International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 07 | July - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Predicting Heart Disease Using Machine Learning

Prajath H N[1], Prof Swetha C S[2]

[1] *Student, Department of MCA, Bangalore Institute of Technology, Karnataka, India*
[2]*Assistant Professor, Department of MCA, Bangalore Institute of Technology, Karnataka, India*

Abstract—According to a recent WHO research, there is an increase in heart-related disorders, which account for 17.9 million fatalities per year. As the population expands, recognizing and treating these disorders becomes more difficult. However, technological improvements have allowed Machine Learning (ML) techniques to have a substantial impact on the healthcare sector. The goal of this study is to develop an ML model to predict cardiac illness based on a Kaggle benchmark dataset that includes 14 different metrics linked with the condition. We used a variety of machine learning approaches, including Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Logistic Regression, and decision trees. Our investigation also looked at the correlations between the dataset's distinct attributes. The findings show that K-Nearest Neighbors outperformed the other machine learning algorithms in terms of prediction accuracy and efficiency.
The medical professionals at their clinic may find this model useful as a decision support-system.

I. INTRODUCTION

The heart is responsible for regulating blood flow through the veins. Blood travels throughout the body, supplying critical nutrients, oxygen, metals, and other vital chemicals, making it an important part of the circulatory system. Heart failure can result in serious health issues and even death. Unhealthy lifestyle choices, such as smoking, drinking alcohol, and eating high-fat meals, can raise the risk of heart disease. The World
Machine learning (ML) is advancing rapidly in the healthcare sector, assisting in disease diagnosis, drug discovery, and image classification. It provides substantial benefits to hospital administration, medical professionals, and treatment facilities. Modern medical tools are crucial for the early detection and prediction of heart disease. ML algorithms are key in developing new models that enable early diagnosis and treatment by analyzing data and revealing hidden patterns. This study employed various ML techniques, including logistic regression, k-nearest neighbors, support vector machines, decision trees, random forests, and extreme gradient boosting, to assess model performance on two heart disease datasets. Grid search cross-validation was used to optimize training and testing performance and to find the best parameters for heart disease prediction. While hyperparameter tuning can effectively predict patient outcomes and heart disease in larger, comprehensive datasets, it is less effective with smaller datasets. Using training and testing statistical data enhances the performance and accuracy of ML algorithms in predicting heart disease.

II. LITERATURE SURVEY

In order to diagnose and ascertain whether a person has heart disease, this study offers a prediction model. The study evaluates the accuracy of several machine learning methods, such as SVM, Random forest, Naive Bayes classifier, and Logistic regression, applied to a dataset gathered from a particular location in order to develop an accurate model for predicting cardiovascular sickness.

The paper "Machine Learning Classification Techniques for Heart Disease Prediction," authored by Maryam I. AI-Janabi, Mahmoud H. Qutqut, and Mohammad Hijjawi in 2018, explores various techniques such as Naïve Bayes, ANN, DT, KNN, SVM, and Hybrid Approach. The study compares K-Nearest Neighbor (KNN) with other data mining classification algorithms to achieve higher accuracy for predicting heart disease .

"An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection" was created by Ashir Javed, Shjie Zhou, et al. in 2017. In this work, the random forst model is used to diagnose cardiovascular disease, and the random-search algorithm (RSA) is used to identify factors. This model is primarily optimized to be used with an algorithmic grid search tool. There are 2 types of experiments used to predict cardiovascular disease. The first experiment only develops a RF model; the second experiment develops a random forest model based on Random-Search Algorithm. Compare to the traditional random forest model, our methodology is less complicated and more efficient.[8]

Senthilkumar Mohan, Chandrasegar Thirumalai, presented "Effective Heart Disease Prediction Using Hybrid ML Techniques," which was an effective hybrid machine learning methodological technique. The hybrid approach combines the linear method with random forest. For the reason of prediction, the dataset and attribute subsets were gathered. The pre-processed knowledge set of some qualities. [9].

