Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.776740

PREDICTION OF CORONARY DISEASES USING MACHINE LEARNING (PCDML)

M K Sri Ranga Sai [1], Dr K V Satyanarayana [2], Y V Nagesh Meesala[3]
Raghu Engineering College, Visakhapatnam, India

Abstract: *Heart diseases which are usually known as cardio vascular diseases are a wide range of conditions that affect the heart. These Cardiovascular diseases (CVDs) kill about 20.5 million people every year. It is also the primary cause for death worldwide over the past few decades. It is the need of the moment to obtain a precise and reliable approach to obtain an early diagnosis of the disease by automating the task and thus carrying out effective management. Many researchers used several data mining techniques to help medical professionals diagnose heart disease. However, using data mining can reduce the number of tests required. In order to reduce the number of deaths from heart disease, you must have a fast and effective detection technique. Early prediction can help people change their lifestyle. It also ensures proper medical treatment if needed. In order to reduce the number of deaths from heart disease, a rapid and effective detection technique is needed. The proposed work predicts the possibilities of heart diseases by implementing different data mining techniques such as logistic regression, nearby K nearest decision trees, support vector machine. Therefore, this article presents a comparative study analysing the performance of different machine learning algorithms. In this paper, a web based system which will predict the possibility or chance of person to get Heart Disease based on certain basic factors like cholesterol, diabetes, smoking etc. In this paper, a web based system is developed to predict thepossibility of getting a coronary heart disease. The test results verify that the Support Vector Machine achieved a maximum accuracy of 86.76% against other implemented ML algorithms.*
Keywords: Machine Learning, Health care, Cardio Vascular Diseases, Classification, Logistic regression, K-nearest neighbours, Decision trees, Support vector machine.

INTRODUCTION: The main objective of our paper is prediction of heart condition within the next 10 years by exploitation totally different data processing tools. Heart is a crucial organ of the material body. It pumps blood to each a part of our anatomy. As per the World Health Organization, Cardio Vascular Disease (CVD) is the prime explanation for mortality worldwide.

According to World Health Organisation, heart related diseases are responsible for taking 17.7 million lives every year, 31% of all global deaths. In India too, heart-related diseases have become the leading cause of mortality. Heart diseases have killed 1.7 million Indians in 2016, according to the 2016 Global Burden of Disease Report, released on September 15,2017 [1].

Generally in the United States, in every 40 seconds a person dies with heart attack. This clearly indicates the need to concentrate more on heart related issues. Consequently, this led to an expenditure of over $200 billion per year in the United States alone. This also going to increase at a high rate in the upcoming years as well.

Coronary heart disease usually includes the group of diseases called stable angina, unstable

angina, sudden cardiac death etc..The main symptoms of heart disease are chest pain, bloating, swollen legs, breathing issues, fatigue and irregular heart beat rhythm. The factors that cause heart disease are age, overweight, stress, unhealthy diet and smoking and many other factors [2].

Unhealthy diet, excess alcohol consumption also has a greater risk on functioning of our heart. Early detection of heart diseases are often challenging and difficult to determine in most of the cases. Thus, the increase in the computer-aided detection helped to analyze and understanding the diagnoses in the medical field as well.

Today the major challenge is the efficient and accurate prediction of these diseases. One capable was is the Machine Learning that is training and testing with the help of python and its libraries. Data mining is the process that is used to extract useful data from a large set of raw data.

Machine Learning is nothing but the subfield of data that can handle large datasets efficiently. Thus, in the medical field machine learning can help a lot in the prediction and detection of the diseases.

Various Machine Learning algorithms such as Logistic regression, K-nearest neighbours, Decision trees and Support vector machine are used in this research by comparing these techniques. The paper also mentions scope of future research and different advancement possibilities.
Machine learning is a part of Artificial Intelligence, that can learn by itself and from past experiences and thus capable of making decisions and any predictions. Initially, we train the classification algorithm with the dataset as the input, then the model learns from the data and able to find the patterns from the dataset. Afterwards, by testing with new data it can predict to which class data belongs to.

