IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 7, July 2022

DOI: 10.17148/IJARCCE.2022.11786

Cardiovascular Disease Prediction Model using Data Mining Classifiers

Uma K[1], M Hanumanthappa[2]

Research Scholar, Department of Computer Science and Applications, Bangalore University, Bangalore, India[1]

Professor, Department of Computer Science and Applications, Bangalore University, Bangalore, India[2]

Abstract: Cardiovascular disease is the primary cause of death in the nation. Though the data available in the health field is vast, there is still a need to develop a supporting decision system to maintain, analyse, and knowledge evaluation. One such technique that can address such a problem is data mining. Data mining techniques can help to classify whether a patient has heart disease or not. This paper explores the different classification techniques for heart disease prediction. Logistic Regression, Support Vector Machine, Naïve Bayes, Nearest Neighbor, and Decision Tree methods are applied. Build the model to predict new data, and various measures have been taken to assess the classifiers' performance, including accuracy, recall, precision, and F1 score.

Keywords: Data Mining, Heart disease, Data pre-processing, Classification Techniques.

I. INTRODUCTION

Hospitals and medical centers face the challenge of providing quality services at affordable prices. Effective treatment and accurate diagnosis are essential components of quality service. It is unacceptable when poor clinical decisions are made. Clinical tests should be conducted in hospitals at the lowest possible cost. Information and decision support systems can assist them in achieving these results. A hospital information system manages patient data in most hospitals today. These systems usually generate numbers, text, graphics, and charts. Unfortunately, these data are not used frequently to support clinical decision-making. It is estimated that these data contain a great deal of hidden information. This raises a crucial question: "How do we transform data into valuable information for practitioners to make clever clinical findings?" This is the primary motivation for this research. Information systems in hospitals support patient billing, inventory management, and simple statistics. The use of decision support systems in hospitals is limited. Data are hidden in databases often provide more knowledge than doctors can comprehend through intuition and experience. Medical costs are excessive due to this practice, leading to unwanted biases, errors, and errors, affecting the quality of care patients receive. This is a good suggestion because data modeling and analysis tools, such as data mining have the potential to create a knowledge-rich environment.

Data mining is to identify relevant and intelligible patterns in the massive volume of data, whether in healthcare or business. Specific knowledge outlines assist in estimating information trends and the associated decision-making process. Data mining can be applied in the healthcare industry to minimize overheads by enhancing productivity, better patient safety in terms of carefulness, and, most considerably, more lives saved [2]. Predictive treatment, customer relationship management, uncovering the pattern of scams and misuse in the medical sector, healthcare management, and monitoring the success of particular medications have all been effective with data mining in healthcare. Every hospital collects many patient data during admission and treatment times. These data are stored in the form of Electronic Health records. From these records, data mining helps physicians discover the disease pattern earlier and prescribe better treatments. Data mining can help identify the best treatment for specific diseases by comparing the causes and symptoms of diseases. Other data mining applications are recommended for effective treatment associating the various side-effects of treatment and finding the most effective medicine for particular illnesses.

Two prominent data mining examples in the medical field are discussed below: Measuring treatment effectiveness: Healthcare data mining evaluates health indications, reasons, and treatment options to regulate the best treatment option for a particular medical problem. For example, one can compare patients treated with different medications, choose an effective treatment strategy, and minimize costs. Consistent application of this data mining technique, on the other hand, provides standardized treatment procedures for certain diseases, making the care and management process much faster.

Heart disease is now the top reason of death in many countries. The term "heart disease" encompasses problems in the heart. Coronary artery disease (CAD) is the most frequent type of heart disorder, which can increase the risk. Heart disease can be a "silent" heart attack, or heart failure occurs [1] with no sign of symptoms. There are various types of heart diseases. Although inevitable overlap exists, each disease has its symptoms and treatments [3].

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 7.39 ⁜ Vol. 11, Issue x, Month 2022
DOI: 10.17148/IJARCCE.2022.11xx

Below mentioned are a few types of heart diseases:

- Cardiovascular disease
- Heart defect from birth
- Arrhythmia
- Cardiomyopathy with dilated blood vessels
- Acute Myocardial Infraction
- Cardiac Arrest
- Cardiomyopathy with hypertrophy
- Prolapse of the mitral valve
- Aortic stenosis

And the factors that can lead to the risk of heart disease: Family history of heart disease, Consumption of Tobacco, Low nutrition intake, High pulse, Cholesterol, Hypercholesterolemia, Fatness, Physical lethargy.