An algorithmic accuracy-focused technique was developed by Archana Singh and Rakesh Kumar. The authors took it into consideration as a parameter for analyz the algorithmic performance. The dataset that is utilized for testing and training determines how accurate the ML algorithms are.

They discovered that KNN was the good algorithm after analyzing them using a dataset that included factors like age, cholesterol, kind of chest discomfort, resting, and more. They employ the following methodologies: Knearest Neighbor, SVM, Decision Tree, and Linear Regression. One disadvantage is that it takes a lot of data sets because it employs a lot of attributes.[10]

III. DATASET DESCRIPTION

There are 10908 individual data in the dataset. The dataset consists of 14 columns, each of which is explained as follows:
1. Age: indicates the person's age.
2. Sex: Uses the following format to indicate the person's gender:
1 = masculine
0 indicates a female
3. Type of chest pain: This shows the person's type of chest pain in the following format:
1 = normal angina 2 = non-typical angina
Non-anginal pain (3 =). 4 is an asymptotic
4. Resting Blood Pressure: This shows a person's resting blood pressure reading in millimeter-hours (mmHg).
5. Serum Cholesterol: This shows the serum cholesterol as milligrams per deciliter.
6. Fasting Blood Sugar: 120 mg/dl is compared to an individual's fasting blood sugar measurement.
If the blood sugar level when fasting is greater than 120 mg/dl, then: 1 (true) else: 0 (not true)

1. Restecg: shows electrocardiographic data at rest. 0 indicates normal.
1 = possessing aberrant ST-T waves Left ventricular hypertrophy is equals to 2.
2. Max heart rate attained: shows the highest heart rate that a person has attained. 9. Angina induced by exercise:

1 = indeed
0 indicates no
10. Exercise-induced ST depression in reltion to rest: shows the value, can be a float or an integer.
11. Peak workout ST section: 1 = inclining upward
2 is flat.
3 = sloping downward
12. Fluoroscopy-colored number of main vessels (0–3); shows value as an integer or float.
13. Thal: exhibits thalassemia:

3 is typical.
6 = corrected flaw
7 is a reversible error

10. Heart disease diagnosis: Indicates whether or not the patient has heart disease:
0 indicates no illness
1 = Hypercholesterolemia 2 = Cardiovascular illness 3 = Cardiovascular illness
4 = Syncope

```
df = pd.read_csv('HDP.csv',sep=',',encoding="utf-8")
df.head(15)
```
0.0s

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
1	67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
2	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
3	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
4	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
5	56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
6	62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
7	57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
8	63	1	4	130	254	0	2	147	0	1.4	2	1	7	2
9	53	1	4	140	203	1	2	155	1	3.1	3	0	7	1
10	57	1	4	140	192	0	0	148	0	0.4	2	0	6	0
11	56	0	2	140	294	0	2	153	0	1.3	2	0	3	0
12	56	1	3	130	256	1	2	142	1	0.6	2	1	6	2
13	44	1	2	120	263	0	0	173	0	0.0	1	0	7	0
14	52	1	3	172	199	1	0	162	0	0.5	1	0	7	0

Figure 1. Dataset

Let's examine the ageof range those who have the illness and those who do not. Here, num = 1, 2, 3, and 4 indicate patient have heart disease related to coronary artery, peripheral artery, carotid artery, and arrhythmia, respectively, and num = 0 indicates the patient is has no heart illness at all.

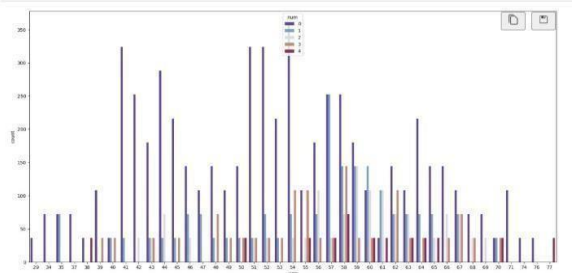


Figure 2. Values count in age feature

Let's now examine the age and gender distribution within each num class.

