Applying Ensemble Learning Algorithm in Early Prognosis of Heart Illness

Lijetha.C. Jaffrin[1], Dr. J. Visumathi[2]

[1]Assistant Professor, Department of IT, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India
[2]Professor, Department of CSE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, India
lijethacjaffrin@veltech.edu.in[1], drvisumathij@veltech.edu.in[2]

Article History: Received: 10 November 2020; Revised: 12 January 2021; Accepted: 27 January 2021; Published online: 05 April 2021

Abstract: Medical diagnosis and treatment of diseases are the key elements of machine learning algorithms nowadays. To find similarities between various diseases, machine learning algorithms are used. Many people are now dying due to sudden heart attacks. Predicting and diagnosing heart disease is a daunting aspect faced by physicians and hospitals around the world. There is a need to foreknow whether or not a person is at risk of heart syndrome in advance, in order to minimize the number of deaths due to heart disease. In this field, machine learning algorithms play a very significant role. Many researchers are carrying out their research in this field to create software that can assist doctors to make decisions about cardiac illness prognosis. In this paper, Random Forest and AdaBoost ensemble Machine Learning Procedures are used in advance to predict heart disease. The datasets are handled in python programming by means of Anaconda Spyder IDE to validate the machine learning algorithm.

Keywords: Heart disease, Machine learning, Random Forest, AdaBoost

1. Introduction

In underdeveloped, emerging and even developed countries, heart illness is the greatest cause of demise, lakhs of people die every year because of this. An assessed 17 million individuals expire annually from CVDs, especially cardiac arrest and aneurysm. Smoking, which raises the threat of expiring from heart illness and Stroke by 2 or 3 times, can be linked to a large number of these deaths. Other key risk factors that raise individual risks for cardiovascular disorders are physical inactivity and unhealthy diet. The mortality rate can definitely be reduced by early detection and accurate prediction of this disease. There is a need for clinical history and causes that contribute to heart diseases for the successful diagnosis and prediction of heart diseases. Machine learning can be used in the medical field for the diagnosis, identification and prediction of different diseases. Using its numerous machine learning techniques and hybrid models, machine learning plays a vital role in predicting and retaining critical heart disease data. The core drive of this paper is to include tools at an early stage for doctors to diagnose heart disease. Random Forest and AdaBoost ensemble Machine Learning Procedures are used to predict heart disease.

Ensemble methods are techniques in machine learning that combines many basic models to construct one prime prognostic model. Ensemble techniques are meta-algorithms that incorporate many techniques of machine learning into a single predictive model to minimize uncertainty. The algorithm can be any algorithm for machine learning, such as logistic regression, decision tree, etc. These models are called "base models" when they are used as inputs for ensemble methods. Three general methods for merging the forecasts from various prototypes include bagging, bias (boosting) or enhance predictions (stacking). Bagging generates several prototypes from diverse down samples of the training data records. Boosting builds several prototypes which studies to correct the forecast in accuracies in the previous model. To combine predictions, stacking is used to construct numerous prototypes and basic measurements such as computing the mean. The benefits of the Ensemble technique are that it is an established way of improving the model's accuracy and works in most situations, making the model more robust and stable, ensuring good performance in most scenarios on the test cases.

A supervised ensemble learning algorithm that is used for both classifications and regression issues is Random Forest. But it is, however, used mostly for classification issues. There is a clear connection between the number of trees in the forest and the results that can be obtained. A forest is made up of trees, more trees mean robust forests. Similarly, on data samples, the random forest algorithm produces decision trees and then gets the prediction from each of them and ultimately selects the best solution by voting. It uses a variety of decision trees and predicts the more correct outcome in the case of regression and voting in the case of classification by averaging. It is an ensemble strategy that is better than a single decision tree since, by averaging the outcome, it

avoids over-fitting. The two main principles that refer it as random is that when constructing trees, a random sample of training data collection and Random subsets of features that are considered when nodes are separated.

A Boosting algorithm is Adaptive Boosting, or more generally recognized as AdaBoost. The method of correcting its predecessor is used by this algorithm. It pays more attention to the previous model in the fitted training situations. Therefore, the emphasis is more on the complicated cases than the others in each new predictor. On different weighted training data, it suits a sequence of weak students. This begins by predicting the original set of data and gives each observation equal weight. If the first learner's prediction is inaccurate, then it gives greater weight to results that have been wrongly predicted. As an iterative process, it continues to add learneruntil the number of models or accuracy has reached a cap. AdaBoost uses decision stamps but any machine learning algorithm if it accepts weight on the training data set can be used. For both grouping and deterioration complications, AdaBoost algorithms can be used.

