Effective Heart Disease Prediction using Machine Learning and Data Mining Techniques

Muhammad Zeeshan Younas

Department of Computer Science, Capital University of Science and Technology, Islamabad, Pakistan

---

***Abstract*--** Nowadays, heart disease is one of the prevailing main causes of morbidity and mortality. It is a hot health topic in our daily life, and heart disease treatment is very complicated. It is one-third of all deaths globally, stroke and heart disease. They both are globally the biggest killer, and their diagnosis availability is infrequent, especially in developing countries. This paper contains a framework based on some machine learning and data mining classification techniques on the heart disease dataset. There is no operational use of the data produced from the hospitals. Some convinced tools are used to extract the facts from the database to recognize the heart. This work is done by using Cleveland heart disease dataset that is sourced from the "UCI Machine Learning (ML) repository" to test and analyze on some various supervised ML and data mining techniques, some different attributes associated with causing of cardiovascular heart disease age, sex, chest pain type, chol, thal, etc. We will use these respective data to a model that will predict whether the patient has heart disease or not. This paper discussed the results of the modern techniques and will be used to predict the results for heart disease by summarizing some current research. The proposed method works best result in 86.89% accuracy by using a logistic regression algorithm.

***Keywords*- Machine Learning, Classification Techniques, Prediction, Data Mining, Heart Disease, Python Programming.**

I. INTRODUCTION

Data mining is a process that is used for mining information or knowledge from a huge database. It is an essential and significant step for discovering knowledge from existing databases. Data mining's primary task is that extract the hidden information and knowledge from the vast database. It is identified as Knowledge Discovery in Database (KDD). It is an important process where some common data mining techniques are used to extract the data arrangement. Data mining's technique helps to organizations to gain knowledge-based information. It includes understanding the business, data preparation, evaluating the data, and deployment. Its techniques work very rapidly and can find large amounts of data with the short passage of time. More likely, sometimes, it is referred to as knowledge discovery in databases. Suppose we use some professional and proficient computerized systems that are based on data mining and machine-learning algorithms. In that case, they can help us for achieving clinical assessments or diagnoses to minimize heart disease risk.

Machine learning is self-restraint that deals with programming, and it learns automatically and improves with experience. Bayesian and data mining analysis is trending, adding the demand for machine learning. Data mining has four different main techniques: cluster, Regression, Classification, and association rules. Classification is a fundamental technique in data mining. We can get the future outcome and predict the data based on historical data available in a database. The dataset can be classified into two categories through the classification technique, namely Yes and No. This method can achieve relevant and essential information for data and easily classify our data into different classes. "Data mining is the method for determining potentially useful arrangements through huge data sets and a large amount of database or metadata. It comes from different data sources, it may be sorted in various data warehouses and data mining sorting techniques" [2]. Knowledge Discovery in Database (KDD) is used for data integration and cleaning, data discovery patterns, Knowledge Presentation, data selection, and data transformation. Healthcare association produces broad data to mark the factual decisions.

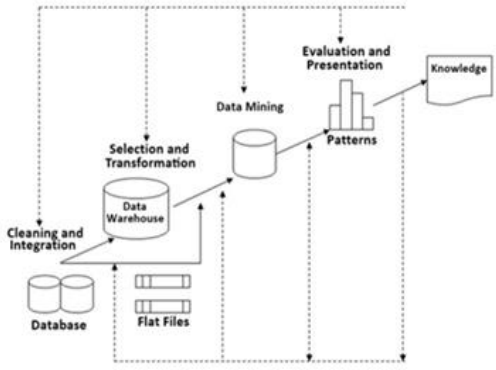

Figure 1. Process of knowledge discovery in data.

Data mining whole process is based on some various steps for Extracting respective knowledge. Data cleaning in data mining is how we can remove noise and corrupt or inaccurate records from data. We can prepare correct and complete data for data analysis by eliminating duplication in data through the data cleaning process. This data is usually not helpful when it comes to data analysis. The data cleaning process helps ensure that respective information is matched with the field and ensures data selection and transformation. The data transformation process is used to transform the data in a

proper way required by data mining procedures. The pattern evaluation is used to represents knowledge based on different measures of interest that are given. We can use other heart disease patients' data collected after some diagnosis analysis and utilize the experience and knowledge of several specialists split with the same symptoms of coronary heart diseases. Complete and correct data helps the diagnosis analysis of patients for providing efficient treatment.

