Ensembling Coalesce of Logistic Regression Classifier for Heart Disease Prediction using Machine Learning

Shakila Basheer, Rincy Merlin Mathew, M. Shyamala Devi

Abstract: ***In today's modern world, the world population is affected with some kind of heart diseases. With the vast knowledge and advancement in applications, the analysis and the identification of the heart disease still remain as a challenging issue. Due to the lack of awareness in the availability of patient symptoms, the prediction of heart disease is a questionable task. The World Health Organization has released that 33% of population were died due to the attack of heart diseases. With this background, we have used Heart Disease Prediction dataset extracted from UCI Machine Learning Repository for analyzing and the prediction of heart disease by integrating the ensembling methods. The prediction of heart disease classes are achieved in four ways. Firstly, The important features are extracted for the various ensembling methods like Extra Trees Regressor, Ada boost regressor, Gradient booster regress, Random forest regressor and Ada boost classifier. Secondly, the highly importance features of each of the ensembling methods is filtered from the dataset and it is fitted to logistic regression classifier to analyze the performance. Thirdly, the same extracted important features of each of the ensembling methods are subjected to feature scaling and then fitted with logistic regression to analyze the performance. Fourth, the Performance analysis is done with the performance metric such as Mean Squared error (MSE), Mean Absolute error (MAE), R2 Score, Explained Variance Score (EVS) and Mean Squared Log Error (MSLE). The implementation is done using python language under Spyder platform with Anaconda Navigator. Experimental results shows that before applying feature scaling, the feature importance extracted from the Ada boost classifier is found to be effective with the MSE of 0.04, MAE of 0.07, R2 Score of 92%, EVS of 0.86 and MSLE of 0.16 as compared to other ensembling methods. Experimental results shows that after applying feature scaling, the feature importance extracted from the Ada boost classifier is found to be effective with the MSE of 0.09, MAE of 0.13, R2 Score of 91%, EVS of 0.93 and MSLE of 0.18 as compared to other ensembling methods.***

Index Terms*: *Machine Learning, Classification, MAE, MSE, MSLE, EVS and R2 Score.

I. INTRODUCTION

For enabling the patients to treat in advance, the prediction of particular disease is necessary. Generally the heart disease is affected for the elderly people as per the survey. The heart disease prediction may help the immobile chronically ill patients to undergo the medical treatment in advance so as to save their life. As the symptoms needed to identify the heart disease is minimum, we are in need of predicting software to identify the existence of the heart disease. The doctors are unable to predict the disease by the lack of future symptoms of the patients. This requires the use of machine learning techniques for forecasting the future symptoms of the patients.

The paper is organized in such a way that Section 2 deals with the related works. Proposed work is discussed in Section 3 followed by the implementation and Performance Analysis in Section 4. The paper is concluded with Section 5.

II. RELATED WORK

A. Literature Review

The health care industry is now focusing on using the appropriate technology in the prediction of health diseases. The diseases like Loco motor disorders and Heart diseases are using the technology for the detection of existence of the disease. This helps the doctors in diagnosing to forecast the disease in advance [1]. Heart is the significant limb which is equal to brain in the survival of human body. Heart smacks the blood and move to all the organs of the body for functioning. The prediction of heart disease still remains as a difficult task in the medical field. The patient's medical data is stored in the database and they are subjected to data mining techniques like artificial neural Network, Decision tree, Fuzzy Logic, K-Nearest Neighbor, Naive Bayes and Support Vector Machine [2].

When the person is suffering from heart disease, the heart will not be able to drive the desirable blood to other parts of the body. The non invasive based machine learning algorithms are used to diagnose the heart disease [3]. The different constraints of the heart and the body can be used to forecast the heart disease [4].

The analysis was done for the male patients for the heart disease prediction using data mining techniques. They have used the three data mining techniques like Naive Bayes, Artificial neural network, and J48 decision tree [5].

Revised Manuscript Received on October 10, 2019

Shakila Basheer, Assistant Professor, Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdul Rahman university, Riyadh-11671, Saudi Arabia

Rincy Merlin Mathew, Lecturer, Department of Computer Science, College of Science and Arts, Khamis Mushayt, King Khalid university, Abha, Asir, Saudi Arabia.

M. Shyamala Devi, Associate Professor, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

A overall technological review on various feature selection, feature extraction methods, classification methods and the performances parameters are examined for predicting the wine quality [6]-[22].

III. PROPOSED WORK

In our proposed work, machine learning algorithms are used to predict the disease of Heart disease data set. Our contribution in this paper is folded in two ways.

(i) Firstly, the important features are extracted for the various ensembling methods like Extra Trees Regressor, Ada Boost Regressor, Gradient Booster Regressor, Random Forest Regressor and Ada Boost Classifier.