2. Literature Survey

In order to reliably envisage the existence of heart illness for a specific sufferer, this paper [1] analyzes different ensemble approaches Bagged Tree, Random Forest, and AdaBoost with the variable sub categorization process - Particle Swarm Optimization (PSO). Experimental findings indicate that the highest precision was obtained by Bagged Tree and PSO. The authors in this paper improved the group classifier for PSO grouping as subcategory selection function for heart illness prognosis. The suggested framework utilizes the UCI repository's main Stalog dataset. To delete irrelevant and incomplete data, the data collection is pre-processed. In addition, the prevailing characteristics for different performance indicators are checked on the ensemble model.

The authors in this paper [2] is analyzing data set by carrying out data justification and preprocessing procedures, visualization of data exploration and model training, building classification prototype and performance evaluation of supervised machine learning procedures with grouping report, defining contingency matrix and grouping data with precedence. The primary aim is building model of prognostic analytics to analyze the different phases of heart patients through collaborative learning techniques such as Bagging, Boosting and Voting to increase precision of procedures that are inefficient. The product of these ensemble methods is evaluated and the one that proves to boost accuracy is considered and displayed using a GUI.

An enhanced machine learning method for prognosis of heart illness possibility is suggested in this paper [3]. The procedure comprises arbitrarily segregating the data records using a mean-based slicing strategy into smaller subsets. By classification and regression tree, the different partitions are then modeled (CART). A similar group is formed by precision-based weighted aging classifier group from the various CART models, which is an alteration of weighted aging classifier ensemble (WAE). The strategy guarantees the achievement of optimum efficiency. Supplementary machine learning procedures and related academic tasks were outperformed by experimental findings on the Cleveland and Framingham datasets. The enhanced efficiency of proposed ensemble learning approach is further confirmed by the recipient operational characteristic curves. The findings showed that suggested ensemble can efficiently foresee the possibility of heart disease.

This paper [4] suggested the use of a K-NN, SVM, MK-NN and CART (Decision Tree Algorithm) ensemble classifier to effectively predict heart disease. The performance and efficiency of the classifier algorithms and ensemble are evaluated. The findings show that the system suggested for determining the occurrence or non-existence of heart disease was more effective. The ensemble classifier predicts cardiac disease more reliably from these algorithms.

In this paper [5], for precise coronary heart illness analysis and result estimates, a progressive ensemble machine learning system is developed using an adaptive Boosting procedure. Four different data sets for the diagnosis of heart illness were added to established ensemble learning grouping and extrapolation prototypes, including datasets from Cleveland Clinic Foundation (CCF), the Hungarian Institute of Cardiology (HIC), the Long Beach Medical Center (LBMC) and the Switzerland University Hospital (SUH). Research findings revealed that model accuracies for CCF, HIC, LBMC and SUH were improved by built ensemble learning grouping and forecast prototypes. Therefore, diagnoses of coronary heart illness resulting from established collaborative learning and forecast prototypes are accurate, useful and benefit sufferers worldwide, particularly those from developing countries and regions where diagnostic specialists in heart illness are few.

This paper [6] explored the practice of hybrid ensemble prototype in which accurate ensemble is proposed than simple collaborative prototypes, leading to better results than other models of prediction of heart disease. A dataset enclosing 278 records of SPECT heart illness is used to assess performance of proposed model by

obtaining enhanced classification accuracy, sensitivity and specificity after relating the prototype over data, indicating appropriate performance of the suggested hybrid ensemble model compared to the simple ensemble model and other advanced models.

The core objective of this paper [7] is to practice machine learning algorithms to build a Heart Illness Prognosis Framework to foresee the existence of heart illness. Extreme Learning Machine (ELM) is the novel group of Single-Hidden Layer Feed Forward Neural Network (SLFN), which is theoretically easy, quick to implement and has been documented to suffer from over-fitting. It is reported that these traditional classification methods have some shortcomings in their diagnostic capability. This work suggested an Ensemble ELM to rectify these problems in ELM. Accordingly, the proposal to create an ensemble of many training predictors using different groups of random factors where each predictor's factors are modified based on unique criteria and then make conclusions for testing models by majority voting with ensemble. In order to foresee the presence of cardiovascular syndrome, experimental findings indicate that the proposed approach works well and offers improved classification accuracy.