This paper aims to identify and categorize some important feature classification using data mining and ML techniques to predict cardiovascular heart disease and supervised ML algorithms. The regression and classification model is the main model that is used in supervised machine learning, and this research work is based on some classifications models. This methodology predicts and compares the following two main machine learning classification and data mining classification algorithms, logistic regression, and naive Bayes classification to compare and confuse matrix. Cleveland dataset is selected, and this dataset is gathered from the UCI ML repository. These models are performed by using Python Programming Language. This paper discussed the results of the current technique and predicted the result for heart disease. Additionally, the experiment results compare the accuracy achieved by these algorithms and evaluated results by various respective authors.

II. HEART DISEASE: AN OVERVIEW

World Health Organization (WHO, 2017) every year, heart disease is becoming the cause of approximately 17.7 million deaths worldwide. It is one-third of all deaths in the whole world. Stroke and heart disease both are the global biggest killers. Suppose we enable these techniques for the medical diagnosis center. In that case, it will be more beneficial and minimize the overall cost after associating various data mining techniques for showing their appropriateness results [3]. Coronary Artery Disease (CAD) is the primary cause and widespread kind of cardiovascular heart disease. Coronary Artery Disease happens when the coronary arteries become narrowed, and the blood supply to the heart muscles is not enough.

Heart Disease is the most important and major cause of death worldwide nowadays. Coronary arteries are the structure or a network that is used for oxygen supply rich blood from the entire heart muscle. It may cause swear pain and heart attack. "We required very professional medical specialists for this cure because a diagnosis of the heart disease is not easy" [1]. Nowadays, heart disease is a hot health topic in our daily life, this type of disease that a cause of heart failure and affects the human heart and blood vessels. "A heart is the most important organ in body structure. For instance, if its working is not properly, it will become a cause and damage the other organ of the human body like coronary arteries, brain, kidney, etc. This risk factor of heart disease is increased by High blood pressure bp, Unhealthy diet, Smoking, High cholesterol, Diabetics, Consuming immense alcohol, coronary infection, being overweight, hypertension" [2].

Symptoms of Heart attack:

- Shortness of breath
- Pain may travel to the left or right arm or neck
- Fatigue
- Rapid or irregular heartbeat
- Cold sweat and unsteadiness
- Coughing or wheezing

Types of cardiovascular disease:

- Inherited heart disease.
- Heart attack.
- Stroke and more.
- Coronary artery disease (CAD) Coronary artery becomes narrowed
- Vascular disease (blood vessel disease).

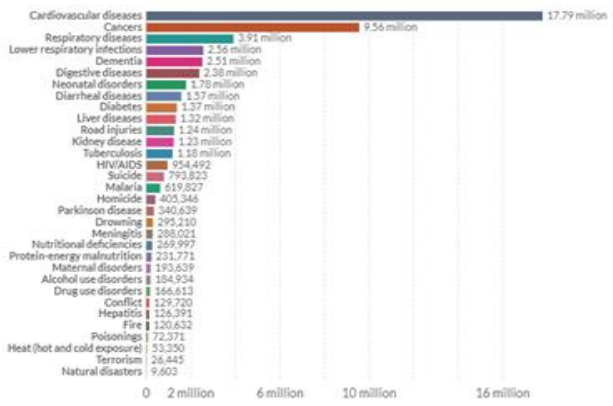

Figure 2. Number of Deaths by Cause Worldwide 2017, World Health Organization (WHO) 2018.

III. LITERATURE REVIEW

H. Benjamin et al. [4], in their work on the "supervised machine learning" concept used to find the predictions of heart disease, have used the following data mining classification algorithms for analysis and prediction, namely, Naïve Bayes, Random Forest, and Decision Tree. They have proposed by experimental results and proved that Random Forest gives better result performance as compare to Naïve Bayes and Decision tree. In this research work, the dataset is sourced from the data source StatLog for creating heart disease prediction.

