Effective Cardiovascular Disease Prediction System using Ensemble Machine Learning Methods

Dr.P.Deepika
Associate Professor, Department of Artificial Intelligence and Machine Learning,
Dr. N.G.P. Arts and Science College, (Achyut Tiwari., 2023)
Coimbatore, Tamilnadu, India.
emailid: mscdeepika65@gmail.com

Abstract

Cardiovascular Disease (CVD) remains a leading global health concern. Accurate prediction models are crucial for early detection and risk stratification. This paper evaluates the prediction results of ensemble learning methods with traditional linear method and additive regression approaches for predicting CVD. In this paper, the evaluation of ensemble model performance is experimented and the produced results are compared with results of Linear Regression method and Additive Regression methods. The results presented the Bagging ensemble method was outperform linear and additive regression between CVD risk factors and disease occurrence is expected to provide a significant advantage.
Keywords: Ensemble Learning, Cardiovascular Disease, Prediction, Machine Learning, Bagging

1. INTRODUCTION

The Cardiovascular Disease (CVD) is most serious health issue which leads death around the world. According to the report of World Heart federation there are 20.5 million deaths due to the heart diseases. The early prediction using traditional methods are crucial. There are four types of cardiovascular disease known as failure of heart, coronary disease, cardiomyopathy and congenital heart disease (K.Thenmozhi., 2014). The machine learning methods are advanced which helps to predict the disease and the risk factors. CVD occurs due to various reasons such as pressure level of blood in human body, cholesterol level, irregular pulse rate and unhealthy habits. Tobacco use damages the lining of arteries which leads to atherosclerosis (Sagir A M., 2017). This can reduce the oxygen level in blood which will increase the blood pressure and heart rate. The poorly controlled blood sugar will lead for diabetes which can increase full of fat deposits on the walls of blood vessels. The family history is also influencing the risk of cardiovascular disease. The conditions like choric inflammatory diseases and certain infections can increase the risk of CVD. The sleeping disorder and poor kidney functions are the associative factors of increasing the risk of heart disease (P.Deepika. D. , 2020).

Machine learning techniques are more efficient in the area of healthcare which are used for various functions like medical imaging and diagnostics, predictive analysis, personalized medicine, clinical decision support, operational efficiency, remote monitoring and telehealth, patient monitoring and precision health etc. The techniques of machine learning provide significant growth in healthcare, revolutionizing various aspects of patient care treatment, planning and healthcare management (P.Deepika. A. S., 2018). This work includes review of existing methods, data collection and correlation analysis process for applying ensemble methods for accurate prediction. The results will be compared with existing models and performance evaluation used for presenting the significant impact of ensemble methods on CVD prediction.

2. ENSEMBLE METHODS

Ensemble Methods are effective techniques of machine learning which collect together multiple models to increase the accuracy and robustness of predictions (Abdulla Alqahtani., 2022). These methods are particularly effective in reducing variance, bias and enhancing the overall performance of predictive models. Some ensemble methods are Bagging, Boosting, Stacking and Voting (Vardhan Shorewala., 2021).

2.1 Bagging:

Bagging model consist training diverse versions of the same model on different subsets of the training samples and then aggregating their predictions, typically by averaging or voting (Kaan Uyar., 2017). Bagging is short of Bootstrap Aggregating which can increase the level of prediction accuracy and model robustness. This model aims to reduce the variance and prevent overfitting by training various kinds of model on unlike subset of data to train and then collecting all models prediction (Ibomoiye Domor Mienye., 2020). The reduction of variance can be achieved by training more number of models on different subset of data and leads for stability and reliability of predictions. The common bagging algorithms are Random Forest method and Bagged Decision Tree.