In numerous fields, the advancement of computer science has brought vast opportunities. One of them is machine learning, which is generally used in various domains. At an early stage of human life, machine learning methods are employed to envisage medical conditions. To conduct the experiments, this paper considered Heart illness-related data collection. The most popular ensemble learning algorithms, Random Forest and Support Vector Machine were implemented in this system [8] to construct a classifier model that will predict disease with greater performance and accuracy.

This paper [9] provides a survey and performance analysis of different prototypes based on certain procedures and systems. The researchers have found very common prototypes based on supervised learning procedures like Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Naïve Bayes, Decision Trees (DT), Random Forest (RF) and ensemble prototypes. To systematize the study of complex data, machine learning procedures were implemented on different medicinal data records.

3. System Methodology

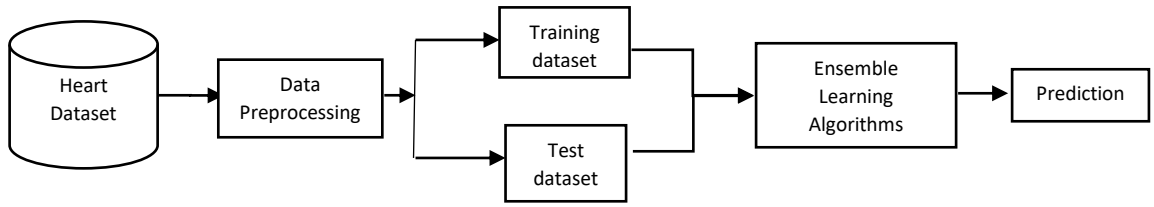

Figure 1. Overview of Ensemble learning algorithm

3.1. Data Preprocessing

Data Collection is the first phase in envisaging the disease. Data collection is a systematic method in which observations or measurements are taken. Data collection helps us to obtain first-hand information and original insights into your research issue, whether you conduct research for business, governmental or academic purposes. In csv format, the dataset for heart disease was taken here. This dataset consists of 14 separate fields.

The second step of disease prediction is the preprocessing of data. In Machine Learning, data preprocessing is an integral step as the quality of data and the valuable knowledge that can be obtained from it directly affects our model's ability to learn; thus, it is extremely necessary to preprocess the information before feeding it into our model. You still have a pre-process stage to work on every time you construct an ML model. So the ML model you are going to develop can be trained on the data in the right way.

The next step here is to divide our dataset into two sets, a training samples and a test samples, after preprocessing the dataset to get better performance. In our training set, machine learning models such as Random Forest and AdaBoost understand any similarities and then test the test dataset models to verify how accurately they can predict. 70 per cent of records are taken as a dataset for preparation and 30 per cent of records are taken as a dataset for study.

The final step in the preprocessing of data is to apply the very critical scaling function. It is a methodology used to standardize the selection of independent data variables or characteristics. Scaling of Functionality transforms all our variables into the same scale.

3.2. Feature Selection

Feature Selection is the method in which you pick certain features that devote most to the prognosis variable or outcome automatically or manually. The features of the heart dataset are taken here which include the time of life, sexual category, chest pain category, hidden pressure level, serum cholesterol, blood sugar level, electrical heart recording, pulse rate, agina, capacity of peak exercise etc.

Table 1. Features of Heart dataset

Feature Name	Feature Description	Data Type
Age group	Patient's time of life	Numerical
sexual category	Sexual category of patient	Nominal
Cp	Chest pain category	Nominal
Trestbps	Hidden pressure level	Numerical
Chol	Lipid disorder in mg/dl	Numerical
Fbs	Blood Glucose level	Nominal
Resting	Electrical heart recording result	Nominal
Thalach	Pulse rate achieved	Numerical
Exang	agina	Nominal
Oldpeak	Exercise stress test related to rest	Numerical
Slope	Capacity of peak exercise	Nominal
Ca	Count of major vessels	Numerical
Thal	Result of stress test	Nominal
Targets	1 or 0	Nominal

4. Performance Evaluation

4.1. About Anaconda Spyder IDE

For data scientists, Anaconda is a data science forum. Spyder is a cross-platform open-source IDE. The Spyder IDE of Python is written completely in Python. It is developed solely for scientists, data analysts, and engineers and is designed by scientists. It is also known as the Scientific Python Creation IDE and has a huge range of remarkable features that highlight customizable syntax, breakpoint availability, interactive execution that enables you to run line, file, cell, working directory configurations, can automatically clear variables (or join debugging), can accomplish cell navigation, functions, lines, etc.