Senthilkumar Mohan et al. [5], Hybrid Random Forest, and novel method by using Linear Model (HRFLM) and their goals to finding the important features by using Machine learning's techniques and increase the performance and accuracy for heart disease prediction. Research work core aims to process raw data through different steps and deliver a new respective novel judgment of heart disease prediction. Their prediction model is presented by various combinations of features and numerous recognized classification methods to increase the accuracy performance result. They have done work on many classification models to predict cardiovascular heart disease and compared their accuracy. They have proposed a comparison with HRFLM. Dataset used by UCI ML repository, and their approach claimed an accuracy level of 88.7%.

Mohammad Shafenoor Amin et al. [6], in their work they suggested data mining classification methods for predicting the heart disease result. The proposed testing was used to classify important features by using data mining techniques. The Cleveland dataset was collected from the UCI ML Repository for heart disease prediction. They have used some data mining techniques, namely SVM, DT, K-NN, LR, Naïve Bayes, vote, and Neural network. They have also performed experiments on another dataset using the UCI Statlog data set to identify the verdicts. The maximum accuracy gain results for heart disease diagnostic system can proficiently predict the danger level of heart disease in the future. Their approach claimed maximum accuracy was accomplished by SVM.

Ching-seh Mike Wu et al. [7], Heart disease is a regular problem globally, and the death rate is very high due to heart diseases and increases day by day. "Nowadays, cardiovascular heart disease (CHD) is the main cause of human deaths in the whole world. In this research work, they have used different classifier data mining techniques. Test dataset scraped from the UCI repository, and there are 13 attributes of the patient. They have tested various experiments. Logistic regression and Naive Bayes predicted the maximum accuracy when they used a huge dataset. Decision Tree and Random Forest give an enhanced result on the small dataset". They have proposed that a random forest shows well accuracy performance than a decision tree.

Jagdeep Singh et al. [8], They performed work by using different association and classification methods to predict the heart disease dataset gathered from the UCI ML Repositories. Apriori and FPGrowth are used here to find some heart disease dataset association rules to predict cardiovascular heart disease (CHD Cleveland dataset used here, a total of 313 occurrences and 13 attributes. Core work of this research work is present to attain high accuracy for earlier diagnoses of (CHD), They have proposed hybrid associative classification using Waikato Environment for Knowledge Analysis (WEKA) environment and claimed that the highest accuracy was achieved by using IBK (Nearest Neighbor) with Apriori associative algorithms.

Nathaniel David Oye et al. [9], "Heart Disease Prediction using Machine Learning and Data Mining Techniques," namely as the Decision Tree, Naive Bayes, and Artificial Neural Network (ANN) to predict heart diseases. They observed that most studies based on the Cleveland heart disease (CHD) data set normally hold 303 occurrences and 13-14 attributes. According to their research work and observation, this dataset is so small and restricted with limited heart disease features. They proposed that there should be a further composite model that joins many geographical data sources to maximize the precision of predicting the primary trending of heart disease.

Abhishek Rairikar et al. [10], they have suggested a well-organized method for predicting heart disease, they applied different data mining techniques, namely KNN, Decision trees (DT), and Naive Bayes (NB). They built an effective method for Diagnosing heart attack results through GUI form. They proposed from results that KNN provides better accuracy than Naive Bayes and Decision tree.

IV. METHODOLOGY

A. Data Source

This research work dataset is sourced from the UCI ML repository. The following four databases in UCI ML Repository are Switzerland, Hungary, Cleveland, and the VA LB. Cleveland database is mostly used here in this research because the Cleveland database is the most useable database by ML researchers and with complete records. The dataset contains 303 instances, and this database contains 76 attributes with the suitable 14 clinical parameters [14]. There are total 14 attributes, but 1 attribute is used as the projected attribute for heart disease. A dataset's clinical attribute is referred to as tests related to heart disease, i.e., chest pain (cp) type, blood pressure (bp), blood sugar level, electrocardiographic result, etc. All attributes and features with their descriptions values are shown in Table 1. After data pre-processing, data has been converted from Numeric to Nominal. In Figure 4, the percentage of patients who have not heart disease is 45.54%, and the percentage of patients having heart disease is 54%.46.