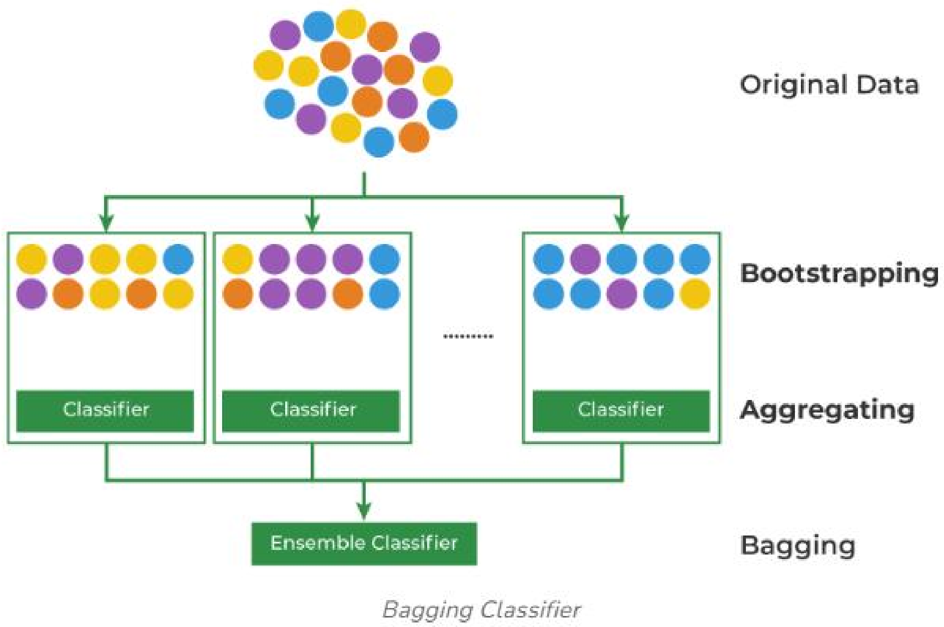

Fig:1 Bagging Method

The figure.1 (geeksforgeeks, 2024) explores the bagging classifier working process with the original dataset.

2.2 Boosting:

Boosting is used to train the models consecutively with each new model focusing on the errors done by the existing last one. The concluding prediction is a weighted sum of the predictions produced by all models (Mohammed Nasir Uddin., 2021). This combines various weak learners for creating a strong learner. There are three principles for Boosting method such as Sequential Learning, Combining Models and Error Reduction (S.Sasikala., 2017). The common Boosting algorithms are AdaBoost (Adaptive Boosting), Gradient Boosting Machine (GBM), XGBoost (Extreme Gradient Boosting), LightGBM and CatBoost (Jothi Prakash V and Karthikeyan N K., 2021). By using these models the improve accuracy can be achieved. These models are also helpful for reducing the Bias and variance and mode flexible for execution.

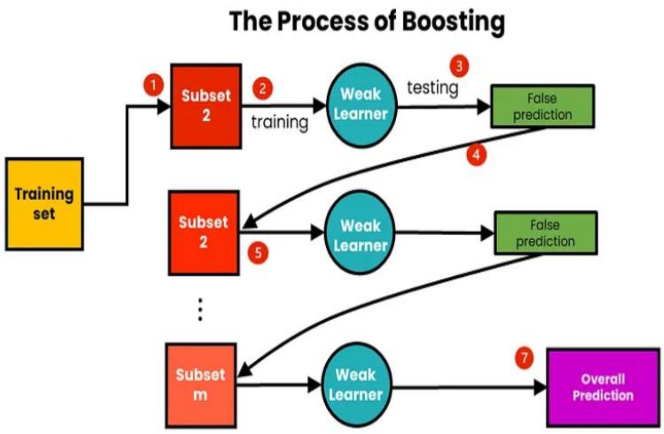

Fig: 2 Boosting Method

The figure.2 (Soni, 2023)explains Boosting method using training samples for prediction process.

2.3 Stacking:

Stacking involves training multiple base models which are in different types and then using another model. This other model is known as meta-model which is used for combining their prediction results (Karadeniz T., 2021). The prediction performance can be improved by using another model called meta-model to model-1 to learn how to combine the predictions of the base model. This allows different types of models as base learners which enables the ensemble to capture diverse patterns in the data for providing flexibility (Krishnamoorthy Natarajan, 2024). Stacking can reduce the risk of overfitting that might occur if a single model were used.