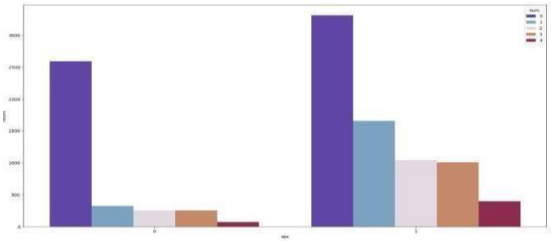


Figure 3. Distribution of the data in sex parameter

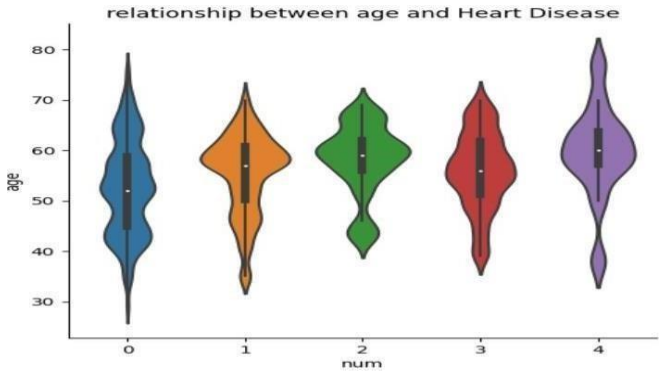
```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
num         0
dtype: int64
```

Figure 4. Checking null values in data

Since there are no null values in dataset, we can use it directly to describe the building process..

Let's divide the data into the train and test sets now. I divide the data for the project is into an 80:20 ratio. In other words, 80% of the data is used for training, while 20% is used for testing.

Density curves are used in violin plot to show the distributions of numerical data for one or more groups. The approximate frequency of data points in each region is reflected in the width of each curve. Densities are sometimes accompanied by a of superimposed chart, like a box plot, to give more details. As seen in Figure 6, the violin plots are generated for age and num, with age
on the y-axis and num on thex-axis.



The graphs that show the association between two variables in the data set are called scatter plots. It displays data points either as a Cartesian system or as a two-dimensional plane. Plotting the dependent variable on the Y-axis corresponds to the independent variable, or attribute, on the X-axis. As seen in figure. 40, this pattern represents age and heart- disease kinds.

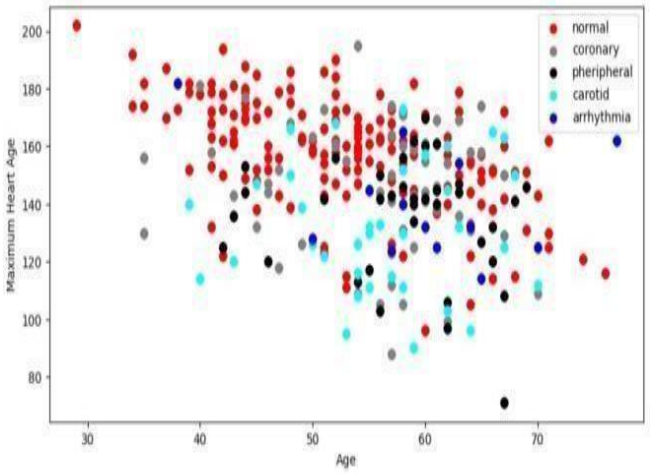
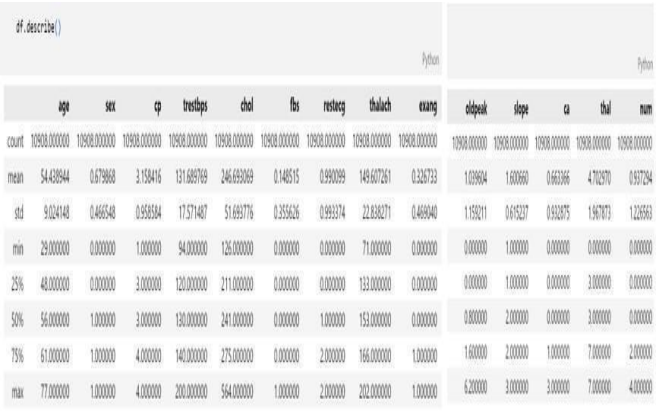


