5 basic classifiers Decision Tree (CART and CTREE), Random Forest, Support Vector SVM and k-NN of data mining approaches are compared with the novel ensemble learning approach Bagging and Boosting for the classification tasks. The analysis is implemented on two different datasets i.e. Binary class and Multi-class. All such classifiers have been modelled with different training-testing partitions to find out the best classifier in terms of accuracy.

2. RELATED WORK

It is found that there are number of techniques that are used to analyse large volume datasets are not very efficient for performing the tasks as some of them are fast but they had to compromise with the accuracy [3]. Some techniques result in good accuracy but took more execution time. In 2013, the researchers analysed 14 different classification algorithms and found no one classifiers outperformed all others in terms of the accuracy when applied to the number of datasets [4]. Researchers also highlighted that there are no classifiers available in the literature that can classify binary, multi-class and multi-label classification at the same time [5]. They proposed a novel online universal classifier based on an extreme learning machine and found that the performance was almost uniform in datasets of all classification types.

Authors Seyed Hossein Nourzad and Anu Pradhan presented 2 ensemble methods Bagging and AdaBoost for binary and multi-class classification to improve the accuracy and they were able to achieve for the binary classification, the accuracy up to 98.9% and for the multi-class classification, the accuracy is 94.6% [6]. Overall, the performance of ensemble models was found higher than that of base classifiers [7] [8].

Dewiani, Armin Lawi et al. in 2019 proposed a combined technique of Ensemble Bagging and Support Vector Machine (SVM) to improve single classification performance to detect fraud in a firm. They achieved the highest accuracy of 89.95% [9].

In 2020, Authors Isaac Kofi Nti et al. provide a wide-ranging comparison of various ensemble methods such as bagging, boosting, stacking, and blending for predicting stock-market indices. It was found that although stacking and blending achieved higher accuracy but due to their higher training and testing time, they are computationally expensive as compared to Decision Tree by boosting and bagging [10].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 03, March-2022

After reviewing the different approaches that we have discussed above, it is found that, for some kind of problems, there is a need to design a new system that can provide some changes in the existing methodologies or to design a new methodology that can deal with the different classes of the dataset as well as different training-testing partitions to increase the performance. For classification tasks, Decision tree and Random Forest algorithms are widely used because both are easy to understand and implement in various attributes. ID3, C4.5, and C5.0 were the most frequently used Decision tree.

3. PROPOSED METHODOLOGY

With this research, a proposed methodology is presented with some new Decision tree techniques such as CART, CTREE, Random Forest, SVM, k-NN including advance ensemble techniques such as Bagging and Boosting so that comparison can be made with the basic and advance classifiers too.

For any Data Mining classification tasks, Confusion matrix plays an important metrics for performance evaluation, which result the accuracy in terms of percentage. It is found that the accuracy of a model on a given test set is the percentage of test set tuples that are correctly classified by the model. The methodology for training and testing each of the above model consists of the following steps:

Steps for Proposed Methodology

Step1a: Loading of Binary and Multi-class datasets.
Step1b: Counting of total number of observations.
Step1c: Install relevant packages, and libraries.

Step 2: Selecting Input variables and the Target variable.

Step 3a: Dividing data into Training/Testing set.
Step 3b: Counting the total number of training & testing observations.

Step 4: Building the model using Training dataset.

Step 5: Testing the model using Testing dataset

Step 6a: Model Evaluation is based on building of the Confusion Matrix.
Step 6b: Compute Accuracy for both base classifiers as well as ensemble classifiers and then compare them across.

The above tasks for model building are implemented using one of the Big Data analytical tool known as System R. System R & RStudio tools are used for implementing various data mining techniques because it can perform data manipulation, analysis, machine learning tasks and data visualization [11] operations. It is widely used by the researchers, data miners, and statisticians on a high dimensional pattern extraction system. It is used to analyse the effectiveness of different machine learning algorithms with the fact that it has several in-built machine learning packages, libraries and methods that are directly associated with each task [12].

