A MINI PROJECT REPORT

ON

"Heart Disease Prediction Model Using MachineLearning"

Submitted in the partial fulfilment of the requirements for the degree

of

BACHELOR OF ENGINEERING IN COMPUTER ENGINEERING

By

1) **01_PRATHAMESH AHIR**
2) **15_SHUBHAM POKALE**
3) **17_KAUSHAL SAWANT**
4) **22_FURQUAN THOKAN**

UNDER THE GUIDANCE OF

Dr. Anjali Dadhich

Department of Computer Engineering
Saraswati College of Engineering, Kharghar, Navi Mumbai

University of Mumbai
2021-22

Saraswati College of Engineering, Kharghar

Vision:

To be universally accepted as autonomous center of learning inEngineering Education and Research.

Mission:

- To educate students to become responsible and quality technocrats tofulfil society and industry needs.
- To nurture student's creativity and skills for taking up challenges in allfacets of life.

Department of Computer Engineering

Vision:

To be among renowned institution in Computer Engineering Education and Research by developing globally competent graduates.

Mission:

- To produce quality Engineering graduates by imparting quality training, hands on experience and value education.
- To pursue research and new technologies in Computer Engineering and across interdisciplinary areas that extends the scope of Computer Engineering and benefit humanity.

- To provide stimulating learning ambience to enhance innovative ideas, problem solving ability, leadership qualities, team-spirit and ethical responsibilities.

DEPARTMENT OF COMPUTER ENGINEERING PROGRAM EDUCATIONAL OBJECTIVE'S

1. To embed a strong foundation of Computer Engineering fundamentalsto identify, solve, analyze and design real time engineering problems as a professional or entrepreneur for the benefit of society.
2. To motivate and prepare students for lifelong learning & research to manifest global competitiveness.
3. To equip students with communication, teamwork and leadership skills to accept challenges in all the facets of life ethically.

DEPARTMENT OF COMPUTER ENGINEERING

PROGRAM OUTCOMES

1. Apply the knowledge of Mathematics, Science and Engineering Fundamentals to solve complex Computer Engineering Problems.
2. Identify, formulate and analyze Computer Engineering Problems and derive conclusion using First Principle of Mathematics, Engineering Science and Computer Science.
3. Investigate Complex Computer Engineering problems to find appropriate solution leading to valid conclusion.
4. Design a software System, components, Process to meet specified needs with appropriate attention to health and Safety Standards, Environmental and Societal Considerations.
5. Create, select and apply appropriate techniques, resources and advance Engineering software to analyze tools and design for Computer Engineering Problems.
6. Understand the Impact of Computer Engineering solution on society and environment for Sustainable development.
7. Understand Societal, health, Safety, cultural, Legal issues and Responsibilities relevant to Engineering Profession.
8. Apply Professional ethics, accountability and equity in Engineering Profession.

9. Work Effectively as a member and leader in multidisciplinary team fora common goal.
10. Communicate effectively within a Profession and Society at large.
11. Appropriately incorporate principles of Management and Finance inone's own Work.
12. Identify educational needs and engage in lifelong learning in aChanging World of Technology.

DEPARTMENT OF COMPUTER ENGINEERING

PROGRAM SPECIFIC OUTCOME

1. Formulate and analyze complex engineering problems in computer engineering (Networking/Big data/ Intelligent Systems/Cloud Computing/Real time systems).
2. Plan and develop efficient, reliable, secure and customized application software using cost effective emerging software tools ethically.

SARASWATI Education Society's

SARASWATI College of Engineering

Learn Live Achieve and Contribute

Kharghar, Navi Mumbai - 410 210.

(Approved by AICTE, recg. By Maharashtra Govt. DTE, Affiliated to Mumbai University)

PLOT NO. 46/46A, SECTOR NO 5, BEHIND MSEB SUBSTATION, KHARGHAR, NAVI MUMBAI-410210

Tel.: 022-27743706 to 11 *Fax: 022-27743712 * Website: www.sce.edu.in

CERTIFICATE

*This is to certify that the requirements for the mini project report entitled "**Heart Disease Prediction using Machine Learning**" have been successfully completed by the following students:*

Roll numbers	Name
01	Prathamesh Ahir
15	Shubham Pokale
17	Kaushal Sawant
22	Furquan Thokan

In partial fulfillment of Sem –VI, **Bachelor of Engineering of Mumbai University in Computer Engineering** of Saraswati College of Engineering, Kharghar during the academic year 2021-22.

Internal Guide
Dr. Anjali Dadhich

External Examiner

Mini Project Co-Ordinator
Dr. Anjali Dadhich

Head of Department
Prof. Sujata Bhairnallykar

DECLARATION

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included. I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

1. Prathamesh Ahir
2. Shubham Pokale
3. Kaushal Sawant
4. Furquan Thokan

Date:

ACKNOWLEDGEMENT

After the completion of this work, words are not enough to express feelings about all those who helped us to reach goal.

It's a great pleasure and moment of immense satisfaction for us to express my profound gratitude to **MiniProject Guide**, **Dr. Anjali Dadhich** whose constant encouragement enabled us to work enthusiastically. His perpetual motivation, patience and excellent expertise in discussion during progress of the project work have benefited us to an extent, which is beyond expression.

We would also like to give our sincere thanks to **Prof. Sujata Bhairnallykar, Head of Department**, and **Dr. Anjali Dadhich, MiniProject co-coordinator** from Department of Computer Engineering, Saraswati college of Engineering, Kharghar, Navi Mumbai, for their guidance, encouragement and support during a project.

I am thankful to **Dr. Manjusha Deshmukh, Principal,** Saraswati College of Engineering, Kharghar, Navi Mumbai for providing an outstanding academic environment, also for providing the adequate facilities.

Last but not the least we would also like to thank all the staffs of Saraswati college of Engineering (Computer Engineering Department) for their valuable guidance with their interest and valuable suggestions brightened us.