II. CLASSIFICATION TECHNIQUES

The classification technique is commonly used in disease diagnostics to categorize the dataset into several classes.
The supervised learning method is used to create a classifier model. It describes a set of class labels, and the model is built using available data, such as classified examples. These samples include a range of characteristics. This model was created via supervised learning, with the classified instances serving as training data. The classification technique is supplied with training examples and is allowed to extract specific knowledge eventually. After the classifier development, the model is tested based on the accuracy of the results it provides during the testing process.

A. Logistic Regression (LR)

An approach for estimating the likelihood of a given variable is known as logistic regression. True or false is the most common binary response in logistic regression models. For classification problems, logistic regression is an effective analyzer. It is a multiclass statistical method that can be generalized from binary classification.

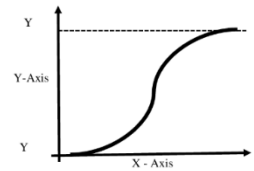


Figure.1. Logistic Regression

B. K-Nearest Neighbor (KNN)

This method is based on the idea that because the neighbor is adjacent in the feature set, it is more likely to be comparable to the item being categorized and so belong to the same class. The nearest neighbor algorithms work as follows,
Step 1: Implement the KNN using the dataset provided. Training and test data must be loaded in the first stage only.
Step 2: Locate the nearest data point's value, i.e., K. (any integer).
Step 3: Using Euclidean or Hamming distance methods, calculate the distance between each data point in the test data.

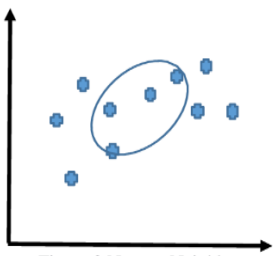


Figure.2.Nearest Neighbour

C. Naïve Bayes (NB)

A Naïve Bayesian classification method uses the Bayes theorem for classification. It is an eager learning algorithm. Since it does not wait for test data to learn, it can classify the new instance faster.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue x, Month 2022

DOI: 10.17148/IJARCCE.2022.11xx

D. Support Vector Machine (SVM)

The SVM method seeks a hyperplane that distinguishes between data points in an N-dimensional space to divide the two data points. One can choose from a variety of hyperplane. The goal is to find a plane where the distance between data points from both classes is the shortest. Raising the margin distance provides some feedback, formulating new data points simpler to classify.

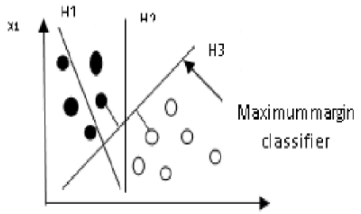


Figure. 3. Support Vector Machine

E. Decision Tree (DT)

A DT is a hierarchal structure where each internal node represents an attribute testing, each split denotes a test conclusion, and each leaf node represents a class label.

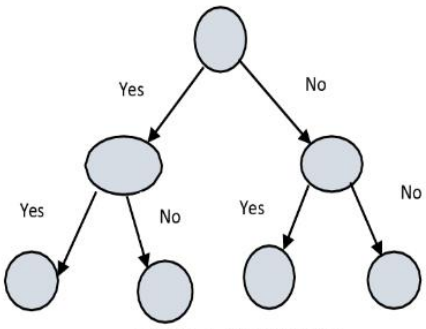


Figure. 4. Decision Tree

III. RELATED WORK

Many researchers have used data mining techniques to identify and forecast cardiac disease.

Sarath Babu et al. [2017] have worked on the diagnosis of heart disease by data mining techniques. The authors selected the various data mining techniques, such as genetic, k-means, MAFIA algorithms, and decision tree classification to diagnose heart disease earlier. A genetic algorithm is used to select essential attributes from many attributes in the dataset. K-means is applied to calculate the low-risk and high-risk patient groups, the MAFIA algorithm extracts the best item set from the dataset, and the decision tree is used to construct the classification tree.