4. ENSEMBLE MODEL APPROACH USING BAGGING AND BOOSTING

Ensemble models are now standard in data mining because they perform extremely well on large and complex dataset. It is a technique where multiple models generally called weak learners are trained to solve the same problem and combined to get better results. The ensemble method can reduce classification errors effectively, and is believed to perform well compared to the use of a single classifier.
Compared to an individual classifier, where they only learn and train a set of data only, using ensemble classifiers, they can learn and train the various data generated from the original dataset and the results will build a set of hypotheses from the data trained and produce better accuracy.

Several Ensemble classification techniques have been developed such as Bagging, Boosting and Stacking. However, this study focuses on Ensemble Bagging and Boosting techniques. Both approaches are used to improve the performance of simple models and reduce overfitting of more complex models. With these approaches, a set of weak learners are combined to create a strong learner that obtains better performance than a single one.

4.1 BAGGING

Bagging is a kind of ensemble method, also called *Bootstrap Aggregating* which combines Bootstrapping and *Aggregation* to form one ensemble model. It is mostly used to reduce the variance in a model. In this approach, a number of bootstrap samples is selected from the training set, and after applying the bootstrapping process, the noisy observations are reduced and even eliminated from the training sets. Therefore, these sets will provide the classifiers with a better behaviour compared with the original set. This makes bagging technique really useful to build a better Classifier when there are noisy rows in the training set.
In this research work, we have implemented the concept of bagging using a system R package “adaBag” to achieve high classification accuracy.

4.2 BOOSTING

Boosting is another kind of ensemble method mostly used to reduce the bias in a model. It shows the ability to significantly enhance the prediction accuracy of the weak learner algorithm. This method combines a set of weak learning algorithms to build a model with better prediction outcomes. Due to its low error rate and performing excellently in noise data set, it has gained a lot of attention among the machine learning techniques [13], which we have implemented using a system R package “adaBoost” to achieve high classification accuracy.

5. EXPERIMENTAL RESULTS AND ANALYSIS

A number of comparisons have been made for the evaluation of performance analysis in view to check the scalability too in various circumstances and the results obtained are visualized in the form of bar graphs.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 03, March-2022

Two biomedical datasets Multi-class IRIS and Binary-class Breast Cancer are collected from the UCI repository (University of California at Irvine), which is a collection of databases that are used by machine learning communities for the development and testing of classification algorithms

5.1 Result comparison of Existing Vs Proposed Approach

Authors Gopala Krishna Murthy et. al, in their paper evaluated a comparative performance analysis on various breast-cancer datasets by using 14 different classifiers [14]. The results indicated that none of the classifiers outperformed all others in terms of accuracy when applied to all the data sets. Further, the authors recommended that researchers should try their dataset on a set of classifiers and then select the best one**.**

With this research, the work focuses on finding the correct classifier that works better on diverse datasets, various classes of datasets (binary and multi-class) as well as by using the different training-testing partition (%) split (Table 1).

Table 1: Accuracy (%) of different models using multiple training-testing partitions (%) using the Proposed Approach [15]

DATA MINING MODELS – ACCURACY in %							
Dataset	Training-Testing Partition (%)	Decision Tree		RF	SVM	KNN	Best Accuracy found in partition
		CART	CTREE				
IRIS (Multi Class)	60-40	16.67	16.67	16.67	16.67	16.67	(70, 30)
	70-30	**86.67**	**91.11**	**75.56**	53.33	46.67	
	80-20	83.33	90	73.33	**80**	**80**	
Breast Cancer (Binary Class)	60-40	95.71	98.21	97.86	97.5	98.21	(80, 20)
	70-30	94.76	98.1	97.14	98.1	98.1	
	80-20	**97.86**	**98.57**	**99.29**	**98.57**	**99.29**	

The results are compared with one of the research studies which includes Breast Cancer Microarray Gene Expression Dataset (Table 2). According to the authors, the results indicated that none of the classifiers outperformed all others in terms of accuracy when applied to all the data sets that they have used. Mostly they found the accuracy between 70% to 80%. Further, the authors recommended that researchers should try their dataset using different algorithms/classifiers and then select the best one**.**

