Heart Disease Prediction Using Machine Learning

Manoj Mannem – 700725556

Kinoree Meda - 700732653

Abstract

In various fields around the world, machine learning is used. There are no exceptions in the healthcare sector. Machine learning can be crucial in determining whether or not there will be locomotor abnormalities, heart ailments, and other conditions. If foreseen far in advance, such information can offer crucial intuitions to doctors, who can then modify their diagnosis and approach per patient. Heart disease cases are quickly rising every day, thus it's crucial to predict any potential illnesses in advance. This diagnosis is a challenging task that requires accuracy and efficiency. The primary focus of the project is show about the patients, given certain medical characteristics, are more likely to suffer heart disease. Using the patient's medical history, we developed a system to determine if a heart disease diagnosis is likely or not for the patient. To forecast and categorize the patient with heart disease, we employed a variety of machine learning methods, including K-Nearest Neighbors(KNN), Naive Bayes, Support Vector Machine(SVM), Logistic Regression, Decision tree and Random Forest. The regulation of how the model can be utilized to increase the precision of heart attack prediction in any individual is done in a very helpful way. The proposed model's accuracy in predicting signs of having a heart illness in a certain individual is quite satisfactory as we are using many different machine learning models to display it.

We analyze existing classifiers, which can provide better accuracy and predictive analysis. The Given heart disease prediction system enhances medical care and reduces the cost. This initiative has provided us with a wealth of information that can be used to predict who will get heart disease.

Keywords: KNN, SVM, Naive Bayes, Decision Tree, Random Forest, Logistic Regression

Introduction:

The World Health Organization states that cardiovascular diseases (CVDs) are the leading cause of death worldwide, accounting for more deaths each year than all other causes combined. In 2016, 17.9 million deaths worldwide were attributed to CVDs, or 31% of all fatalities. Heart attacks and strokes are to blame for 85% of these fatalities[1]. Heart disease has already become a severe threat to many families around the world, especially in those impoverished nations, due to the high mortality rate and high cost of surgery. It is important for people to understand how different types of human characteristics relate to the risk of developing heart disease. In order to plan ahead or prevent heart disease, a robust model is useful and meaningful in identifying the types of people who are most likely to develop the condition.

More people suffer from heart failure disorders than from other autoimmune diseases today. Cardiovascular diseases (CVDs) impede blood flow via the blood arteries and have an impact on the heart. Heart disease (heart attack), cerebrovascular illnesses (strokes), congestive heart failure, and several other pathologies are examples of chronic ailments in CVD. Around 17 million people worldwide die from CVDs each year, and since the COVID-19 epidemic, the death rate from heart disease has gone up.

https://github.com/manojchowdary005/ML-FINALPROJECT

Machine learning is a technique for manipulating and extracting implicit knowledge about data that was either unknown or known in the past. The area of machine learning is extremely broad and diversified, and its application and breadth are growing daily. Machine learning uses a variety of classifiers from supervised, unsupervised, and ensemble learning to predict outcomes and assess the accuracy of a dataset. By using a person's medical history, our initiative can identify those who are most likely to be diagnosed with a cardiac condition.

A number of risk factors for manual heart disease prediction may include lack of physical activity, bad eating patterns, or even alcohol use. For the purpose of predicting heart disease, preexisting conditions, age, the severity of chest discomfort, the results of blood tests, and many other parameters can be computationally combined.

Using machine learning technologies, a data-driven strategy can undoubtedly aid in the prediction of cardiac disease given the clearly specified parameters and the growth of data science. A prediction model can be suggested for the early detection of cardiac disease, improved diagnosis, and high-risk patients, and decision-making is improved for additional treatment and prevention.

As recent advancements in medical healthcare are seen. Massive amounts of data about heart illness have been gathered by the healthcare system, and these data have been combined to produce datasets that include various medical traits or parameters like age, sex, blood pressure, cholesterol, chest type, and so forth. There are roughly 13 to 15 different medical parameters in each dataset. These datasets are now available for analysis and information extraction. Therefore, by using machine learning algorithms on this vast amount of data to extract features (information/medical parameters), we can anticipate cardiac disease at its early stages.

After using different machine learning methods on characteristics extracted from datasets, such as Logistic Regression, Naive Bayes, Support Vector Machines, K closest neighbors (KNN), etc., we can categorize whether a person has cardiovascular disease or not.