TABLE 1. ATTRIBUTES AND DESCRIPTION OF THE DATASET FROM UCI CLEVELAND DATASET

S. NO.	Attribute name	Description	Type
1	Age	Patients age in years	Numeric
2	Sex	1 for male 0 for female	Nominal
3	Cp	Chest pain type	Nominal
4	Trestbpd	Resting blood pressure: 92-200	Numeric
5	Chol	Serum cholesterol in mg/dl	Numeric
6	Fbs	Fasting blood sugar level 1 if true 0 if false	Nominal
7	Restecg	Resting electrocardiographic results in 3 values; Value 0: normal Value 1: having ST-T wave abnormality Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria	Nominal
8	Thalach	Maximum heart rate achieved	Numeric
9	Exang	Exercise induced angina (1 for yes and 0 for no)	Nominal
10	Oldpeak	ST depression induced by exercise 71-202	Numeric
11	peakSlope	The slope of the peak exercise ST segment Value 1: up sloping Value 2: flat Value 3: down sloping	Numeric
12	Ca	Number of major vessels (0-3) colored by fluoroscopy	Numeric
13	Thal	The heart status is described with 3 values. Value 3: normal Value 6: fixed defect Value 7: reversible defect	Nominal
14	Disease	It represents the diagnosis of heart disease with 5 values. 0 meaning absence 1-4 indicate presence of heart disease	Nominal

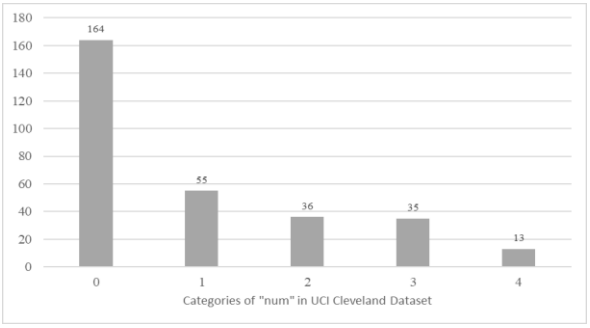

Figure 3: Distribution of "numbers" in UCI ML Cleveland dataset.

B. Architecture Diagram:

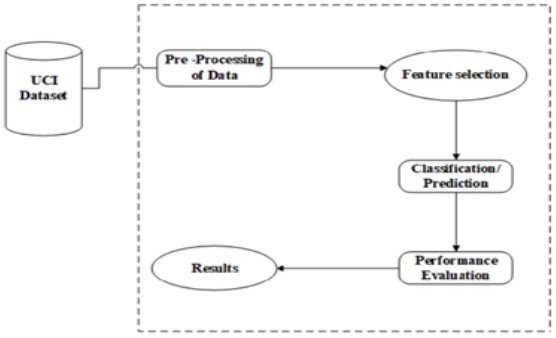

Figure 4. Experiment workflow with UCI dataset.

C. Description of Algorithms

The following six data mining classification techniques used here, namely Decision Tree, SVM, K-NN, Naive Bayes, Logistic Regression, and Random Forest, are used to analyze the dataset.

a) Decision Tree:

It is based on a supervised learning technique and decision tree used for classification and regression models. This is a very common algorithm used for classifications [13]. The decision tree aims to generate a model for predicting the value of an objective variable. Decision tree flowchart-like structure helps us decide by learning some simple decision rules. Here each node acts as a test case for some features and each leaf node provides the outcome and it shows definite results like true or

false and 1 or 0 etc.

b) *K-Nearest Neighbors (KNN):*

It is a supervised ML algorithm that is used for regression and classification problems. It is normally used for classification predictive problems. KNN helps us to classify the data into various groups. KNN implementation is very simple, but it is a lazy learning algorithm and creates no earlier supposition.

c) *Support Vector Machine (SVM):*

The main motive of SVM is to discover the hyperplane which divides the two classes of the dataset it is an ML algorithm and is used to categorize the dataset. SVM sorts the data into one of two categories. A hyperplane is divided into two classes with the maximum distance it is known as ideal hyperplane of the form f(x) = (wt x + b), It will show low performance if the given data is noisy.

d) *Random Forest:*

It is a cooperative learning method for classification it consists of many decision trees based on the parent's tree and Integrates all of them to get the best results. It can handle a huge amount of data easily and efficiently work on large data. For instance a given data, X = { x1, x2, ….. , xn} with reactions to Y = { x1, x2, ….. , xn} and it recurrences the getting from b = 1 – B.