Table 2: Comparison of Proposed Vs Existing research study for Binary Class Breast Cancer dataset

S. No.	1	2
Study/ Reference Paper	Gopala Krishna Murthy, Nagaraju Orsu, Bharath Kumar Pottumuthu and Suresh B. Mudunuri, "Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification".	Our Proposed Approach
Classification Approach Used	Supervised	Supervised
Classification Techniques Used	Applied to 14 classifiers such as J48, Simple CART, Random Forest, AD Tree etc.	Decision Tree (CART and CTREE), Random Forest, SVM and KNN
Big Data Analytics Tools	Weka	RStudio, System R
Dataset used	**Breast Cancer (Binary Class)**	

Our approach compared the result on their dataset (with 286 instances and 60-40 training testing partition %) with 2 basic data mining classifiers CART and Random Forest.

Table 3: Accuracy (%) Comparison of Proposed Vs Existing research study for various Classifiers

Accuracy (%) Comparison under different Training Testing partitions (%) for Breast Cancer Dataset			
	Partition /Classifiers	**CART**	**Random Forest**
Existing Study: Reference 1 Approach	**60-40**	70.3	97.2
Our Proposed Approach applied to 3 partitions	**60-40**	95.71	97.86
	70-30	94.76	97.14
	80-20	97.86	99.29

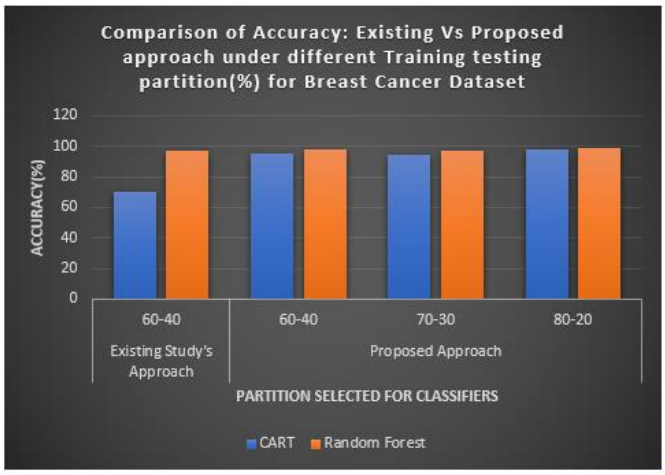

Figure 1: Accuracy Comparison: Existing Vs Proposed approach under different Training testing partition (%) for Breast Cancer Dataset

It is found that the accuracy extremely improves for Simple CART (Decision tree) for all the training testing partitions as compared to (60-40) of the existing research study (Table 3, Figure 1). And for the Random Forest classifier, the performance is approximately the same or increasing. Overall, for both classifiers, the (80-20) partition achieves a higher accuracy rate.

5.2 Accuracy comparison between Basic Classifiers Vs. Ensemble Classifiers for Binary as well as Multi-class Classification

A Comparison of Basic Data Mining Models Vs. Ensemble Models for Binary class as well as Multi-class classification accuracy under various training-testing partitions (%), can be seen in Figure 2 and Figure 3 as shown below.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 03, March-2022

Table 4: Accuracy Comparison of Basic Classifier Vs. Ensemble Classifier

		DATA MINING MODELS – ACCURACY in %						
		Basic Models					Ensemble Models	
Dataset	Training & Testing Partition	Decision Tree(DT)		Random Forest	Support Vector Machine	k-Nearest Neighbour	Bagging	Boosting
	(%)	CART	CTREE	(RF)	(SVM)	(KNN)	(AdaBag)	(AdaBoost)
IRIS (Multi-Class)	(60, 40)	16.67	16.67	16.67	16.67	16.67	88.34	88.34
	(70, 30)	**86.67**	**91.11**	**75.56**	53.33	46.67	97.78	91.12
	(80, 20)	83.33	90	73.33	**80**	**80**	96.67	93.34
Breast-Cancer (Binary Class)	(60, 40)	95.71	98.21	97.86	97.5	98.21	98.57	96.79
	(70, 30)	94.76	98.1	97.14	98.1	98.1	98.09	96.67
	(80, 20)	**97.86**	**98.57**	**99.29**	**99.29**	**98.57**	97.12	97.86

The accuracy of bagging classifier is implemented with R package/library adaBag whereas, boosting is implemented with adaBoost package/library (Table 4).