After comparing them, we can identify the optimal algorithm that accurately predicts heart disease. Different algorithms will yield varying degrees of accuracy. Our project's primary goal is to increase the accuracy of heart disease prediction.

Computational statistics, which emphasizes using mathematical optimization to give methodologies, theory, and application domains to tackle industrial, social, business, and medical problems in the real world, is closely related to machine learning.

Unsupervised learning and supervised learning are the two main divisions. In supervised learning, an algorithm creates a mathematical model using a set of data that includes the required inputs and outputs. The approach creates a mathematical model from a collection of data that just comprises inputs and no desired output labels in unsupervised learning.

Since our goal is to estimate the likelihood of having heart disease based on how the body physically functions. Given the expected inputs and outputs, supervised learning is unquestionably the best option. In this project we are using decision tree, logistic regression, SVM, Naïve Bayes in supervised learning to predict a person's likelihood of developing heart disease based on their bodily functions.

Regression which showed a good accuracy in comparison to the previously used classifier such as naive bayes etc.

Motivation:

The primary reason for conducting this study is to propose a model for predicting the development of heart disease. The provision of high-quality services at reasonable prices is a significant problem for healthcare institutions (hospitals, clinics). Effective therapy delivery and accurate patient diagnosis are prerequisites for providing quality care. It is unacceptable to make poor clinical decisions since they can have disastrous results. The price of clinical tests must be kept to a minimum by hospitals. They can accomplish these goals by utilizing the proper computer-based information and/or decision support technologies. Today, the majority of hospitals use hospital information systems to manage their patient data or healthcare. Huge volumes of data in the form of statistics, text, charts, and images are frequently produced by these systems. Sadly, very few clinical decisions are supported by these findings. These data contain a wealth of secret knowledge that has mostly gone unexplored. How can we transform data into information that will help healthcare professionals make wise clinical decisions? is a crucial topic that is raised. The primary driving force behind this study is this. Additionally, the goal of this project is to determine the optimum classification method for detecting cardiac disease in a patient. Six classification algorithms, namely Naive Bayes, Decision Tree, Random Forest, SVM, Linear Regression and KNN are employed at various levels of evaluations in a comparative study and analysis to support this work. Although these machine learning methods are widely utilized, predicting cardiac disease is a crucial task requiring the highest level of accuracy. Consequently, a variety of levels and assessment strategy types are used to evaluate the three algorithms. This will enable scientists and medical professionals to create a better.

Main Contributions & Objectives:

This project's goal is to determine, depending on the patient's medical characteristics—such as gender, age, chest discomfort, fasting blood sugar level, etc.—whether they are likely to be diagnosed with any cardiovascular heart illnesses. A dataset containing the characteristics and medical background of the patient is chosen from the UCI repository. We make a prediction about the patient's potential for heart disease using this dataset. We categorise a patient based on 14 medical characteristics to determine whether they are likely to develop a heart condition in order to anticipate this. Six algorithms—KNN, Decision Tree, SVM, Naive Bayes, Random Forest Classifier, and Logistic Regression—were used to train these medical characteristics. Random Forest is the most effective algorithm here, providing accuracy of 90.16% followed by Logistic Regression and Naïve Bayes with an accuracy of 85.25%. Finally, we categorize patients according to whether they are at risk of developing a heart condition or not. This method is also incredibly economical.

Related Work:

The leading cause of death in the modern world is heart disease. The accuracy obtained by Umair Shafique et al.[1] using decision tree, naive bayes, and neural network algorithms for data mining was 82% for naive bayes and 78% for decision tree. WEKA machine learning software was used by the authors. Sabarinathan Vachiravel et al. [2] devised a decision machine learning system to predict cardiac disease and used decision trees to reach 85% accuracy.

The dataset utilized in the study by Vikas Chaurasia et al. [3] was retrieved from a UCI laboratory and contains 14 different attributes, only 11 of which were used to predict heart disease using Naive Bayes and Decision Tree machine learning methods. They

employed the WEKA tool, which helped them reach accuracy of 82% for Naive Bayes and 84% for decision trees. N. Komal Kumar, G. Sarika Sindhu, et al. [4] suggested machine learning techniques for cardiac disease prediction, including Random Forest, Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). Their maximum accuracy was 85% while utilizing the Random Forest algorithm, 74% when using Logistic Regression, and 77% when using SVM.