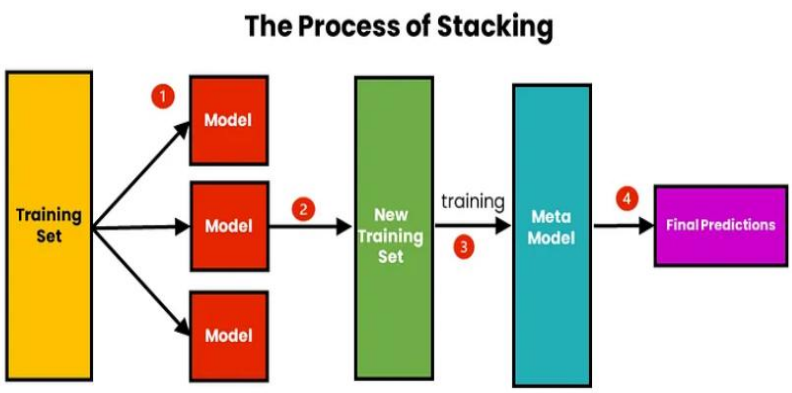

Fig:3 Stacking Method

The figure.3 (Soni, 2023) presents Stacking method using training set which incorporates various models to produce the prediction result.

2.4 Voting:

Voting ensembles model is used for combining the predictions of number of models by taking a majority vote for classification or average the values used for regression. It is an effective ensemble technique which combines the predictions from multiple models to make final decisions (Runchan Li., 2019). The major idea is to influence the strengths of various models for improving overall performance and robustness. This method is suitable for classification tasks and also for regression. These techniques work based on two principles such as Multiple Models and Combining Predictions (Achyut Tiwari., 2023). Three voting types are in prediction process such as hard voting method, soft voting method and Averaging.

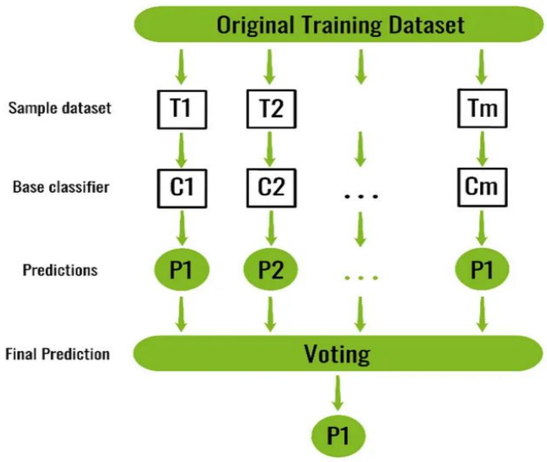

Fig:4 Voting Method

The figure.4 (Kumar, 2023)explains voting method using collected dataset and to provide the prediction result using selected voting method.