After analysing Binary Class dataset, it is found that in most of the models, highest accuracy achieved when the partition was (80-20) % and probably basic classification models perform better than the ensemble models (Figure 2). The best accuracy is found in RF (basic model) and AdaBoost (ensemble model).

Similarly, after analysing Multi-Class dataset, it is found that in most of the models, highest accuracy achieved when the partition was (70-30) % and ensemble models perform better than the basic classification models (Figure 3). The best accuracy is found in CTREE (basic model) and AdaBag (ensemble model).

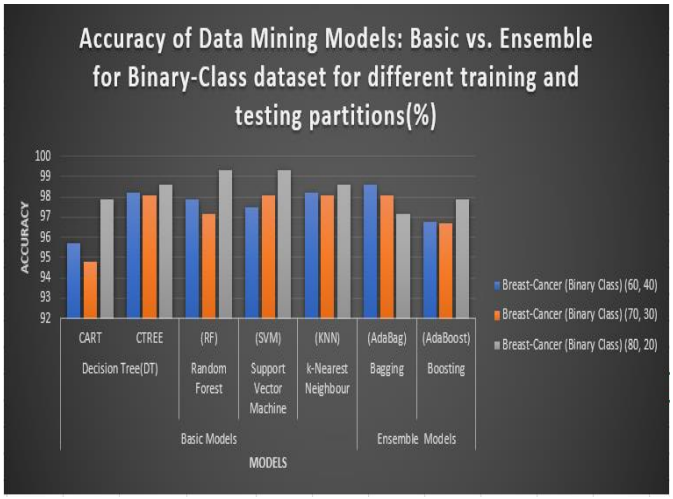

Figure 2: Accuracy comparison of Basic Vs. Ensemble Data Mining Models for Binary Class

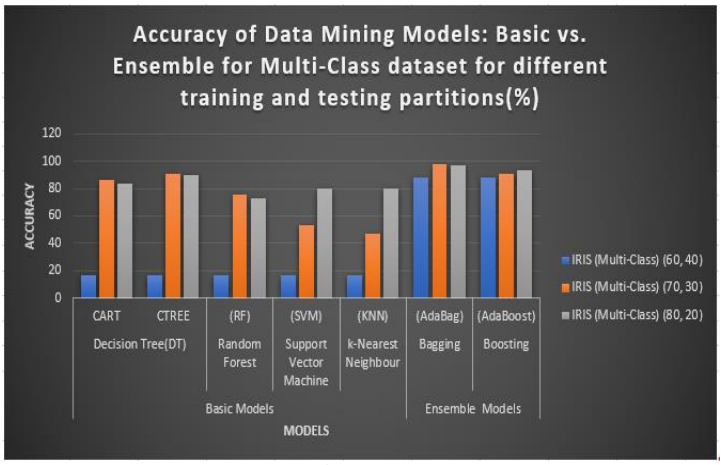

Figure 3: Accuracy comparison of Basic Vs. Ensemble Data Mining Models for Multi-class

5.3 Accuracy comparison between Decision tree and Bagging

After comparing the performances of Decision Tree and Bagging Ensemble classifiers, it is found that the accuracy of Ensemble method increases (Figure 4) for both the Multi-class and Binary class datasets for all the partitions, whereas it was found that the result compared for the table, the result was uniform for both of the classifiers., which motivate us to do such kind of scalability analysis.