1. Prathamesh Ahir.
2. Shubham Pokale
3. Kaushal Sawant
4. Furquan Thokan

ABSTRACT

Heart disease, alternatively known as cardiovascular disease, encases various conditions that impact the heart and is the primary basis of death worldwide over the span of the past few decades. It associates many risk factors in heart disease and a need of the time to get accurate, reliable, and sensible approaches to make an early diagnosis to achieve prompt management of the disease. Data mining is a commonly used technique for processing enormous data in the healthcare domain. Researchers apply several data mining and machine learning techniques to analyze huge complex medical data, helping healthcare professionals to predict heart disease. This research paper presents various attributes related to heart disease, and the model on basis of supervised learning algorithms as LogisticRegression, decision tree, and random forest algorithm. It uses the existing dataset from the Cleveland database of UCI repository of heart diseasepatients. The dataset comprises 303 instances and 76 attributes. Of these 76 attributes, only 14 attributes are considered for testing, important to substantiate the performance of different algorithms. This research paper aims to envision the probability of developing heart disease in the patients.

Keywords: Machine Learning , Heart Disease , Prediction.

Table of Contents

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1. GENERAL:

In day-to-day , life many factors affect a human heart. Many problems are occurring at a rapid pace and new heart diseases are rapidly being identified. In today's world of stress Heart, being an essential organ in a human body which pumps blood through the body for the blood circulationis essential and its health is to be conserved for a healthy living. The healthof a human heart is based on the experiences in a person's life and is completely dependent on professional and personal behaviors of a person. There may also be several genetic factors through which a type of heart disease is passed down from generations. According to the World Health Organization, every year more than 12 million deaths are occurring

worldwide due to the various types of heart diseases which is also known by the term cardiovascular disease. The term Heart-disease includes many diseases that are diverse and specifically affect the heart and the arteries of a human being. Even young aged people around their 20-30 years of lifespan are getting affected by heart diseases. The increase in the possibility of heart disease among young may be due to the bad eating habits, lack of sleep, restless nature, depression and numerous other factors such as obesity, poor diet, family history, high blood pressure, high blood cholesterol, idle behavior, family history, smoking and hypertension. The diagnosis of the heart diseases is a very important and is itself the most complicated task in the medical field. All the mentioned factors are taken into consideration when analyzing and understanding the patients by the doctor through manual check-ups at regular intervals of time. The symptoms of heart disease greatly depend upon which of the discomfort felt by an individual. Some symptoms are not usually identified by the common people. However, common symptoms include chest pain, breathlessness, and heart palpitations. The chest pain common to many types of heart disease is known as angina, or angina pectoris, and occurs when a part of the heart does not receive enoughoxygen. Angina may be triggered by stressful events or physical exertion and normally lasts under 10 minutes. Heart attacks can also occur as a resultof different types of heart disease. The signs of a heart attack are similar to angina except that they can occur during rest and tend to be more severe.

The symptoms of a heart attack can sometimes resemble indigestion. Heartburn and a stomach ache can occur, as well as a heavy feeling in the chest. Other symptoms of a heart attack include pain that travels through the body, for example from the chest to the arms, neck, back, abdomen, or jaw, lightheadedness and dizzy sensations, profuse sweating, nausea and vomiting. Heart failure is also an outcome of heart disease, and breathlessness can occur when the heart becomes too weak to circulate blood. Some heart conditions occur with no symptoms at all, especially in older adults and individuals with diabetes. The term 'congenital heart disease' covers a range of conditions, but the general symptoms include sweating, high levels of fatigue, fast heartbeat and breathing, breathlessness, chest pain. However, these symptoms might not develop until a person is older than 13 years. In these types of cases, the diagnosis becomes an intricate task requiring great experience and high skill. A risk of a heart attack or the possibility of the heart disease if identified early, can help the patients take precautions and take regulatory measures. Data mining refers to the extraction of required information from huge datasets in various fieldssuch as the medical field, business field, and educational field. Machine learning is one of the most rapidly evolving domains of artificial intelligence. These algorithms can analyze huge data from various fields, one such important field is the medical field. It is a substitute to routine prediction modelling approach using a computer to gain an understanding of complex and nonlinear interactions among different factors by reducing the errors in predicted and factual outcomes. Data mining is exploring huge datasets to extract hidden crucial decision- making information from a collection of a past repository for future analysis. The medical field comprises tremendous data of patients. These data need mining by various machine learning algorithms. Healthcare professionals do analysis of these data to achieve effective diagnostic decision by health- care professionals.

Medical data mining using classification algorithms provides clinical aid through analysis. It tests the classification algorithms to predict heart disease in patients. Data mining is the process of extracting valuable data and information from huge databases. Various data mining techniques such as regression, clustering, association rule and classification techniques like, decision tree, random forest and are used to classify various heart disease attributes in predicting heart disease. A comparative analysis of the classification techniques is used. In this research, dataset from the UCI repository is taken. The classification model is developed using classification algorithms for prediction of heart disease.

1.2. OBJECTIVE AND PROBLEM STATEMENT:

Many hospital information systems are designed to support patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but they are largely limited. They can answer simple queries like "What is the average age of patients who have heart disease?", "How many surgeries had resulted in hospital stays longer than 10 days?", "Identify the female patients who are single, above 30 years old, and who have been treated for cancer." However, they cannot answer complex queries like "Given patient records on cancer, should treatment include chemotherapy alone, radiation alone, or both chemotherapy and radiation?", and "Given patient records, predict the probability of patients getting a heart disease." Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge-rich data hidden in the database.

This practice leads to unwanted biases and errors which affects the quality of service provided to patients. It was proposed that integration of clinical decision support with computer-based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome. This suggestion is promising as data modelling and analysis tools, e.g., data mining, have the potential to generate acknowledge -rich environment which can help to significantly improve the quality of clinical decisions. Most hospitals today employ some sort of hospital information systems to manage their healthcare or patient data.