(ii) Secondly, the highly importance features of each of the ensembling methods is filtered from the dataset and it s fitted to logistic regression classifier to analyze the performance.

(iii) Thirdly, the same extracted important features of each of the ensembling methods are subjected to feature scaling and then fitted with logistic regression to analyze the performance.

(iv) Fourth, the Performance analysis is done with the performance metric such as Mean Squared error (MSE), Mean Absolute error (MAE), R2 Score, Explained Variance Score (EVS) and Mean Squared Log Error (MSLE).

A. System Architecture

The overall design of this paper work is shown in Fig. 1

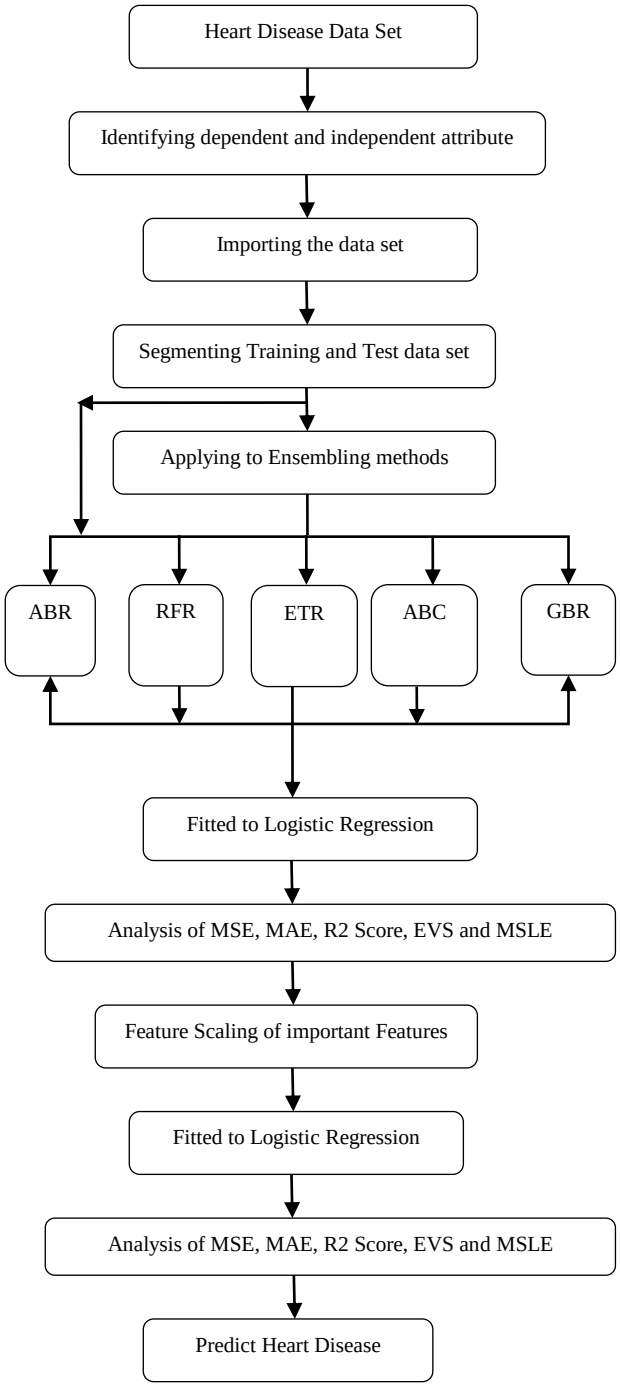


Fig. 1 System Architecture (ABR - Ada Boost Regressor, RFR - Random Forest Regressor, ETR - Extra Trees Regressor, GBR - Gradient Booster Regressor and ABC-Ada Boost Classifier).

IV. IMPLEMENTATION AND PERFORMANCE ANALYSIS

A. Heart Disease Prediction for Feature Extraction

The Heart Disease dataset extracted from UCL ML Repository is used for implementation with 13 independent attribute and 1 diagnosis dependent attribute. The dataset consists of 779 individual's data. The attribute are shown below.

1. Age
2. Sex
3. Chest-pain type
4. Resting Blood Pressure
5. Serum Cholestrol
6. Fasting Blood Sugar
7. Resting ECG
8. Max heart rate
9. Angina
10. ST depression
11. Peak exercise ST segment
12. Number of major vessels
13. Thal
14. Diagnosis of heart disease - Dependent Attribute

The portrayal of the dependent attribute is shown in Fig. 2.

Chest Pain Format	Description
1	Typical Angina
2	Atypical Angina
3	Non - Anginal Pain
4	Asymptotic

Fig. 2. Chest Pain Type Distribution

The expansion of the dataset attributes shown in Fig. 3.