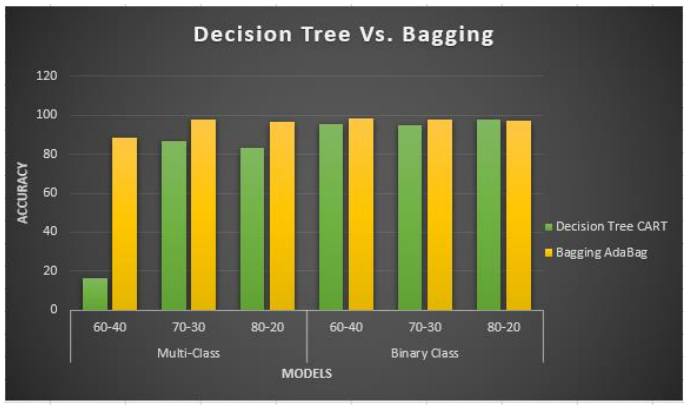

Figure 4: Decision tree Vs. Bagging

5.4 Accuracy comparison between Random Forest and Boosting

But when compared with Random Forest and Boosting Ensemble classifiers with the same dataset, it is found that Boosting in Multi-Class dataset achieve higher accuracy than Random forest and for binary class datasets, the accuracy is almost same (Figure 5).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 03, March-2022

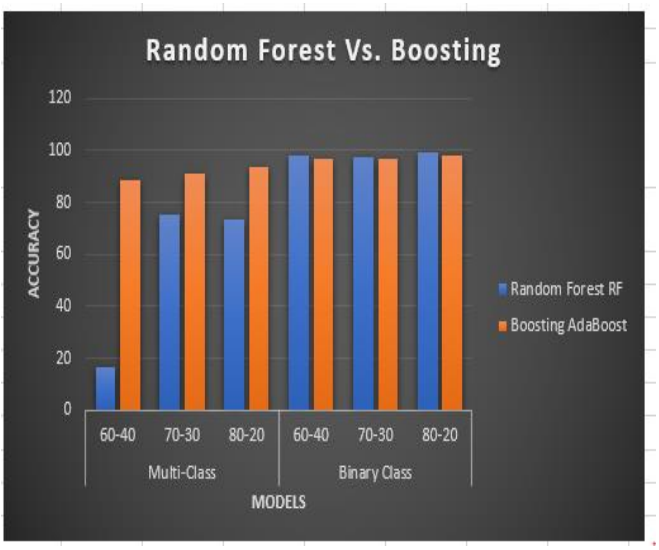

Figure 5: Random Forest Vs. Boosting

5. CONCLUSION

The above implementation focused on finding the best-suited algorithm for both binary class and multi-class classification tasks so that it can be further applied to any kind of Big Data Analytics. It is observed from the above analysis related to Classification task that, for the Binary class classification, Random Forest (RF) algorithm performs better due to the highest accuracy 99.29% applied to (80, 20) % partition on the train-test dataset. However, for the Multi-class classification task, Bagging performs well with both (70, 30) % and (80, 20) % partition with the highest accuracy 97.78% and 96.67% respectively. Moreover, it is observed that the Bagging performs well for all types of training-testing splits of the datasets (Table 5).

With this research work, we conclude that for any kind of model's performance whether it is basic or advance, we should not completely depend upon a particular algorithm and a fixed training-testing partitions as normally the model can be build using 70-30 partition %. We can implement with various partitions too. In future, it is recommended that the researches can try their algorithms with different types of datasets and with different training-testing partition %.

ACKNOWLEDGEMENT

I would like to thank Vivekananda Institute of Professional Studies (VIPS), GGSIP University, Delhi for providing such a positive and healthy work environment to complete this research paper.

