International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 13 • Issue 1 pp. 14-23 Sept 2021 - March 2022 www.csjournals.com

Prediction of Cardiovascular Disease using Machine Learning Algorithms and Ensemble Techniques

Aalapi Bhave, Prof. Siddharth Gaikwad
Department of Computer Engineering and IT, College of Engineering, Pune, India

Abstract: Heart disease is a critical disease and a large number people are suffering from this disease all over the world. There are various factors that lead to patients suffering from heart disease and it is very important that heart disease is diagnosed as early as possible and the patient receives timely treatment for the same. This paper is about Predicting Cardiovascular disease using Machine Learning Algorithms and ensemble methods. The dataset for cardiovascular disease from Kaggle is utilized for this research which has 70000 patient records. This paper presents results by using five machine learning algorithms namely Decision Tree Algorithm, Random Forest, AdaBoost, Light Gradient Boosting Machine and XGBoost for the prediction of Cardiovascular disease. Further, Majority Voting ensemble technique by using soft voting and hard voting is used and the how the system performs is evaluated in both cases. The results showed that out of the five machine algorithms used, Light Gradient Boosting Machine is the best performing algorithm in terms of accuracy. Also, an improvement in accuracy was obtained by using the Majority Voting ensemble technique by using soft voting. The implementation of the project is in the Python programming language.
Keywords: AdaBoost, Cardiovascular disease, Decision Tree, Light Gradient Boosting Machine, Majority Voting Ensemble Technique, Random Forest, XGBoost.

Introduction

Cardiovascular disease is heart disease or the blood vessel disease. Cardiovascular diseases can be broadly classified into 6 types namely the coronary heart disease where arteries of the heart are unable to supply required oxygenated blood to the heart, cerebrovascular disease which affects the blood vessels supplying blood to the brain, peripheral arterial disease which results in blocking of the vessels that carrying blood from the heart to the legs as a result of which there is reduced blood flow to the limbs, rheumatic heart disease in which the heart valve is damaged due to rheumatic fever which is a response to a bacterial infection due to streptococcal bacteria, congenital heart disease which occurs due to defects at birth that affect the normal way the heart functions and DVT and pulmonary embolism in which clots of blood developed in the veins of the legs can move to lungs and result in blockage of the pulmonary arteries in the lungs. Strokes and Heart attacks are serious conditions. They are mainly caused by blocking of blood vessels due to fat deposits on the walls of arteries involved in supply of blood to the heart muscle or the brain. Strokes happen due to bleeding of a vessel in the brain or due to clots of blood. Some of the most important causes of heart disease are lifestyle related and include tobacco consumption, improper diet, physical inactivity and alcohol consumption. Cardiovascular diseases can also be caused by stress and hereditary factors. It is also necessary to treat patients suffering from hypertension, diabetes and high blood lipids to prevent heart disease from developing in people with these conditions. The risk of heart disease can be minimized by addressing the lifestyle factors. Heart attack can be recognized by symptoms such as pain in the chest, left shoulder, elbows, arms, jaw or back or breaking of cold sweat with pain or discomfort in the jaw, neck or back accompanied or shortness of breath with chest discomfort. Cardiovascular diseases are a premier reason of death globally as recorded by the WHO [1]. It is critical to diagnose CVD at an early stage so it can be treated effectively and damage can be reduced to a minimal.

Early detection of Cardio Vascular disease is vital in order to receive timely treatment for the same. We can predict the possibility of having Cardio Vascular disease using Machine Learning algorithms. Machine learning can process huge data quantities and identify trends which are not possible to identify for the human eye. Today, there is ever increasing amounts of data in healthcare and machine learning can be employed to process data accurately with efficiency. Machine learning algorithms can do their job without human intervention ensuring automated prediction and decision making.