Figure 6. Scatter plot



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
count	10908.000000	10908.000000	10908.000000	10908.000000	10908.000000	10908.000000	10908.000000	10908.000000	10908.000000	10908.000000	10908.000000	10908.000000	10908.000000	10908.000000
mean	54.438944	0.679868	3.158416	131.689769	246.693069	0.148515	0.990099	149.607261	0.326733	1.039604	1.600660	0.663366	4.712870	0.937294
std	9.024748	0.466548	0.960564	17.571487	51.693776	0.355626	0.993374	22.838271	0.469040	1.159211	0.615237	0.932875	1.967873	1.226563
min	29.000000	0.000000	1.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000
25%	48.000000	0.000000	3.000000	120.000000	211.000000	0.000000	0.000000	132.000000	0.000000	0.000000	1.000000	0.000000	3.000000	0.000000
50%	56.000000	1.000000	3.000000	130.000000	241.000000	0.000000	1.000000	153.000000	0.000000	0.800000	2.000000	0.000000	3.000000	0.000000
75%	61.000000	1.000000	4.000000	140.000000	275.000000	0.000000	2.000000	166.000000	1.000000	1.600000	2.000000	1.000000	7.000000	2.000000
max	77.000000	1.000000	4.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	3.000000	3.000000	7.000000	4.000000

Figure 7. Described the data

IV. MATERIALS AND METHODS

Classification, a supervised learning process, predicts outcomes using existing data. This research proposes a method to diagnose heart disease based on classification algorithms. The train dataset is divided into a training set and a test set in an 80:20 ratio, and individual classifiers are trained on the train dataset. The efficacy of these classifiers is then evaluated using the test dataset. Following section explains the operation of each classifier.

A. Logistic Regression

Logistic regression is widely acknowledged as one of the best machine learning methods for binary classification due to its simplicity and adaptability to a wide range of problems
. It operates on a categorical dependent variable, where the dependent variables are binary, indicating outcomes such as 1 or 0, pass or fail, etc. Ordinal logistic regression is used for multiple ordered categories, while multinomial logistic regression is used when there are several outcomes for each variable. The logistic function is expressed as follows, where \(a \) is the function's input. and \(e \) is Euler's number. The ROC curve represents the

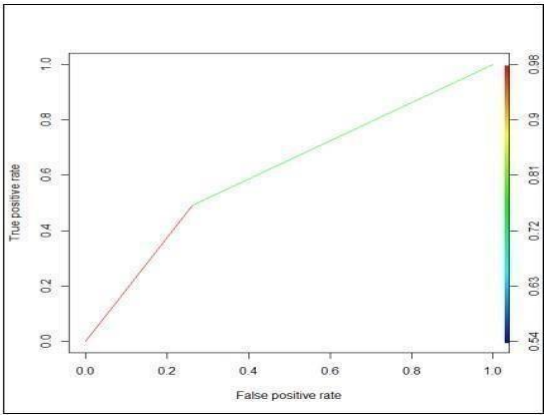


Figure 8. Logistic regression graph

B. Navies Bayes

It is the greatest ML classification algorithms that makes use the Bayesian-algorithm is the Navies Bayes classifier.

The classification algorithm Naives Bayes is veryscalable, necessitating linear variables in the problem description as predictor variables [17]. It is competitive with SVM and comparable for regression and classification. It indicates the patients' specialization in relation to the illness. It provides the likelihood that an event will occur as well as the of each in probables put characteristic for the predictable state. The possiblity of a conclusion A given some data or observation B, when there is a dependency relationship between A and B, is known as a conditional probability. This probability is written as P (A|B), where P(A) represents the likely hood of event A, P(B) represents the likelihood of event B, and P(B|A) represents the likely hood of event B provided that event A has occurred.

