A Project Report

on

Cardiovascular disease prediction using Machine Learning

Submitted in partial fulfillment of the

requirement for the award of the degree of

Bachelor of Technology in Computer Science and Engineering

Under The Supervision of
Dr. Naresh Kumar
Department of Computer Science and Engineering

Submitted By

19SCSE1010324 – KRISHNENDU DAS

19SCSE1010102 – SATYAM KUMAR CHOUDHARY

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

GALGOTIAS UNIVERSITY, GREATER NOIDA, INDIA

DECEMBER - 2021

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING
GALGOTIAS UNIVERSITY, GREATER NOIDA

CANDIDATE’S DECLARATION

I/We hereby certify that the work which is being presented in the project, entitled **“Cardiovascular disease prediction using Machine Learning”** in partial fulfillment of the requirements for the award of the **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING** submitted in the **School of Computing Science and Engineering** of Galgotias University, Greater Noida, is an original work carried out during the period of **JULY-2021 to DECEMBER-2021**, under the supervision of **Dr. Naresh Kumar, Department of Computer Science and Engineering** of School of Computing Science and Engineering, Galgotias University, Greater Noida

The matter presented in the project has not been submitted by me/us for the award of any other degree of this or any other places.

19SCSE1010324 – KRISHNENDU DAS

19SCSE1010102 – SATYAM KR CHOUDHARY

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Supervisor

(Dr. Naresh Kumar)

CERTIFICATE

The Final Thesis/Project/ Dissertation Viva-Voce examination of **19SCSE1010324 – KRISHNENDU DAS, 19SCSE1010102 – SATYAM KR CHOUDHARY** has been held on _ ____and his/her work is recommended for the award of **BACHELOR OF TECHNOLOGY IN COMPUTER SCIENCE AND ENGINEERING**.

Signature of Examiner(s) **Signature of Supervisor(s)**

Signature of Project Coordinator **Signature of Dean**

Date:

Place:

ABSTRACT

The idea is to build a heart disease prediction system using machine learning with python. It is based on predictive modeling. The goal of predictive modeling is to develop a model that makes accurate predictions on new data, unseen during training. The workflow will be, firstly we'll get heart data which will contain various types of health parameters that correspond to a person's healthiness to heart.

After taking the data set, we have to process the dataset since we can't feed raw data directly to the system to make it fit and compatible with our ML algorithm. Then we'll split the data into training and testing data set since we first train our data using training data and then we'll evaluate using testing data. When training data is fed to ML algorithm, we use logistic regression which is used for binary classification to know whether a person is suffering from heart disease or not. Once we train this logistic regression model using train data, we'll do some evaluation to check its performance, after that we'll get a trained logistic regression model. Now we can feed new data to predict whether a person is having heart disease or not.

Table of Contents

Acronyms

SVM	Support Vector Machine
ML	Machine Learning
DT	Decision Tree
KNN	K-Nearest Neighbour
LR	Logistic Regression
RF	Random Forrest

CHAPTER-1

Introduction

The heart is an important part of the human body. It pumps blood to all parts of our body. If it fails to function properly, then the brain and various other organs will stop working, and within a few minutes, that person will die. Lifestyle changes, work-related stress, and poor eating habits contribute to an increase in heart-related disorders. Heart disease is already one of the leading causes of death worldwide.

According to the World Health Organization, cardiovascular diseases kill 17.7 million people each year, 31% of all deaths worldwide. In India too, heart disease has become a leading cause of death. Heart disease killed 1.7 million Indians in 2016, according to the 2016 Global Burden of Disease Report, released September 15, 20177. Heart disease-related diseases increase spending on health care and also reduce individual productivity. Estimates made by the World Health Organization (WHO), suggest that India has lost up to $ 237 billion, from 2005-2015, due to heart-related diseases or heart diseases.

Therefore, accurate predictions of heart disease are very important. Medical organizations, worldwide, collect data on a variety of health-related issues. This data can be processed using a variety of machine learning methods to obtain useful information. But the data collected is huge and, in general, this data can be very noisy. These data sets, too strong for human understanding, can be easily tested using a variety of machine learning techniques. Therefore, these algorithms have been very useful, in more recent times, to accurately predict the presence or absence of heart-related diseases.

A major challenge in heart disease is its diagnosis. It is difficult to predict whether a person has heart disease or not. There are devices available that can detect heart disease, however, they are expensive or they do not succeed in human calculations chances of heart disease [4]. According to a study conducted by the World Health Organization (WHO), for medical purposes, specialists can only predict 67% of heart disease, so there is extensive research in this area. In the land of India, access to good doctors and hospitals in rural areas is very low. 2016 WHO report states that only 58% of doctors have medical degrees in urban areas and 19% in rural areas.

In the USA, a person develops heart disease every 40 seconds, that is, more than one person dies in the USA due to heart disease. in 712 With a death toll of 100,000 people, Turkmenistan also has a very high rate death rate up to 2012. Kazakhstan, on the other hand, has the second-highest rate of deaths from heart disease. India is counted 56th in the series. Research also shows that, in the 30-69 years, 1.3 million cardiovascular deaths, 0.9 million (68.4%) caused by heart disease, and 0.4 million (28.0%) by stroke.

CHAPTER-2

LITERATURE SURVEY

Much work has been done to predict heart disease using the UCI Machine Learning data set. Different levels of accuracy are achieved using different data mining techniques described as follows. Avinash Golande and et. al. Research was conducted to study the Decision Tree, K, N, N, and K-Means algorithms that could be used for classification, and its accuracy was compared. The study concludes that the accuracy obtained by the Decision Tree was significantly higher than previously thought.

T. Nagamani, et al. have proposed a program that uses data mining techniques and the MapReduce algorithm. The accuracy obtained according to this paper in the 45 cases of the test set, was greater than the accuracy obtained using a standard non-standard neural network. Here, the accuracy of the algorithm used has been improved due to the use of dynamic schema and linear scaling.