3. LITERATURE REVIEW

Abdulla Alqahtani et al. (Abdulla Alqahtani., 2022)presented an ensemble learning model which is used for predicting Cardiovascular Diseases (CVD). The proposed ensemble model was based on deep learning and machine learning methods for predicting the heart diseases. The experiment showed that the six classification algorithms were used for CVD prediction. The random forest method was used for extracting features from dataset. The proposed method indicated four machine learning methods were combines such as Random Forest (RF), Decision Tree, Extreme Gradient Boosting and K-Nearest Neighbour(KNN). The experiment was done by using the publicly available dataset. The Stacking Ensemble method was used and the results are compared. The Deep Learning methods were combined for ensemble process to predict the CVD. The results highlighted the proposed Machine learning method reached highest accuracy of 88.70 among all existing methods they were experimented. Vardhan Shorewala presented an early risk prediction of heart failures and CVD with the use of proposed ensemble model. The experiment work used the ensemble method which combined KNN, Naïve Bayes and Binary Logistic Classification methods. The Bagging models, Boosting models and Stacking models were used on the dataset. The comparative analytical study was done on the produced results. The average accuracy of Boosted model is 73.4%, Bagging model accuracy is 1.96% higher than the traditional methods. The stacked model combined Random Forest method, KNN method and Support Vector Machine methods which achieved 75.1% accuracy in coronary heart disease prediction. K-fold cross validation techniques applied for validation. Ibomoiye Domor Mienye et al. proposed an ensemble learning model which is used for identifying the risk factors of heart disease. The improved model involved the task to randomly partitioning the dataset using approach mean based splitting to produce the subset of dataset. The homogeneous ensemble model is created using different Classification and Regression Models. The experiment was done using Cleveland and Framingham datasets. The proposed model produced 93% accuracy and 91% accuracy for Cleveland and Framingham datasets respectively. Mohammed Nasir Uddin et al. (Mohammed Nasir Uddin., 2021)presented the prediction model for CVD by implementing various Machine learning approach in association with one ensemble method. The proposed method utilized various methods to select the features like Gain Ratio Evaluator, Correlation Attribute Evaluator, Extra Trees Classification and LASSO. The ensemble approach used K-Nearest Neighbour, Random Forest, Gradient Boosting Models and Naïve Bayes. The result showed that the used ensemble model produced 94% accuracy by using publicly available dataset for cardiovascular disease prediction. Jothi Prakash V and Karthikeyan N K (Jothi Prakash V and Karthikeyan N K., 2021)proposed a new approach for feature selection and ensemble technique for effective prediction of heart disease. The main features are selected using evolutionary features selection techniques and the ensemble model approach used for prediction. The Genetic Algorithm with Linear Discriminant Analysis model is combined for attaining high accurateness for prediction. This model achieved 82.1% accuracy on SPECTF dataset and produced 84.95% accuracy on Coronary Heart Disease dataset. Karadeniz T et al. (Karadeniz T., 2021) used ensemble method for heart disease prediction. They used Statlog and Spectf datasets for this work. They used Choas Firefly Attribute Reduction and Fuzzy Logic method for classification. Two new methods were proposed such as Reference Vactor Classifier (RVC) and Shrunk Covariance Classifier (SCC) for prediction analysis. The experiment results are compared and highlighted that the prediction accuracy was 88.7% on Spectf dataset and 88.8% on Statlog dataset. Bhanu Prakash Doppala et al. (Bhanu Prakash Doppala., 2022) produced an ensemble model for accurate prediction of heart disease. Three different datasets used for the experiment such as Mendeley Data Center, IEEE DataPort and Cleveland Dataset. The proposed ensemble model combined Naïve Bayes method, Support Vector Method, Random Forest Method and XGBoost method. The Voting Classifier method is used for identifying heart disease on the dataset. The proposed model achieved 96.75%, 93.39% and 88.24% accuracies on Mendeley Data Center, IEEE DataPort and Cleveland Datasets respectively.

4. DATA PROCESSING AND ANALYSIS

4.1 Description of Dataset

The dataset used for this work is collected from KAGGLE repository namely Cleveland. This dataset contains 303 observations with 13 attributes and 1 class label. All these attributes indicating various parameter value which can influence the occurrences of Cardiovascular diseases (Mishra, Mallick, Tripathy, Bhoi, & González-Briones, 2020). This helps to find the non-invasive diagnosis of Cardiovascular Diseases.

S. No	Field	Description	Range and Values
1	Age	Age of the patient	0 -100 years
2	Sex	Gender of patient	0-1 (1: Male, 0: Female)
3	Chest pain	Type of chest pain	1-4 (1: typical angina, 2: atypical angina, 3: non-anginal, 4: asymptotic)
4	Resting Blood pressure	Blood pressure during rest	Mm Hg
5	Cholesterol	Serum Cholesterol	Mg/dl
6	Fasting Blood Sugar	Blood Sugar content before food intake if> 120 mg/dl	0-1 (0: false, 1: true)
7	ECG	Resting Electrocardiographic results	0-1 (0: normal, 1: having ST-T wave)
8	Max heart Rate	Maximum Heart Beat rate	Beast/min
9	Exercise induced angina	Has pain been induced by exercise	0-1 (0: No, 1: Yes)
10	Old Peak	ST depression induced by exercise relative to rest	0-4
11	Slope of peak exercise	Slope of the peak exercise ST Segment	1-3 (1: up sloping, 2: flat, 3: down sloping)
12	ca	Number of vessels colored by Fluoroscopy	0-3
13	Thal	Defect Type	3: Normal 6: fixed defect 7: reversible defect
14	Num	Diagnosis of heart disease	(0: <50% narrowing, 1:>50% narrowing)