S.No	Attribute	Description
1.	Age	Age of the individual.
2.	Sex	Gender of the individual with the following format: 1 = male 0 = female.
3.	Chest-pain type	Type of chest-pain as in Table. 2.
4.	Resting Blood Pressure	Resting blood pressure value in mmHg (unit)
5.	Serum Cholestrol	serum cholestrol in mg/dl (unit)
6.	Fasting Blood Sugar	Fasting blood sugar value of an individual with 120mg/dl. If fasting blood sugar > 120mg/dl then : 1 (true) else : 0 (false)
7.	Resting ECG	0 = normal 1 = having ST-T wave abnormality 2 = left ventricular hypertrophy
8.	Max heart rate	Max Heart Rate
9.	Angina	1 = yes 0 = no
10.	ST depression	Value (integer or float).
11.	Peak exercise ST segment	1 = Upsloping 2 = Flat 3 = Downsloping
12.	Number of major vessels	Values (0-3) colored by flourosopy
13.	Thal	displays the Thalassemia : 3 = Normal 6 = Fixed Defect 7 = Reversable Defect
14.	Diagnosis of heart disease	0 = Absence 1,2,3,4 = Present.

Fig. 3. Heart Disease Dataset Schema

B. Performance Analysis

The association among the parameters of the heart disease dataset is shown as a histogram relationship notation in the fig 4. The variance between the each of the attributes of the heart disease dataset is depicted as a correlation matrix and is shown in fig 5. The distribution of the target variable of the heart disease dataset is shown in fig 4. The important features are extracted for the various ensembling methods like Extra Trees Regressor, Ada Boost Regressor, Gradient Booster Regress, Random Forest Regressor and Ada Boost Classifier.

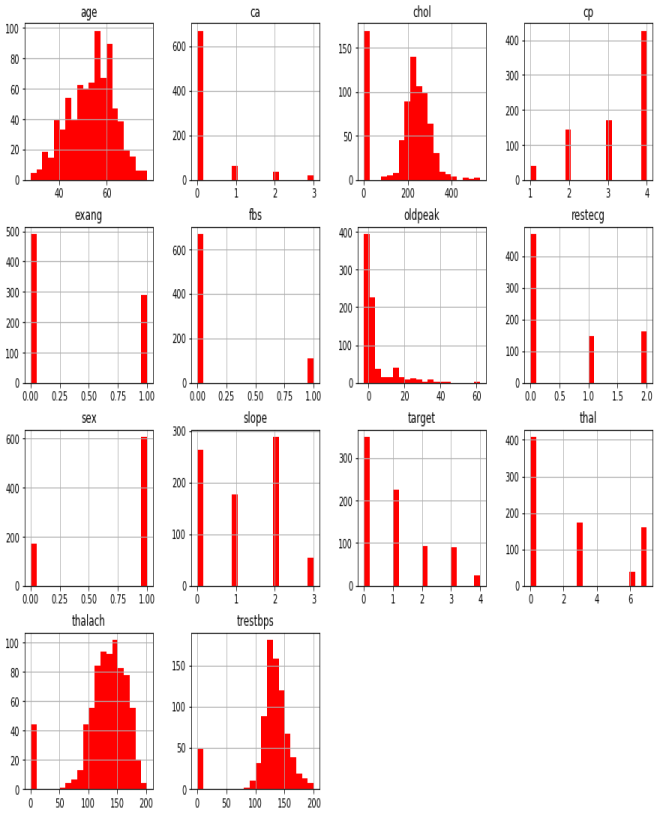


Fig. 4. Histogram Relation of the heart disease data set

Fig. 5. Correlation Matrix of Heart Disease Data Set

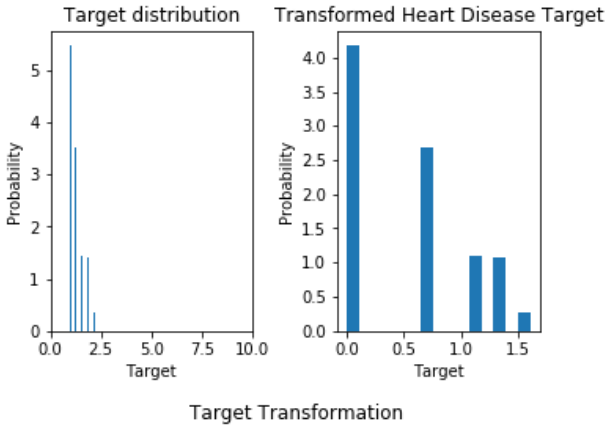


Fig. 6. Target Distribution of Heart Disease Data Set

The highly important attributes are extorted from the Ada boost regressor and the relationship is shown as a bar graph in Fig. 6. and the values of each of the parameters is shown in fig. 7.