Anjan Nikhil Repaka et al. [2019] have worked on heart disease prediction using the Naïve Bayesian technique. The researchers proposed an architecture to perform the processes, including data collection and pre-processing, and applied the naïve bayesian classification method. The proposed technique is more efficient than other prevailing techniques used before. The proposed method gives 89.77% accuracy towards the predict heart disease.

Pratiksha Shetgaonkar et al. [2021] use data mining techniques to predict heart disease is discussed. The authors have chosen better and more efficient classification techniques to predict the heart diseases, such as neural network, NB, and decision tree, in terms of accuracy. The dataset selected 678 records with 14 attributes. Finally, conclude that accuracy with 81.83% (neural network), 85.01% (naïve Bayes), and 98.54% (decision tree).

Barbara Martins et al. [2021] have used data mining techniques to work on cardiovascular disease prediction. The researchers use the CRISP-DM methodology to predict the outcome. Five classifiers, namely Optimized Random Forest, Decision Tree, and Deep Learning, were applied in RapidMiner and Weka software tools with a split validation method and achieved an accuracy of 73.02% and 73.54%, 71.91%, and 71.94%, respectively. Finally then, applied a defined threshold to get the best result.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue x, Month 2022

DOI: 10.17148/IJARCCE.2022.11xx

IV. PROPOSED METHODOLOGY

Figure.5 depicts the proposed methodology for predicting the outcome of cardiovascular disease.

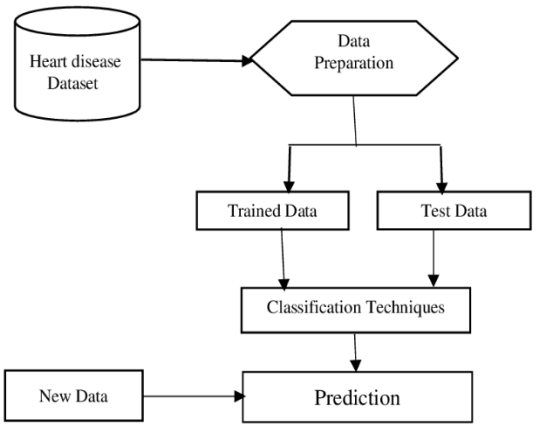


Fig.1. Methodology.

The research provides heart disease prediction with different data collected online. Two other datasets lead the experimentation with unlike features. The first dataset Framingham with 4238 records and 14 columns, is used for the experiment. Another Cleveland dataset with 303 records and 13 columns. In the Framingham dataset, overall rows with missing values are 582 entries; however, these are removed because they represent only 14% of the entire dataset. After the selection dataset, data is prepared for mining with pre-processing techniques. The dataset is split into trained and test data, respectively. Five Classifications such as SVM, Decision Tree, Naive Bayes, Logistic Regression, and Nearest Neighbor are applied to the dataset to predict heart disease. Finally, new data is tested using data mining application with classification techniques. The following table consists of the attributes of the dataset.

SL No	Attributes	Meaning
1	Gender	Male or Female
2	Age	Patient age
3	Current Smoker	Presently smoking or not
4	Cigs Per Day	Average cigarettes per day
5	BP Meds	Medicated under BP
6	Prevalent Stroke	History of stroke
7	Prevalent Hyp	Hypertension condition
8	Diabetis	Diabetics
9	Tot Chol	Blood cholesterol level
10	Sys BP	Systolic blood pressure
11	Dia BP	Diastolic blood pressure
12	BMI	Body Mass Index
13	Heart Rate	Patient Heart Beat
14	Glucose	Blood glucose level

Table.1. Framingham data attributes

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue x, Month 2022

DOI: 10.17148/IJARCCE.2022.11xx

SL No.	Attributes	Meaning
1	age	Patient age
2	Sex	Male or Female
3	Cpt	Type of chest pain
4	restbp	BP in rest
5	chol	Cholesterol level
6	Fbs	Blood sugar level in fasting
7	restecg	ECG in resting
8	thalach	Maximum heart rate
9	exang	Angina induced by exercise
10	oldpeak	ST depression
11	slope	ST segment
12	Ca	Vessels
13	thal	Thalassemia