Table:1 Cleveland Dataset Description (Oumaima Terrada., 2020)

The dataset details are represented in table view for insight view about features and the description. The range and values for each feature is presented in this table.The collected dataset has 303 records with no missing values. The real-world data contains a great deal of noisy and partial data. To remove these errors and produce reliable factors, the collected data must be pre-processed. The collected data typically contains noisy and null values which are used to remove or modify categories, get rid of duplicated values, handling null values and outliers and etc (Ahmad, Munir, Bhatti, Aftab, & Raza, 2017). It is mandatory to scale and normalize the data in order to specify the range of value. Data transformation is used for transforming the raw data into format or structured data which is more suitable for building a model and making data exploration process.

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
56	1	3	130	256	1	2	142	1	0.6	2	1	6	2
44	1	2	120	263	0	0	173	0	0	1	0	7	0
52	1	3	172	199	1	0	162	0	0.5	1	0	7	0
57	1	3	150	168	0	0	174	0	1.6	1	0	3	0
48	1	2	110	229	0	0	168	0	1	3	0	7	1
54	1	4	140	239	0	0	160	0	1.2	1	0	3	0
48	0	3	130	275	0	0	139	0	0.2	1	0	3	0
49	1	2	130	266	0	0	171	0	0.6	1	0	3	0
64	1	1	110	211	0	2	144	1	1.8	2	0	3	0
58	0	1	150	283	1	2	162	0	1	1	0	3	0
58	1	2	120	284	0	2	160	0	1.8	2	0	3	1
58	1	3	132	224	0	2	173	0	3.2	1	2	7	3
60	1	4	130	206	0	2	132	1	2.4	2	2	7	4

Fig:5 Sample dataset- A Snapshot (Original Work)

4.2 Correlation Coefficient of Attributes

The correlation coefficient is used to measure the linear relationship of two attributes in the collected dataset (P.Deepika. S. , 2020). The significant features are having high correlation with target attribute. Pearson Correlation is one of the methods which is used for finding absolute value or threshold for selecting features.

$$r = \frac{Covar(x,y)}{\sqrt{Var(x)Var(y)}}$$

$$Covar(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n}$$

$$Var(x) = \frac{\sum(x-\bar{x})^2}{n}$$

$$Var(y) = \frac{\sum(y-\bar{y})^2}{n}$$

r : *Linear Correlation*

Covar : *Covariance*

Var : *Variance*

Here the value of r indicates strength of two variables. This can be presented between 1 to -1 where the perfect positive correlation is addressed using 1 and perfect negative correlation is addressed using -1. The value 0 indicates there is no correlation between the variables (Shimpi, Shah, Shroff, & Godbole, 2017).

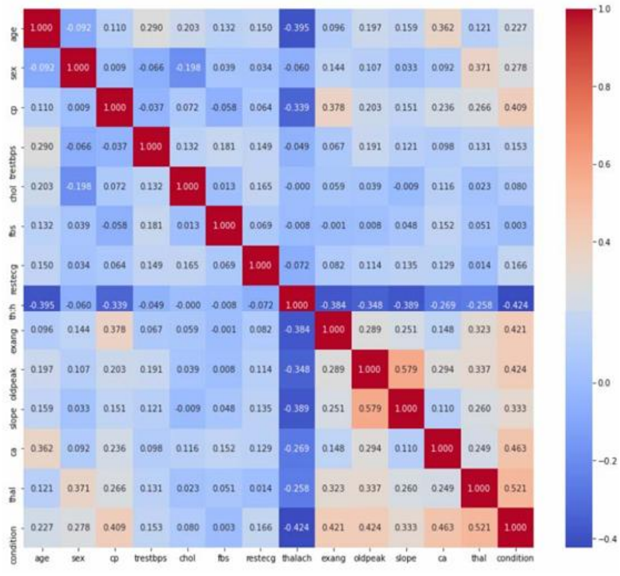

Fig:6 Cleveland CVD dataset Correlation Matrix

4.3 Histogram of Dataset

A histogram chart is used to present the unveils the distribution on numerical values. This indicates the range of values in x-axis and frequency or count of values in y-axis.