4.2. Normalization of dataset and Train-and-Test –Split

As implementation was done using python programming, this test_train_split is imported from model_selection library of scikit.

```
fromsklearn.linear_model import LogisticRegression
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
y_pred = logreg.predict(X_test)
```

4.3. Visualizing the results

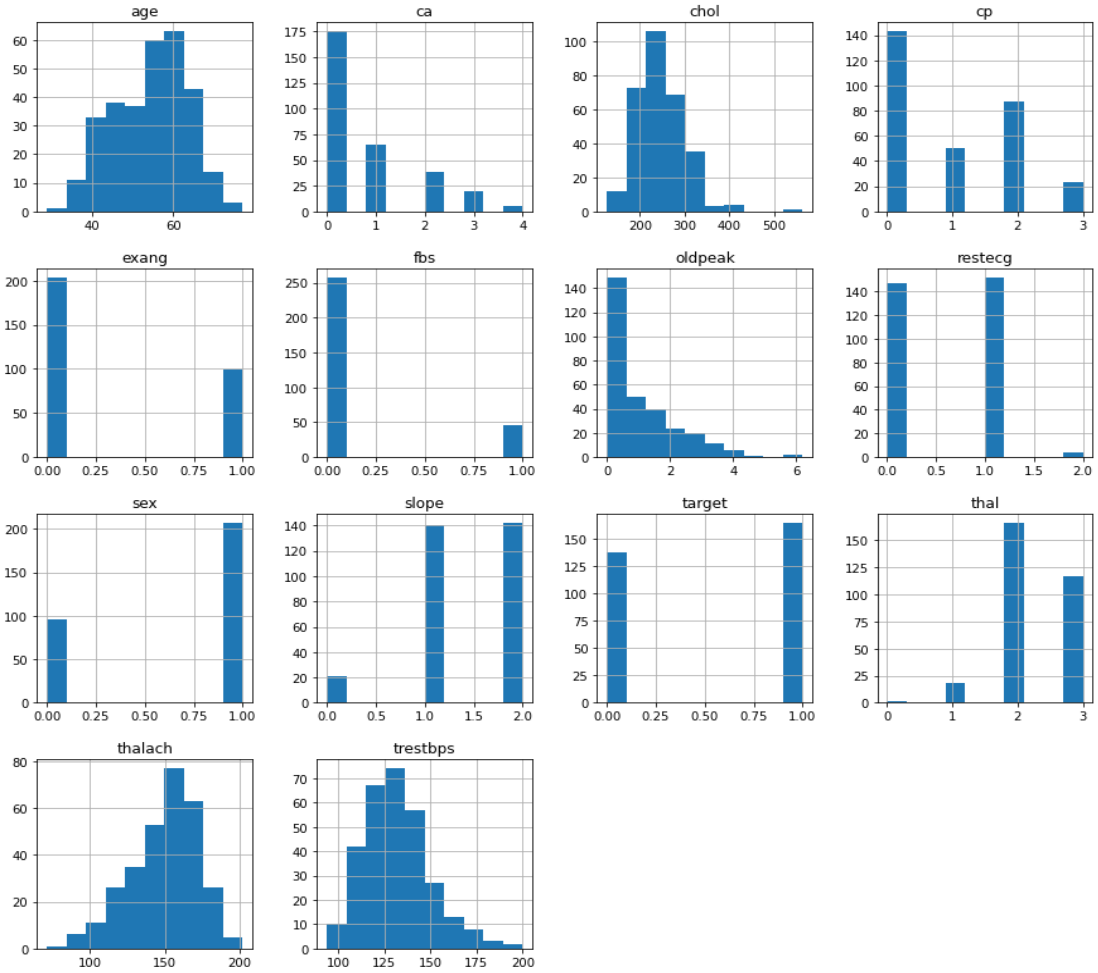

Figure 2. Visualization of features

4.4. Accuracy Comparison

A Contingency table is a particular table layout that enables an algorithm to visualize its output, normally a supervised learning method. The occurrences in a projected class are represented by each tuple of matrix, while each column represents the occurrences in a real group. The term comes from the fact that it creates easy way to see whether two groups are confused by the method. It is a particular type of contingency table with two dimensions actual and predicted and similar classes in both dimensions. In predictive analysis, there are two rows and two columns in the Contingency table that measures the sum of false positives, false negatives, true positives, and actual negatives. To calculate the accuracy of procedures such as Random Forest and AdaBoost algorithms, contingency table or Error matrix is used here.