Fahd Saleh Alotaibi designed an ML model that compares five different algorithms. A Rapid Miner tool has been used which has resulted in higher accuracy compared to MATLAB and Week. In this study, the accuracy of Decision Tree, Logistic Regression, random forest, Naive Bayes, and SVM classification algorithms were compared. The decision tree algorithm has the highest accuracy.

Anjan Nikhil Repaka, ealtl., Proposed a system in a research paper that uses NB (Naïve Bayesian) data classification techniques and the AES (Advanced Encryption Standard) algorithm for secure data transmission.

The Emergence of Artificial Intelligence in the field of health science has encouraged numerous research intended to reduce the death rate by applying different data mining techniques. Efficient Heart Attack Prediction by extracting significant patterns from the dataset was proposed by Shantakumar B. Patil et al. K-means clustering algorithm was used weightage of each item was calculated using the MAFIA algorithm. Based on the calculated weightage, patterns with greater values than the threshold were considered for prediction. Prediction of Heart Disease using a 15 attributes dataset with data mining techniques like ANN, Time Series, Clustering Rules, Association Rules were proposed by Jyoti Soni et al. Increasing the accuracy by reducing the data size by applying the genetic algorithm was proposed in the paA computer-aided aided system for diagnosis and prediction was proposed by R. Chitra et al. Neural Network with preprocessed and normalized data with feature reduction was considered for heart disease classification in the paper. Research by experimenting with various algorithms like j48, SIMPLE CART, and reptree was proposed by Hlaudi Daniel Masethe et al. The prediction rate is compared and the best method was proposed in the paper.

The widespread data mining classification techniques like ANN, fuzzy logic, Neural Networks, Decision trees, data mining Genetic Algorithms, and the Nearest Neighbor method were presented by G. Purusothaman et al. Applied hybrid data mining methods were proposed in the paper. The importance of Big Data Analytics for predicting, preventing, and treating chronic diseases was discussed in the paper by Cheryl Ann Alexander et al. The idea of IoT, cloud computing technologies in the medicinal field was proposed in the paper. Improving heart attack prediction using feature selection was proposed by Headey Takci et al. Twelve classification methods and four feature selection algorithms were used for the prediction. Model accuracy, processing time, and ROC analysis were used for compute-based.

An IoT-based application that will work for prediction was proposed by Fizar Ahmed et al. An effective heart disease prediction system was proposed by Poornima Singh et al.Algorithms used were backpropagation(BP) algorithm.was considered for heart disease classification in the paper. Research by experimenting with various algorithms like j48, SIMPLE CAR, Tandd reptree was proposed by Hlaudi Daniel Masethe et al. The prediction rate is compared and the best method was proposed in the paper. The widespread data mining classification techniques like ANN, fuzzy logic, Neural Networks, Decision trees, data mining genetic Algorithm and Nearest Neighbour method were presented by G. Purusothaman et al. Appliedhybrid data mining methods were proposed in the paper importance of Big Data Analytics for predicting, preventing and treating chronic diseases were discussed in the paper by Cheryl Ann Alexander et al.

The idea of IoT, cloud computing technologies in the medicinal field was proposed in the paper. Improving heart attack prediction using feature selection was proposed by Headey Takci et al. Twelve classification methods and four feature select algorithms were used for the prediction. Model accuracy, processing time, and ROC analysis were used for comparison. An effective heart disease prediction system was proposed by Poornima Singh et al. Algorithms used were MLPNN were backpropagation(BP) algorithm. There are various subjects available that focus on the heart predicting the disease when the diagnosis is amusing different data mining techniques.

According to research conducted by the research team in a research paper, the decision separator show shows very good performance compared to everything else models in which the performance of the tested principles of category accuracy. In a research paper, the focus is onto develop a program to assist the medical profession in evaluating a patient's risk of hearing-based patients patient clinical data. The prognosis for heart disease is ais done using machine learning where the applied parameters are presenting, gender, blood pressure, heart rate, diabetes, Hyper cholesterol, Body Mass Index (obesity). ANN, KNN, k-means, as well as K-medoids algorithms the trained in the Cleveland Database of Heart Disease. Body Mass Index (BMI) is one of the most important factors that cause Heart Disease.

Some show research based on the recognition of BMI at 2.3 Million Adolescents and their Cardiovascular Death in Old Age. Here the data is measured from 1967in 2010 entitled BMI Increase in adolescence increases the risk of death as death average. A study conducted focused on results of high-fat mass index and BMI at risk of various cardiovascular conditions, which that high fence that high BMI increases the risk of aortic valve stenosis. In heart failure patients, the relationship between BMI and BMI Death U-shaped with a BMI of 32 to 33 kg / m2as very low. Also, an increase in BMI in adolescence contributes to an increased risk of stroke in adults and, ICH in men. Some indicate the effect of BMI on the risk of heart disease.

Evidence about the bout relationship when excess profit on BMI bet between childhood old age and increased risk of CVD (Cardiovascular Diseases). An author performed a cardiac risk analysis using Deep How to learn. Use patient data internally in Taichung areas in Taiwan. Autoencoder and Softmaxthey were used for feature extraction and separation. their test was able to predict the presence of cardiovascular disease with the help of outpatient inappropriate data environmental monitoring data. Similar natural data temperature, heart rate, etc. was collected with the help of wearable item seems. As their database grows during the day, for future predictions researchers ap erth e, authors proposed an end-to-system designing and implementing a cardiovascular predictor Bayesian algorithm. The method includes eps such as data collection, user registration, and segregation with the help of the Naive Bayes algorithm. They also followed the 80:20 division training and model testing, respectively. AES (Enhanced Encryption was introduced to transfer secure data are they use a web-based application to collect and store data they were able to get good accuracy from the naive Bayesyes classifier.

The authors in a research paper use the Convolutional neural network(CNN) for predicting disease risk. They follow a common way to do inclusive analysis step steps such as 1. Data Collection, 2. Database Production data cleaning and Data Entry, 4. Performance Planning using KNN a naive naive Bayes (Heart Disease) Predictability via CNN. Together with CNN, they played again disease risk prediction using different Machine Learning Algorithms Naive Bayes, KNN to measure ratings proposed trial. Model training, use separate training and test division ratio like 80:20, 70:30, etc. respectively. They were able to get good accuracy in the sing CNN model.
This place contains the ongoing activities of endless anticipation as well preventable diseases using machine learning stages. Juan examined the clinical study stages of the machine hire ass far as their authorization and accuracy, t olive tree, random forest, vector support machine, neural network, and depletion of the material used in the experiment.