There are 3 main categories of machine learning namely Supervised, Unsupervised and Reinforcement learning. Supervised learning works on labelled data. The model is trained on a labelled dataset which has both dependent and independent variables. Supervised learning is further categorized into classification and Regression. Classification is used when the output variable has defined labels or discrete values whereas Regression is used when the output variable has continuous values. Unsupervised learning works on the data that is not labelled. It detects patterns in the data. It is further classified into clustering which finds grouping in the data and association which is used to discover rules.

In order to train and test the machine learning model, in this project, the Cardio Vascular dataset available on Kaggle was used which has 70,000 patient records and 12 attributes which are considered as risk factors for heart disease or will be useful in classifying heart disease patients from normal patients [2]. This research paper aims to find the best approach in terms of accuracy for the various algorithms and ensemble methods implemented.

Literature Review

There have been a variety of machine learning approaches and techniques that have been used in the detection of heart diseases on various different datasets.

Wada Mohammed Jinjri et al have used this same cardiovascular disease dataset [3] but different machine learning models were implemented. The following algorithms have been used to train the model namely Naïve Bayes, K-NN, Logistic Regression, SVM and Decision Tree. SVM was the algorithm which gave the best accuracy with an accuracy of 72.6 %.

Dr Poonam Ghuli et al have used the dataset available in UCI machine learning repository for heart diseases [4]. The UCI Cleveland dataset was used to carry out the research. The paper compares DT, Logistic Regression, Naïve Bayes and Random Forest algorithms. Random Forest gave the highest accuracy of 90.16% which is higher than the other algorithms implemented. The values for Precision, Recall and F measure have also been calculated for all the algorithms and for Random Forest these values are 0.937, 0.882 and 0.909. The dataset used has 303 instances and 14 attributes. Compared to this dataset, the dataset used in the proposed work in this paper is much larger with 70000 instances compared to the traditional Cleveland dataset and hence a much realistic model can be trained and built.

Jialong Zhou has worked on Cardiovascular Disease Diagnosis Based on Stacking Technology [5]. The work has been done on the same dataset that has been used in this project. Using Pearson Correlation Coefficient, it has been determined that age, weight, cholesterol and cardio attributes are highly correlated. Stacking Technology has been used with SVM, Logistic Regression and BP Neural network. The stacked model has given an improvement in performance as compared to the individual algorithms. The maximum accuracy obtained using the stacked model is 72.4 %.

The research paper of Devansh Shah at al [6] presents supervised learning algorithms such as K-NN, Decision Tree, Naïve Bayes and Random Forest. It uses UCI Cleveland database of heart disease patients. KNN at a value of k=7 gave the highest accuracy for the training dataset. For the testing dataset, Naïve Bayes algorithm gave the highest accuracy of 88.15. In the future, the authors would like to extend their work by using more models. Compared to the Cleveland dataset of 303 instances, the dataset used in the proposed work in this paper is much larger with 70000 instances and hence a much realistic and robust model can be trained and built.

JaoujaMaiga1, Gilbert GutabagaHungilo and Pranowo have calculated Body Mass Index (BMI) as an attribute and added it to the features [7]. It has been derived from the weight and height attributes. 10-fold cross validation has been performed at the modeling stage. K-NN, Naïve Bayes, Random Forest and Logistic Regression have been used. From the study Random Forest achieved accuracy of 73% which is the highest.

Bárbara Martins et al worked-on Data Mining for Cardio Vascular Disease Prediction [8]. 5 classifiers were applied, namely Decision Tree, Rule Induction, Random Forest, Optimized Decision Tree and Deep Learning. Optimized Decision Tree was the best performing algorithm. It gave a value of 73.5%, for accuracy and highest values for precision, sensitivity, specificity, and AUC.

Jatin Gupta worked on SVM, Decision Tree, KNN and Random Forest in the research paper [9]. The results showed SVM to have the highest accuracy of 73%. SVM gave a better performance than the other models for the parameters of accuracy, AUC score and F1 score. The research used the same dataset that has been used in this project.