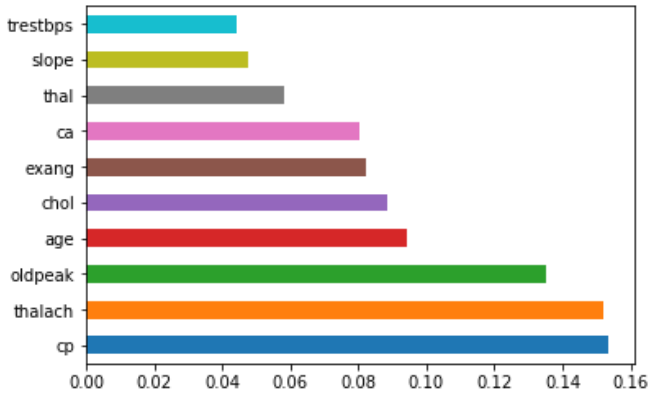


Fig. 7 Feature Importance Distribution of Ada Boost Regressor for Heart Disease Data Set

The obtained feature importance of the Ada boost regressor and Random forest regressor is shown in fig 8.

AdaBoost_featureimportances - Seri

Index	0
age	0.0944392
sex	0.0270572
cp	0.15352
trestbps	0.0442271
chol	0.0883613
fbs	0.0117385
restecg	0.0251511
thalach	0.15226
exang	0.0824403
oldpeak	0.135159
slope	0.0474269
ca	0.0801188
thal	0.0581013

RandomForestRegressor_feature_imp

Index	0
age	0.138552
sex	0.0156574
cp	0.213343
trestbps	0.0738932
chol	0.101258
fbs	0.014783
restecg	0.032417
thalach	0.142988
exang	0.0364828
oldpeak	0.0713496
slope	0.0553604
ca	0.0446399
thal	0.0592757

Fig. 8 Feature Importance of Ada Boost Regressor and Random Forest Regressor of Heart Disease Data Set

The highly important attributes are extorted from the Random Forest regressor and the relationship is depicted as a bar graph and is shown in Fig. 9.

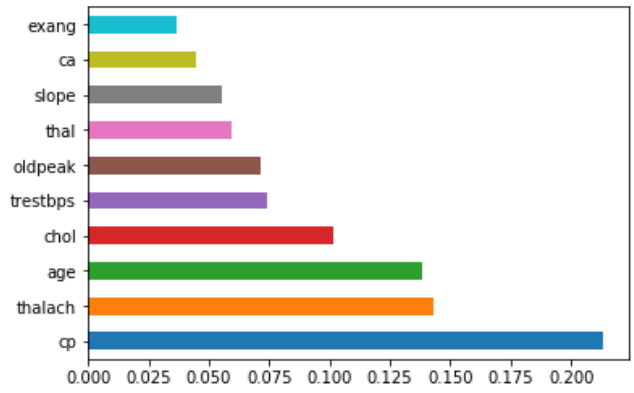


Fig. 9. Feature Importance Distribution of Random Forest Regressor for Heart Disease Data Set

The highly important attributes are extorted from the Extra trees regressor and the relationship is depicted as a bar graph and is shown in Fig. 10.

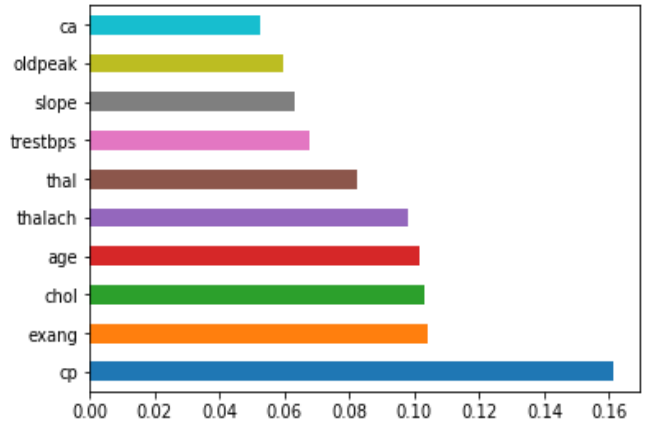


Fig. 10. Feature Importance Distribution of Extra Trees Regressor for Heart Disease Data Set

The highly important attributes are extorted from the Gradient Boost regressor and the relationship is depicted as a bar graph and is shown in Fig. 11.

