VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

SET5001 - SCIENCE, ENGINEERING AND TECHNOLOGY PROJECT - II

TITLE

HEART DISEASE PREDICTION USING ML DEPLOYED ON AWS

Done by,

22MCA0137- KANISHK SHARMA

22MCA0174 - UDDASHAY KUMAR GUPTA

Under The Guidance Of

PROF.

Abstract

Machine Learning is used across many ranges around the world. The healthcare industry is no exclusion. Machine Learning can play an essential role in predicting presence/absence of locomotors disorders, Heart diseases and more. Such information, if predicted well in advance, can provide important intuitions to doctors who can then adapt their diagnosis and dealing per patient basis. We work on predicting possible Heart Diseases in people using Machine Learning algorithms. In this project we perform the comparative analysis of classifiers like decision tree, Naïve Bayes, Logistic Regression, SVM and Random Forest and we propose an ensemble classifier which perform hybrid classification by taking strong and weak classifiers since it can have multiple number of samples for training and validating the data so we perform the analysis of existing classifier and proposed classifier like Ada-boost and XG-boost which can give the better accuracy and predictive analysis.

Keywords: SVM; Naive Bayes; Decision Tree; Random Forest; Logistic Regression; Adaboost; XG-boost; python programming; confusion matrix; correlation matrix

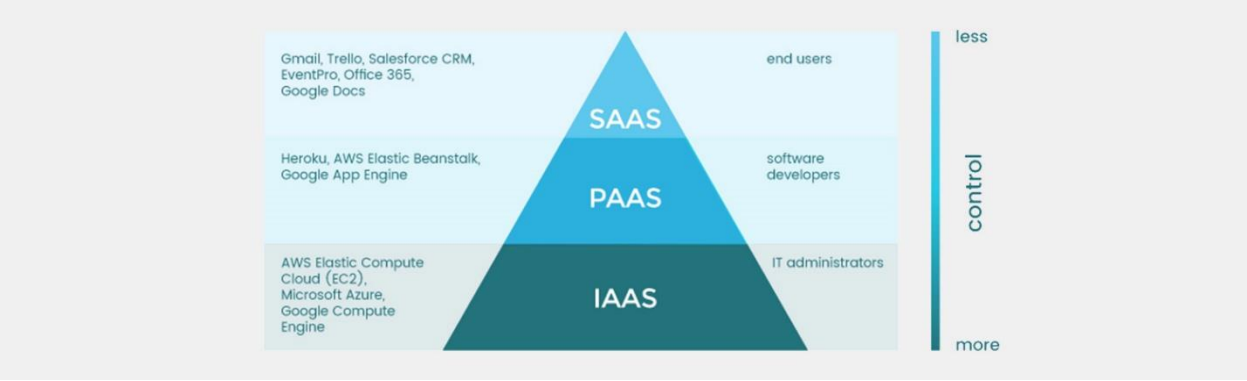
Introduction

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of data analysis. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduces the complications. Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using machine-learning algorithm. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

Cloud Computing

Cloud computing is the delivery of on-demand computing services over the internet. These services can include servers, storage, databases, networking, software, and more. Cloud computing offers many benefits, including scalability, flexibility, cost savings, and increased security.

There are three main types of cloud computing: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS). IaaS provides the basic building blocks of computing infrastructure, such as virtual machines and storage, while PaaS provides a platform for developers to build, deploy, and manage applications. SaaS provides complete software applications that are delivered over the internet, such as email and office productivity software.



Cloud computing is used by individuals and organizations of all sizes, from small startups to large enterprises. It allows users to access computing resources on an as-needed basis, without having to invest in expensive hardware or manage complex infrastructure. Cloud computing providers, such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform, offer a wide range of services that can be customized to meet the specific needs of each user.

AWS and its Services

Amazon Web Services (AWS) is a cloud computing platform that provides a wide range of services for building and running applications in the cloud. AWS was launched in 2006 and has since become one of the most widely used cloud computing platforms in the world, with millions of customers in over 190 countries.

Some of the key services offered by AWS include:

1) Compute Services: AWS provides a variety of compute services, including Amazon Elastic Compute Cloud (EC2), which provides scalable compute capacity in the cloud, and AWS Lambda, which allows users to run code in response to events without provisioning or managing servers.

2. Storage Services: AWS provides a variety of storage services, including Amazon Simple Storage Service (S3), which provides scalable object storage for data, and Amazon Elastic Block Store (EBS), which provides persistent block storage for EC2 instances.