V. PROPOSED METHOD

Following assorted earlier studies [1, 6, 4, 11], various authors have discussed predicting the significant features of heart disease prediction by using different machine learning and data mining techniques. We proposed a Logistic regression machine learning technique for heart disease prediction of significant features. This proposed model for the heart disease prediction method is introduced for deep learning algorithms and perspectives. After pre-processing the dataset Logistic Regression, a data mining classification technique was applied here by using the Sklearn library to analyze the score. Implementation of the Naïve Bayes method of getting accuracy results, and this classification results section done by using Python. Finally, at the end, compare the Comparing Model and Confusion Matrix results. Firstly we imported the data that contain different variable like gender, age, cp(chest pain), sex, slope, target, etc. After the accessibility of the data, we created a predictive model based on the Logistic Regression algorithm. This classified data based on various organized features of heart disease patients. Create a Logistic Regression model with the help of temporary variables and used the sigmoid function for graphical representation classified dataset.

a) Logistic Regression:

The logistic regression is a classification algorithm of Machine Learning (ML). It is used to predict the probability of the dependent variable. It also provides high accuracy and here first of all data should be imported and then it can be trained for prediction. It is logistic regression, and it is also presented by sigmoid functions which help to show a good representation of the graph. The dependent variables are the binary variables that hold the coded data as 1(good, yes, pass, etc.) or 0 (bad, no, fail, etc.).

b) Naïve Bayes:

It is a supervised classifier algorithm of ML that categorizes the dataset. It is used to classify data into predefined classes. It uses conditional probability to classify the test dataset and this model applies Bayes rules by independent features. Naïve Bayes classifier finds the probability of each feature. It requires a small number of datasets and fasts to predict the class of test data. It does not show results if the features are correlated [14].

VI. RESULTS AND DISCUSSIONS

The proposed system has a Cleveland Heart Disease dataset which is used to classified whether the patients have heart disease or not according to their features. The overall records in the dataset are distributed into two category training and testing data sets. Logistic regression can also be considered a sigmoid function. The sigmoid function is usually used to denote precisely the logistic function; by using this, we can easily represent our graphs and charts any real value to the range 1 or 0. The proposed system is applied to this data and tries to create an accurate model that predicts (Data exploration and reading data) if the patients have this disease or not. Figure 7 shows the whole process of fixing the targets of heart disease. The percentage of patients having heart disease is 54.46%, and the percentage of patients who have not heart disease is 45.54%, where 0 shows the absence and 1 shows the presence of heart disease. Cleveland Heart Disease dataset is also classified into two categories for male and female patients. Male patients are 68.32%, and females are 31.68%.