In the literature, when feature factor engineering and feature selection are used, results are improved, both in classification and speculation. Dun et al. [34] experimented with various machine learning and in-depth learning techniques for diagnosing heart disease and also performed hyperparameter tuning to increase the accuracy of the results. Neural networks achieved a high accuracy of 78.3 percent, while other models were retrofitting objects, SVM, and integration methods such as Database Forest, etc. To reduce cardiovascular features, Singh et al. [35] used standardized analysis to exclude inaccurate features; a binary classifier as an advanced learning machine for slow filling and maximizing training speed and method of measurement used in all of this was Fisher. The accuracy achieved was 100 percent for heart disease. Classification of Arrhythmias was developed by Yaghouby et al. [36] on varying heart rates. The multilayer perceptron network was used to create distinction and 100 percent accuracy was achieved by reducing features or Gaussian Discrimination Analysis. Asl, et al. [37] used discriminatory Gaussian analysis in reducing HRV signal characteristics to 15, and 100 percent accuracy was achieved using the SVM separator.

S.Nashif, Md. Rakib Raihan, Md. Rasul Islam, Mohammad Hasan Imam [38] proposed a cloud-based cardiac forecasting system to detect nearby heart disease with machine learning algorithms. A real monitoring system designed using Arduino, allows you to feel every 10 seconds various parameters such as temperature, blood pressure, heart rate. In this study, SVM proves that it performs well with more than 95% accuracy.

In HD Prediction 302 the events were compared and evaluated by seven machine learning algorithms for example Naïve Bayes, Decision Tree, K-Nearest Neighbor, Multilayer perceptron, Radial basis function, one integrated student, and SVM. The researchers conducted a case study and led to an SVM approach that worked well [39]. SVM methods are also used in patients with diabetes in the diagnosis of HD [40].

CHAPTER 3

EXISTING SYSTEM

Heart disease has even been emphasized as the killer silence with no obvious symptoms is leading to death. The pre-existing system operates in-depth reading and data mining sets. Clinical diagnosis is important, however complex, a role that needs to be done effectively and accurately done. Appropriate computer-based data and decision-making assistance should be helped to reduce the cost of clinical trials. Data mining is the use of computer diagnostic techniques patterns and fidelity to data sets. Moreover, with the emergence of data mining in the last two decades, there is a great opportunity for computers to build and differentiate different attributes or specific categories. Understanding the dangerous components associated with heart disease allows medical services specialists to identify high-risk patients with heart disease. Statistical analysis reveals risk factors associated with heart diseases such as age, blood pressure, volume cholesterol, diabetes, high blood pressure, heart disease family history, obesity and lack of exercise, fasting blood sugar, etc.

CHAPTER 4

PROPOSED PLAN

We are very much focused on the Logistic Regression model where, from the data we have, it is divided into different structured data based on the patient's heart characteristics. From the availability of data, we should build a model that predicts patient disease using a logistic regression algorithm. First, we must import data sets read data sets, data must contain different variables such as age, gender, sex, chest pain, slope, target. Data should be checked for information to be verified. Create temporary variables and build a retrospective model. Here, we use the sigmoid function that helps to represent the image of the split data. By using retrospect, the accuracy is increased compared to the previous work done on the existing system. A web-based machine learning program trained by the UCI database. The user enters his or her specific medical information to obtain a predictor of that user's heart disease. The algorithm will calculate the risk of heart disease. The result will be displayed on the web page itself. Thus, reducing the cost and time required to predict the disease.

Data format plays an important role in this application. During loading, the user data application will check its appropriate file format and if not for each requirement it will then MAKE a dialog box.

There will be the following four algorithms to be used:

a. Vector Support Machine (SVM)
b. Decision Tree
c. Naïve Bayes Algorithm
d. Logistic Regression
e. KNN Algorithm
f. Random Forrest

The effectiveness of these algorithms is described in future sections.

Algorithms are trained using a data set obtained from the University of California, Irvine.75% of the data input sets were used for training, and the remaining 25% to test algorithm accuracy. In addition, further steps have been taken to improve algorithms by improving accuracy. These steps include cleaning the database and pre-processing the data.

The algorithms were judged based on their accuracy and it was found that SVM was the third most accurate at 64.4% efficiency. Therefore, it has been chosen for the larger application.

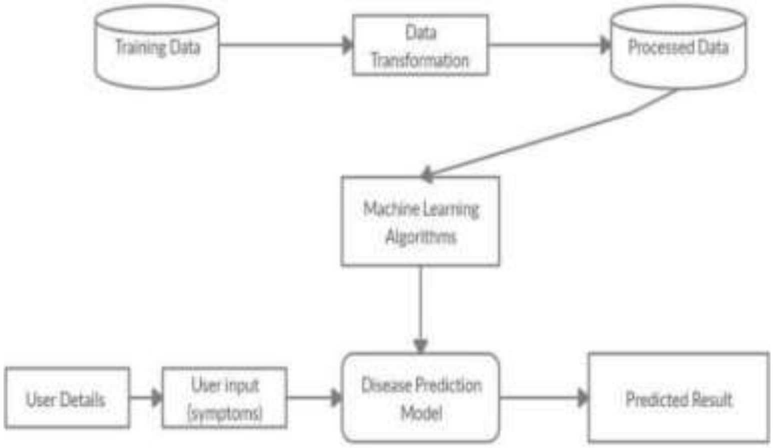
The main application is a web application that accepts various parameters from the user as installed and calculates the result. The result is displayed along with the accuracy of the prediction.

Inputs: Data set, User Data
Outputs: Prediction of heart disease is successfully done.