Table.2. Cleveland data attributes

V. EXPERIMENTATION RESULT AND DISCUSSION

The experiment was conducted using the following techniques such as Nearest Neighbor (KNN), and Naive Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT).
The study aims to discover the most effective data mining methods for predicting heart disease on a different dataset. An experiment was done on the heart disease dataset to find the best prediction system. Various classification techniques are used to see which ones provide the most accurate results for heart disease prediction. The result achieves accuracy for five distinct classification algorithms and different metrics used to evaluate the performance.
Accuracy: The percentage of all successful predictions divided by the number of samples is used to calculate a classifier's accuracy. If the accuracy of the classifier is satisfied, it can be used on upcoming item sets for which the class label is unknown.

$$\text{Accuracy} = \frac{\text{Number of exact likelihoods}}{\text{Total number of likelihoods}}$$

ROC-AUC: A ROC (Receiver Operating Characteristics) curve is a possibility curve. AUC (Area Under the Curve) is a statistical curve. ROC, on the other hand, is a measure of interpretability. It shows how the model can differentiate between classes.
Recall: The capability of a model to discover all the significant samples in a given dataset. Statistically, recall is also defined as number of correct predicted records divided by the number of correct predicted records in addition to the number of wrong predicted records.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

Precision is the no of positive class forecasts that truly belong to the positive class. Mathematically, In addition to the number of false positive values, *accuracy* is defined as the total number of correct predicted records divided by the number of correct predicted records.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

F1 score: The F1-score uses the harmonic mean (HM) of a classifier's precision and recall to establish a single measure.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue x, Month 2022

DOI: 10.17148/IJARCCE.2022.11xx

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Methods	Accuracy (%)	ROC_CUV	Recall	Precision	F1
Logistic Regression	89.18	0.73	0.06	0.64	0.11
SVM	86.14	0.62	0.02	0.50	0.04
Nearest Neighbor	84.41	0.60	0.08	0.28	0.12
Naïve Bayes	83.96	0.71	0.18	0.35	0.23
Decision Tree	77.30	0.56	0.26	0.27	0.27

Table.3. Performance Evolution of Framingham dataset.

From the above table values, LR method shows better accuracy than other techniques, i.e., 89.18% and SVM, KNN, NB, and DT, and classification techniques achieve 86.14 %, 84.41 %, 83.96 %, and 77.30%, respectively. SVM classifier also gives a good result after logistic regression is 86.14%, with only 0.37% difference accuracy. NB also performs well after LR and SVM classifier, and Decision Tree gives the least accuracy compared to other techniques, i.e., 77.30%.

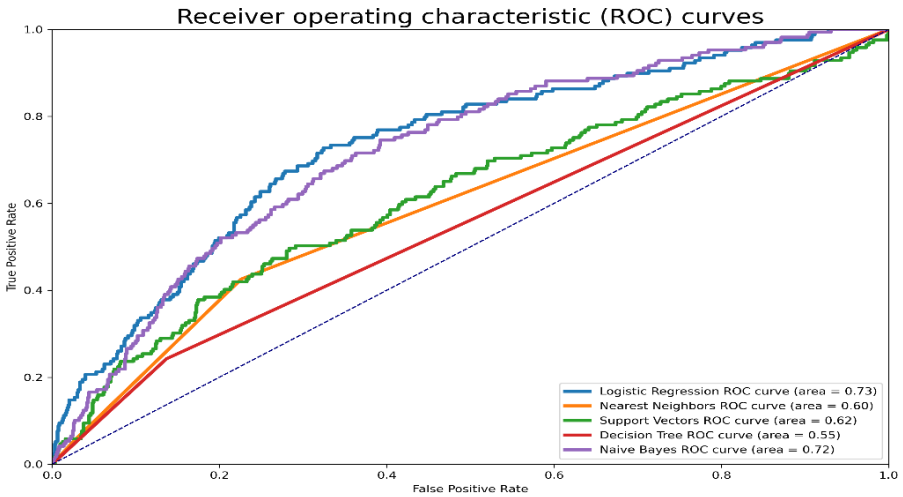


Figure.6. ROC curves for Framingham dataset

Methods	Accuracy (%)	ROC_CUV	Recall	Precision	F1
Logistic Regression	90.00	0.94	0.79	0.94	0.86
Naïve Bayes	82.00	0.85	0.82	0.82	0.82
SVM	87.00	0.73	0.87	0.87	0.87
Nearest Neighbour	89.00	0.90	0.88	0.89	0.88
Decision Tree	77.00	0.76	0.77	0.77	0.77