Malkari Bhargav et al. [5] advocated utilizing various ML approaches to detect cardiac problems. Datasets were gathered from the UCI ML repository. There are 14 such criteria in all, including age, blood pressure, and cholesterol. Using ANN, he reached the greatest accuracy of 96%. The lowest accuracy he achieved was 68% using KNN, followed by 88% using Logistic regression, 83% using Random Forest, 83% using Decision Tree, and 70% using SVM. Using ML models, Gayatri Ramamoorthy et al. [6] were able to predict cardiac illness. The authors' accuracy scores ranged from 83% for KNN to 65% for SVM and 80% for Naive Bayes, with KNN receiving the greatest accuracy score.

Apurb Rajdhan et al's [7] analysis of cardiovascular illness used ML algorithms such decision trees, logistic regression, random forests, and nave bayes, and they achieved accuracy of 81%, 85%, 90%, and 85%, respectively. Deep learning and machine learning techniques were utilized by Hana H. Alalawi et al. [8] to diagnose heart illness utilizing a mix of two datasets that were gathered from Kaggle and Cleveland dataset for heart. It resulted in a 92% accuracy rate for Random Forest. He obtained the corresponding accuracy values for Naive Bayes 83%, ANN 77%, KNN 71%, Logistic regression 75%, and SVM 72%. A model for predicting heart disease that makes use of different combinations of characteristics was created by J. Maiga et al. Different classification techniques, including KNN, naive bayes, and random forest, were applied. With Random Forest, the authors' maximum accuracy was 73%. Because they didn't do feature scaling and normalization on the data, they weren't able to attain good accuracy.

A. Lakshmanrao et alresearch .'s[10] for the diagnosis of heart disease included a number of data mining and gradient boosting methods. For managing unbalanced datasets, the authors used a variety of sampling approaches. From Kaggle, a dataset titled "Framingham heart disease" was gathered. 4220 patient records are included in the dataset, which comprises 15 characteristics. They employed many algorithms for boosting, such as Adaboost and Gradient boosting, and they got accuracy of 78% and 88%, respectively. However, they only achieved an accuracy of 61% for Naive Bayes and 66% for Logistic Regression.

They used numerous machine learning (ML) algorithms in their study, and the accuracy of Logistic Regression was 80%, SVM was 83%, Random Forest was 86%, Decision tree was 86%, and neural network was 84%. Several machine learning (ML) methods, including SVM, Naive Bayes, and MLP, were compared by Hossam Meshref et al. And also selected specific features from the datasets, for which they obtained varying degrees of accuracy when choosing various features, such as when they selected all 14 features from the datasets, for which they obtained the highest accuracy using nave bayes of 81%, but for which they obtained the highest accuracy using SVM when they selected only particular features. They concluded from their research that the most crucial step in increasing the precision of machine learning algorithms is feature selection.

Proposed Framework:

The collection of data and selection of the most crucial attributes is the first step in the system's operation. The relevant data is then preprocessed into the format needed. After that, the data is split into training and testing data. The algorithms are used, and the training data is used to train the model. By testing the system with test data, the correctness of the system is determined. The modules listed below are used to implement this system.

1. Collection of Dataset
2. Selection of attributes
3. Data Pre-Processing
4. Balancing of Data
5. Disease Prediction

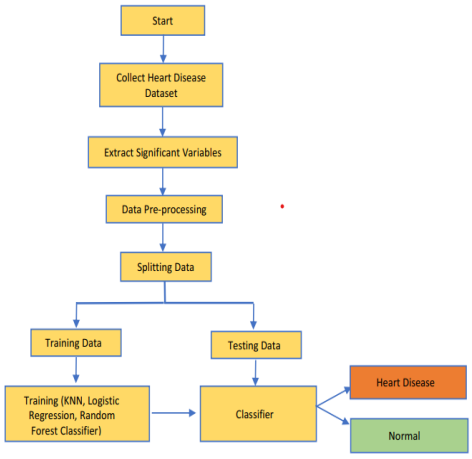


Fig - Proposed model

1. Collection of dataset:

For our algorithm to forecast cardiac disease, we must gather a dataset. Following dataset collection, we divided the dataset into training and testing data. The training dataset is used to develop prediction models, while the testing dataset is used to assess those models. In this project, 30% of the data are used for testing and 70% of the data are utilized for training. Heart Disease UCI was the dataset utilized for this project. There are 76 attributes in the dataset; 14 of them are utilised by the system.