3. Database Services: AWS provides a variety of database services, including Amazon Relational Database Service (RDS), which provides managed database services for MySQL, PostgreSQL, Oracle, and SQL Server, and Amazon DynamoDB, which provides a fast and flexible NoSQL database service.

4. Networking Services: AWS provides a variety of networking services, including Amazon Virtual Private Cloud (VPC), which allows users to provision a private, isolated section of the AWS cloud, and AWS Direct Connect, which provides a dedicated network connection between an on-premises data center and the AWS cloud.

5. Machine Learning Services: AWS provides a variety of machine learning services, including Amazon SageMaker, which provides a fully-managed platform for building, training, and deploying machine learning models, and Amazon Rekognition, which provides computer vision services for image and video analysis.

6. Security and Identity Services: AWS provides a variety of security and identity services, including AWS Identity and Access Management (IAM), which allows users to manage access to AWS services and resources, and AWS Key Management Service (KMS), which provides a secure and scalable key management service.

These are just a few of the many services offered by AWS. AWS also provides a wide range of tools and services for managing and deploying applications in the cloud, including AWS CloudFormation, AWS Elastic Beanstalk, and AWS CodeDeploy.

Literature Survey

With growing development in the field of medical science alongside machine learning various experiments and researches has been carried out in these recent years releasing the relevant significant papers.

[1] Purushottam ,et ,al proposed a paper "Efficient Heart Disease Prediction System" using hill climbing and decision tree algorithms .They used Cleveland dataset and preprocessing of data is performed before using classification algorithms. The Knowledge Extraction is done based on Evolutionary Learning (KEEL), an open- source data mining tool that fills the missing values in the data set.A decision tree follows top-down order. For each actual node selected by hill-climbing algorithm a node is selected by a test at each level. The parameters and their values used are confidence. Its minimum confidence value is 0.25. The accuracy of the system is about 86.7%.

[2] Santhana Krishnan. J ,et ,al proposed a paper "Prediction of Heart Disease Using Machine Learning Algorithms" using decision tree and Naive Bayes algorithm for prediction of heart disease. In decision tree algorithm the tree is built using certain conditions which gives True or False decisions. The algorithms like SVM, KNN are results based on vertical or horizontal split conditions depends on dependent variables.

But decision tree for a tree like structure having root node, leaves and branches base on the decision made in each of tree Decision tree also help in the understating the importance of the attributes in the dataset. They have also used Cleveland data set. Dataset splits in 70% training and 30% testing by using some methods. This algorithm gives 91% accuracy. The second algorithm is Naive Bayes, which is used for classification. It can handle complicated, nonlinear, dependent data so it is found suitable for heart disease dataset as this dataset is also complicated, dependent and nonlinear in nature. This algorithm gives an 87% accuracy.

[3] Sonam Nikhar et al proposed paper " Prediction of Heart Disease Using Machine Learning Algorithms" their research gives point to point explanation of Naïve Bayes and decision tree classifier that are used especially in the prediction of Heart Disease. Some analysis has been led to think about the execution of prescient data mining strategy on the same dataset, and the result decided that Decision Tree has highest accuracy than Bayesian classifier.

[4] Aditi Gavhane et al proposed a paper "Prediction of Heart Disease Using Machine Learning", in which training and testing of dataset is performed by using neural network algorithm multi-layer perceptron. In this algorithm there will be one input layer and one output layer and one or more layers are hidden layers between these two input and output layers. Through hidden layers each input node is connected to output layer. This connection is assigned with some random weights. The other input is called bias which is assigned with weight based on requirement the connection between the nodes can be feedforwarded or feedback.

[5] Avinash Golande et al, proposed "Heart Disease Prediction Using Effective Machine Learning Techniques" in which few data mining techniques are used that support the doctors to differentiate the heart disease. Usually utilized methodologies are k-nearest neighbour, Decision tree and Naïve Bayes. Other unique characterization-based strategies utilized are packing