Table 5: Final Accuracy Comparison in multiple circumstances

Training – testing Partition (%)	Classification Techniques Used	Accuracy (%)	Highest Accuracy (%)
(80, 20) As selected for the Binary-Class dataset	CART	97.86	**99.29 when used with Random Forest (RF) classifier**
	CTREE	98.57	
	RF	**99.29**	
	SVM	98.57	
	K-NN	98.57	
	Bagging	97.12	
	Boosting	97.86	
(70, 30) As selected for the Multi-Class dataset	CART	86.67	**97.78 when used with Bagging classifier**
	CTREE	91.11	
	RF	75.56	
	SVM	53.33	
	K-NN	46.67	
	Bagging	**97.78**	
	Boosting	91.12	

REFERENCES

[1] A. Jha, M. Dave and S. Madan, "A Review on the Study and Analysis of Big Data using Data Mining Techniques", International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol6, Issue 3, Jan 2016.

[2] Anupama Jha, Meenu Dave and Supriya Madan, "PERFORMANCE ANALYSIS OF TWEETS USING HADOOP PIG AND HIVE: A COMPREHENSIVE STUDY", Journal of Advanced Research in Dynamical & Control Systems, Scopus Indexed, ISSN 1943-023X, Vol. 10, 11-Special Issue, July 2018.

[3] Hlaudi Daniel Masethe and Mosima Anna Masethe: "Prediction of Heart Disease using Classification Algorithms", Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II WCECS, San Francisco, USA, 22-24 October, 2014.

[4] Gopala Krishna Murthy, Nagaraju Orsu, Bharath Kumar Pottumuthu and Suresh B. Mudunuri: "Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification", International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No.5, 2013.

[5] Meng Joo Er, Rajasekar Venkatesan and Ning Wang; "An Online Universal Classifier for Binary, Multiclass and Multi-label Classification", https://www.researchgate.net/ publication/307636344, pp 1-6, 2016.

[6] Seyed Hossein Nourzad, Anu Pradhan: "Binary and Multi-Class Classification of Fused LIDAR-Imagery Data Using an Ensemble Method", Construction Research Congress 2012 © ASCE 2012.

[7] Nikita Joshi, Shweta Srivastava; "Improving Classification Accuracy Using Ensemble Learning Technique (Using Different Decision Trees)", IJCSMC, Vol. 3, Issue. 5, pp.727 – 732, 2014.

[8] Mohamad Amin Pourhoseingholi, Sedigheh Kheirian and Mohammad Reza Zali; "Comparison of Basic and Ensemble Data Mining Methods in Predicting 5-Year Survival of Colorectal Cancer Patients"; doi: 10.5455/aim.2017.25.254-258, ACTA INFORM MED. 2017; 25(4): pp. 254-258, 2017.

[9] Dewiani , Armin Lawi, Muhammad Idris Rifai Sarro and Firman Aziz; "Classification of Firm External Audit Using Ensemble Support Vector Machine Method", ICOST, Makassar, Indonesia Copyright, EAI DOI 10.4108/eai.2-5-2019.2284605, May 02-03, 2019.

[10] Isaac Kofi Nti, Adebayo Felix Adekoya and Benjamin Asubam Weyori, "A comprehensive evaluation of ensemble learning for stock-market prediction", Journal of Big Data, 7:20, https://doi.org/10.1186/s40537-020- 00299-5, Springer Open access, 2020.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 03, March-2022

[11] Dalgaard P.: "Introductory Statistics with R". Springer, NewYork, 2002
[12] J H Maindonald, "R for Data Analysis and Graphics: Introduction, Code and Commentary", Centre for Mathematics and Its Applications, Australian National University, ©J. H. Maindonald 2000, 2004, 2008.
[13] Ma Y. & Ding X. (2003). Robust Real-Time Face Detection Based on Cost Sensitive Adaboost Method. In Proc. The International Conference on Multimedia and Expo, 465 - 473.
[14] Gopala Krishna Murthy, Nagaraju Orsu, Bharath Kumar Pottumuthu and Suresh B. Mudunuri: "Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification", International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No.5, 2013.
[15] Anupama Jha, Meenu Dave and Supriya Madan: "Comparison of Binary Class and Multi-Class Classifier Using Different Data Mining Classification Techniques", International Conference on Advancements in Computing & Management, Hosting by SSRN. April' 2019.