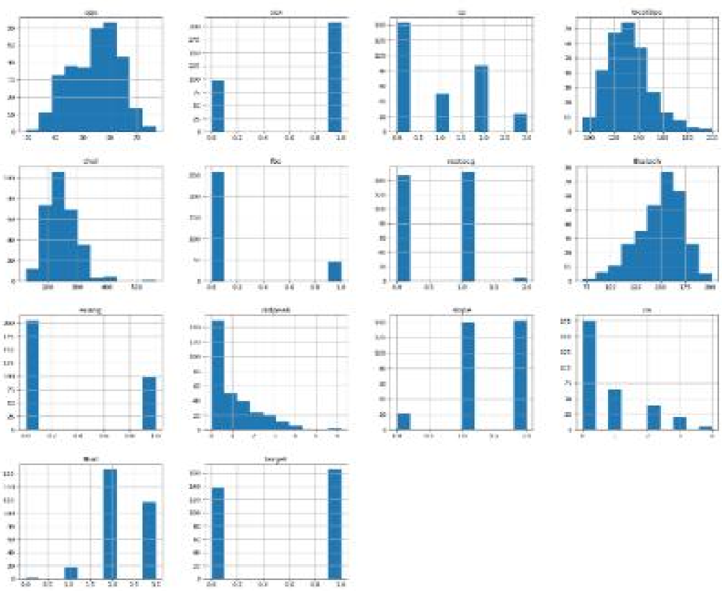

Fig:7 Histogram of Cleveland CVD dataset

5. METHODS

5.1 Ensemble Machine Learning Method

Classification is one of the machine learning methods which is used to process data and dividing into groups according to the correlation between various data points. Classification is employed to predict cardiovascular disease (P.Deepika, 2022). While there are many machine learning models are there, the recommended approach can make use of any of the following models or methodologies. In order to predict cardiovascular disease three methods are used such as Bagging, Additive Regression and Linear Regression (Ayad E. Korial., 2024).

5.2 Bagging

The bagging model is formally known as Bootstrap Aggregating which consist the creation of multiple models and combine all the models to produce the overall results. The data points are randomly selected and trained with multiple models (Tanjim Mahmud., 2023). The main focus of this model is to reduce the variance and to avoid the overfitting issues.

5.3 Additive Regression

This is a statistical technique which is applied for modelling the relationship between dependent attributes and multiple independent attributes. The discussed method follows non-linear relationship that is not similar with sequential regression.

5.4 Linear Regression

Linear Regression is applied to find the relationship between a dependent feature and one or more independent features. This model mainly focused the finding the factors influencing the dependent variable.

6. EXPERIMENTAL RESULTS AND DISCUSSION

All the discussed above techniques are experimented using WEKA platform. The following figure presents the execution time for all three models for prediction process.

Fig:8 Model Building time Comparison Chart

6.1 Performance Evaluation

The execution results for Bagging method, additive regression and Linear regression are compared using the performance indicators like correlation coefficient, Root Mean Squared Error, Mean Absolute error, Root Relative Squared Error and Relative Absolute Error.

Performance Indicators	Bagging	Additive Regression	Linear Regression
Correlation Coefficient	0.8399	0.7704	0.7571
Mean Absolute Error	0.5435	0.5996	0.6107
Root Mean Squared Error	0.7084	0.7841	0.8013
Relative Absolute Error	53.5624	59.097	60.1886
Root Relative Squared Error	57.7569	63.9336	65.33

Table:2 Performance Analysis for Bagging, Additive Regression and Linear Regression

The Root Mean Squared Error provides meaningful results for understanding average magnitude. The Relative Absolute Error is used to calculate the average percent of error. It calculates the absolute difference between the actual outcome with the predicted results. This makes the error relative to the size of actual value. The Root Relative Squared Error is used to analyze efficiency of used regression models. This considered both magnitude of errors and natural range of data used for prediction.

6.2 Performance Comparison

The following chart presents the performance comparison of Bagging, Additive Regression and Linear Regression methods.