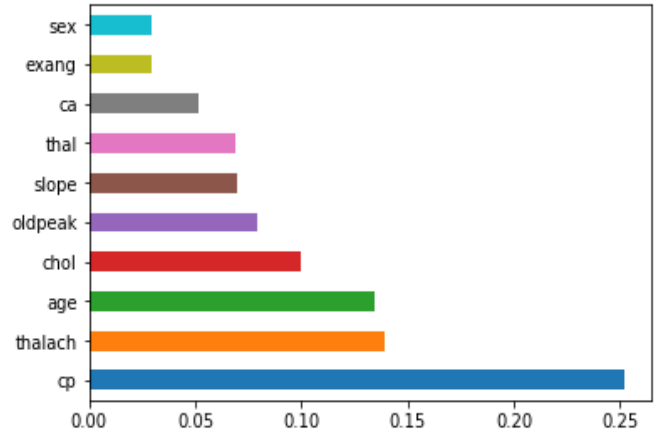
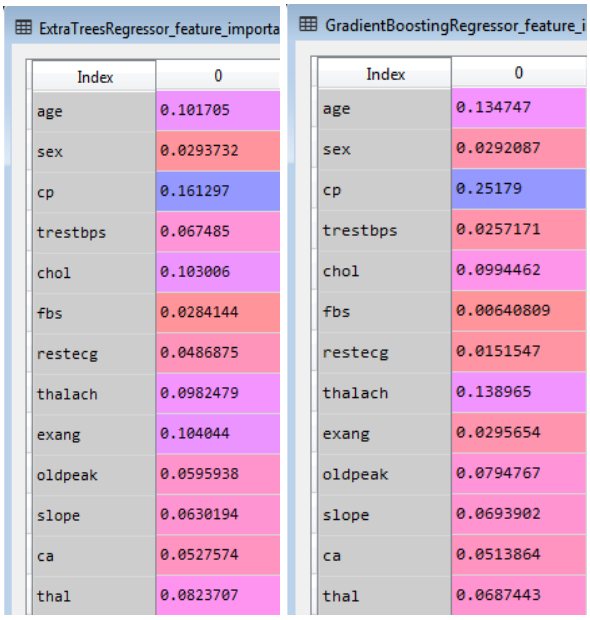


Fig. 11. Feature Importance Distribution of Gradient Boost Regressor for Heart Disease Data Set

The obtained feature importance of the Extra trees regressor and Gradient Boost regressor is shown in fig 12.



ExtraTreesRegressor_feature_importa

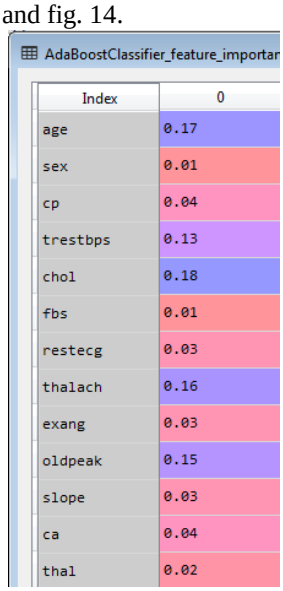
Index	0
age	0.101705
sex	0.0293732
cp	0.161297
trestbps	0.067485
chol	0.103006
fbs	0.0284144
restecg	0.0486875
thalach	0.0982479
exang	0.104044
oldpeak	0.0595938
slope	0.0630194
ca	0.0527574
thal	0.0823707

GradientBoostingRegressor_feature_i

Index	0
age	0.134747
sex	0.0292087
cp	0.25179
trestbps	0.0257171
chol	0.0994462
fbs	0.00640809
restecg	0.0151547
thalach	0.138965
exang	0.0295654
oldpeak	0.0794767
slope	0.0693902
ca	0.0513864
thal	0.0687443

Fig. 12. Feature Importance of Extra Trees Regressor and Gradient Boost Regressor of Heart Disease Data Set

The obtained feature importance of the Ada Boost Classifier is shown in fig 13 and fig. 14.



AdaBoostClassifier_feature_importar

Index	0
age	0.17
sex	0.01
cp	0.04
trestbps	0.13
chol	0.18
fbs	0.01
restecg	0.03
thalach	0.16
exang	0.03
oldpeak	0.15
slope	0.03
ca	0.04
thal	0.02

Fig. 13. Feature Importance of Ada Boost classifer of Heart Disease Data Set

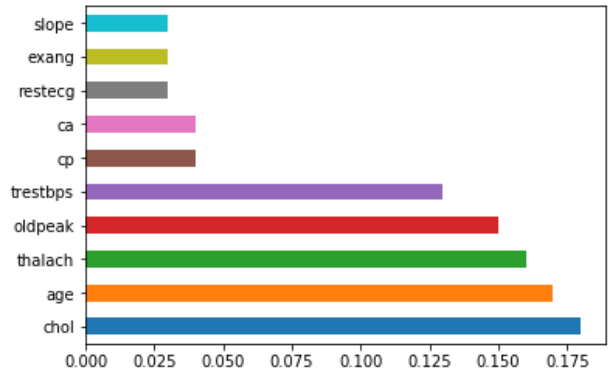


Fig. 14. Feature Importance Distribution of Ada Boost Classifer for Heart Disease Data Set

The extracted important features of each of the ensembling methods are subjected to feature scaling and then fitted with logistic regression to analyze the performance of metrics like Mean Squared error, Mean Absolute error, R2 Score, Explained Variance Score and Mean Squared Log Error and is shown in the Table 1 -Table 2.