SYSTEM ARCHITECTURE

Disease prediction used to study machine predicts the presence of a user-targeted disease for a variety of symptoms and information that the user provides such as glucose levels, plasma levels, and many other such common details with indicators. System diagnostic planning using machine learning involves multiple data sets so we will compare user characteristics and predict it, and then the data sets are reconfigured into smaller sets and subsequently segmented based on subsequent

classification algorithms. for predicting all sick input from the user mentioned above. Then when future user information and processed information are fully compiled they are compiled within the system's speculative model and ultimately predict illness. The design diagram may be a clear illustration of a group of learning ideas, which are part of the AN design, as well as the principles, components, and elements.



Here are some of the things this system can do.

1. Signing / Information
2. Disease Prediction

Signing / Information:
If the user has successfully opened the application he or she must select the symbols according to the drop-down menu provided.

Disease prediction:
A predictable model predicts human disease and provides an outcome as you may have a disease, not a disease, depending on the symptoms the user has incurred.

CHAPTER 5

METHODOLOGIES

A. The Logistic regression

This algorithm uses a line equation with independent predictions to predict value. The predicted value can be anywhere between negative and negative endpoints. We need algorithm output to be class variable, i.e. 0-no, 1-yes. Therefore, we press the output of the line number into the range [0, 1]. To reduce the predicted value between 0 and 1, we use the sigmoid function.

Logistic Regression becomes a method of separation only when the decision limit is presented in the image. The threshold value setting is the most important aspect of Logistic regression and is dependent on the separation problem itself.

Based on the number of categories, Logistic decline can be categorized as:

1. Binomial: target variables may have only 2 possible variants: “0” or “1” which may have to “win” vs “lose”, “pass” vs “fail”, “dead” vs “alive”, etc.

2. Multinomial: targeted variants may have 3 or more potential singles (i.e. species have no dose significance) such as “disease A” vs “disease B” vs “disease C”.

3. Ordinal: deals with targeted variables with ordered categories. For example, test results can be categorized as: “very bad”, “bad”, “good”, “very good”. Here, each category can be assigned points such as 0, 1, 2, 3.

B. Vector Support Machine

Vector support machine (SVM) is a supervised learning method that analyzes data used for classification and retransmission analysis. It is provided with a set of training data, marked as part of one of the two phases, the SVM training algorithm, and builds a model that provides new models in one phase or another, making it a viable binary line divider. The SVM model is a model of models such as space points, drawn so that the models of the different categories are separated by a clear gap as wide as possible. New models are then designed in the same space and predicted to belong to a category based on which side of the gap they fall into. Points are divided based on the top flight that separates them.

C. Decision Tree

The goal of using the Decision Tree is to create a training model that can use to predict class or amount of flexibility by studying simple decision rules based on prior data (training data). We compare root attribute values with the record attributes. On a comparative basis, we follow a branch corresponding to that number and jump to the next location.

D. KNN Algorithm

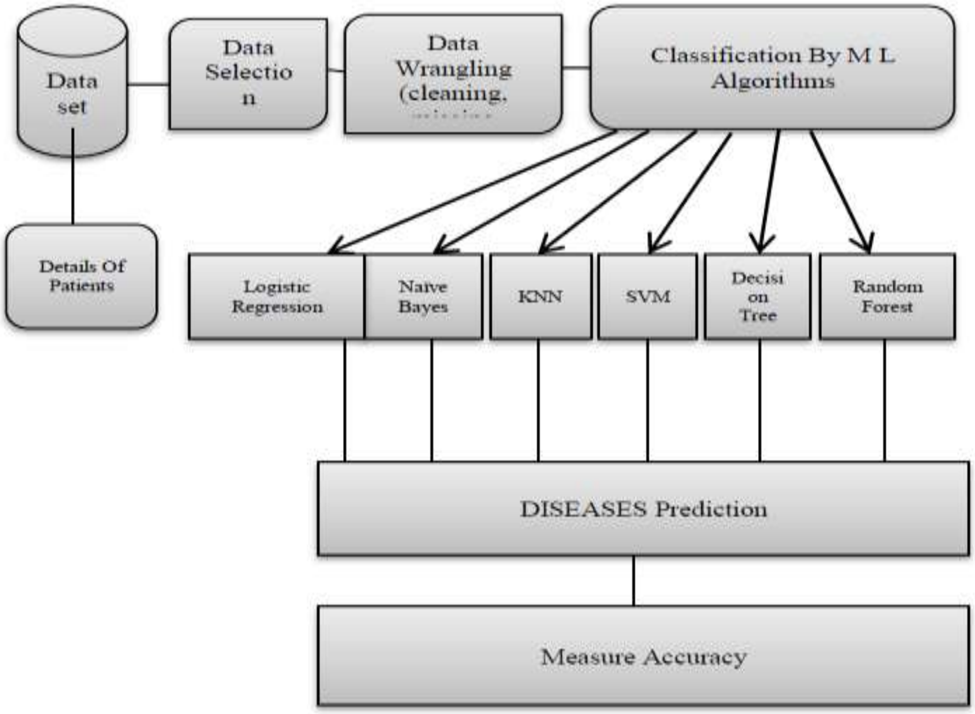
KNN is an algorithm for classification that does not monitor learning. It separates a business that relies on a close neighbor. KNN can be a widely used method used as a separator and multi-sectoral retrievals such as image processing, data processing, pattern recognition, and various applications. The output effect of the algorithmic system depends on the adjacent K-class neighborhood by finding the K-variety of training points closest to a particular character and considering the votes between the K-item.

E. Naive Bayes

Naive Bayes is a family of probabilistic algorithms that take advantage of probability theory and Bayes' Theorem to predict the tag of a text (like a piece of news or a customer review). They are probabilistic, which means that they calculate the probability of each tag for a given text, and then output the tag with the highest one. The way they get these probabilities is by using Bayes' Theorem, which describes the probability of a feature, based on prior knowledge of conditions that might be related to that feature.