These systems typically generate huge amounts of data which take the formof numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making. There is a wealth of hidden information in these data that is largely untapped.

CHAPTER 2
METHODOLOGY

2.1 ALGORITHMIC DETAILS:

Use Case Data Flow Diagrams: -

2.1.1 BLOCK DIAGRAM:

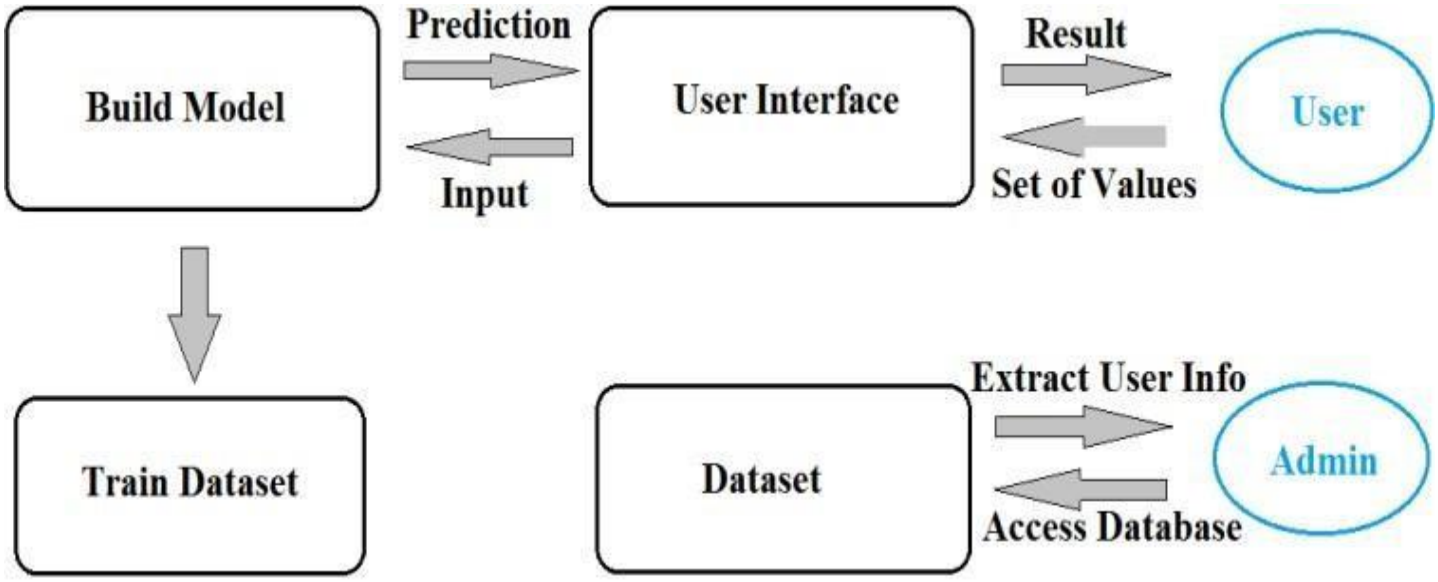
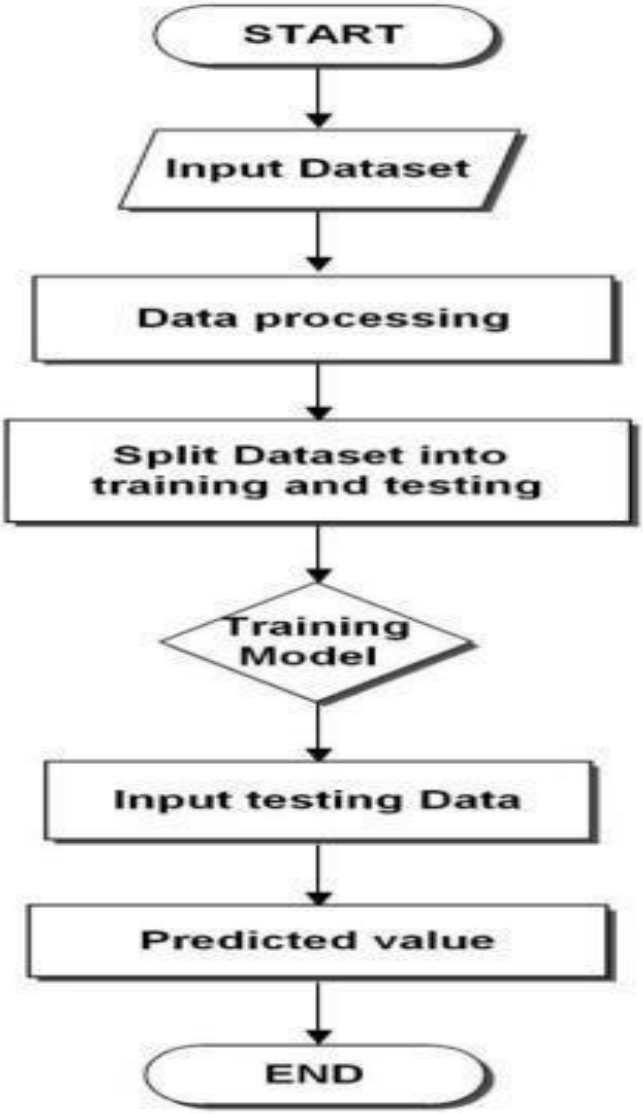


Fig. 2.1.1 Block Diagram

2. FLOW CHART



2.1.2 Flowchart

2.2. HARDWARE AND SOFTWARE REQUIREMENTS:

This describes the hardware and software requirements of the system

2.2.1. Hardware Requirements:

- Intel core i3 10th generation is used as a processor because it is mostlyavailable in every computer and we can run our pc for long time. We can keep developing the project anytime we want.
- RAM 4 GB is used as it'll provide fast and smooth reading andwriting.