C. SVM

SVM is a machine learning method that can be applied to applications involving classification and regression.. Its efficiency compared to other algorithms makes it a popular choice for classification. With this method, a hyperplane plotted as a coordinate represents each attribute in the dataset. The process of classification involves locating the hyperplane that divides several classes. Using this hyperplane as a guide, SVM builds a non-probabilistic binary linear classifier that divides fresh instances into one of the classes.

D. K-Nearest Neighbour

One technique for supervised classification is the K- Nearest Neighbor algorithm. Objects are categorized based on their closest neighbor. This type of learning is instance- based. Euclidean distance isused to calculate an attribute's distance from its neighbors. It employs a sets of designated points and applies them to the marking of an extra point. Based on their shared characteristics, the data are grouped together. The K-NN algorithm can be easily implemented without requiring the creating the model or other assumptions. This approach is flexible and can be applied to search, regression, and classification tasks. K-NN is the simplest method, however its accuracy is impacted by features that are irrelevant and noisy.[6]

E. Decision tree

An approach for classification that works with both numerical and categorical input is the Decision tree.It constructs tree-like structures that simplify the implementation and analysis of data. The algorithm divides data into 2 or more similar subsets based on key indicators. It calculates the entropy of each attribute and then splits the data, focusing on predictors that offer the maximum information gain or minimum entropy

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

$$IG(Y,X) = E(Y) - E(Y|X)$$

The outcomes are simpler to read and understand. Because it examines the dataset in a tree-like graph, this algorithm performs more accuratly than other algorithms.

However, the data may be overclassified, with only 1 attribute checked at a time for decision-making.[6]

F. RF(Random fores)t

The Randomforest classifier is a method of Supervised learning used to regression and classification problems.It employs ensemble learning by combining multiple classifiers to address complex problems and enhance model performance. Random forest aggregates numerous decision trees, To increase predicting accuracy, each forecaster averages their results after being educated on several dataset subsets. The ultimate output of a random forest algorithm is not determined by a single decision tree, but rather by aggregating predictions from all decision trees and averaging them to determine the majority vote for classification problems or the average of all forecasts for regression issues. Accuracy is improved and overfitting is less likely when there are more trees in the forest.

Block Diagram

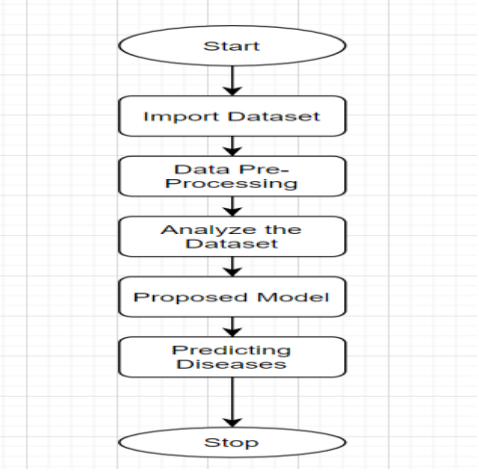


Figure 9. Block Diagram

V. PERFORMANCE ANALYSIS

In this project, various machine learning techniques— such as Svm, Naives Bayes, Decision tree, Random forest, Logistic regression, and KNN—are emploHeart Disease Risk Assessment Tool to forecast heart diseaseUCI dataset, which consists of 14 attributes. Age, gender, kind of chest discomfort, serum cholesterol, blood pressure at rest, blood sugar levels while fasting, maximum heart rate, slope, and number of main vessels are among the criteria that go into making the forecast. Each algorithm's performance is evaluated, and the most accurate algorithm is chosen to
forecast heart disease. Numerous measures are used in the evaluation, such as recall, accuracy, precision, confusion matrix, and F1-score.

The ratio of accurate predictions to all inputs in the dataset is known as accuracy.