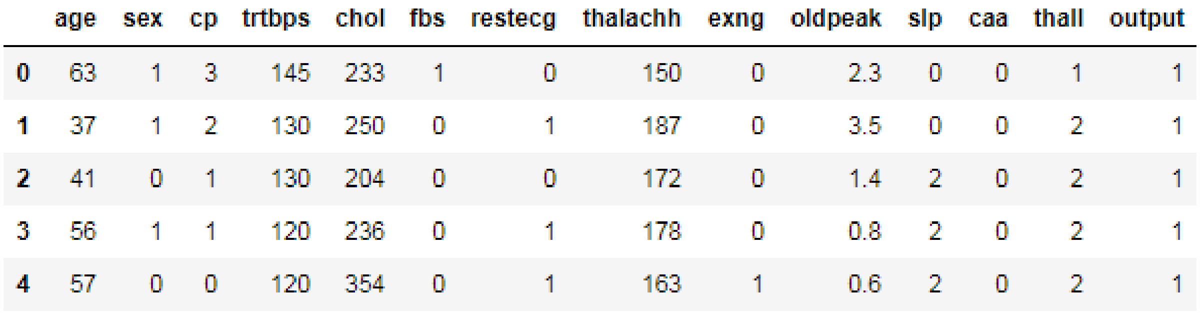
F. Random Forrest Algorithm

The RF algorithm is monitored primarily based on learning. It is used as a classifier in many fields. By using this more trees make the forest. If we have an additional number of trees then it creates high accuracy. It is also used for retrofitting. but it works best when you split the work. It may also exceed the values placed in the wrong place.



DATA COLLECTION

The Cleveland Heart Database contains 303 episodes with 76 symptoms, but only 14 features are considered the most relevant in the experimental study. And they stand like this:



	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
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

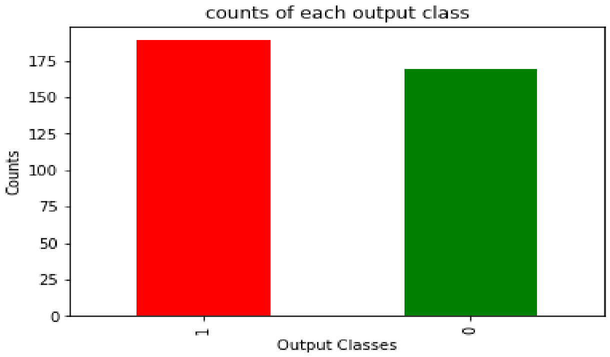
ELEMENT SELECTION

The elements of the data set are the assets of the website used for analysis and forecasting. There are many factors such as age, gender, inclination, and many more that are displayed in the table above for system analysis. Some of the elements are shown below:

SNO	EXPLANATION	ELEMENTS
1	PATIENT's AGE	Age
2	MALE, FEMALE	Sex
3	CHEST PAIN	Cp
4	REST BLOOD PRESSURE	Trtbps
5	CHOLESTEROL	chol
6	FASTING BLOOD SUGAR	Fbs
7	REST ELECTROCARDIOGRAPH	Restecg
8	MAX HEART RATE	Thalachh
9	EXERCISE_INDUCED ANGINA	Exang
10	ST. DEPRESSION	Oldpeak
11	SLOPE	Slp
12	NO. OF VESSELS	Caa
13	THALASSEMIA	Thall
14	OUTPUT(Heart disease Patient	output

PRE-PROCESSING OF DATA

Pre-processing is required to achieve the desired result from the machine learning algorithms. Another ML algorithm does not support missing values in this we have to control empty values from raw data. Another data set attribute has been found that is not helpful in predicting as a city of education etc. Sample Fig. below shows a green color [0] bar representing a non-sick patient and a red bar [1] depicts a heart patient.



ATTRIBUTES USED

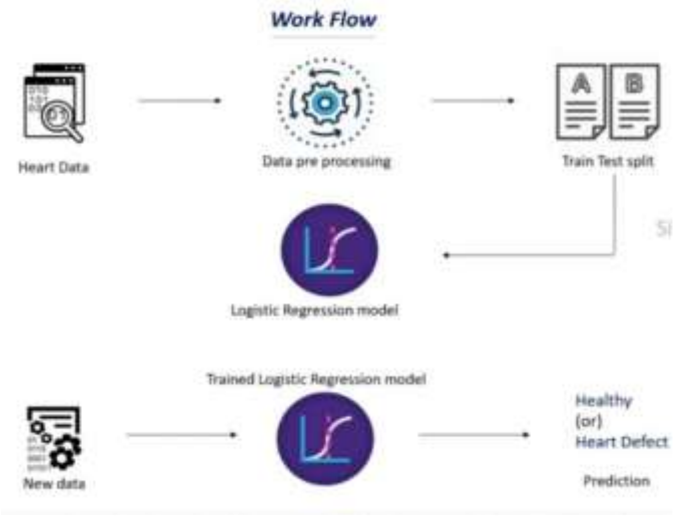
The attributes which are used in this research purpose are described as follows and for what they are used or resemble:

1. Age—age of the patient in years, sex—(1 = male; 0 = female).
2. Cp—chest pain type.
3. Trestbps—resting blood pressure (in mm Hg on admission to the hospital). The normal range is 120/80 (if you have a normal blood pressure reading, it is fine, but if it is a little higher than it should be, you should try to lower it. Make healthy changes to your lifestyle).
4. Chol—serum cholesterol shows the number of triglycerides present. Triglycerides are another lipid that can be measured in the blood. It should be less than 170 mg/dL (may differ in different Labs).
5. Fbs—fasting blood sugar larger than 120 mg/dl (1 true). Less than 100 mg/dL (5.6 mmol/L) is normal, and 100 to 125 mg/dL (5.6 to 6.9 mmol/L) is considered prediabetes.
6. Restecg—resting electrocardiographic results.
7. Thalach—maximum heart rate achieved. The maximum heart rate is 220 minus your age.
8. Exang—exercise-induced angina (1 yes). Angina is a type of chest pain caused by reduced blood flow to the heart. Angina is a symptom of coronary artery disease.
9. Oldpeak—ST depression induced by exercise relative to rest.
10. Slope—the slope of the peak exercise ST segment.
11. Ca—number of major vessels (0–3) colored by fluoroscopy.
12. Thal—no explanation provided, but probably thalassemia (3 normal; 6 fixed defects; 7 reversible defects).
13. Target (T)—no disease = 0 and disease = 1, (angiographic disease status).