LITERATURE SURVEY

Machine Learning techniques, models in CVD prediction

KetutAgung Enrico et.al [3] a system was proposed by him for heart disease prediction using KNN algorithm with an accuracy of 81.85%. Using KNN, with increase of number of parameters the performance decreases and it considers 90% of data for training which is computationally expensive.

Himanshustal [4] briefly discussed about large and small data set of heart diseases prediction. They shared that small data set take minimum time for training as well as testing and performed prediction using SVM and KNN algorithm. It also discussed about prediction of heart diseases and prove that some algorithms of machine learning does not perform better for accurateness predication.

Avinash Golande and et. al. [5] studies different ML algorithms that could be used for classification of heart disease. Research was carried out to study Decision Tree, KNN and K-

Means algorithms that can be used for classification and their accuracy were compared. This research concludes that accuracy obtained by Decision Tree was highest further it was inferred that it can be made efficient by combination of different techniques and parameter tuning.

Lakshmana Rao et al,[6]Heart Disease Prediction in elements like diabetes, current smoker, high cholesterol, etc.. contribute So, it is difficult to distinguish heart disease. Different systems in data mining have been utilized to discover the seriousness of heart disease among people. Machine learning makes rationale dependent on chronicled information.
Stephen F. Wenget.al, [7] studied application of various machine learning algorithms to improve the prediction in cardiovascular diseases. They showed that different Machine Learning algorithms are highly successful for improving the accuracy of prediction in CVD diseases, but it needs more patient records. So the more the data, the better the results.
Marjia Sultana et.al, [8] traversed various datasets for Heart disease illness and determined the usage of various Machine Learning algorithms with them. Obviously, datasets are to be pre processed before applying any Machine Learning algorithms. They also suggest the various features that plays important for accuracy determination.
Prajakta Ghadge et al. [9] researched an intelligent heart attack prediction system. The main intention of this article was to find a system that uses big data and data mining modelling techniques. Therefore, the system can identify any hidden knowledge and pattern in the vast data.
AH Chen et al. [10] presented a heart disease prediction system which can help doctors predict the art disease status with the help of clinical data. The C language is used an artificial neural networks for classification and prediction of heart disease. Programming languages like C and C# are used to develop system and with an accuracy of 80%.

Web based CVD

A web based system is developed which consist of efficient method in AI to predict the possibility of a person to get coronary heart disease. The basic factors that both systems consider for prediction are gender, age, cholesterol, blood pressure, diabetes, smoking, family history and physical activity [11].

This paper describes that the proposed system provides a web based interface to the user with machine learning system to predict the chance of occurring Coronary Heart Disease. Initially, the webpage with form will appear where user can enter the details about the basic risk factors of Coronary Heart Disease [11]. Using these details, system will predict the result based on the rules defined. Here, Fig 1 shows the Block diagram of Proposed System.

User can give details about the basic risk factors of CHD. These details are used to predict the result by checking it with the model created when the dataset is trained by using Machine Learning algorithm [11].

The following are some of the devices through which a user can enter his details.

a) Device 1: A mobile phone that has an active internet connection, can be used to enter the

details of the patient i.e, Age, Gender, Blood Pressure, BPM etc.. A user can go through the website and there's no need to login or register to it. He can just enter the details that are needed and click on submit to get the results. A user can go back and reset by clicking the Home button.

b) Device 2: A user can also enter his details from a laptop/desktop with internet connection. The process is same as in mobile phone and laptop as well. So a user can enter the corresponding details by visiting the website.