In the paper presented by R.JanePreethaPrincy et al [10], the results indicated that the Decision Tree predicted the CVDs better than Logistic Regression, Naive Bayes, SVM , KNN and Random Forest based models. Feature reduction was performed. Decision Tree gave the best result showing an accuracy value of 73%. The same cardiovascular disease dataset from Kaggle was used as the one used in this project. The future work involved using ensemble methods to build a better prediction model.

Jan Carlo T. Arroyo et al worked on An Optimized Neural Network Using Genetic Algorithm for Cardiovascular Disease Prediction [11]. In this study, the use of the genetic algorithm to optimize the parameters for an ANN in predicting cardiovascular disease was implemented. The use of GA had improved the performance of ANN by 5.08 percentage points against the prediction accuracy of the lone ANN. Further, the GA-ANN prediction model outperformed the other models based on ANN, Logistic

Regression, DT, Random Forest, SVM and K-NN giving an accuracy of 73.43%. The research used the same dataset that has been used in this project.

Experimental Setup

A. Dataset attribute information

This paper uses the Cardiovascular disease dataset from Kaggle. The dataset consists of 70000 patients' data with total of 12 features out of which 11 features are input features and one target (output) attribute which shows the if the person is suffering from cardiovascular disease or not. Table 1 contains dataset attributes along with their description. All the features have been taken at the time of medical examination.

Table 1: Dataset Attributes or Features

S.No	Attribute Name	Type of Feature	Description	Value
1	age	Objective	Age	int (days)
2	height	Objective	Height	int (cm)
3	weight	Objective	Weight	float (kg)
4	gender	Objective	Gender	categorical code
5	ap_hi	Examination	Systolic blood pressure	int
6	ap_lo	Examination	Diastolic blood pressure	int
7	cholesterol	Examination	Cholesterol level	1: normal, 2: above normal, 3: well above normal
8	gluc	Examination	Glucose	1: normal, 2: above normal, 3: well above normal
9	smoke	Subjective	Smoking	binary
10	alco	Subjective	Alcohol intake	binary
11	active	Subjective	Physical activity	binary
12	cardio	Target Variable	Presence or absence of cardiovascular disease	binary

For the purpose of data analysis correlation was calculated between the dataset attribute and the Target variable. Systolic BP, Diastolic BP and age were found to have maximum correlation with the target variable. Table 2 depicts the correlation between the dataset attribute and the target variable.

Table 2: Correlation between the dataset attribute and the target variable

Attribute Name	Correlation Value
age	0.241594
gender	0.005750
height	-0.016081
weight	0.178388
ap_hi(Systolic BP)	0.428310
ap_lo(Diastolic BP)	0.340261
cholesterol	0.221133
gluc	0.088872
smoke	-0.017171
alco	-0.009274
active	-0.037357

In order to find the correlation between each of the attributes, the correlation matrix is printed as shown in the figure below in Figure 1.