calculation, Part thickness, consecutive negligible streamlining and neural systems, straight Kernel selfarranging guide and SVM (Bolster Vector Machine).
[6] Lakshmana Rao et al,proposed “Machine Learning Techniques for Heart Disease Prediction” in which the contributing elements for heart disease are more. So, it is difficult to distinguish heart disease.To find the seriousness of the heart disease among people different neural systems and data mining techniques are used.
[7] Abhay Kishore et alproposed “Heart Attack Prediction Using Deep Learning” in which heart attack prediction system by using Deep learning techniques and to predict the probable aspects of heart related infections of the patient Recurrent Neural System is used. This model uses deep learning and data mining to give the best precise model and least blunders. This paper acts as strong reference model for another type of heart attack prediction models
[8] Senthil Kumar Mohan et al, proposed “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques” in which their main objective is to improve exactness in cardiovascular problems. The algorithms used are KNN, LR, SVM, NN to produce an improved exhibition level with a precision level of 88.7% through the prediction model for heart disease with hybrid random forest with linear model (HRFLM).
[9] Anjan N. Repaka et al, proposed a model stated the performance of prediction for two classification models, which is analyzed and compared to previous work. The experimental results show that accuracy is improved in finding the percentage of risk prediction of our proposed method in comparison with other models.
[10] Aakash Chauhan et al, proposed “Heart Disease Prediction using Evolutionary Rule Learning”. Data is directly retrieved from electronic records that reduce the manual tasks. The number of services are decreased and shown major number of rules helps within the best prediction of heart disease. Frequent pattern growth association mining is performed on patient’s dataset to generate strong association.

Proposed System

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is preprocessed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. The accuracy of the system is obtained by testing the system using the testing data. This system is implemented using the following modules.

1.) Collection of Dataset
2.) Selection of attributes
3.) Data Pre-Processing
4.) Balancing of Data
5.) Deployment of the model and testing Disease Prediction

1.) Initially, we collect a dataset for our heart disease prediction system. After the collection of the dataset, we split the dataset into training data and testing data. The training dataset is used for prediction model learning and testing data is used for evaluating the prediction model. For this project, 70% of training data is used and 30% of data is used for testing. The dataset used for this project is Heart Disease UCI. The dataset consists of 76 attributes; out of which, 14 attributes are used for the system.

Dataset details:

- Of the 76 attributes available in the dataset,14 attributes are considered for the prediction of the output.
- Heart Disease UCI : https://archive.ics.uci.edu/ml/datasets/Heart+Disease

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target	
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1	
3	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1	
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1	
5	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1	
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1	
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1	
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1	
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1	
10	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1	
11	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1	
12	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1	
13	48	0	2	130	275	0	1	139	0	0.2	2	0	2	1	
14	49	1	1	130	266	0	1	171	0	0.6	2	0	2	1	
15	64	1	3	110	211	0	0	144	1	1.8	1	0	2	1	
16	58	0	3	150	283	1	0	162	0	1	2	0	2	1	
17	50	0	2	120	219	0	1	158	0	1.6	1	0	2	1	
18	58	0	2	120	340	0	1	172	0	0	2	0	2	1	
19	66	0	3	150	226	0	1	114	0	2.6	0	0	2	1	
20	43	1	0	150	247	0	1	171	0	1.5	2	0	2	1	
21	69	0	3	140	239	0	1	151	0	1.8	2	2	2	1	
22	59	1	0	135	234	0	1	161	0	0.5	1	0	3	1	
23	44	1	2	130	233	0	1	179	1	0.4	2	0	2	1	
24	42	1	0	140	226	0	1	178	0	0	2	0	2	1	
25	61	1	2	150	243	1	1	137	1	1	1	0	2	1	
26	40	1	3	140	199	0	1	178	1	1.4	2	0	3	1	
27	71	0	1	160	302	0	1	162	0	0.4	2	2	2	1	
28	59	1	2	150	212	1	1	157	0	1.6	2	0	2	1	

Figure: Dataset Attributes

2). Selection of attributes

Attribute or Feature selection includes the selection of appropriate attributes for the prediction system. This is used to increase the efficiency of the system. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are selected for the prediction. The Correlation matrix is used for attribute selection for this model.

3.) Pre-processing of Data

Data pre-processing is an important step for the creation of a machine learning model. Initially, data may not be clean or in the required format for the model which can cause misleading outcomes. In pre-processing of data, we transform data into our required format. It is used to deal with noises, duplicates, and missing values of the dataset. Data pre-processing has the activities like importing datasets, splitting datasets, attribute scaling, etc. Preprocessing of data is required for improving the accuracy of the model.

4.) Prediction of Disease

Various machine learning algorithms like SVM, Naive Bayes, Decision Tree, Random Tree, Logistic Regression, Ada-boost, Xg-boost are used for classification. Comparative analysis is performed among algorithms and the algorithm that gives the highest accuracy is used for heart disease prediction.