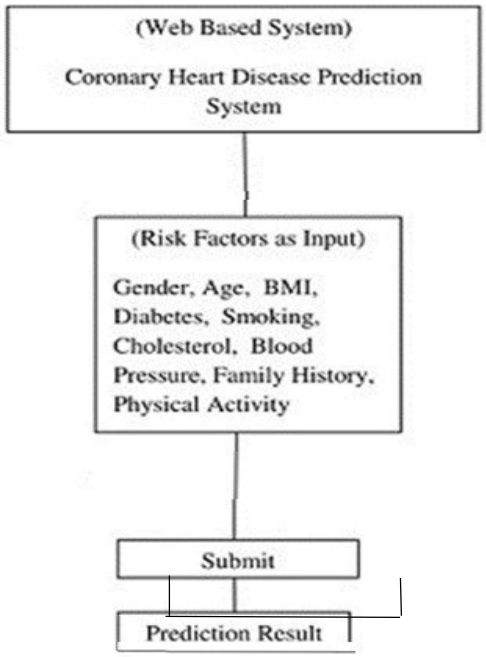

Figure 1:Block Diagram of Proposed System

Web based Implementation:

Coronary Illness hazard prediction System is a web based system created using web scripting language machine learning system to predict the possibility of occurring CHD. There is no need of login or sign up for user to check their status and user need not give any of their personal details. Any user can use the system directly [11].

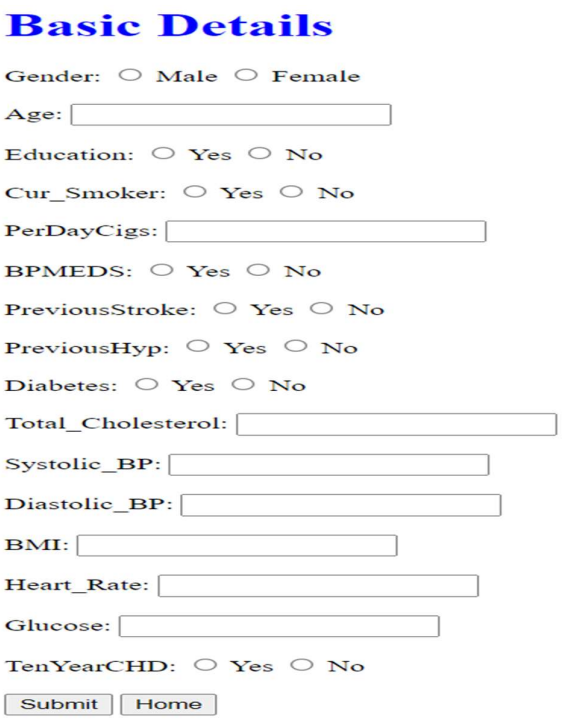

Figure 2 Input given to system

Initially, the user have to give all details about the basic risk factors that is asked for, in that page. Then they have to click the "Submit" button after giving the details that are required. When the click "Submit" button the server directs the details to the machine learning model that is trained and tested. Then the system check the input value with the model of machine learning system and predicts the result which will be displayed back to the user .If they want to go back to main page, then they can click on 'Home' button.

Implementation of ML Algorithms for prediction of CVD

Figure 3: Result

Aim: To predict with the outcome of whether a patient ought to be determined to have a coronary illness in the next ten years. The goal of this classification is to predict whether the patient has 10-year risk coronary heart disease in the future. It includes over 15 attributes. A variable from each attribute is a potential risk factor.

Attributes-
Gender: male or female(Nominal)
Age: Patient's Age
Education: Is he/she currently pursuing education
Cur_Smoker: whether patient is a current smoker or not
PerDayCigs: cigarettes smoked per day
BPMEDS: blood pressure medication of the patient
PreviousStroke: patient previous stroke information
PreviousHyp: whether patient was hypertensive or not
Diabetes: whether the patient had diabetes
Total_Cholesterol: total cholesterol level
Systolic_BP: systolic blood pressure
Diastolic_BP: diastolic blood pressure
BMI: Body Mass Index
Heart_Rate: Heart Rate of the patient
Glucose: glucose level of the patient

Target variable to predict:
Risk of coronary heart disease(10 years) – {1:'yes', 0:'no'}
Here, Table 1 shows the rows and columns of the dataset

Table 1, Rows and Columns of the dataset

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Information Analysis: Correlation is nothing but an indication about the changes between two variables. The relationship of factors utilizing positive Correlation Matrix in Figure 2. Using this we can see there is a positive connection between diabetes and glucose levels in the body. This looks good because the more is the glucose in the body leading to diabetes.