Table.4. Performance Evolution of Cleveland dataset.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue x, Month 2022

DOI: 10.17148/IJARCCE.2022.11xx

From the above table values, LR method shows better accuracy than other techniques, i.e., 90% and NB, DT, SVM, and KNN, and classification techniques achieve 82 %, 77 %, 87 %, and 89%, respectively. The Nearest Neighbor classifier also gives a good result after logistic regression is 89%. The decision tree gives the least accuracy compared to other techniques, i.e., 77%.

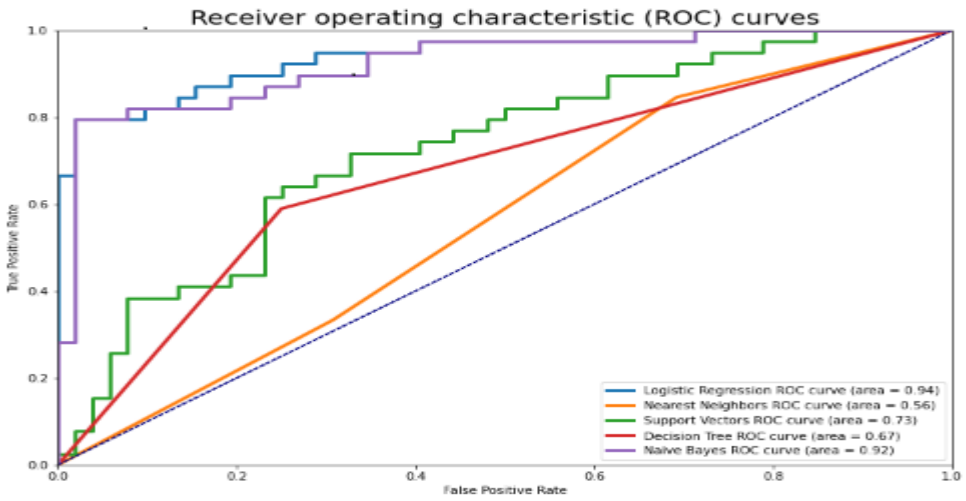
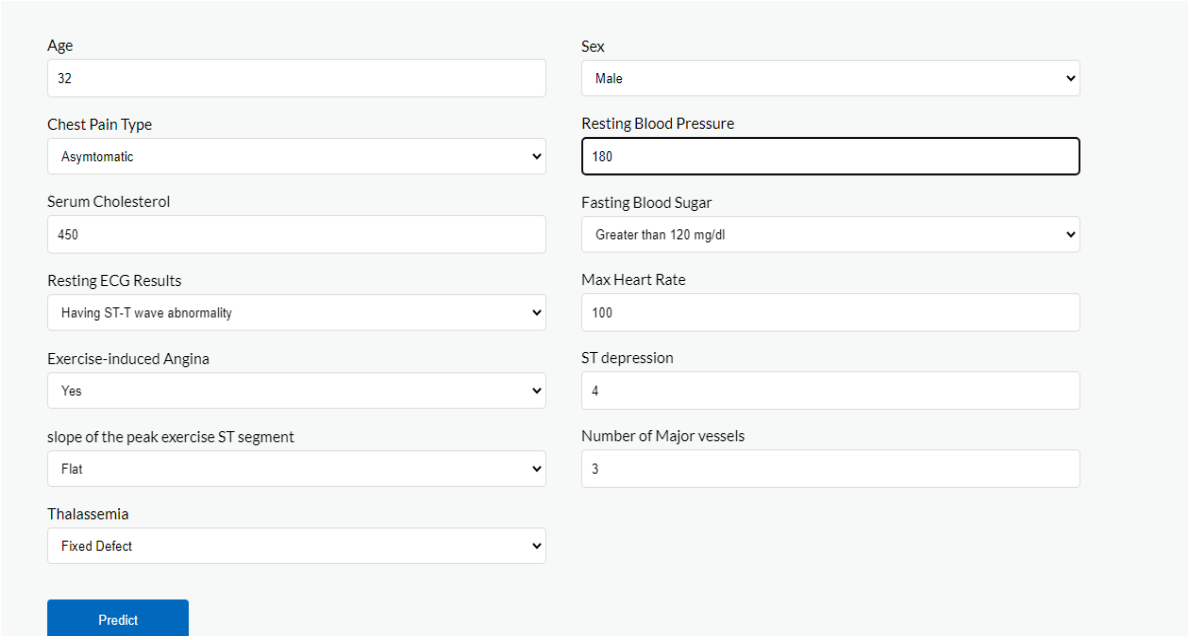
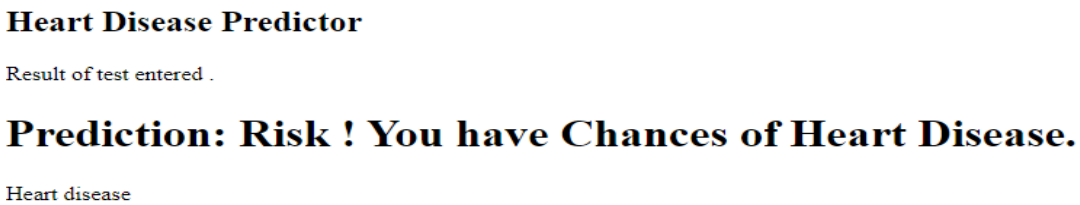