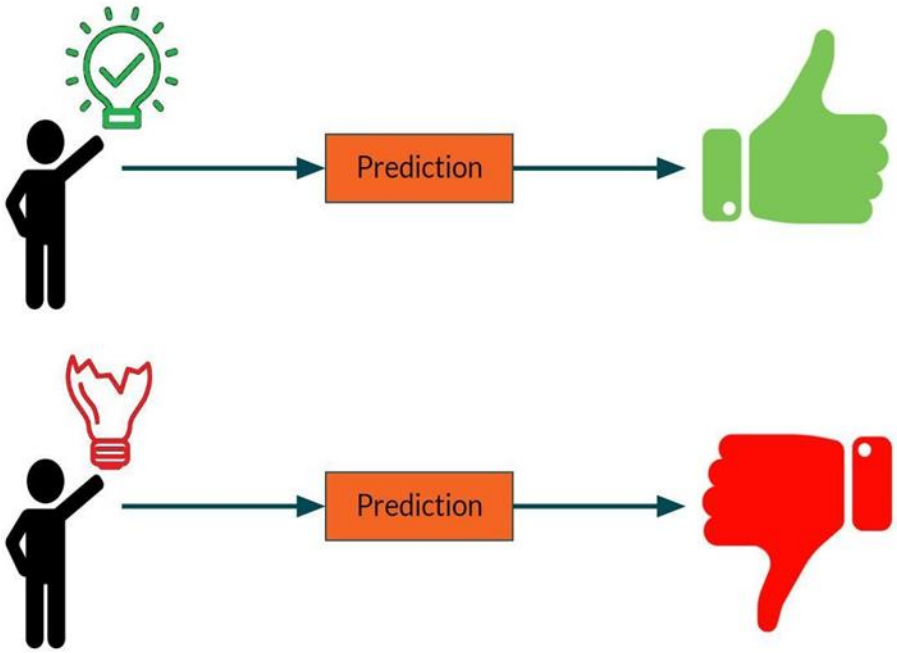


Figure: Prediction of Disease

5.) Deployment Of Machine Learning Model in AWS Using Amazon SageMaker

Building, training, and deploying machine learning models can all be done with one integrated set of tools thanks to Amazon SageMaker, a fully managed service. The procedures for deploying an ML model using Amazon SageMaker are as follows: We Develop our model and make sure our machine learning model is performance-optimized by training and testing it on a dataset. Create an inference script to specify the inputs and outputs of our model and package it using a framework like TensorFlow, Scikit-Learn, or MXNet. Then we upload it to Amazon S3: Put our packed model in a bucket on Amazon S3.Build a SageMaker model on the Amazon SageMaker console by indicating the location of our model data in Amazon S3.Endpoint configuration creation: Provide the resources to be used for hosting the model in an endpoint configuration, including the kind and number of instances. Launch the model: Create an Amazon SageMaker endpoint with the endpoint settings we created in s to deploy the model. Then we Analyzed the model: Once the endpoint has been constructed, we may use the SageMaker runtime API to make predictions to test the model .Machine learning model deployment and management are made simple with a number of capabilities offered by Amazon SageMaker. To optimize our model's hyperparameters, utilize SageMaker's automatic model tuning function. SageMaker's automatic scaling feature lets us automatically change the number of instances needed to serve predictions based on incoming traffic.

Diagram for the deployment

Fig. FlowChart to deploy the project using AWS Sagemaker

This diagram shows how we prepare and package our machine learning model before uploading it to Amazon S3. The next step is to build a SageMaker model that references the Amazon S3 location of the model data. After building the model, we specify the resources we'll use to host it, including the type of instance and the number of instances, in an endpoint configuration. With the endpoint settings, we create an endpoint to distribute the model. Once the endpoint has been constructed, test our model by making predictions using the SageMaker runtime API.