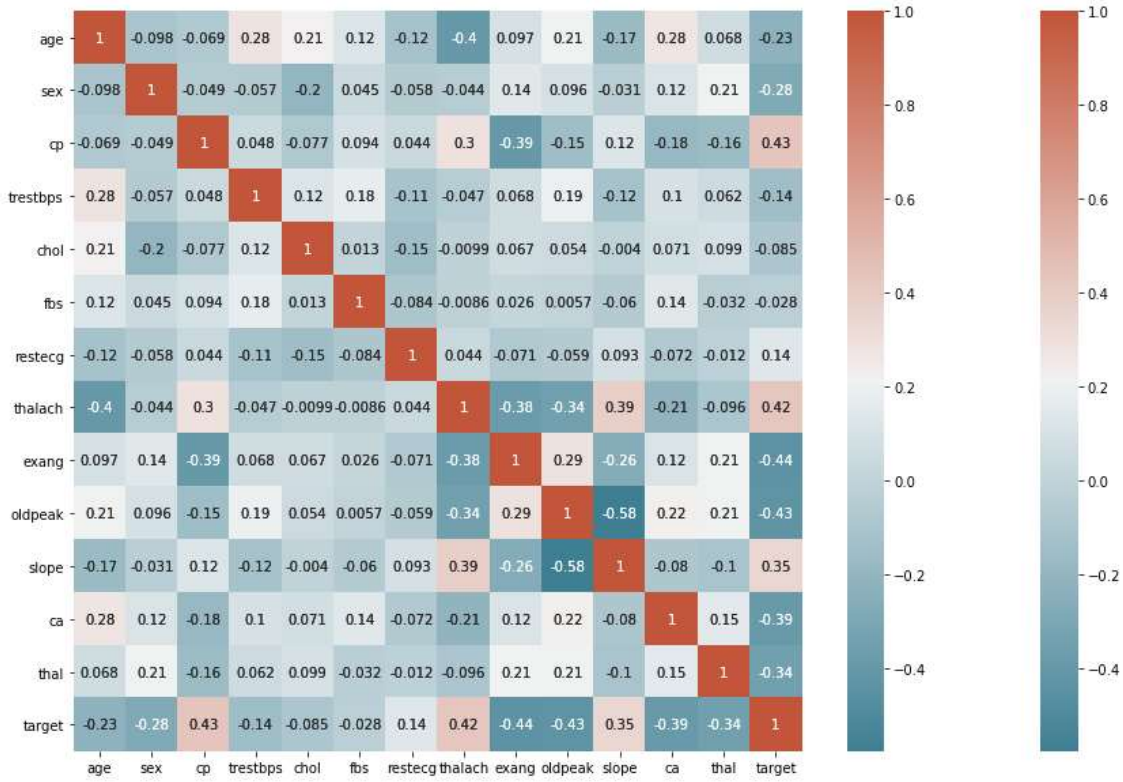

Figure 4: Information is emphatically or adversely associated with our indicator(target).