CHAPTER 6

FIGURES AND FLOWCHART

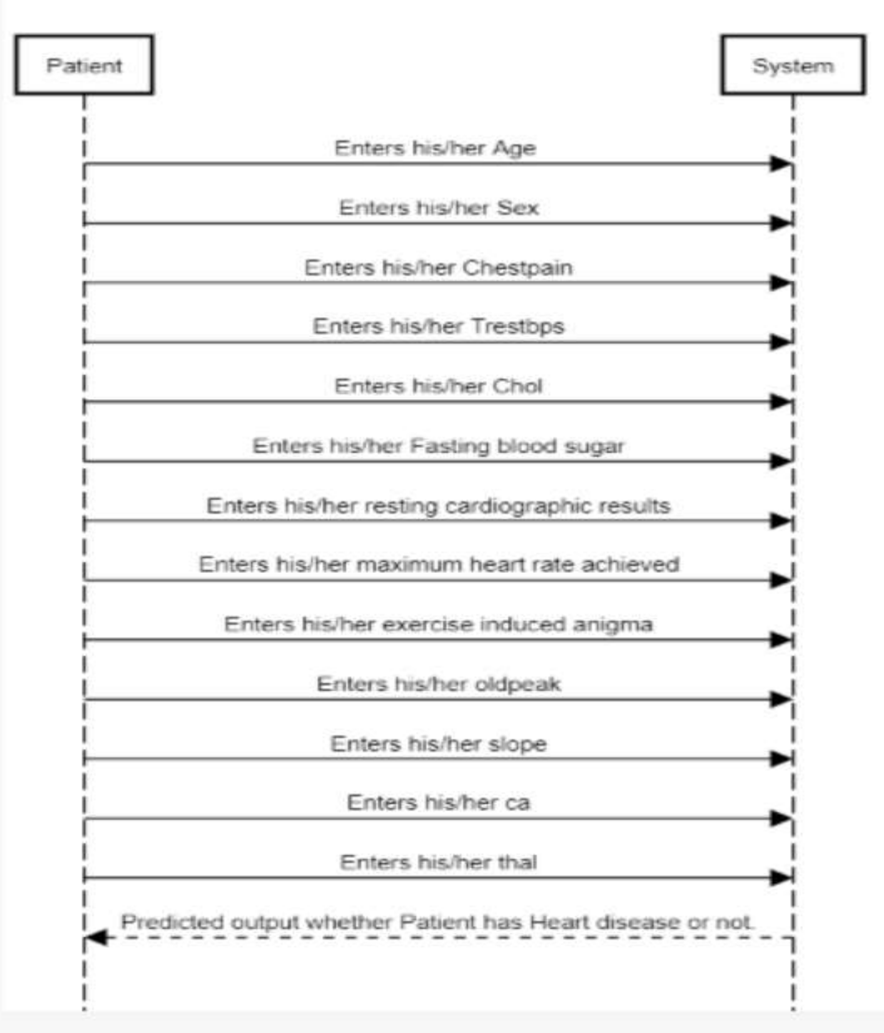
FLOWCHART of Logistic Regression Model



A sample flowchart of the model

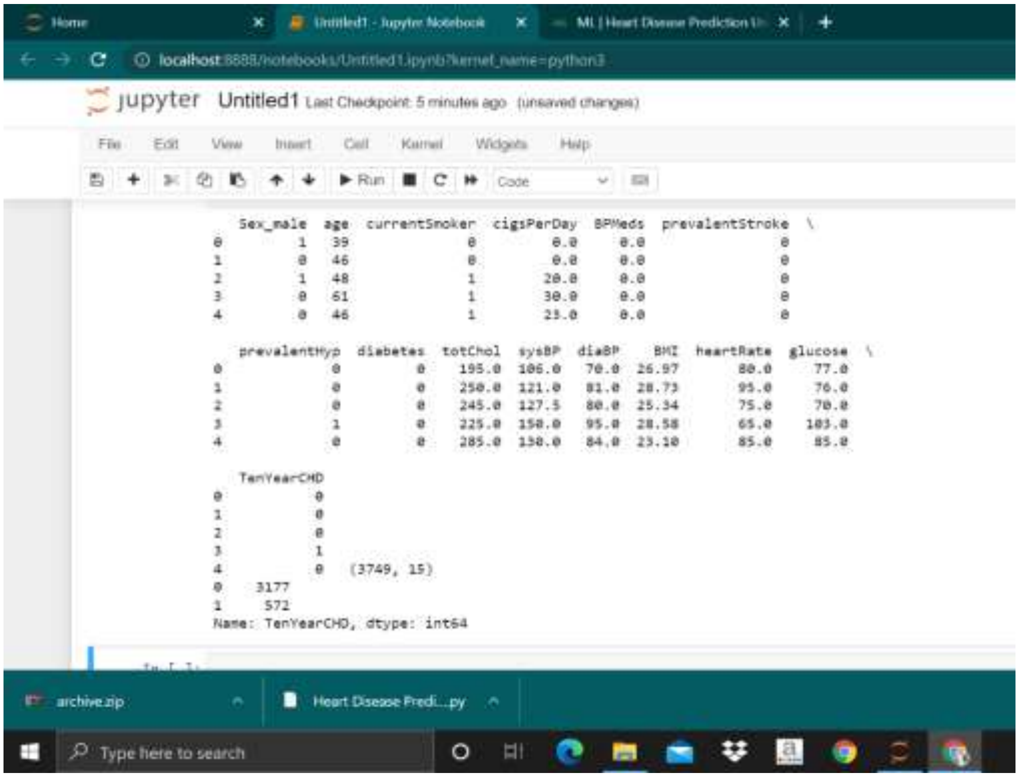
SEQUENCE DIAGRAM

Sequence diagrams are relationship graph that describes how tasks are performed. Sequential diagrams provide chronological representations and show the sequence of relationships using a precise drawing axis showing when messages are received and how often. Here, the system interacts with the patient. It takes the required data from the user/patient, gives the system the required questions, and sends the output to the generated system. Questions about a patient's health statement, which predicts whether or not a patient has a heart condition.