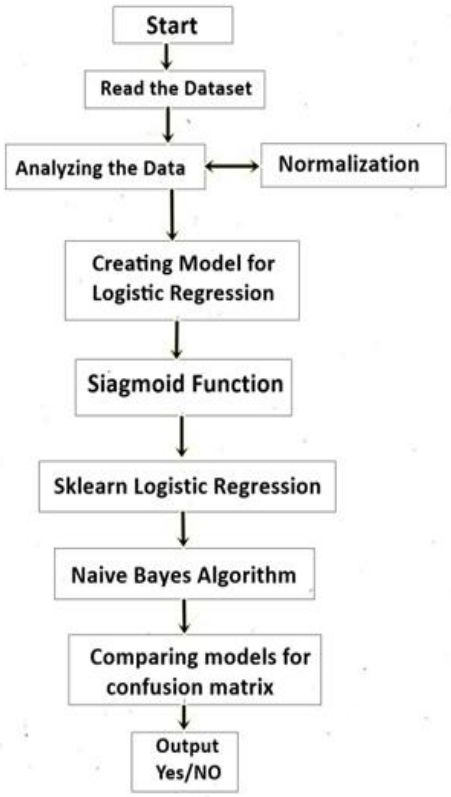

Figure 5. Implementation of Logistic Regression Algorithm

```
['heart.csv', 'heartDiseaseAndAges.png', 'k.py']
    age  sex  cp  trestbps  chol  fbs  ...  exang  oldpeak  slope  ca  thal  target
0    63    1   3       145   233    1  ...      0      2.3      0   0     1       1
1    37    1   2       130   250    0  ...      0      3.5      0   0     2       1
2    41    0   1       130   204    0  ...      0      1.4      2   0     2       1
3    56    1   1       120   236    0  ...      0      0.8      2   0     2       1
4    57    0   0       120   354    0  ...      1      0.6      2   0     2       1
5    57    1   0       140   192    0  ...      0      0.4      1   0     1       1
6    56    0   1       140   294    0  ...      0      1.3      1   0     2       1
7    44    1   1       120   263    0  ...      0      0.0      2   0     3       1
8    52    1   2       172   199    1  ...      0      0.5      2   0     3       1
9    57    1   2       150   168    0  ...      0      1.6      2   0     2       1
10   54    1   0       140   239    0  ...      0      1.2      2   0     2       1
11   48    0   2       130   275    0  ...      0      0.2      2   0     2       1
12   49    1   1       130   266    0  ...      0      0.6      2   0     2       1
13   64    1   3       110   211    0  ...      1      1.8      1   0     2       1
14   58    0   3       150   283    1  ...      0      1.0      2   0     2       1
15   50    0   2       120   219    0  ...      0      1.6      1   0     2       1
16   58    0   2       120   340    0  ...      0      0.0      2   0     2       1
17   66    0   3       150   226    0  ...      0      2.6      0   0     2       1
18   43    1   0       150   247    0  ...      0      1.5      2   0     2       1
19   69    0   3       140   239    0  ...      0      1.8      2   2     2       1
20   59    1   0       135   234    0  ...      0      0.5      1   0     3       1
21   44    1   2       130   233    0  ...      1      0.4      2   0     2       1
22   42    1   0       140   226    0  ...      0      0.0      2   0     2       1
23   61    1   2       150   243    1  ...      1      1.0      1   0     2       1
24   40    1   3       140   199    0  ...      1      1.4      2   0     3       1
```