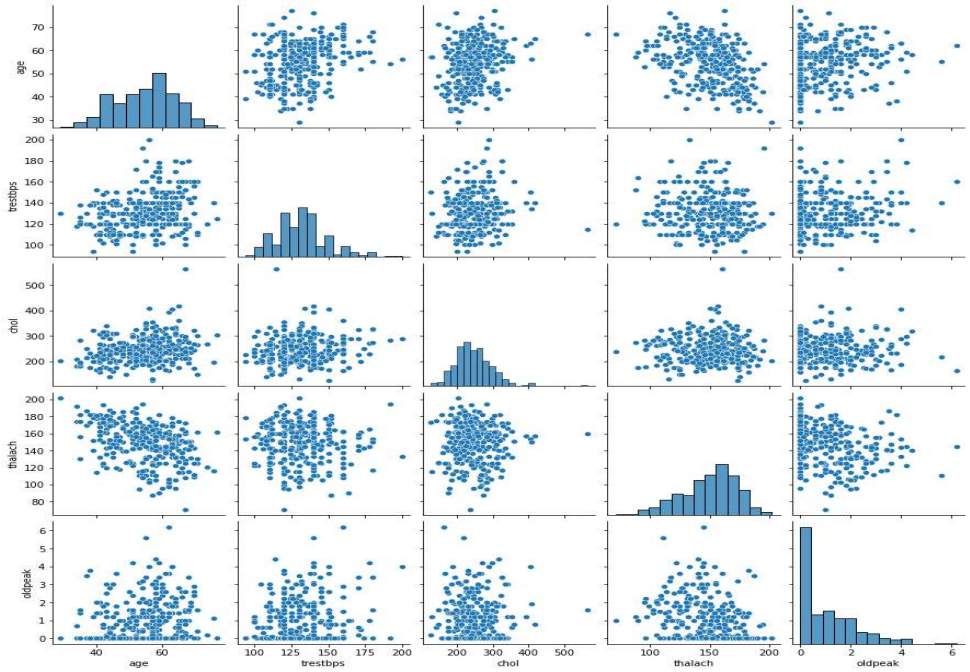

Figure 5: Plot we make a more modest pair plot with just the consistent factors, to jump further into the connections

Pairplots are likewise an extraordinary method to quickly see the relationships between all factors.

Modelling and training: Demonstrating different algorithms on the aimed information which gives the most important elevated precision. We will analyse the precision of comparative on different machine learning algorithms.

Model 1: Logistic Regression

The target variable is selected and the probability of the variable occurrence is predicted by implementing the logistic regression algorithm. Logistic Regression is a classification algorithm mostly used for binary classification problems. The data coded of the target variable is in binary nature (0 or 1) [2]. Here, table 2 shows the result.

Table 2: Accuracy with Logistic Regression.

	precision	recall	f1-score	support
0	0.77	0.67	0.71	30
1	0.71	0.81	0.76	31
accuracy			0.74	61
macro avg	0.74	0.74	0.74	61
weighted avg	0.74	0.74	0.74	61

Logistic regression is a simple yet very effective classification algorithm so it is commonly used for many binary classification tasks.

MODEL 2: K-NN (K-NEAREST NEIGHBORS)

The K-Means splits the data into N groups. Each one is determined by a centroid, which is randomly generated artificial data, to represent the entire group[12]. The following visualization represents the centroids. Then each sample is allocated in the group of the nearest centre. **K-means** is vastly used for clustering in many data science applications, especially useful if you need to quickly discover insights from **unlabeled data. W**e see how to use k-Means for customer segmentation[13].

Table 3: Accuracy with K-NN

	precision	recall	f1-score	support
0	0.78	0.7	0.74	30
1	0.74	0.81	0.77	31

accuracy			0.75	61
macro avg	0.76	0.75	0.75	61
weighted avg	0.76	0.75	0.75	61

Model 3: Support Vector Machine

They are amazingly adaptable managed AI calculations which are utilized both for characterization and relapse. SVM models will categorize new text after being fed sets of named training data for eachgroup [2].Here, table 3 shows the report of Support Vector Machine.

Table 4: Accuracy with SVM

	precision	recall	f1-score	support
0	0.78	0.7	0.74	30
1	0.74	0.81	0.77	31
accuracy			0.75	61
macro avg	0.76	0.75	0.75	61
weighted avg	0.76	0.75	0.75	61

MODEL 4: NAIVES BAYES CLASSIFIER

It is used in constructing the classifiers in simple manner which are based on bayes theorem, the collection of the classifier algorithm will take the features which are independent to one another. The result is shown in the below table 5.