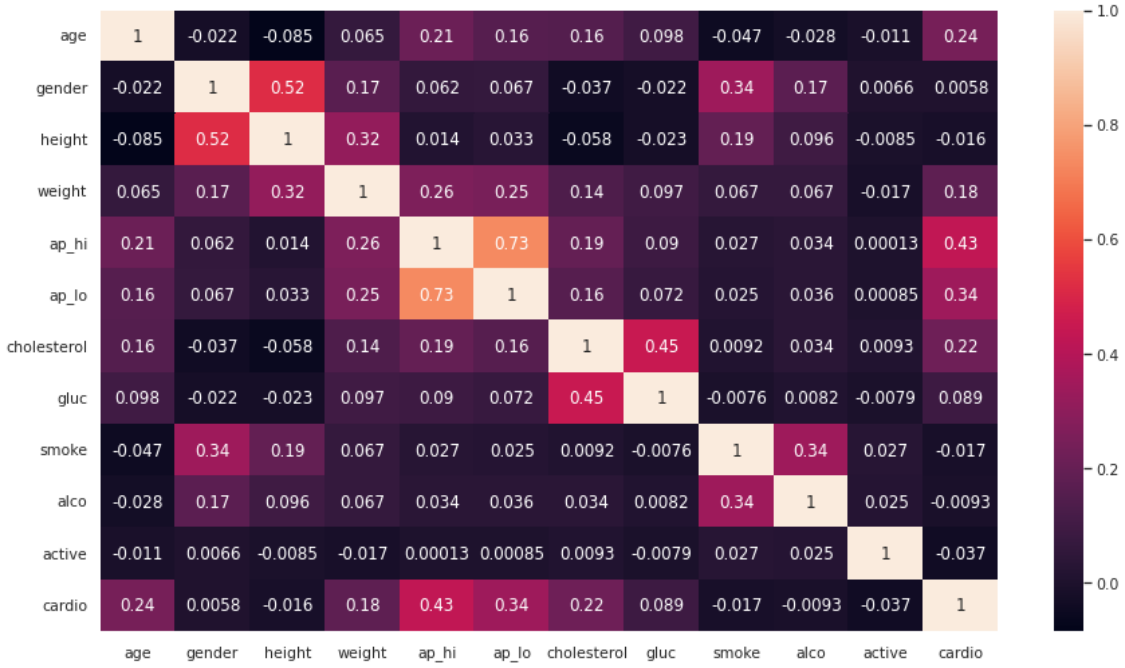

Figure 1. Cross correlation values for the attributes.

B. Data Pre-processing

The dataset contained outliers for some features. Age is in the range of 29 to 65 years which is in the proper range. For the height attribute, rows with values of height less than 125 cms or more than 210 cms have been removed. Some values for blood pressure also fall beyond what the range of blood pressure can be. For systolic blood pressure, the rows which contain values above 300 or below 60 have been removed. For diastolic blood pressure, the rows with values above 160 or less than 40 have been removed. The rows for which diastolic BP value is higher than systolic BP have also been removed as systolic BP should be higher than diastolic BP. Then, the dataset is checked for missing and null values. However, no missing or null values are found for the remaining records of the dataset. The dataset contains an attribute called id which contains a unique identifier for every patient. This attribute is not required for training the model and hence this attribute is dropped from the dataset. The entropy of the dataset and the information gain for each attribute with respect to the target attribute is also calculated. All the values for information gain are between 0 and 1 which shows that they are in the acceptable range. The values for information gain are as shown in Table 3.

Entropy of the dataset: 0.9998378109487849

Table 3: Information Gain values for every attribute with respect to target attribute

Attribute	Information Gain with respect to target variable
age	0.14888051025290716
gender	3.0508437168119684e-05
height	0.00128607503107514
weight	0.026153194956143877
Systolic BP	0.1691382080412337
Diastolic BP	0.1033766209562812
cholesterol	0.03672199662055775
gluc	0.00594919362374835
smoke	0.0002377456705777714
alco	7.708729087296806e-05
active	0.0009677414917852456

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 13 • Issue 1 pp. 14-23 Sept 2021 - March 2022 www.csjournals.com

Machine Learning Techniques

A. Decision Tree Classifier

Decision Tree classifier is a supervised learning model which is used in Classification as well as Regression problems. This model is based on a tree-like structure. Decision tree(DT) has a root node, branches and leaf nodes. In the decision tree algorithm, the criteria used was entropy. Decision trees are simple to understand, implement, visualize and implicitly perform feature selection. Decision trees may sometimes suffer from the problem of overfitting.

B. Random Forest Classifier

In case of continuous values in the dataset, Random Forest can be used for regression whereas in case of categorical values it can be used for classification. Random Forest consists of a number of DTs. Each decision tree produces its output and the final output is based on majority-based voting for classification and averaging method for regression. The criteria used for Random Forest is entropy whereas the number of estimators is 100. This combination gave the best result for Random Forest on this dataset. The working of random forest can be visualized from Figure 2.