Accuracy = $(TP+TN)/(TP+FP+FN+TN)$ Confusion Matrix- It outputs a matrix that represents the system's overall performance.

Confusion matrix of random forest as show in fig.10

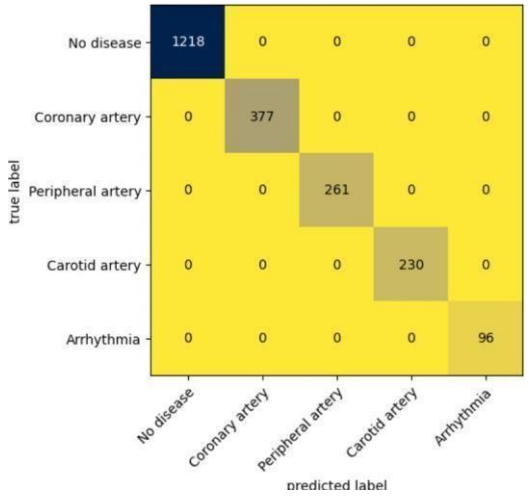
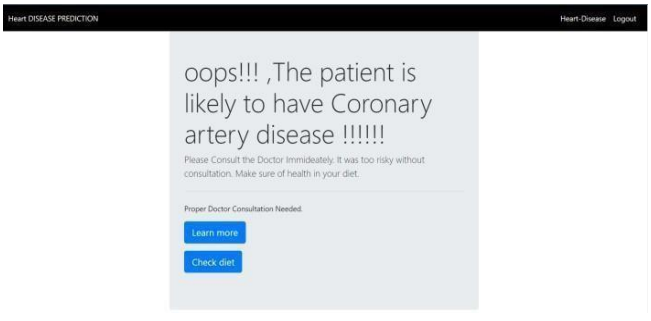


Fig 10.Confusion matrix

Correlation Matrix: A correlation matrix is used in the ML to pick features. It represents the dependencies between several attributes



	precision	recall	f1-score	support
0	0.87	0.87	0.87	1218
1	0.34	0.34	0.34	377
2	0.39	0.36	0.37	261
3	0.41	0.45	0.43	230
4	0.19	0.19	0.19	96
accuracy			0.64	2182
macro avg	0.44	0.44	0.44	2182
weighted avg	0.64	0.64	0.64	2182

Figure 12. Precision, Recall and f1- score

VI. RESULT AND DISCUSSION

Using machine learning techniques for both training and testing, we discovered that Random Forest performs more accurately than the other algorithms. Each algorithm's confusion matrix is utilized to calculate accuracy, and the accuracy formula uses the values of TP, TN, FP, and FN. With a rate of 98%, it was found that the Extreme Random forest achieves the maximum accuracy.