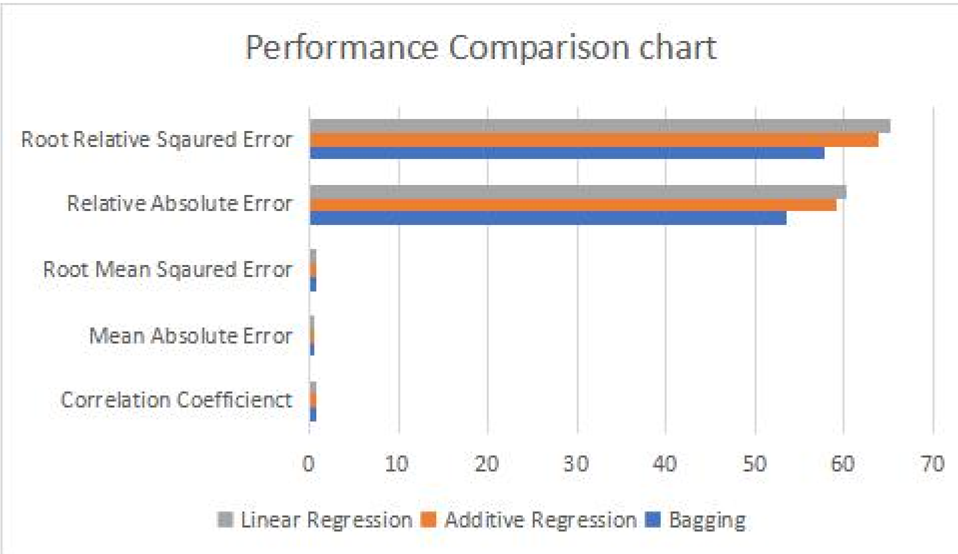

Fig:9 Performance Comparison of Ensemble model with traditional methods.

The Comparison results highlighted that the Correlation Coefficient for bagging methods is 0.8399 and the model building time is comparatively less than other models. The Mean Absolute Error value for bagging is 0.5435 which is less than other two methods. The Mean Absolute Error is used to represent the average between predicted values and the actual values. The lower Mean Absolute Error indicates the more appropriate prediction using the model. Likewise, the Root Mean Squared Error, Relative Absolute Error, Root Relative

Squared Error values are comparatively low with bagging results when compared to additive regression and Linear regression methods.

7. CONCLUSION

In this work three different machine learning approaches executed for the predicting the cardiovascular disease. The used methods are effective and produced effective prediction results. The performance indicators are used to analyze the efficiency of Ensemble method and other machine learning methods. The Cleveland dataset with 303 records used for the prediction analysis. The Bagging method is employed and the performance was compared with Additive Regression and Linear Regression methods. The used regression models and ensemble methods are evaluated using various performance indicators such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Relative Absolute Error and Mean Absolute Error (MAE). The produced outcome of each model is analyzed and compared. The comparison results highlighted that the Bagging method is more suitable for the prediction of cardiovascular disease than the linear regression and additive regression methods.

References:

Bibliography

Abdulla Alqahtani., S. A. (2022). Cardiovascular Disease Detection using Ensemble Learning. *Computational Intelligence and Neuroscience*, Vol-2022, Issue-1.

Achyut Tiwari., A. C. (2023). Ensemble Framework for cardiovascular disease prediction. *Computers Biology and medicine*, Vol-146.

Ahmad, T., Munir, A., Bhatti, S., Aftab, M., & Raza, M. (2017). Survival analysis of heart failure patients: A case study. *PLoS ONE* .

Ayad E. Korial., I. I. (2024). An Improved Ensemble-Based Cardiovascular Disease Detection System with Chi-Square Feature Selection. *Computers*, Vol-13, issue-6.

Bhanu Prakash Doppala., D. B. (2022). A Reliable Machine Intelligence Model for Accurate Identification of Cardiovascular Disease using Ensemble Techniques. *Journal of Healthcare Engineering*, vol-2022.

geeksforgeeks. (2024, 2 12). Retrieved 7 2, 2024, from https://www.geeksforgeeks.org/ml-bagging-classifier/

Ibomoiye Domor Mienye., Y. S. (2020). An Improved Ensemble Learning Approach for the prediction of Heart Diseases. *Informatics in Medicine Unlocked*, Vol-20.