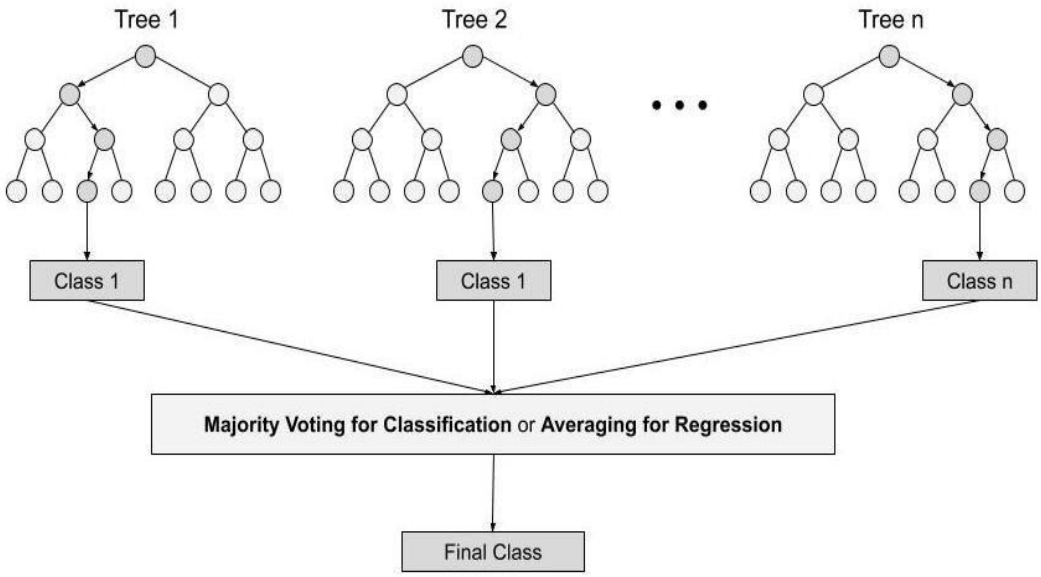

Figure 2. Random Forest.

C. AdaBoost

AdaBoost is boosting technique in machine learning which is also called as Adaptive Boosting. AdaBoost generally uses Decision tree algorithm with one level which is called as Decision Stumps.

This algorithm builds the model and gives equal weight to all the data points initially. The formula to calculate the sample weights is

$w(xi,yi) = 1/N$ where $i = 1,2,3,\ldots..n$.

Higher values of weight are assigned to points that are wrongly classified. Consequently, points with greater weights are given more importance in the next model. Models will be trained till a lower error is obtained. The decision trees that are constructed are executed in a sequential manner. The test data will pass through all the decision trees and which class has the majority depending on that the prediction will be made.

D. XGBoost

XGBoost represents the Extreme Gradient Boosting. It is designed to be highly efficient, portable and flexible and is an optimized distributed gradient boosting library. It uses the approach of parallel tree boosting and is often used to improve the accuracy and computing time.

E. LightGBM

LightGBM is developed by Microsoft. It has been applied on this dataset. LightGBM is known to improve the accuracy. It is based on two techniques namely the Gradient based one side sampling and exclusive feature bundling. LightGBM engages in leaf-wise splitting of the tree in contrast to other boosting algorithms that engage in growing the tree level wise. Without hurting accuracy, speed for training framework is improved.

F. Majority Voting Ensemble Technique

The majority voting ensemble technique is mostly employed for classification. Voting can be hard or soft. In this research we have used the Voting Classifier using AdaBoost, LightGBM and XGBoost as estimators. The results of both hard voting and soft voting are compared.

Methodology

The cardio vascular disease dataset consisting of 70000 instances and a total of 12 attributes goes through Data Pre-processing and Data Cleaning Phase as explained above in the paper. Then the entropy of the dataset and information gain for each attribute with reference to the target variable is calculated. The dataset is then split into training data and testing data and ML algorithms are applied. Then majority Voting ensemble is applied using hard voting as well as soft voting and the results are analyzed and compared.