Table 5: Accuracy with SVM

	precision	recall	f1-score	support
0	0.79	0.73	0.76	30
1	0.76	0.81	0.78	31
accuracy			0.77	61
macro avg	0.77	0.77	0.77	61
weighted avg	0.77	0.77	0.77	61

The main idea behind K-NN is that the value or class of a data point is determined by the data points around it.

Model 5: Decision Trees

They use the tree like structures in the model to decide on the outcomes which contain control statements. The impact of the decision will also be calculated with the probability of estimating

the result. The aim is to learn basic decision rules from data features to build a model that predicts the value of a target variable [2].

Table 6: Accuracy with Decision Trees

	precision	recall	f1-score	support
0	0.68	0.7	0.69	30
1	0.7	0.68	0.69	31
accuracy			0.69	61
macro avg	0.69	0.69	0.69	61
weighted avg	0.69	0.69	0.69	61

MODEL 6: RANDOM FOREST

It is an algorithm used for classification and the underlying process involves in selecting the sub part from the given data with the selection of tree structure from many. It takes the normal from the selected subpart to improve the correctness of the calculated result, the following table 10 shows the accuracy obtained.

Table 7: Accuracy with Random Forest

	precision	recall	f1-score	support
0	0.88	0.7	0.78	30
1	0.76	0.9	0.82	31
accuracy			0.8	61
macro avg	0.82	0.8	0.8	61
weighted avg	0.81	0.8	0.8	61

MODEL 7: XGBOOST

It is an algorithm which is implemented to make decisions for classifying the given data records (here patients details). It supports all programming languages and is portable with remote storage, the following table 11 shows the accuracy obtained.

Table 8: Accuracy with XGBoost

	precision	recall	f1-score	support
0	0.84	0.7	0.76	30
1	0.75	0.87	0.81	31
accuracy			0.79	61

macro avg	0.79	0.79	0.78	61
weighted avg	0.79	0.79	0.79	61

Making the Confusion Matrix

Result: It gives a score that shows how supportive each component was in our model

		Original Values	
		+ve (1)	-ve (0)
	+ve (1)	23	7
Estimated prediction of values	-ve (0)	8	23
		0.75098361	

0.754098361 is the exactness score

Instructions to deduce information from confusion matrix:

The total measure of positive cases are 23 from the information and 3 is the total measure of the negative cases. The other values of the matrix 7 and 8 are the mistakes which are measured and resemble the error. Subsequently, on the off chance that we compute the exactness depends on the correctness of the result (i. e actual positive and negative cases).

Exactness = (TruePositive + TrueNegative) / (TruePositive + TrueNegative + FalsePositive + FalseNegative).

Exactness = (23+23) / (23+23+7+8) = 0.80 = 75**%** accuracy

Highlight Importance

This section gives the importance factor which is causing the disease with the score comparison of the selected variables. The accuracy of the model mostly depends on the selection of the factors and the ranges of the values which a variable assigned with. The calculated result of the each factor leads to selection and inclusion of the factor in the model if the result is low then factor is omitted if it is showing desirable results it is included. The correlated factors of the included factor are examined to consider them for including them in the decision.

Result: Score how accommodating each element was in our model

Highlight: 0, Score: 0.01959
Highlight: 1, Score: 0.12968
Highlight: 2, Score: 0.01703
Highlight: 3, Score: 0.01205
Highlight: 4, Score: 0.05312
Highlight: 5, Score: 0.00861
Highlight: 6, Score: 0.00343
Highlight: 7, Score: 0.01830
Highlight: 8, Score: 0.00930
Highlight: 9, Score: 0.12227
Highlight: 10, Score: 0.14504
Highlight: 11, Score: 0.11663
Highlight: 12, Score: 0.12935
Highlight: 13, Score: 0.09710
Highlight: 14, Score: 0.11848

Conclusion:

The heart related disease detection, prediction and medication aspects are always challenging and crucial for patient. It is always risk if the heart patients are not treated in time and if they are far from their home town or if the regular doctored is not available. The Support vector machine was the best performing model in terms of accuracy and the F1 score. Its high AUC shows that it has a high true positive rate and the SMOTE technique helped in improving the models sensitivity by balancing the dataset, this is when compared to the performance metrics of other models on different notebooks on the same dataset. Furthermore, the advancements and availability of internet enables users to access irrespective of their economic and social factors.