The Contingency Table to measure the accuracy of Random Forest algorithm for the given dataset is as follows

```
Confusion Matrix:  [[25  8]
 [ 8 35]]


Precision:  [0.75757576 0.81395349]
Recall:     [0.75757576 0.81395349]
Fscore:     [0.75757576 0.81395349]
Support:    [33 43]
14 14 12 0.881578947368421
```

Figure 3. Contingency Table for Random Forest algorithm

The accuracy in foreseeing the heart illness using Random Forest procedure is 88%

The Contingency Table to measure the accuracy of AdaBoost algorithm for the given dataset is as follows

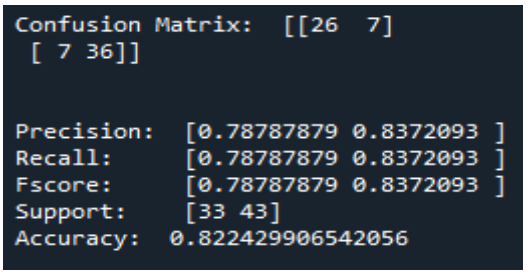

Figure 4. Contingency Table for AdaBoost algorithm

The accuracy in foreseeing the heart disease using Random Forest procedure is 82%.

By comparing the Contingency table of Random Forest and AdaBoost procedures, the accuracy of Random Forest procedure is more than AdaBoost algorithms. Thus it was concluded that Random Forest algorithm predicts the presence of disease accurately than AdaBoost algorithm.

5. Conclusion

Heart is a vital part of the human body and predicting cardiac disease in advance is also a significant challenge for humans. The dataset related to heart illness was taken. The fields in dataset were controlled in python programming by means of Random Forest and AdaBoost machine learning procedure. Accuracy of Random Forest and AdaBoost machine learning procedures depended on the dataset taken. Algorithm assessment was carried out on dataset whose features were shown in Table1. It was concluded that heart illness prediction using Random Forest algorithms was efficient and provided better accuracy than AdaBoost algorithm. Accuracy was estimated using Contingency table or Error matrix of algorithm as in Figure.3 and Figure.4. This paper uses Random Forest and AdaBoost algorithms to identify heart disease predicators. Further, data assessment was carried out in Python by means of Spyder IDE.

References

1. Yekkala.I, Dixit.S, &Jabbar.M. A, Prediction of heart disease using ensemble learning and Particle Swarm Optimization. 2017 International Conference on Smart Technologies for Smart Nation, 2017
2. RutujaGujare, D.Viji, Simran Bhatt, Enhanced Heart Disease Prediction Using Ensemble Learning Methods, International Journal of Advanced Science and Technology, Vol. 29 No. 06 2020
3. IbomoiyeDomorMienye, Yanxia Sun, Zenghui Wang, An improved ensemble learning approach for the prediction of heart disease risk, Informatics in Medicine Unlocked,Volume 20, 2020
4. RamatenkiSateesh Kumar, S.Sameen Fatima, Anna Thomas, Heart Disease Prediction using Ensemble Learning Method, International Journal of Recent Technology and Engineering (IJRTE), Volume-9 Issue-1, May 2020
5. Gopalakrishnan, R., Mohan, A., Sankar, L. P., & Vijayan, D. S. (2020). Characterisation On Toughness Property Of Self-Compacting Fibre Reinforced Concrete. In Journal of Environmental Protection and Ecology (Vol. 21, Issue 6, pp. 2153–2163).
6. ElhamNikookar, EbrahimNaderi Hybrid Ensemble Framework for Heart Disease Detection and Prediction, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 5, 2018
7. R.Subha, K.Anandakumar, A. Bharathi, An Ensemble based Extreme Learning Machine for Cardiovascular Disease Prediction, International Journal of Applied Engineering Research, Volume 13, Number 10 ,2018
8. AkshayJayrajSuvarna, Arvind Kumar M, Ajay Billav, Muthamma K M,GadugSudhamsu, Predicting The Presence of Heart Disease Using Machine Learning, International Journal of Computer Science and Mobile Computing, Vol. 8, Issue. 5, 119 – 125, May 2019
9. M. Tholkapiyan, A.Mohan, Vijayan.D.S, A survey of recent studies on chlorophyll variation in Indian coastal waters, IOP Conf. Series: Materials Science and Engineering 993 (2020) 012041, 1-6.
10. V.V. Ramalingam, AyantanDandapath, M Karthik Raja, Heart disease prediction using machine learning techniques: a survey, International Journal of Engineering & Technology, 7 (2.8), 684-687, 2018