Results and Analysis

Our main aim is to predict the accuracy for cardio vascular disease and which algorithm gives more accuracy so that in the future models can be used accordingly. The data set is split with a test size of 30% and training data is 70%. Google colab has been used for implementation and sklearn library, LightGBM and XGBoost python packages have been used in the implementation. Confusion Matrix for Decision Tree, Random Forest, AdaBoost, LightGBM and XGBoost are as shown in Figures 3,4,5,6,7 respectively.

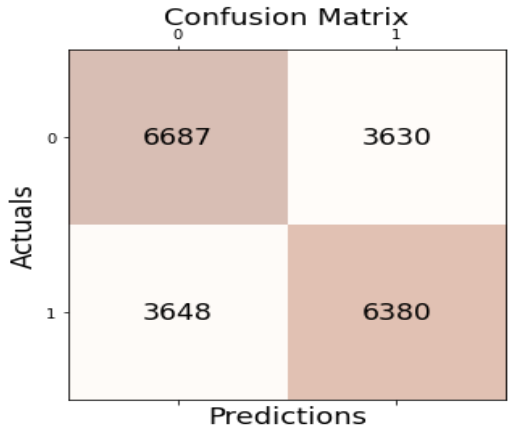

Figure 3. For Decision Tree.

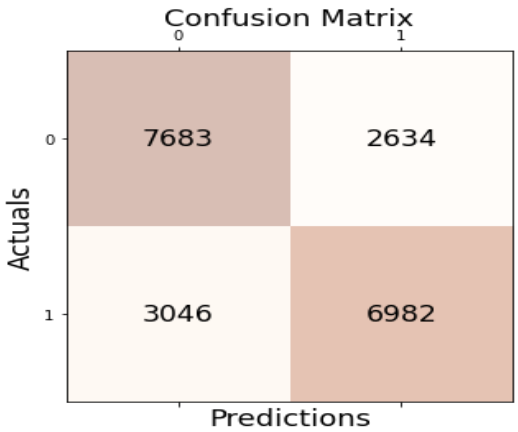

Figure 4. For Random Forest

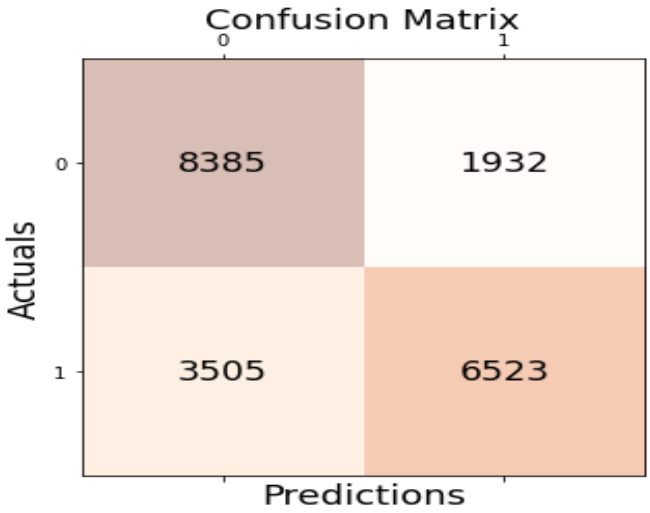

Figure 5. For AdaBoost

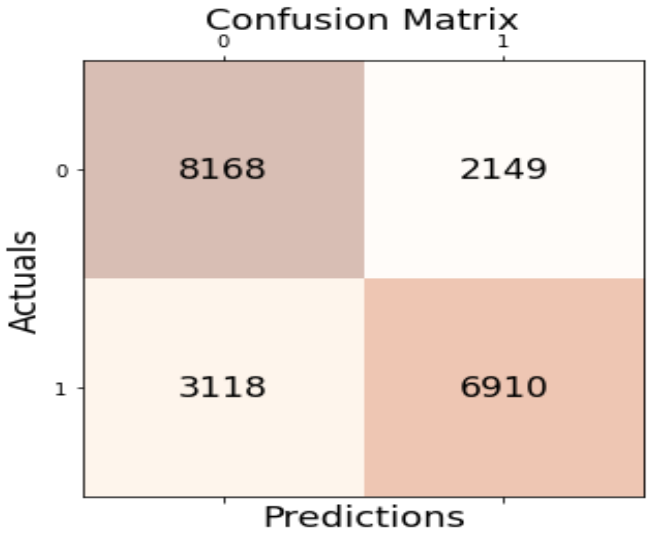

Figure 6. For LightGBM

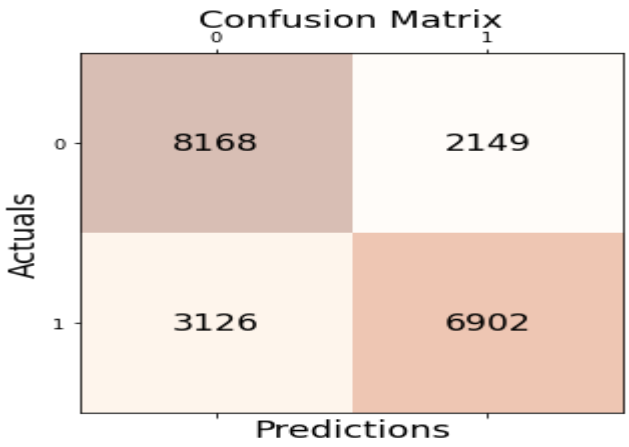

Figure 7 For XGBoost

The Accuracy represents the number of correctly classified data instances over the total number of data instances and the formula for accuracy is given by:

Accuracy = (TP+TN)/(TP+FP+FN+TN)

The Precision is positive predictive value while performing the classification of data instances and it can be calculated as:

Precision = (TP)/(TP+FP)

The Recall is a measure of the sensitivity or true positive rate and can be calculated as:

Recall = (TP)/(TP+FN)
where
TN is True Negative
TP is True Positive
FP is False Positive
FN is False Negative

The formula for F1 score is given by:

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

The computed values of Accuracy, Precision, F1 score and Recall for the 5 machine learning algorithms are given in Table 4:

Table 4: Accuracy, Recall, Precision, F1 score values