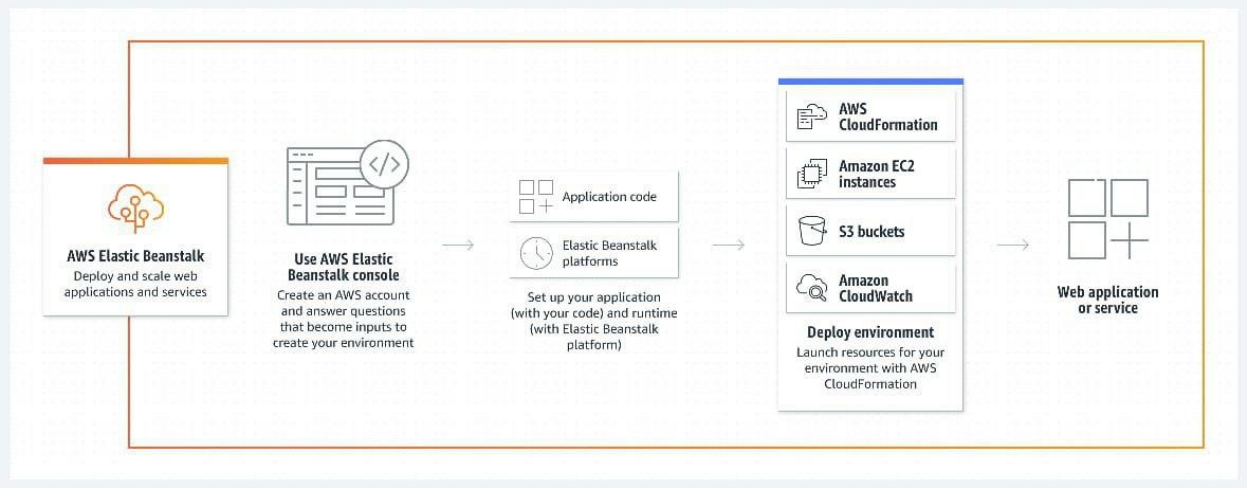


Fig. Features and advantages of AWS Elastic Beanstalk

EXPERIMENTAL ANALYSIS

SYSTEM CONFIGURATION

Hardware requirements:

Processer : Any Update Processer

Ram : Min 4GB

Hard Disk : Min 100GB

Software requirements:

Operating System : Windows family

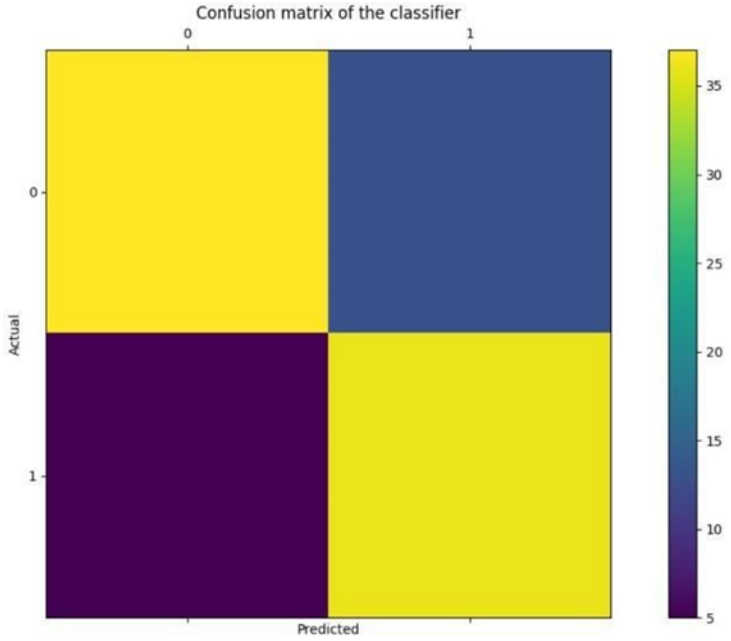
Technology : Python3.7

IDE : Jupiter notebook

PERFORMANCE ANALYSIS

In this project, various machine learning algorithms like SVM, Naive Bayes, Decision Tree, Random Forest, Logistic Regression, Adaboost, XG-boost are used to predict heart disease. Heart Disease UCI dataset, has a total of 76 attributes, out of those only 14 attributes are considered for the prediction of heart disease. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are considered for this project. The accuracy for

individual algorithms has to measure and whichever algorithm is giving the best accuracy, that is considered for the heart disease prediction. For evaluating the experiment, various evaluation metrics like accuracy, confusion matrix, precision, recall, and f1-score are considered. Accuracy- Accuracy is the ratio of the number of correct predictions to the total number of inputs in the dataset. It is expressed as: Accuracy = (TP + TN) /(TP+FP+FN+TN) Confusion Matrix- It gives us a matrix as output and gives the total performance of the system.



Where TP: True positive FP: False Positive FN: False Negative TN: True Negative 59 Correlation Matrix: The correlation matrix in machine learning is used for feature selection. It represents dependency between various attributes.