Table. 1 Analysis of R2, MAE and MSE metrics for various Ensembling methods before Feature Scaling

Ensembling Methods	Fitting to Logistic Regression Before Feature Scaling		
	MSE	MAE	R2 Score
AdaBoost Regressor	0.19	0.13	0.79
Random Forest Regressor	0.17	0.12	0.81
Extra Trees Regressor	0.06	0.09	0.84
Gradient Boosting Regressor	0.08	0.11	0.78
Ada Boost Classifier	0.04	0.07	0.92

Table. 2 Analysis of EVS and MSLE methods for various Ensembling methods before Feature Scaling

Ensembling Methods	Fitting to Logistic Regression Before Feature Scaling	
	EVS	MSLE
AdaBoost Regressor	0.84	0.32
Random Forest Regressor	0.79	0.27
Extra Trees Regressor	0.85	0.29
Gradient Boosting Regressor	0.78	0.24
Ada Boost Classifier	0.86	0.16

The same extracted important features of each of the ensembling methods are subjected to feature scaling and then fitted with logistic regression to analyze the performance of metrics like Mean Squared error, Mean Absolute error, R2 Score, Explained Variance Score and Mean Squared Log Error and is shown in the Table 3 -Table 4.

Table. 3 Analysis of R2, MAE and MSE metrics for various Ensembling methods after Feature Scaling

Ensembling Methods	Fitting to Logistic Regression after Feature Scaling		
	MSE	MAE	R2 Score
AdaBoost Regressor	0.26	0.26	0.76
Random Forest Regressor	0.24	0.24	0.84
Extra Trees Regressor	0.14	0.18	0.79
Gradient Boosting Regressor	0.19	0.20	0.81
Ada Boost Classifier	0.09	0.13	0.91

Table.4 Analysis of EVS and MSLE metrics for various Ensembling methods after Feature Scaling

Ensembling Methods	Fitting to Logistic Regression after Feature Scaling	
	EVS	MSLE
AdaBoost Regressor	0.86	0.27
Random Forest Regressor	0.82	0.36
Extra Trees Regressor	0.87	0.47
Gradient Boosting Regressor	0.81	0.49
Ada Boost Classifier	0.93	0.18

The performance metrics analysis is depicted as a pictorial representation and is shown in the fig. 15 – fig. 19.