2.2.2. Software Requirements:

Jupyter Notebook:

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaningand transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more.

Python:

Python is an easy to learn, powerful programming language. You can use Python when your data analysis tasks need to be integrated with web apps or if statistics code needs to be incorporated into a production database. Being a full-fledged programming language, Python is a great tool to implement algorithms for production use. There are several Python packages for basic data analysis and machine learning. In this free course, you will Learn about two popular packages in Python: NumPy and Pandas. These are the essential foundational packages that are required for basic data manipulation.

Libraries Used:

Pandas:

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

NumPy:

NumPy can be used to perform a wide variety of mathematical operations on arrays. It adds powerful data structures to Python that guarantee efficient calculations with arrays and matrices and it supplies an enormous library of high-level mathematical functions that operate onthese arrays and matrices.

Matplotlib:

Matplotlib is an amazing visualization library in Python for 2Dplots of arrays. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram.

Scikit learn:

Scikit-learn is a machine learning library for Python. It features various algorithms like support vector machine, random forests, and k- neighbors, and it also supports Python numerical and scientific libraries like NumPy and SciPy.

CHAPTER 3
IMPLEMENTATION AND RESULTS

3.1 IMPLEMENTATION:

Data Source: For this study, dataset from UCI Machine learning repository is used. It comprises a real dataset of 303 examples of data with 14 various attributes (13 predictors; 1 class) like blood pressure, type of chest pain, electrocardiogram result, etc. In this research, we have used four algorithms to get reasons for heart disease and create a model with the maximum possible accuracy.

Data Pre-processing : The real-life information contains large numbers with missing and noisy data. These data are pre-processed to over- come such issues and make predictions vigorously. Cleaning the collected data usually has noise and missing values. To get an accurate and effective result, this data needs to be cleaned in terms of noise and missing values are to be filledup. Transformation it changes the format of the data from one form to another to make it more comprehensible. It involves smoothing, normalization, and aggregation tasks.

Logistic Regression Logistic regression is a process of modelling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on. The primary difference between linear regression and logistic regression is that logistic regression's range is bounded between 0 and 1.

3.1.1 Types of Logistic Regression:

1. Binary Logistic Regression
2. Multinomial Logistic Regression
3. Ordinal Logistic Regression

We are using first type i.e., Binary Logistic Regression. where p(x)/(1 - p(x)) is termed odds, and the left-hand side is called the logit or log-odds function. The odds are the ratio of the chances of success to the chances of failure. As a result, in Logistic Regression, a linear combination of inputs is translated to log(odds), with an output of

$$P(X) = \frac{e^{\beta 0+\beta 1X}}{1+e^{\beta 0+\beta 1X}}$$ Equation no 1

This is the Sigmoid function, which produces an S-shaped curve. It always returns a probability value between 0 and 1. The Sigmoid function is used toconvert expected values to probabilities. The function converts any real number into a number between 0 and 1. The mathematically sigmoid function can be.

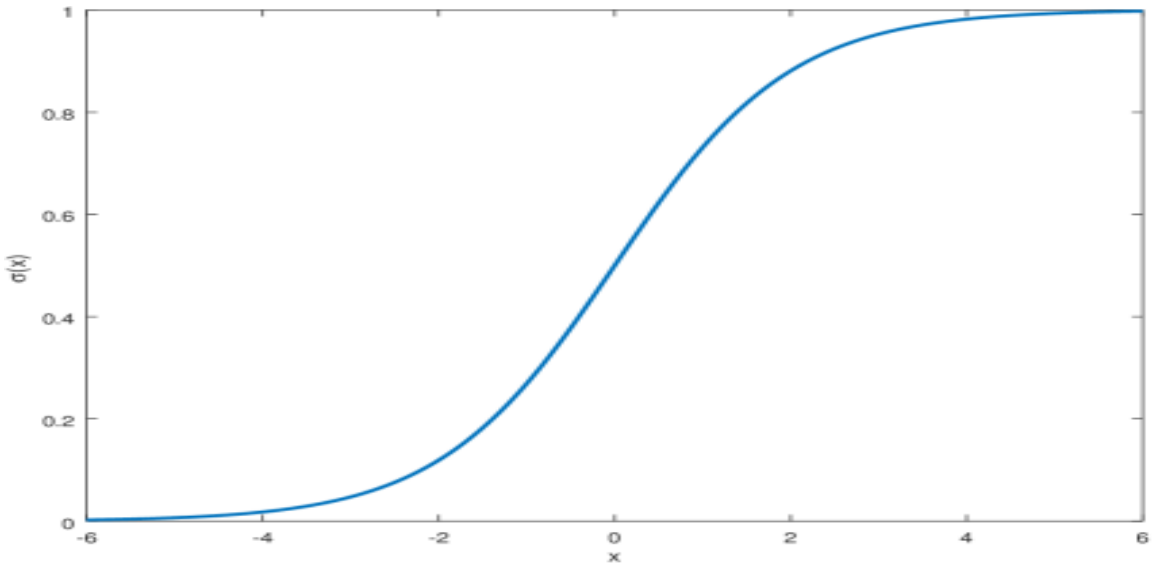


Figure 3.1 Graphical Representation of Sigmoid Function.

3.1.2 DECISION TREE:

Decision tree is a classification algorithm that works on categorical as wellas numerical data. Decision tree is used for creating tree-like structures.
Decision tree is simple and widely used to handle medical dataset. It is easyto implement and analyze the data in tree-shaped graph. The decision tree model makes analysis based on three nodes.

- Root node: main node, based on this all other nodes functions.
- Interior node: handles various attributes.
- Leaf node: represent the result of each test.
- This algorithm splits the data into two or more analogous sets based on the most important indicators. The entropy of each attribute is calculated and then the data are divided, with predictors having maximum information gainor minimum entropy:

- $(S) = \sum_{i=1}^{c} -Pi \log_2 Pi$ ……. Equation no 2

- The results obtained are easier to read and interpret. This algorithm has higher accuracy in comparison to other algorithms as it analyzes the dataset in the tree-like graph. However, the data may be over classified and only oneattribute is tested at a time for decision-making.