Precision- It is the ratio of correct positive results to the total number of positive results predicted by the system. It is expressed as: Recall-It is the ratio of correct positive results to the total number of positive results predicted by the system. It is expressed as: F1 Score-It is the harmonic mean of Precision and Recall. It measures the test accuracy. The range of this metric is 0 to 1.

PERFORMANCE MEASURES

The highest accuracy is given by XG-boost

```
Accuracy of svm: 0.8021978021978022
Accuracy of naive bayes: 0.7692307692307693
Accuracy of logistic regression: 0.7912087912087912
Accuracy of decision tree: 0.7582417582417582
Accuracy of random forest: 0.7912087912087912

Majority Voting accuracy score:  0.7912087912087912
Weighted Average accuracy score:  0.8131868131868132
Bagging_accuracy score:  0.8021978021978022
Ada_boost_accuracy score:  0.7362637362637363
Gradient_boosting_accuracy score:  0.8131868131868132
```

RESULT

After performing the machine learning approach for training and testing we find that accuracy of the XG-boost is better compared to other algorithms. Accuracy is calculated with the support of the confusion matrix of each algorithm, here the number count of TP, TN, FP, FN is given and using the equation of accuracy, value has been calculated and it is concluded that extreme gradient boosting is best with 81% accuracy and the comparison is shown below. TABLE: Accuracy comparison of algorithms Algorithm Accuracy

Algorithm	Accuracy
XG-boost	81.3%
SVM	80.2%
Logistic Regression	79.1%
Random Forest	79.1%
Naive Bayes	76.9%
Decision Tree	75.8%
Adaboost	73.6%

TABLE 1: Accuracy Table

CONCLUSION AND FUTURE WORK

Heart diseases are a major killer in India and throughout the world, application of promising technology like machine learning to the initial prediction of heart diseases will have a profound impact on society. The early prognosis of heart disease can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. The number of people facing heart diseases is on a raise each year. This prompts for its early diagnosis and treatment. The utilization of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity and patients. In this paper, the seven different machine learning algorithms used to measure the performance are SVM, Decision Tree, Random Forest, Naïve Bayes, Logistic Regression, Adaptive Boosting, and Extreme Gradient Boosting applied on the dataset.

The expected attributes leading to heart disease in patients are available in the dataset which contains 76 features and 14 important features that are useful to evaluate the system are selected among them. If all the features taken into the consideration, then the efficiency of the system the author gets is less. To increase efficiency, attribute selection is done. In this n features have to be selected for evaluating the model which gives more accuracy. The correlation of some features in the dataset is almost equal and so they are removed. If all the attributes present in the dataset are taken into account, then the efficiency decreases considerably.

All the seven machine learning methods accuracies are compared based on which one prediction model is generated. Hence, the aim is to use various evaluation metrics like confusion matrix, accuracy, precision, recall, and f1-score which predicts the disease efficiently. Comparing all seven the extreme gradient boosting classifier gives the highest accuracy of 81%.

REFERENCES

[1] Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-8

[2] Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8.

[3] Ordonez C (2006). Association rule discovery with the train and test approach for heart disease prediction. IEEE Transactions on Information Technology in Biomedicine, 10(2), 334-43.

[4] Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9.

[5] Bashir S, Qamar U & Javed M Y (2014, November). An ensemble-based decision support framework for intelligent heart disease diagnosis. In International Conference on Information Society (i-Society 2014) (pp. 259-64). IEEE. ICCRDA 2020 IOP Conf. Series: Materials Science and Engineering 1022 (2021) 012072 IOP Publishing doi:10.1088/1757-899X/1022/1/012072 9

[6] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2014). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e005025.

[7] Ganna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2013). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33(9), 2267-72.

[8] Jabbar M A, Deekshatulu B L & Chandra P (2013, March). Heart disease prediction using lazy associative classification. In 2013 International Mutli- Conference on Automation, Computing,Communication, Control and Compressed Sensing (iMac4s) (pp. 40- 6). IEEE.

[9] Brown N, Young T, Gray D, Skene A M & Hampton J R (1997). Inpatient deaths from acute myocardial infarction, 1982-92: analysis of data in the Nottingham heart attack register. BMJ, 315(7101), 159-64.

[10] Folsom A R, Prineas R J, Kaye S A & Soler J T (1989). Body fat distribution and self-reported prevalence of hypertension, heart attack, and other heart disease in older women. International journal of epidemiologyy, 18(2), 361-7.