Jothi Prakash V and Karthikeyan N K. (2021). Enhanced Evolutionary Feature Selection and Ensemble Method for Cardiovascular Disease Prediction. *Interdisciplinary Sciences: Computational Life Sciences*, Vol-13, issue-3.

K.Thenmozhi., P. (2014). Heart Disease Prediction using Classification with Different Decision Tree Techniques. *International Journal of Engineering Research and General Science*, Vol-2, Isse-6.

Kaan Uyar., A. I. (2017). Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks. *Procedia Computer Science*, Vol-120.

Karadeniz T., T. G. (2021). Ensemble Methods for Heart Disease Prediction. *A New Generation Computing*.

Krishnamoorthy Natarajan, V. V. (2024). Efficient Heart Disease Classification Through Stacked Ensemble with Optimized Firefly Feature Selection. *International Journal of Computational Intelligence Systems*, Vol-17, issue-174.

Kumar, R. (2023, 8 12). *Medium*. Retrieved 7 2, 2024, from https://medium.com/@ranjankumar_29097/votingclassifier-3f85ba8e4580

Mishra, S., Mallick, P., Tripathy, H., Bhoi, A., & González-Briones, A. (2020). Performance evaluation of a proposed machine learning model for chronic disease datasets using an integrated attribute evaluator and an improved decision tree classifier. *Application Science*, Vol-10, .

Mohammed Nasir Uddin., R. K. (2021). An Ensemble method based Multilayer Dynamic System to predict Cardiovascular Disease using Machine Learning Approaches. *Informatics in Medicine Unlocked*, vol-24.

Oumaima Terrada., S. H. (2020). Supervised Machine Learning Based Medical Diagnosis Support System for Prediction of Patients with Heart Disease. *Advances in Science, Technology and Engineering Systems Journal*, Vol. 5, No. 5.

P.Deepika, S. (2022). Data Classification pertaining to heart disease using hybrid chicken swarm optimization with Artificial Neural Network. *International Journal of Mechanical Engineering,*, Vol-7.

P.Deepika., A. S. (2018). A Noval Classification and Prediction Algorithm for Heart Disease Identification. *International Journal of Engineering Science Invention*, Vol-7, issue-2.

P.Deepika., D. (2020). Enhanced model for Prediction and Classification for cardiovascular disease using Decision tree with particle swarm optimization. *2020 4th International Conference on Electronics, Communication and Aerospace technology (ICECA)*.

P.Deepika., S. (2020). Effective Anticipation of Cardiovascular Disease using OPT-ANN Deep Learning Techniques. *Solid State Technology*, Vol-63.

Runchan Li., S. S. (2019). Multilevel Risk Prediction of Cardiovascular Disease based AdaBoost+RF Ensemble Learning. *IOP Conference Series: Materials Science and Engineering,*, Vol-533.

S.Sasikala., S. S. (2017). Anticipating the chronic kidney disorder using performance optimization in AdaBoost and Multilayer perceptron. *Imperial Journal of Interdisciplinary Research*.

Sagir A M., S. S. (2017). A novel adaptive neuro fuzzy inferences system based classification model for heart Disease Prediction. *Pertanika Journal of Science and Technology*, Vol-25, no.1.

Shimpi, P., Shah, S., Shroff, M., & Godbole, A. (2017). A machine learning approach for the classification of cardiac arrhythmia. *In Proceedings of the International Conference on Computing Methodologies and Communication, ICCMC 2017, Erode, India*.

Soni, B. (2023, 4 28). *Medium*. Retrieved 7 2, 2024, from https://medium.com/@brijesh_soni/understanding-boosting-in-machine-learning-a-comprehensive-guide-bdeaa1167a6

Tanjim Mahmud., A. B. (2023). An Improved Framework for Reliable Cardiovascular Disease Prediction Using Hybrid Ensemble Learning. *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*.

Vardhan Shorewala. (2021). Early Detection of Coronary Heart Disease using Ensemble Techniques. *Informatics in Medicine Unlocked*, Vol-26.