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Decision Tree	64.22	63.7	63.6	63.7
Random Forest	72.08	72.6	69.6	71.1
AdaBoost	73.28	77.1	65	70.6
LightGBM	74.11	76.3	68.9	72.4
XGBoost	74.07	76.3	68.8	72.4

The values of Accuracy of Majority Voting ensemble are as shown in Table 5.

Table 5: Accuracy of Majority Voting ensemble

	Algorithms used in ensemble	**Type of voting**	**Accuracy**
Majority Voting Ensemble	AdaBoost, LightGBM, XGBoost	Soft	74.17
Majority Voting Ensemble	AdaBoost, LightGBM, XGBoost	Hard	74.03

From the values in the Table 4, we can see that the accuracy of Random Forest is much higher than the Decision Tree. The Boosting algorithms give a higher accuracy than both Decision Tree and Random Forest. The LightGBM is the best performing algorithm in terms of accuracy with a value of 74.11%. This value of accuracy given by the LightGBM is higher than the accuracies reviewed in the literature survey section for this particular dataset. The second table gives us the values for Majority Voting Ensemble method. An ensemble of AdaBoost, LightGBM and XGBoost gives an accuracy of 74.17% by using soft voting which is higher than the accuracy values returned by all the three algorithms individually. By using hard voting, the same ensemble returns an accuracy of 74.03 which is higher than the accuracy of AdaBoost alone but lower than the values of LightGBM and XGBoost when run individually. Thus, we can see that Majority Voting ensemble by soft voting improves the accuracy over AdaBoost, LightGBM and XGBoost when run individually and when all algorithms are run individually, LightGBM returns the best value for accuracy.