Figure 6. Reading Heart disease data set

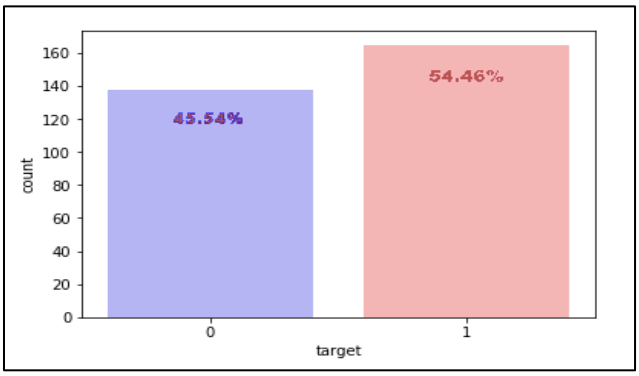

Figure 7. Fixing the respective targets about heart disease

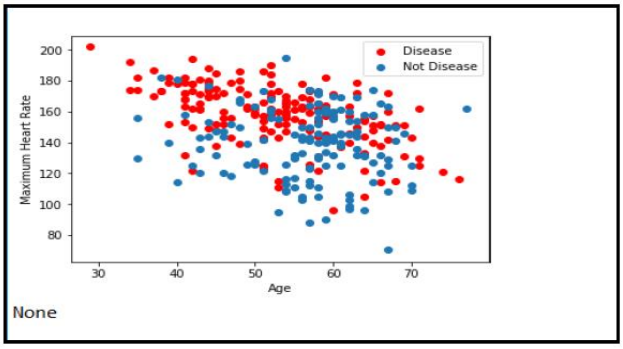

Figure 8. Classifying the heart disease

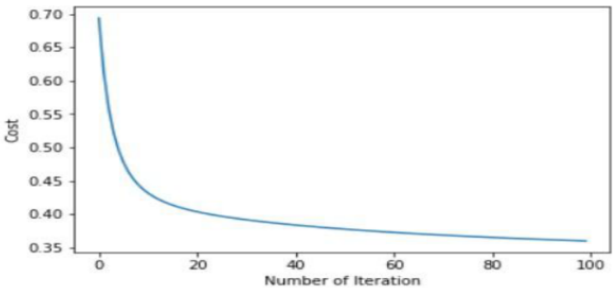

Figure 9. Accuracy output in logistic regression algorithm

A confusion matrix is normally used to measure the performance of respective algorithms, and it is also used to measure the performance of classification. There are some following the most basic terms for the confusion matrix.

TP (True Positive): It means the amount of records relates to yes, they have the disease.

TN (True Negative): it means the amount of the record relates to no. They don't have the disease, we predicted no.

FP (False Positives): We predicted yes, but they don't actually have the disease.

FN (False Negative): We predicted no, but they actually do have the disease.

Table 2: Confusion Matrix

	Predicted Negative(0)	Predicted Positive(1)
Actual Negative(0)	TP	FP
Actual Positive(1)	FN	TN

Table 03: Confusion matrix for Logistic Regression

	Class 0	**Class 1**
Class 0	188	22
Class 1	11	91

Training Set

	Class 0	**Class 1**
Class 0	32	9
Class 1	3	17

Testing Set

a. Logistic Regression accuracy for training set ((118+91)/(11+22+118+91))*100 = **86.89%**

b. Logistic Regression accuracy for the testing set ((32+17)/(3+9+32+17))*100 = **80.32%**

VII. EVALUATION RESULTS

The proposed method works best result in 86.89% accuracy by using a logistic regression algorithm. This work is done by some different steps shown in Figure 4. The heart disease prediction model's accuracy developed using 14 significant attributes that are defined in Table 1. And table 2 summarized the accuracy results of specific classification models that are obtained from proposed methods and by other various authors. Experiment works display that the heart disease prediction model developed using the identified significant features and the best-performing technique (Logistic regression) achieves an accuracy of 86.89% in heart disease prediction.

TABLE 4. Accuracy comparison on the heart diseases dataset by various authors with proposed model.

Author	**Techniques & Methodology**	**Accuracy %**
H.Benjamin Fredrick David et.al [4]	Random Forest	81
S Anitha et.al [11]	KNN	76.67
Senthilkumar Mohan et.al [5]	Decision Tree	78.69
M.A.Jabbar et.al [1]	Naïve Bayes	78.56
M. Shafenoor Amin et.al [6]	SVM	80.98
Proposed Method	Logistic Regression	**86.89**

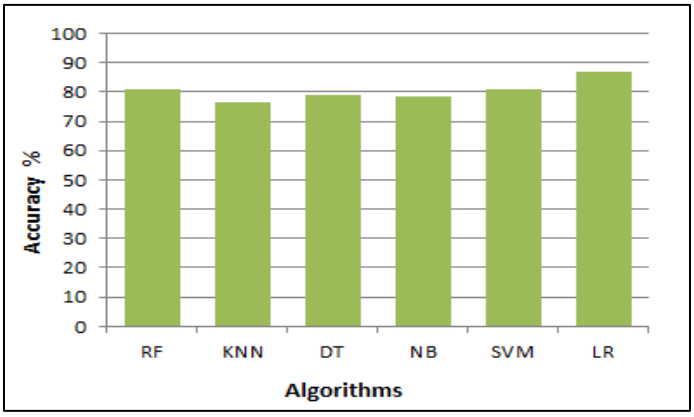

Figure 10. Accuracy Representation in Graph with their respective algorithms

VIII. CONCLUSION

In this research work, various machine learning & data mining classification techniques are used to analyze and predict heart disease accuracy. Heart disease nowadays is a hot health topic in our daily life and it is one-third of all deaths in the whole world. The proposed system applied on this data and tried to create an accurate model that predicts (Data exploration and reading data) if the patients have this disease. The core purpose of this work is the prediction of

heart disease with a maximum amount of accuracy. Dataset is collected from UCI ML repository. Cleveland database is used here for heart disease prediction. After pre-processing the dataset Logistic Regression, a data mining classification technique was applied here by using the Sklearn library to analyze the score. Proposed approach Logistic regression achieved an accuracy of level of 86.89%.