3.1.3 RANDOM FOREST ALGORITHM.

Random forest algorithm is a supervised classification algorithmic technique. In this algorithm, several trees create a forest. Each individual tree in random forest lets out a class expectation and the class with most votes turns into a model's forecast. In the random forest classifier, the more

number of trees give higher accuracy. The three common methodologies are:

- Forest RI (random input choice)
- Forest RC (random blend)
- Combination of forest RI and forest RC.
- It is used for classification as well as regression task, but can do well with classification task, and can overcome missing values. Besides, being slow toobtain predictions as it requires large data sets and more trees, results are unaccountable.

3.2 RESULT:

Data Set:

This is the UCI repository dataset. It includes 303 rows and 13 attributes .

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
3	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
4	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
5	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
6	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
7	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
8	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
9	44	1	1	120	263	0	1	173	0	0	2	0	3	1
10	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
11	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1
12	54	1	0	140	239	0	1	160	0	1.2	2	0	2	1
13	48	0	2	130	275	0	1	139	0	0.2	2	0	2	1
14	49	1	1	130	266	0	1	171	0	0.6	2	0	2	1
15	64	1	3	110	211	0	0	144	1	1.8	1	0	2	1
16	58	0	3	150	283	1	0	162	0	1	2	0	2	1
17	50	0	2	120	219	0	1	158	0	1.6	1	0	2	1
18	58	0	2	120	340	0	1	172	0	0	2	0	2	1
19	66	0	3	150	226	0	1	114	0	2.6	0	0	2	1
20	43	1	0	150	247	0	1	171	0	1.5	2	0	2	1
21	69	0	3	140	239	0	1	151	0	1.8	2	2	2	1
22	59	1	0	135	234	0	1	161	0	0.5	1	0	3	1
23	44	1	2	130	233	0	1	179	1	0.4	2	0	2	1
24	42	1	0	140	226	0	1	178	0	0	2	0	2	1
25	61	1	2	150	243	1	1	137	1	1	1	0	2	1
26	40	1	3	140	199	0	1	178	1	1.4	2	0	3	1

Importing Libraries

These are the libraries required for the model predictions.

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        from sklearn.model_selection import train_test_split
        from sklearn.linear_model import LogisticRegression
        from sklearn.metrics import accuracy_score
        from random import randrange,uniform
        from sklearn import tree
        from sklearn.tree import export_graphviz
        from sklearn.metrics import accuracy_score
        from sklearn.metrics import confusion_matrix
        from sklearn.ensemble import RandomForestClassifier
        import statsmodels.api as sn
        import hvplot.pandas
        from scipy import stats