Conclusion

In order to determine the best approach to predict cardiovascular disease, different machine learning algorithms and ensemble methods were implemented such as Decision Tree, Random Forest, AdaBoost, LightGBM and XGBoost. Majority voting ensemble technique using hard and soft voting was also implemented. The dataset used in this project is of 70000 samples which makes it a realistic dataset as compared to the smaller UCI Cleveland dataset used in the heart disease prediction which consists of 303 samples. From the values obtained, it can be seen that Random Forest gives a much higher accuracy than the Decision Tree. The Boosting algorithms give a higher accuracy than both Decision Tree and Random Forest. The LightGBM is the best performing algorithm in terms of accuracy with a value of 74.11%. This value of accuracy given by the LightGBM is higher than the accuracies reviewed in the literature survey section for this particular dataset. In our approach, we have introduced ensemble method using AdaBoost, LightGBM and XGBoost on this dataset. An improvement in accuracy is obtained by using the Majority Voting ensemble method using AdaBoost, LightGBM and XGBoost. By using hard voting, the ensemble returns an accuracy of 74.03% which is higher than the accuracy of AdaBoost alone but lower than the values of LightGBM and XGBoost when run individually. By using soft voting, the accuracy obtained is 74.17% which is higher than the accuracies given by AdaBoost, LightGBM and XGBoost when run individually. In the future we plan to further increase the accuracy by using optimization algorithms or techniques and further combination techniques.

References

References are important to the reader; therefore, each citation must be complete and correct. References must be from latest research work and should be arranged as followings.

[1] https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

[3] W. M. Jinjri, P. Keikhosrokiani and N. L. Abdullah, "Machine Learning Algorithms for The Classification of Cardiovascular Disease- A Comparative Study," 2021 International Conference on Information Technology (ICIT), 2021, pp. 132-138, doi: 10.1109/ICIT52682.2021.9491677.

[4] Apurb Rajdhan , Avi Agarwal , Milan Sai , Dundigalla Ravi, Dr. Poonam Ghuli, 2020, Heart Disease Prediction using Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 04 (April 2020).

[5] J. Zhou, "Cardiovascular Disease Diagnosis Based on Stacking Technology," 2021 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), 2021, pp. 706-710, doi: 10.1109/IPEC51340.2021.9421128.

[6] Shah, Devansh & Patel, Samir & Bharti, Drsantosh. (2020). Heart Disease Prediction using Machine Learning Techniques. SN Computer Science. 1. 10.1007/s42979-020-00365-y.

[7] J. Maiga, G. G. Hungilo and Pranowo, "Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using Health Record Data," 2019 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), 2019, pp. 45-48, doi: 10.1109/ICIMCIS48181.2019.8985205.

[8] Martins, B., Ferreira, D., Neto, C. et al. Data Mining for Cardiovascular Disease Prediction. J Med Syst 45, 6 (2021). https://doi.org/10.1007/s10916-020-01682-8

[9] J. Gupta, "The Accuracy of Supervised Machine Learning Algorithms in Predicting Cardiovascular Disease," 2021 International Conference on Artificial Intelligence and Computer Science Technology (ICAICST), 2021, pp. 234-239, doi: 10.1109/ICAICST53116.2021.9497837.

[10] Princy R, Jane Preetha & Parthasarathy, Saravanan & Jose, P. & Lakshminarayanan, Arun Raj & Jegananathan, Selvaprabu. (2020). Prediction of Cardiac Disease using Supervised Machine Learning Algorithms. 570-575. 10.1109/ICICCS48265.2020.9121169.

[11] Delima, Allemar Jhone & Arroyo, Jan Carlo. (2022). An Optimized Neural Network Using Genetic Algorithm for Cardiovascular Disease Prediction. 13. 95-99. 10.12720/jait.13.1.95-99.

[12] https://scikit-learn.org/stable/modules/tree.html

[13] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[14] Sharma, Vijeta & Yadav, Shrinkhala & Gupta, Manjari. (2020). Heart Disease Prediction using Machine Learning Techniques. 177-181. 10.1109/ICACCCN51052.2020.9362842.

[15] https://en.wikipedia.org/wiki/LightGBM

[16] https://lightgbm.readthedocs.io/en/latest

[17] https://xgboost.readthedocs.io/en/stable/index.html