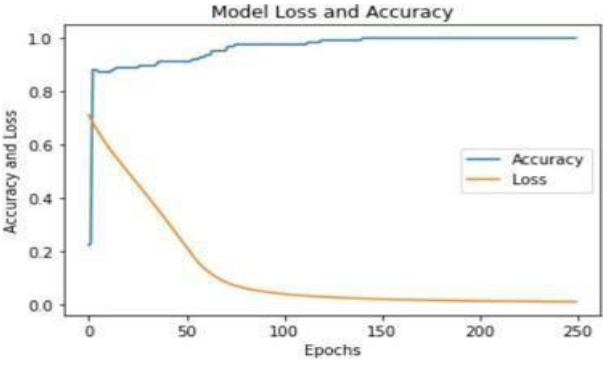


Figure 13. Accuracy and loss graph

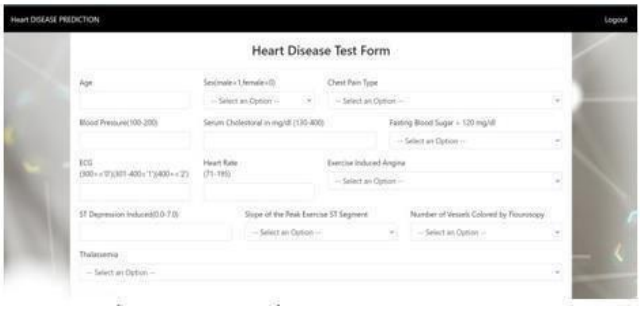


Figure 14. Importing parameters

Figure 15. Result- getying coronary heart disease as a output

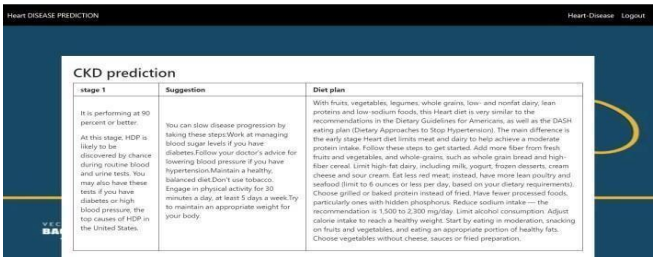


Figure 16. Result- suggestion and diet plan for coronary heartdiseas

VII .CONCLUSION

The heart is a vital organ in the human body, Predicting the Heart disease a critical challenge. Therefore, algorithm accuracy is a key parameter for assessing performance. The effectiveness of ML algorithms largely depends on the quality of the dataset used for training.and testing purposes. When we compare algorithms based on dataset properties and a confusion-matrixes, we find that random forest is the best

VIII. REFERENCES

[1] Ghulab Nabi Ahamad et al. investigated the impact of optimal hyperparameters on the performance of machine learning algorithms for predicting heart

[2] Schmid, J.; Sandhu, S.; Guppy, K.; Lee, S.; Froelicher, V. International use of a novel probability approach for diagnosing coronary artery disease. Am. J. Cardiol. 1989, 64, 304–310.

[3] Ebiaredoh-Mienye S.A., Swart T.G., Esenogho E., and Mienye I.D. A ML approach uses filter-based feature selection disease.

[4] Gayathri, R.; Rani, S.U.; Cepov á, L.; Rajesh, M.; Kalita, K. A comparative analysis of ML models for predicting Mortar compressive strength. Processes:.

[5] : Detrano, R.; Janosi, A.; Steinbrunn, W.; Pfisterer, M.;

[6] Dinesh Kumar G, Santosh Kumar D, Arumugaraj K, Mareeswari V "Prediction of Cardiovascular Disease Using Machine Learning Algorithms" Proceedings of 2018 IEEE International Conferenc on-Current Trends toward Converging-Technologies,Coimbatore,India.

[7].Pooja Anbuselvan, "Heart Disease Prediction using Machine Learning Techniques" Bangalore Institute of Technology Bengaluru, Karnataka, India.

[8].Maryam I. AI-janabi, Mahmoud H. Qutqut, Mohammad Hijjawi. "Machine Learning Classification Techniques for Heart Disease Prediction". International Journal of Engineeringand-Technology-2018.

[9] Ashir Jave, Shijie Zhou, Liao Yongjian, Iqbal Qasim, Adeeb Noor, Redhwan Nour4, Samad Wali, and Abdul Basit, "An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection", IEEE Access 2017.

[10] : Senthilkumar Mohan, Chandrasegar Thirumalai,

[11].Gautam Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", IEEE Access2019. K.Prasanna Lakshmi and Dr. C.R.K.Reddy, "Fast Rule-Based Heart Disease Prediction Using Associative Classification Mining", IEEE International Conference on Computer, Communication, and Control (IC4-2015).

[12].Archana Singh and Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," 2020 IEEE International Conference on Electrical and Electronics Engineering (ICE3), pp. 452-457.

[13].Pranitha Gadde, Gunturu Deepthi, and Cherukuri Shivani. "Heart Disease Prediction Using Machine Learning Algorithms" Department of Computer Science, ANITS, Visakhapatnam, India.