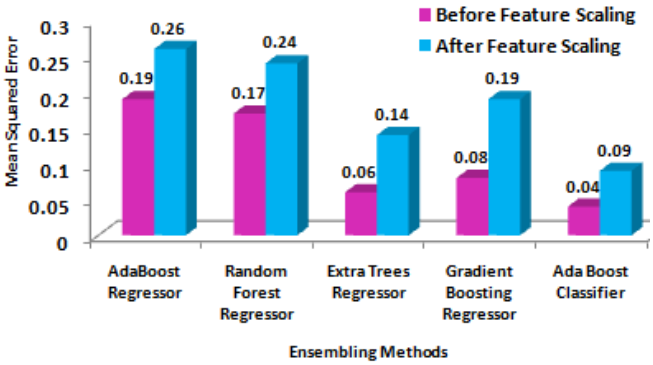


Fig 15. MSE VS Ensembling Methods

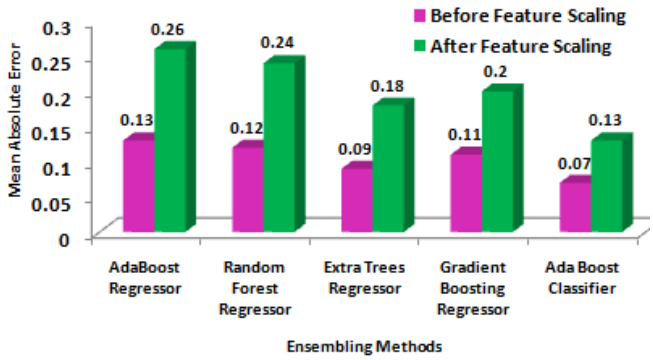


Fig 16. MAE VS Ensembling Methods

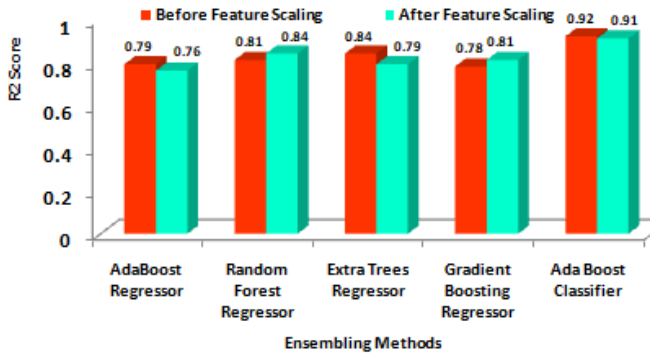


Fig 17. R2 Score VS Ensembling Methods

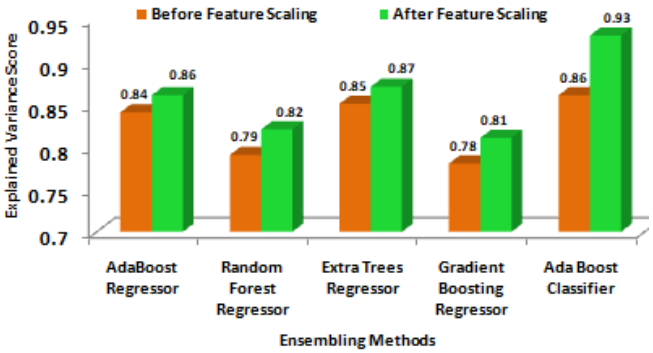


Fig 18. EV Score VS Ensembling Methods

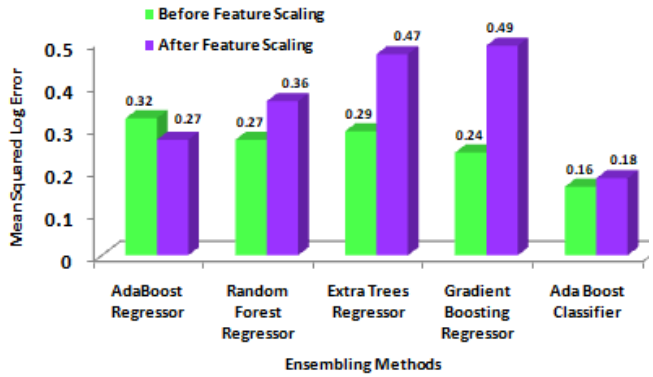


Fig 19. MSLE VS Ensembling Methods

V. CONCLUSION

This paper analyzes the method to predict the heart disease of the heart dataset through the dimensionality reduction thereby improving the accuracy prediction. An attempt is made to implement the ensembling methods for the prediction of the heart disease. The performance of all the ensembling methods is compared before and after applying feature Scaling. Experimental results shows that before applying feature scaling, the feature importance extracted from the Ada boost classifier is found to be effective with the MSE of 0.04, MAE of 0.07, R2 Score of 92%, EVS of 0.86 and MSLE of 0.16 as compared to other ensembling methods. Experimental results shows that after applying feature scaling, the feature importance extracted from the Ada Boost Classifier is found to be effective with the MSE of 0.09, MAE of 0.13, R2 Score of 91%, EVS of 0.93 and MSLE of 0.18 as compared to other ensembling methods.

ACKNOWLEDGEMENT

Thanks to the Deanship of scientific research princess Nourah Bint Abdul Rahman University for supporting. This research was funded by the Deanship of Scientific Research at Princess Nourah Bint Abdul Rahman University through the fast track Research Funding Program.