        %matplotlib inline
        sns.set_style("whitegrid")
        plt.style.use("fivethirtyeight")
```

Data Visualization:

Heart Disease Count This is the histogram plot of the heart disease count. Thisshows how many patients from the dataset have heart disease. On the X axis there are two numbers (1 & 0). 1 represents the person is having heatdisease and 0 represents the person is not having heart disease.

Heart Disease Count:

This is the histogram plot of the heart disease count. This shows how many patients from the dataset have heart disease. On the X axis there are two numbers (1 &0), 1 represents the person is having heat disease and 0 represents the person is not having heart disease.

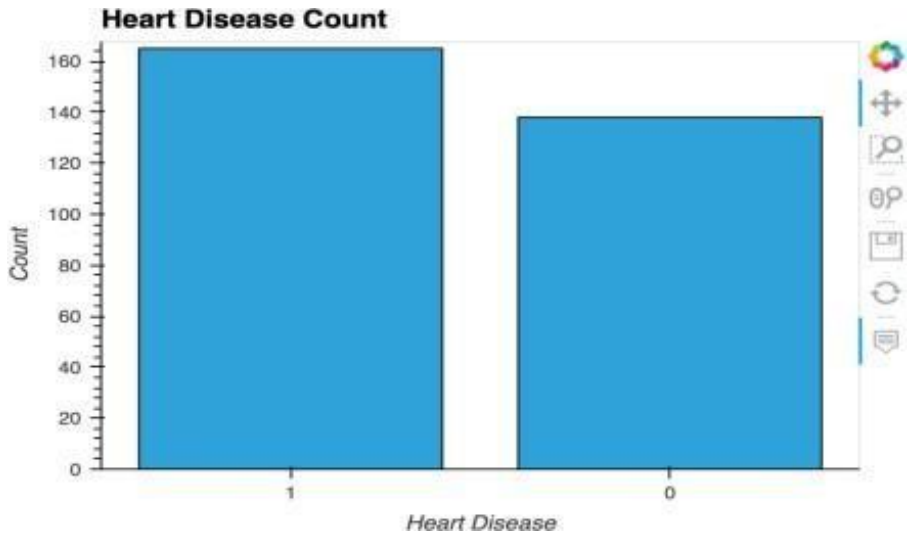


Fig 3.2.1 Heart Disease Count

Heart Disease Detection By Sex:

This is the histogram plot of the patients having heart disease. It is segregated by two plots (1 & 0) and each of them consist of 2 graphs Blue is male and Pink is female. It is seen that heart disease in male is more than females.

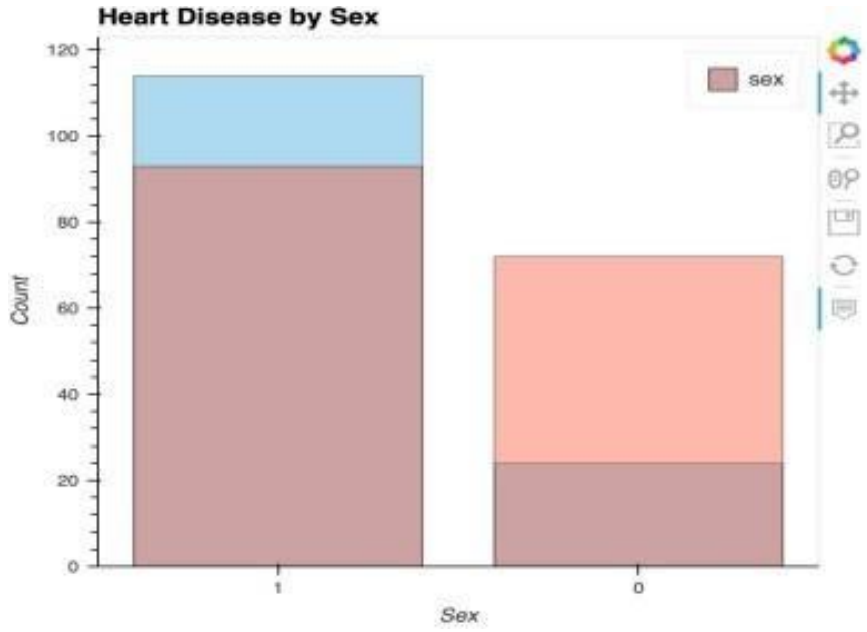


Fig. 3.2.2 Heart Disease by Sex

Heart Disease Detection By Chest Pain:

This is the histogram of heart disease by chest pain. There are 4 types of chest pain, we can see that there are 4 plots with each plot having 2 graphsfor male and female. It is seen that chest pain type (0) is having more patients than others.

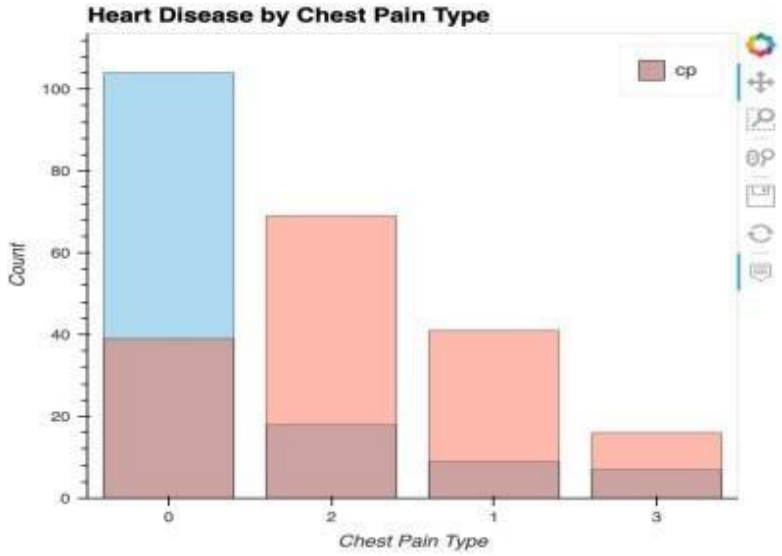


Fig. 3.2.3 Heart Disease by Chest Pain

Heart Disease By Fasting Blood Sugar: Heart Disease By Fasting Blood Sugar In this it is visible that low blood sugar level count are greater than the other. Low blood sugar level is a cause of heart disease.

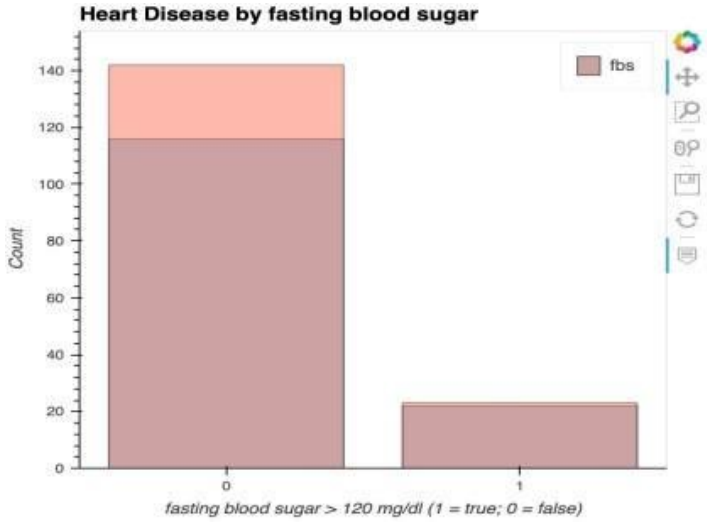


Fig. 3.2.4 Heart Disease by fasting Blood Sugar.

Heart Disease by resting electrocardiography:

In this it is seen that rest ecg count is more than females than men.

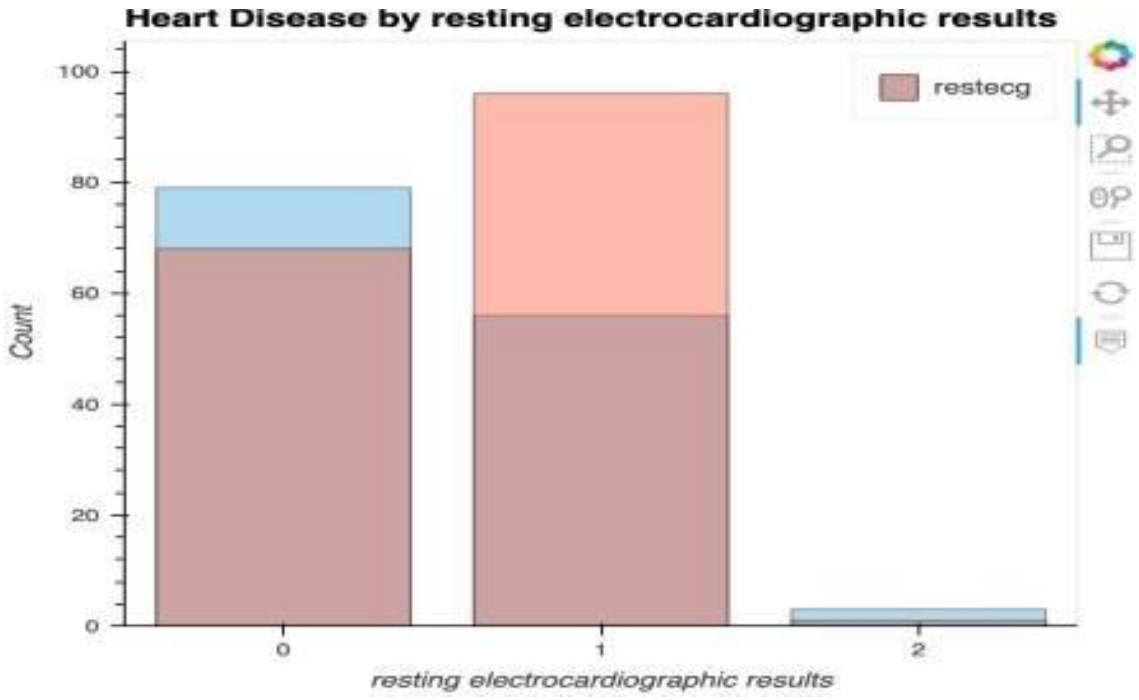


Fig. 3.2.5 Heart Disease by Resting Electrocardiography.

Heart Disease Detection:

Here we can see that there are 9 major attributes which cause heart disease.It is seen that red plot is yes and blue plot is no. So, this analysis is done to understand the prediction of the whole data set those which attributes has major impact on heart disease.

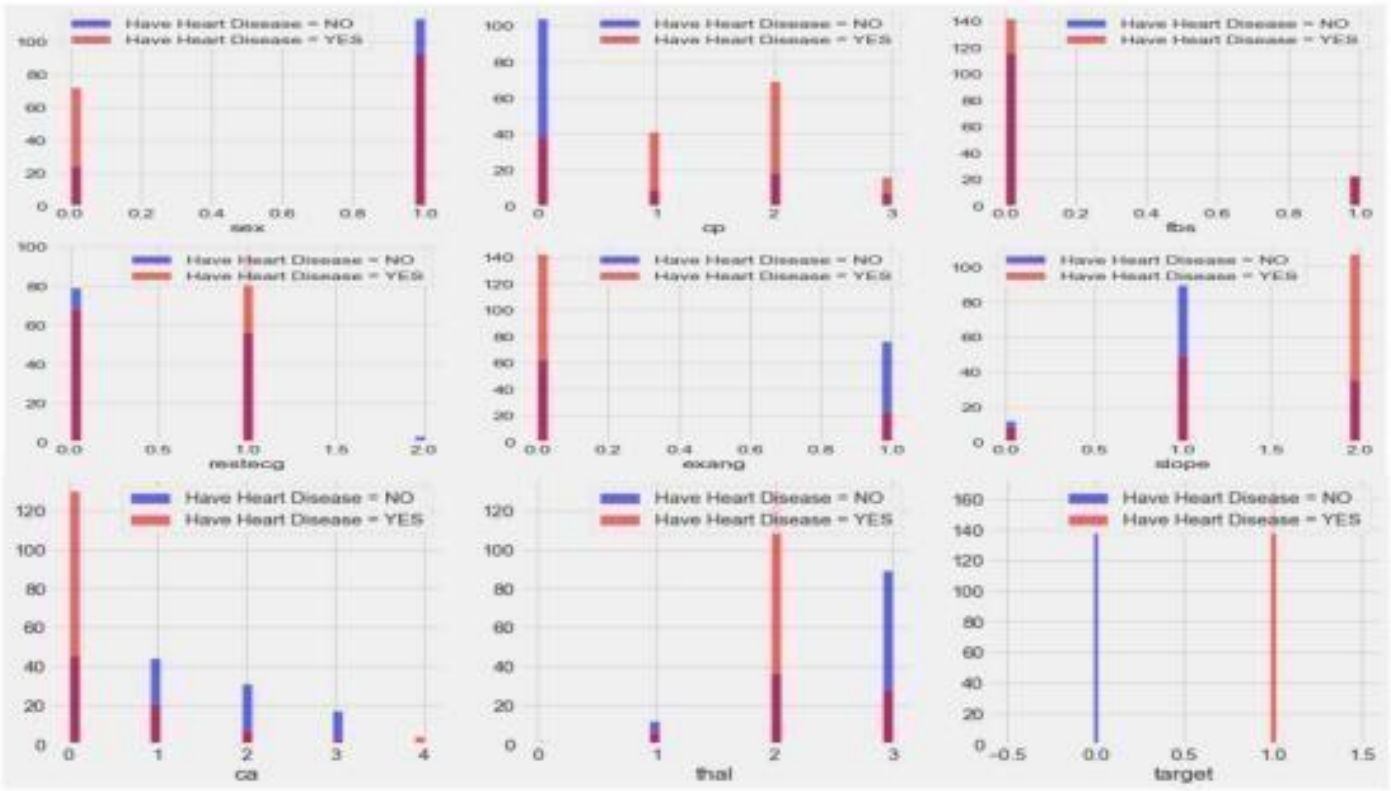


Fig. 3.2.6 Heart Disease Detection.

Logistic Regression