Figure.7.ROC curves for Cleveland dataset

The experiment results show that two datasets with different attributes perform differently through classification techniques. For both datasets, LT gives better accuracy. The data mining application is designed to predict the risk of heart disease for new data.





International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 7.39 ⁜ Vol. 11, Issue x, Month 2022
DOI: 10.17148/IJARCCE.2022.11xx

VI. CONCLUSION AND FUTURE ENHANCEMENT

The main focus of this study explores the application of different classifiers in healthcare, especially in the diagnosis and prediction of heart disease. *Heart disease* is a dangerous condition that can lead to death. Data from online cardiac patients are gathered and used to run an experiment. The following algorithms were used to implement the classification, a data mining technique: logistic Regression, Decision Tree, SVM, KNN, and Naive Bayes. Different criteria were used to evaluate the algorithm's performance, including accuracy, ROC curve, Recall, precision, and F1- score.
According to the experiment, the Logistic Regression algorithm has the highest accuracy, at 89.18% and 90% for Framingham and Cleveland datasets with different attributes. This study demonstrates how accurately and efficiently can be predicted for cardiac diseases with the help of data mining. The results or outcomes of these experiments could aid in making more consistent heart disease diagnoses. Individual procedures are not enough to get the intended result. Ensemble approaches will be used in the future to improve accuracy.

REFERENCES

[1] Adam ,S. Vaughan at al., "Progress Toward Achieving National Targets for Reducing Coronary Heart Disease and Stroke Mortality: A County‐Level Perspective", Journal of the American Heart Association (JAHA), 2021Vol. 10, No. 4.
[2] Hian Chye Koh and Gerald Tan, "Data mining applications in healthcare", Journal of Healthcare Information Management (JHIM), 2005, Spring 2005; 19(2):64-72.
[3] K.Gomathi and D.ShanmugaPriyaa, "Heart Disease Prediction Using Data Mining Classification", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 4 Issue II, February 2016, ISSN: 2321-9653.
[4] Sebastian Raschka and Vahid Mirjalili, "Python Machine Learning",Third Edition, Packt Publishing, 2019.
[5] Sarath Babu et al., "Heart Disease Diagnosis Using Data Mining Techniques", International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2017, 978-1-5090-5686-6/17/$31.00 ©2017 IEEE 750.
[6] Anjan Nikhil Repaka et al., "Design And Implementing Heart Disease Prediction Using Naives Bayesian", Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019), IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8.
[7] Pratiksha Shetgaonkar and Dr. Shailendra Aswale, "Heart Disease Prediction using Data Mining Techniques", International Journal of Engineering Research & Technology (IJERT), 2021, ISSN :2278 0181.
[8] Barbara Martins et al., "Data Mining for Cardiovascular Disease Prediction", Journal of Medical Systems (2021) 45:6, doi.org/10.1007/s10916-020-01682-8.