REFERENCES

1. Samir B Patel," Heart disease prediction using Machine Learning and Data Mining", International journal of computer sciences and Engineering, vol 7,2015,pp.129-137
2. V. V. Ramalingam, "Prediction of Heart Diseases Using Machine Learning", International Journal of Engineering and Technology, vol. 7, no. 2, 2018.
3. Amin Ul Haq ,"A Hybrid Intelligence System Frame Work for Prediction of Heart Disease Using ML", Mobile Information Systems ,article id 3860146,21 pages.
4. Shadman Nashif, "Heart Disease Detection by Using Machine Learning Algorithm and a Real time Cardiovascular Health Monitoring System ", World journal of Engineering and Technology,vol.06 no.04,2018,article id 88650.
5. Poornima Singh ," Effective heart disease prediction system using data mining techniques", International journal of Nano medicine, Issued on 2018, pp 121-124.
6. R. Suguna, M. Shyamala Devi, and Rincy Merlin Mathew, " Customer Churn Predictive Analysis by Component Minimization using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.8, June 2019, pp. 2329-2333.
7. Suguna Ramadass, and Shyamala Devi Munisamy, Praveen Kumar P, Naresh P, "Prediction of Customer Attrition using Feature Extraction Techniques and its Performance Assessment through dissimilar Classifiers", Springer's book series "Learning and Analytics in Intelligent Systems, Springer, , LAIS vol. 3, pp. 613-620, 2019.
8. R.Suguna, M. Shyamala Devi, Rupali Amit Bagate, and Aparna Shashikant Joshi, "Assessment of Feature Selection for Student Academic Performance through Machine Learning Classification", Journal of Statistics and Management Systems, Taylor Francis, vol. 22, no. 4, 25 June 2019, pp. 729-739.
9. R. Suguna, M. Shyamala Devi, Rincy Merlin Mathew, "Customer Segment Prognostic System by Machine Learning using Principal Component and Linear Discriminant Analysis", International Journal of Recent Technology and Engineering, Volume-8 Issue-2, 30 July 2019. pp. 6198-6203.
10. M. Shyamala Devi, Rincy Merlin Mathew, and R. Suguna,"Attribute Heaving Extraction and Performance Analysis for the Prophesy of Roof Fall Rate using Principal Component Analysis", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.8, June 2019, pp. 2319-2323.
11. Shyamala Devi Munisamy, and Suguna Ramadass Aparna Joshi, "Cultivar Prediction of Target Consumer Class using Feature Selection with Machine Learning Classification", Springer's book series "Learning and Analytics in Intelligent Systems, Springer, LAIS vol. 3, pp. 604-612, 2019.
12. M. Shyamala Devi, Rincy Merlin Mathew, R. Suguna, "Feature Snatching and Performance Analysis for Connoting the Admittance Likelihood of student using Principal Component Analysis", International Journal of Recent Technology and Engineering, Volume-8 Issue-2, 30 July 2019.pp. 4800-4807.
13. M. Shyamala Devi, Shefali Dewangan, Satwat Kumar Ambashta, Anjali Jaiswal, Sairam Kondapalli, "Recognition of forest Fire Spruce Type Tagging using Machine Learning Classification", International Journal of Recent Technology and Engineering, Volume-8 Issue-3, 30 September 2019.
14. M. Shyamala Devi, Usha Vudatha, Sukriti Mukherjee, Bhavya Reddy Donthiri, S B Adhiyan, Nallareddy Jishnu, " Linear Attribute Projection and Performance Assessment for Signifying the Absenteeism at Work using Machine Learning", International Journal of Recent Technology and Engineering, Volume-8 Issue-3, 30 September 2019.
15. M. Shyamala Devi, Mothe Sunil Goud, G. Sai Teja, MallyPally Sai Bharath, "Heart Disease Prediction and Performance Assessment through Attribute Element Diminution using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.11, 30 September 2019
16. M. Shyamala Devi, Rincy Merlin Mathew, R. Suguna, " Regressor Fitting of Feature Importance for Customer Segment Prediction with Ensembling Schemes using Machine Learning", International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, pp. 952 – 956, 30 August 2019
17. R. Suguna, M. Shyamala Devi, Rincy Merlin Mathew, "Integrating Ensembling Schemes with Classification for Customer Group Prediction using Machine Learning", International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, pp. 957 – 961, 30 August 2019.
18. Rincy Merlin Mathew, R. Suguna, M. Shyamala Devi, "Composite Model Fabrication of Classification with Transformed Target Regressor for Customer Segmentation using Machine Learning", International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, pp. 962 – 966, 30 August 2019.
19. M. Shyamala Devi, Shefali Dewangan, Satwat Kumar Ambashta, Anjali Jaiswal, Nariboyena Vijaya Sai Ram, "Backward Eliminated Formulation of Fire Area Coverage using Machine Learning Regression", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, 10 October 2019. (Accepted for Publication)
20. M. Shyamala Devi, Ankita Shil, Prakhar Katyayan, Tanmay Surana, "Constituent Depletion and Divination of Hypothyroid Prevalance using Machine Learning Classification", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, 10 October 2019. (Accepted for Publication)
21. M. Shyamala Devi, Shakila Basheer, Rincy Merlin Mathew, "Exploration of Multiple Linear Regression with Ensembling Schemes for Roof Fall Assessment using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, 10 October 2019. (Accepted for Publication)
22. Rincy Merlin Mathew, M. Shyamala Devi, Shakila Basheer," Exploration of Neighbor Kernels and Feature Estimators for Heart Disease Prediction using Machine Learning", International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.12, 10 October 2019. (Accepted for Publication)