```
Train Result:
================================================
Accuracy Score: 86.67%

------------------------------------------------
CLASSIFICATION REPORT:
                    0       1  accuracy  macro avg  weighted avg
precision    0.857143   0.875  0.866667   0.866071      0.866667
recall       0.857143   0.875  0.866667   0.866071      0.866667
f1-score     0.857143   0.875  0.866667   0.866071      0.866667
support     28.000000  32.000  0.866667  60.000000     60.000000

------------------------------------------------
Confusion Matrix:
 [[24  4]
 [ 4 28]]

Test Result:
================================================
Accuracy Score: 82.50%

------------------------------------------------
CLASSIFICATION REPORT:
                    0          1  accuracy  macro avg  weighted avg
precision    0.764706   0.869565     0.825   0.817136      0.827621
recall       0.812500   0.833333     0.825   0.822917      0.825000
f1-score     0.787879   0.851064     0.825   0.819471      0.825790
support     16.000000  24.000000     0.825  40.000000     40.000000

------------------------------------------------
Confusion Matrix:
 [[13  3]
 [ 4 20]]
```

Decision Tree

```
Train Result:
================================================
Accuracy Score: 100.00%

________________________________________________
CLASSIFICATION REPORT:
              0     1  accuracy  macro avg  weighted avg
precision   1.0   1.0       1.0        1.0           1.0
recall      1.0   1.0       1.0        1.0           1.0
f1-score    1.0   1.0       1.0        1.0           1.0
support    28.0  32.0       1.0       60.0          60.0

________________________________________________
Confusion Matrix:
 [[28  0]
 [ 0 32]]

Test Result:
================================================
Accuracy Score: 77.50%

________________________________________________
CLASSIFICATION REPORT:
                   0          1  accuracy  macro avg  weighted avg
precision   0.769231   0.777778     0.775   0.773504      0.774359
recall      0.625000   0.875000     0.775   0.750000      0.775000
f1-score    0.689655   0.823529     0.775   0.756592      0.769980
support    16.000000  24.000000     0.775  40.000000     40.000000

________________________________________________
Confusion Matrix:
 [[10  6]
 [ 3 21]]
```

Random Forest

```
Train Result:
================================================
Accuracy Score: 100.00%

_________________________________________________
CLASSIFICATION REPORT:
                0     1  accuracy  macro avg  weighted avg
precision     1.0   1.0       1.0        1.0           1.0
recall        1.0   1.0       1.0        1.0           1.0
f1-score      1.0   1.0       1.0        1.0           1.0
support      28.0  32.0       1.0       60.0          60.0

_________________________________________________
Confusion Matrix:
 [[28  0]
 [ 0 32]]

Test Result:
================================================
Accuracy Score: 90.00%

_________________________________________________
CLASSIFICATION REPORT:
                  0          1  accuracy  macro avg  weighted avg
precision  0.928571   0.884615       0.9   0.906593      0.902198
recall     0.812500   0.958333       0.9   0.885417      0.900000
f1-score   0.866667   0.920000       0.9   0.893333      0.898667
support   16.000000  24.000000       0.9  40.000000     40.000000

_________________________________________________
Confusion Matrix:
 [[13  3]
 [ 1 23]]
```

Accuracy Scores

	Model	Training Accuracy %	Testing Accuracy %
0	Logistic Regression	86.666667	82.5
1	Decision Tree Classifier	100.000000	77.5
2	Random Forest Classifier	100.000000	90.0

we can that Random Forest Classifier has highest accuracy percentage so we would beconsidering Random Forest for the prediction result.

Prediction

Case1

```
In [31]: input_data = (62,0,0,140,268,0,0,160,0,3.6,0,2,2)

[1]
The Person has Heart Disease
```

If we put a input data according to the given attributes .We will get the result if the person is having heart disease or not, in this case we are getting the result as the person is having heart disease.

Case2

```
In [35]: input_data = (62,0,3,145,233,1,0,150,0,2.3,0,1,1)

[0]
The Person does not have a Heart Disease
```

If we put other values according to the attributes We will get the result, that the person is not having heart disease

CHAPTER 4
CONCLUSION AND FUTURE SCOPE

4.1 CONCLUSION:

The overall aim is to define various data mining techniques useful in effective heart disease prediction. Efficient and accurate prediction with a lesser number of attributes and tests is our goal. In this study, I consider only14 essential attributes. I applied three data mining classification techniques, Logistic Regression, decision tree, and random forest. The data were pre- processed and then used in the model. Logistic Regression, and random forest are the algorithms showing the best results in this model. I found the accuracy after implementing four algorithms to be highest in Random Forest Classifier. We can further expand this research incorporating other data mining techniques such as time series, clustering and association rules, support vector machine, and genetic algorithm. Considering the limitations of this study, there is a need to implement more complex and combination ofmodels to get higher accuracy for early prediction of heart disease.

4.2 FUTURE SCOPE:

The ratio of heart failure patients has been increasing every day. To overcome this dangerous situation and deteriorate the chances of heart failure disease, there is a need of a system that can generate rules or classify the data using machine learning approaches. Therefore, this research discussed, proposed and implemented a machine learning model by combining five different algorithms. Rapid miner is the tool used in this research, which computed the high accuracy than Matlab and Weka tool. In comparison with the previous researches, this study has shown significant improvement and high accuracy than previous work. As far as UCI dataset concerns, the dataset needs to be amplified. As the main limitation in this work is the small size of the dataset. The dataset has limited number of

patient's records; therefore, the dataset was augmented using appropriate techniques. In future, the results indicated that the system can be useful andhelpful for the doctors and heart surgeons for timely diagnoses the chances of heart attack in a patient.

CHAPTER 5
REFERENCE

[1] Dhai Eddine Salhi, Abdelkamel A Kamel Tari: "Using Machine Learning for Heart Disease Prediction", Chapter. February 2021

[2] Ujma Ansari, Jyoti Soni, Dipesh Sharma, Sunita Soni. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction". March 2020 C. Beyene, P. Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques", 118(8):165- 173 · January 2018

[3] Muhammad Usama Riaz, Shahid Mehmood Awan, Abdul Ghaffar Khan, "Prediction Of Heart Disease Using Artificial Neural Network" Prediction_Of_Heart_Disease_Using_Artificial_Neural_Network, October 2018

[4] Komal Kumar Napa, G.Sarika Sindhu, D.Krishna Prashanthi, A.Shaeen Sulthana, "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers",April2020