This model not only increases the confidence of the patient but also enhances the doctor's ability to treat with utmost care. The treatment module of the model is completely based on the quality and availability of internet connection, thus care should be taken to maintain the internet available.

References:

1. Baban.U. Rindhe , Nikita Ahire , RupaliPatil, ShwetaGagare , ManishaDarade , Heart Disease Prediction Using Machine Learning, International Journal of Advanced Research in Science, Communication and Technology, Volume 5, Issue 1, May 2021, ISSN (Online) 2581-9429

2. B Padmajaa ,ChintalaSrinidhib , KothaSindhuc , KalaliVanajad , N M Deepikae , E Krishna RaoPatrof, Early and Accurate Prediction of Heart Disease Using Machine Learning Model, Turkish Journal of Computer and Mathematics Education, Vol.12 No.6 (2021), 4516-4528

3. I KetutAgungEnriko, Muhammad Suryanegara,Dinda Agnes Gunawan, " Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient's health parameters", 2016.

4. Himanshu Sharma and M A Rizvi. (2017). "Prediction of Heart Disease using Machine Learning Algorithms: A Survey" International Journal on Recent and Innovation Trends in Computing and Communication Volume: 5 Issue: 8 , IJRITCC August 2017.

5. AvinashGolande, Pavan Kumar T, Heart Disease Prediction Using Effective Machine Learning Techniques, International Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.
6. A.Lakshmanarao, Y.Swathi, P.SriSaiSundareswar, Machine Learning Techniques For Heart Disease Prediction, International Journal Of Scientific & Technology Research Volume 8, Issue 11, November 2019.
7. Stephen F. Weng, Jenna Reps, Joe Kai1, Jonathan M. Garibaldi, NadeemQureshi, ―Can machine-learning improve cardiovascular risk prediction using routine clinical data?‖, PLOS ONE | https://doi.org/10.1371/journal.pone. 0174944 April 4, 2017
8. M. Sultana, A. Haider, and M. S. Uddin, "Analysis of data mining techniques for heart disease prediction," 2016 3rd Int. Conf. Electr. Eng. Inf. Commun. Technol. ICEEICT 2016, 2017
9. Ghadge, P., Girme, V., Kokane, K. and Deshmukh, P. (2016) Intelligent Heart Attack Prediction System Using Big Data. International Journal of Recent Research in Mathematics Computer Science and Information Technology, 2, 73-77.
10. AH Chen, SY Huang, PS Hong, CH Cheng, and EJ Lin,2011, "HDPS: Heart Disease Prediction System", Computing in Cardiology, ISSN: 0276-6574, pp.557- 560
11. Harish Rajora, Narinder Singh Punn, Sanjay Kumar Sonbhadra , and SonaliAgarwal, Web based disease prediction and recommender system, Indian Institute of Information Technology Allahabad, India.
12. ElliackinFigueiredo , Mariana Macedo And Hugo Valadares Siqueira Swarm Intelligence For Clustering A Systematic Review With New Perspectives On Data Mining,Https://Doi.Org/10.1016/J.Engappai.2019.04.007, Received 10 August 2018; Elsevier Ltd.
13. Abualigah, L.M., Khader, A.T., Al-Betar, M.A., Alomari, O.A., 2017a. Text Feature
Selection With A Robust Weight Scheme And Dynamic Dimension Reduction To Text Document Clustering. Expert Syst. Appl. 84, 24–36.
14. https://red8.com/cloud-storage-in-the-healthcare-industry/
15. https://towardsdatascience.com/ (08 April 2023)