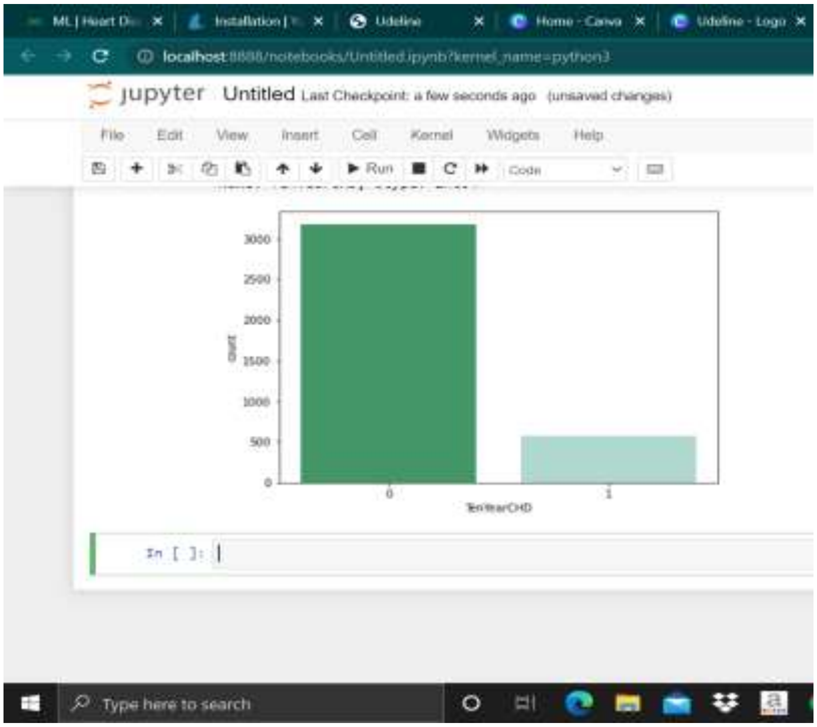

Patient
System
Enters his/her Age
Enters his/her Sex
Enters his/her Chestpain
Enters his/her Trestbps
Enters his/her Chol
Enters his/her Fasting blood sugar
Enters his/her resting cardiographic results
Enters his/her maximum heart rate achieved
Enters his/her exercise induced anigma
Enters his/her oldpeak
Enters his/her slope
Enters his/her ca
Enters his/her thal
Predicted output whether Patient has Heart disease or not