2. Selection of attributes:

The selection of pertinent attributes for the prediction system is part of the attribute or feature selection process. By doing this, the system's effectiveness is improved. The forecast uses a number of patient characteristics, including gender, the type of chest discomfort, fasting blood pressure, serum cholesterol, exang, etc. For this model, attribute selection is done using the correlation matrix.

3. Pre-processing of Data

This is a critical stage in the development of a machine learning model. Data that isn't initially clean or in the model's required format can lead to inaccurate results. Pre-processing involves transforming data into the format we need. It is used to handle the dataset's noise, duplication, and missing values. Activities like importing datasets, partitioning datasets, attribute scaling, etc. are all part of data pre-processing. Preprocessing the data is necessary to increase the model's accuracy.

4. Balancing of Data:

Two methods can be used to balance unbalanced datasets. Both undersampling and oversampling occur.
(a) Under Sampling: Under Sampling reduces the size of the abundant class to balance the dataset. When there is sufficient data, this method is taken into consideration.
(b) Over Sampling: With over sampling, it is possible to balance the dataset by enlarging the size of the small samples. When there is insufficient data, this method is taken into account.

5. Prediction of Disease:

SVM, Naive Bayes, Decision Trees, Random Trees, Logistic Regression are just a few examples of the machine learning algorithms that are used for

categorization. The algorithm that has the highest accuracy is used to forecast cardiac disease after comparative analysis of several algorithms.

Implementation:

Thus, Python programming is the language employed in this project. In the Jupyter notebook of the Anaconda Navigator, we are running Python code.

a) Dataset collection

b) Numpy, Pandas, Scikit-Learn, Matplotlib, and Seaborn libraries were imported.

c) Exploratory data analysis: To gain additional knowledge about the data.

d) Data cleaning and preprocessing: Used the python methods isnull() and isna().sum() to check for null and garbage values.

We performed feature engineering on our dataset during the preprocessing step. Using the get dummies() function of the Pandas library, categorical variables were transformed into numerical variables. There are some categorical variables in both of our datasets.

e) Feature Scaling: In this phase, we apply standardization to our data by utilizing the scikit-learn library's StandardScalar() and fit transform() algorithms.

f) Model selection: We initially distinguished between Xs and Ys.

Our datasets' Xs are their characteristics or input factors, and Ys are their dependent or goal variables, which are essential for disease prediction. Then we divided our X's and Y's into train and test split using the train test split() function of the sklearn library by using the importing model selection function.

We divided our data, setting aside 20% for testing and 80% for training.

g) Using ML models, a confusion matrix of all models was produced.

h) Use of the model that provided the highest level of accuracy.

Data Description:

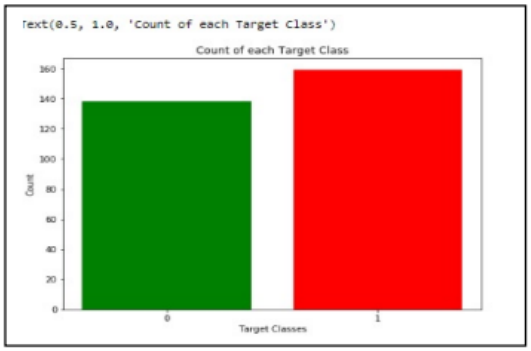
This data set was taken from the webpage for the UCI Machine Learning Repository. It has 76 properties, yet every published experiment only mentions employing a portion of 14. There are 303 observations in total. Additionally, David W. Aha contributed it.