IX. FUTURE SCOPE

In the future, the proposed system with data mining and ML classification algorithm can be used for the prediction of other diseases in the medical field. It provides us good accuracy by observing some different research papers. And there are many strategies to improve this research and address the boundaries of this study. This research work can be extended by conducting the same experiment on a large-scale real-life dataset. This work's coming future scope is the prediction of heart diseases by using innovative techniques and algorithms in minimum time complexity.

REFERENCES

[1] M.A.Jabbar, B.L. Deekshatulu and Priti Chandra, 2015. Prediction of heart disease using Random forest and Feature subset selection, AISC SPRINGER, vol 424, pp187-196.

[2] Mr.Santhana Krishnan.J and Dr.Geetha.S, 2019. Prediction of Heart Disease Using Machine Learning Algorithms, (ICIICT) IEEE, 2019. DOI: 10.1109/ICIICT1.2019.8741465.

[3] Cincy Raju, Philipsy E, Siji Chacko, L Padma Suresh, Deepa Rajan S, 2018. A Survey on Predicting Heart Disease using Data Mining Techniques, IEEE Conference on Emerging Devices and Smart Systems (ICEDSS 2018).DOI 10.1109/ICEDSS.2018.8544333.

[4] H.Benjamin Fredrick David and S. Antony Belcy, 2018. Heart Disease Prediction Using Data Mining Techniques, Ictact Journal On Soft Computing, Volume: 09, Issue: 01. DOI: 10.21917/ijsc.2018.0253.

[5] Senthilkumar Mohan, Chandrasegar Thirumalai and Gautam Srivastava, 2019. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, Computer Science (IEEE Access) Vol 7, pp 81542-81554(2019).DOI 10.1109/ACCESS.2019.2923707.

[6] Mohammad Shafenoor Amin, Yin Kia Chiam and Kasturi Dewi Varathan, 2019. Identification of significant features and data mining techniques in predicting heart disease, Telematics and Informatics, Vol 36, pp 82-93. DOI 10.1016/j.tele.2018.11.007 .

[7] Ching-seh (Mike) Wu, Mustafa Badshah and Vishwa Bhagwat, 2019, Heart Disease Prediction Using Data Mining Techniques, 2nd International Conference on Data Science and Information Technology July 2019, Pp 7–11, DOI 10.1145/3352411.3352413.

[8] Jagdeep Singh, Amit Kamra and Harbhag Singh, 2016. Prediction of Heart Diseases Using Associative Classification, 5th International Conference on Wireless Networks and Embedded Systems (WECON). DOI 10.1109/WECON.2016.7993480.

[9] Lamido Yahaya, Nathaniel David Oye and Etemi Joshua Garba, 2020. A Comprehensive Review on Heart Disease Prediction Using Data Mining and Machine Learning Techniques, American Journal of Artificial Intelligence, Volume 4, Pages: 20-29. DOI: 10.11648/j.ajai.20200401.12.

[10] Abhishek Rairikar, Vedant Kulkarni, Vikas Sabale, Harshavardhan Kale and Anuradha Lamgunde 2017 ,Heart Disease Prediction Using Data Mining Techniques, (IEEE) International Conference on DOI: 10.1109/I2C2.2017.8321771.

[11] S Anitha and NSridevi, 2019. Heart disease prediction using data Mining Techniques, Journal of Analysis and Computation, hal-02196156.

[12] J. Ross Quinlan, 1986. Induction of Decision Trees, Machine Learning, Vol. 1, No. 1, pp. 81-106,.

[13] M.A. Jabbar, 2018. Heart disease prediction system based on hidden naviebayes classifier. International conference on circuits, controls, communications and computing(14C)

[14] Cox, David R., 1958. The regression analysis of binary sequences. Journal of the Royal Statistical Society. Series B (Methodological) 215-242

[15] https://archive.ics.uci.edu/ml/machine-learning databases/heart-disease/