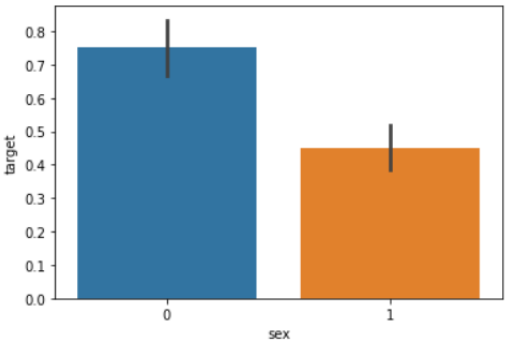
Attribute Information

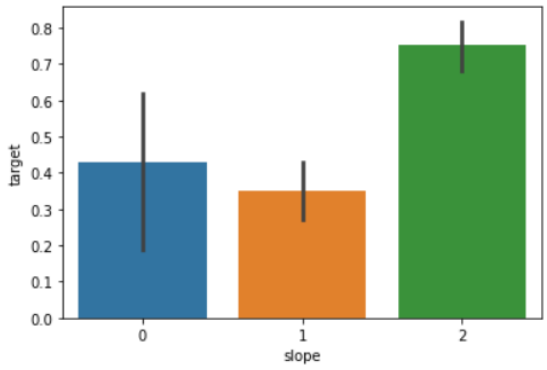
Sr. no.	Parameter / attribute	Information
1	age	Patient's age in years
2	sex	Gender of patient: 0 = Female ; 1 = Male
3	chest pain type	Chest pain type 1 = Typical angina 2 = Atypical angina 3 = Non-anginal pain 4 = Asymptomatic
4	resting bps	Resting blood pressure (mm hg)
5	cholesterol	Serum cholesterol in mg/dl
6	fasting blood sugar	Fasting blood sugar 0 = Less than 120 mg/dl 1 = More than 120 mg/dl
7	resting ecg	Resting electrographic results 0 = Normal 1 = Having ST_T_wave abnormality
8	max heart rate	Maximum heart rate achieved
9	exercise angina	Exercise induced angina 0 = No 1 = Yes
10	oldpeak	Exercise induced ST depression in comparison with rest state
11	ST slope	Slope of exercise ST segment 0 = Normal 1 = Unsloping 2 = Flat 3 = Downsloping
12	target	Has heart disease or not 0 = No 1 = Yes

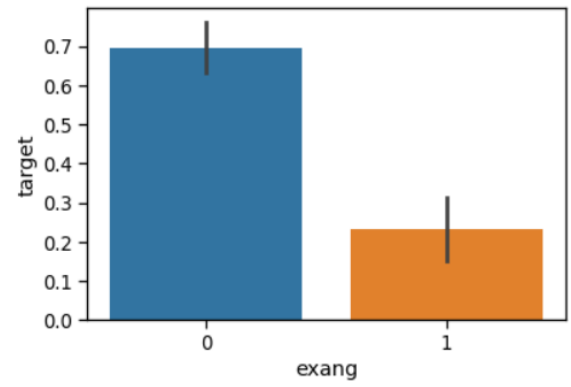
Fig – Dataset Attributes

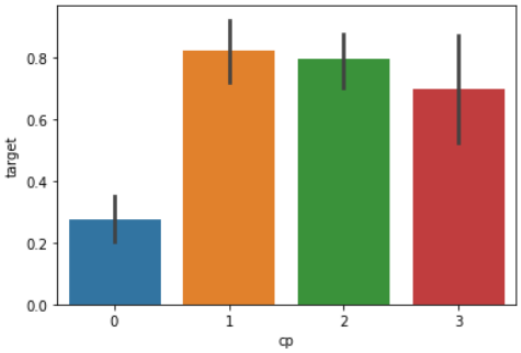
Result/Experimentation & Comparison/Analysis:

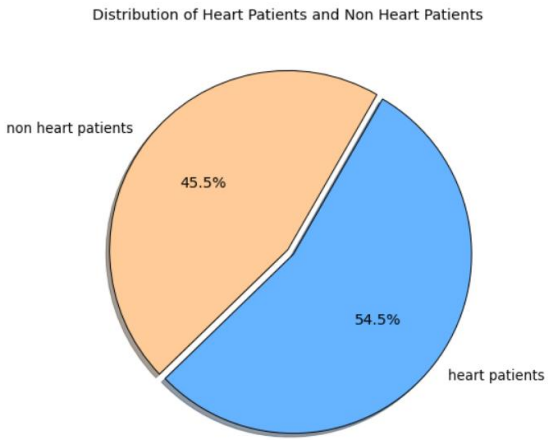













```
Percentage of patience without heart problems: 45.54
Percentage of patience with heart problems: 54.46
```

```
The accuracy score achieved using Logistic Regression is: 85.25 %
The accuracy score achieved using Naive Bayes is: 85.25 %
The accuracy score achieved using Support Vector Machine is: 81.97 %
The accuracy score achieved using K-Nearest Neighbors is: 67.21 %
The accuracy score achieved using Decision Tree is: 81.97 %
The accuracy score achieved using Random Forest is: 90.16 %
```

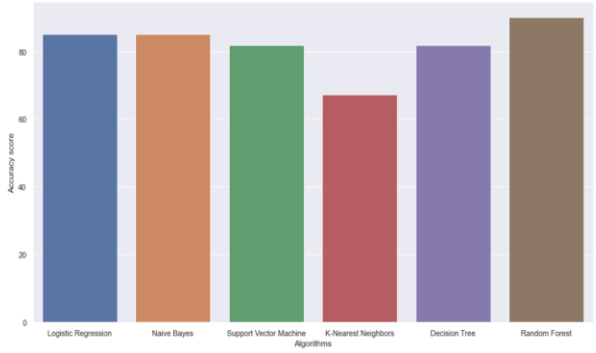


Fig - Comparison b/w the different ML models accuracy

References:

[1] Umair Shafique, Fiaz Majeed, Haseeb Qaiser, Irfan Ul Mustafa “Data Mining In Healthcare For Heart Diseases”, International Journal Of Innovation And Applied Studies Issn 2028-9324 Vol. 10, Issue 4, (2015), Pp. 1312-1322.

[2] V. Sabarinathan “Diagnosis Of Heart Disease Using Decision Tree”, International Journal Of Research In Computer Applications & Information Technology Vol. 2, Issue 6, (2014), Pp. 74-79

[3] Vikas Chaurasia, “Data Mining Approach To Detect Heart Diseses”, International Journal Of Advanced Computer Science And Information Technology (Ijacsit) Vol. 2, Issue 4, 2013, Page: 56-66, Issn: 2296-1739

[4] G. Sarika Sindhu, “Analysis And Prediction Of Cardiovascular Disease Using Machine Learning Classifiers”, International Conference On Advanced Computing & Communication Systems (Icaccs) April 2020.

[5] Malkari Bhargav And J. Raghunath, “A Study On Risk Prediction Of Cardiovascular Disease Using Machine Learning Algorithms”, International Journal Of Emerging Technologies And Innovative Research (Www.Jstor.Org), Issn:2349-5162, Vol.7, Issue 8, Page No.683-688, August 2020

[6] Gayathri Ramamoorthy, “Analysis Of Heart Disease Prediction Using Various Machine Learning Techniques”, International Conference On Artificial Intelligence, Smart Grid And Smart City Applications, (Ais Gsc 2019)

[7] Apurb Rajdhan, “Heart Disease Prediction Using Machine Learning”, International Journal Of Engineering Research & Technology (Ijert)Issn: 2278-0181 Vol. 9 Issue 04, April-2020

[8] Hana H. Alalawi And Manal S. Alsuwat, “Detection Of Cardiovascular Disease Using Machine Learning Classification Models”, International Journal Of Engineering Research & Technology (Ijert) Issn: 2278-0181 Vol. 10, Issue 07, July-2021

[9] J. Maiga, G. G. Hungilo, “Comparison Of Machine Learning Models In Prediction Of Cardiovascular Disease Using Health Record Data,” International Conference On Informatics, Multimedia, Cyber And Information System (Icimcis), 2019, Pp. 45–48

[10] A. Lakshmanarao,Y.Swathi, P.Sri Sai Sundareswarar, 'Machine Learning Techniques For Heart Disease Prediction” International Journal Of Scientific & Technology Research (2019) Vol. 8, Issue 11, November 2019 Issn 2277-8616

[11] Ashok Kumar Dwivedi, “Analysis Of Computational Intelligence Techniques For Diabetes Mellitus Prediction,” Neural Computing Applications, Vol. 13, Issue 3, Pp. 1–9, 2017

[12] Muhammad Saqlain, “Identification Of Heart Failure By Using Unstructured Data Of Cardiac Patients,” (2016) International Conference Parallel Processing. Work, Pp. 426– 431, 2016

[13] Hossam Meshref, “Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach”, International Journal Of Advanced Computer Science And Applications (2019), Vol. 10, Issue 12, 2019

[14] Baban.U. Rindhe, Nikita Ahire, “Heart Disease Prediction Using Machine Learning”, International Journal Of Advanced Research In Science, Communication And Technology (Ijarcet 2021) Vol. 5, Issue 1, May 2021

[15] H. Jayasree, “Heart Disease Prediction System” Journal Of Applied Science And Computations (Jasc 2019) Vol. 1, Issue 6, June/2019 Issn No: 1076-5131

[16] 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

[17] 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.

[18] "Both Blood Pressure Numbers May Predict Heart Disease".Medicalnewstoday.Com

[19]2020,https://www.medicalnewstoday.com/articles/325861.

[20]"Angina(ChestPain)".Www.Heart.Org,2020,https://www.heart.org/en/health-topics/heart-attack/angina-chest-pain.

2020,http://cooleysanemia.org/updates/Cardiac.pdf.

Accessed 14 Mar 2020