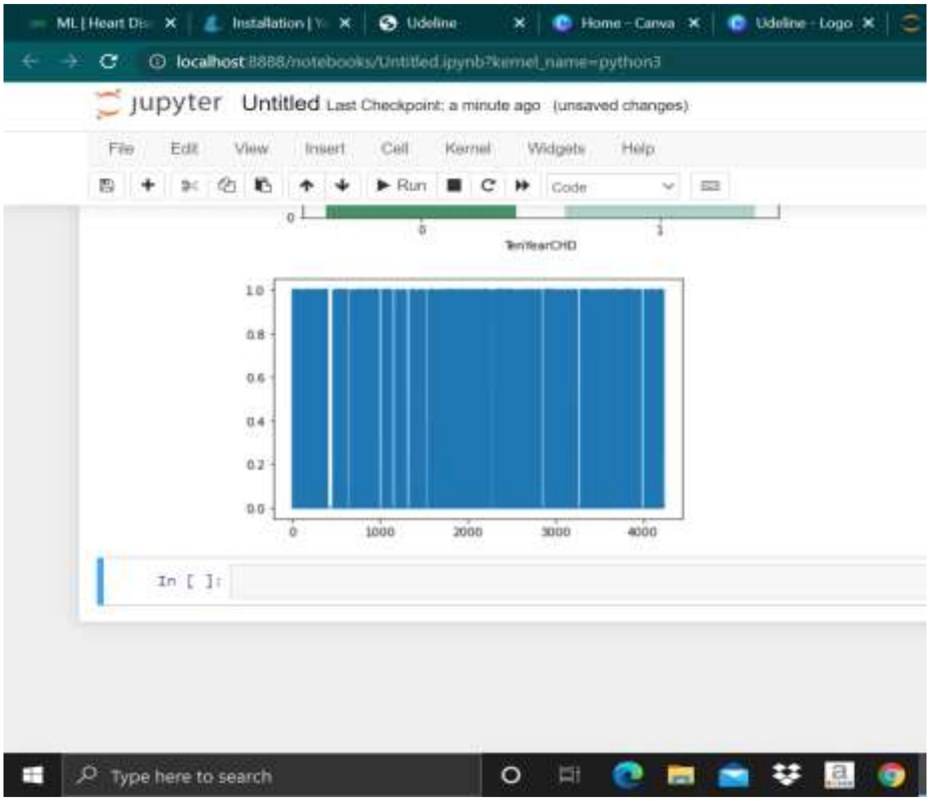
CHAPTER 7

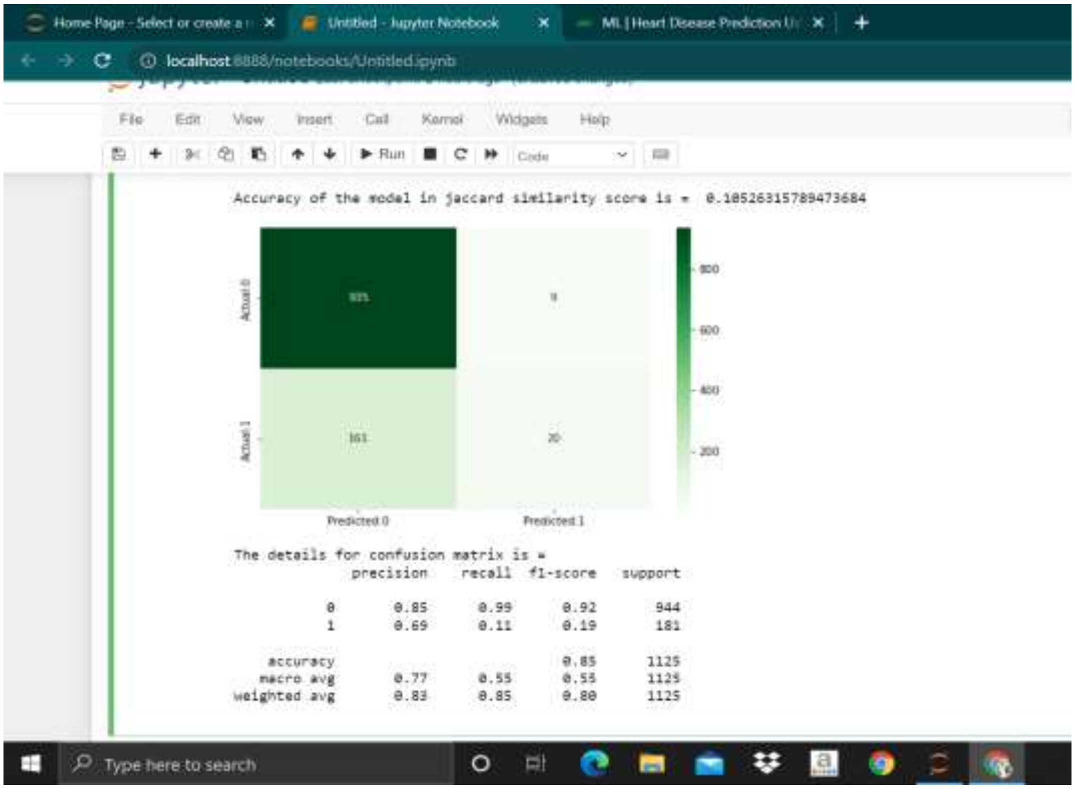
IMPLEMENTATION AND RESULTS


Home
Untitled1 - Jupyter Notebook
ML | Heart Disease Prediction U
localhost:8888/notebooks/Untitled1.ipynb?kernel_name=python3
jupyter
Untitled1 Last Checkpoint: 5 minutes ago (unsaved changes)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run
Code
In [ ]:
archive.zip
Heart Disease Predi....py
Type here to search

```
   Sex_male  age  currentSmoker  cigsPerDay  BPMeds  prevalentStroke  \
0         1   39              0         0.0     0.0                0
1         0   46              0         0.0     0.0                0
2         1   48              1        20.0     0.0                0
3         0   61              1        30.0     0.0                0
4         0   46              1        23.0     0.0                0

   prevalentHyp  diabetes  totChol  sysBP  diaBP    BMI  heartRate  glucose  \
0             0         0    195.0  106.0   70.0  26.97       80.0     77.0
1             0         0    250.0  121.0   81.0  28.73       95.0     76.0
2             0         0    245.0  127.5   80.0  25.34       75.0     70.0
3             1         0    225.0  150.0   95.0  28.58       65.0    103.0
4             0         0    285.0  130.0   84.0  23.10       85.0     85.0

   TenYearCHD
0           0
1           0
2           0
3           1
4           0   (3749, 15)
0    3177
1     572
Name: TenYearCHD, dtype: int64
```


ML | Heart Dis
Installation | Y
Udeline
Home - Canva
Udeline - Logo
localhost:8888/notebooks/Untitled.ipynb?kernel_name=python3
jupyter Untitled Last Checkpoint: a few seconds ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
Run
Code
3000
2500
2000
1500
1000
500
0
count
0
1
TenYearCHD
In [ ]:
Type here to search


ML | Heart Dis
Installation | Y
Udeline
Home - Canva
Udeline - Logo
localhost:8888/notebooks/Untitled.ipynb?kernel_name=python3
jupyter Untitled Last Checkpoint: a minute ago (unsaved changes)
File Edit View Insert Cell Kernel Widgets Help
Run
Code
0
0
1
TenYearCHD
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000
In [ ]:
Type here to search



Home Page - Select or create a n
Untitled - Jupyter Notebook
ML | Heart Disease Prediction U
localhost:8888/notebooks/Untitled.ipynb
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Run
Code
Actual 0
Actual 1
935
9
161
20
Predicted 0
Predicted 1
800
600
400
200
Type here to search

```
Accuracy of the model in jaccard similarity score is =  0.10526315789473684
```

```
The details for confusion matrix is =
              precision    recall  f1-score   support

           0       0.85      0.99      0.92       944
           1       0.69      0.11      0.19       181

    accuracy                           0.85      1125
   macro avg       0.77      0.55      0.55      1125
weighted avg       0.83      0.85      0.80      1125
```

CHAPTER 8

FUTURE WORK

Applying the concept of a newly trained database machine can be used with a more accurate guessing system. Accounts that can be created for each user and then referring to the preferences history of the user's mood can be monitored to see if there is an improvement or if the situation has deteriorated. Now a day's the healthcare industry plays an important role in the treatment of patients' diseases so this is often the case great help in the healthcare industry to inform the user and helpful to the user in the event of his own he does not want to go to the hospital or other clinics, so to include symptoms and everything some useful information within the form of a user who can identify the disease he or she is suffering from therefore the healthcare industry can also find enjoyment in this process by asking for symptoms from the user and log in within the system and in a few seconds, they will tell you the accuracy and arrival to some extent accurate diseases. If the health industry accepts this project it is the job of the doctor they are often reduced and can easily predict a patient's illness. Disease forecast says to provide forecasts for a variety of common and uncommon diseases that, if not re-examined sometimes neglect can turn into a deadly disease and cause many problems for the patient. We can update this project in the future by adding more attributes to the database and more interaction with users and this can be done as an android or ios app. We will fix the system by connecting it to the hospital database.

CHAPTER 9

CONCLUSION

So, finally, we conclude that this disease prediction project is used for machine learning, it is very useful in the daily life of everyone and is very important in health care because they are the ones who use these systems every day to predict patients' illnesses they base their general knowledge on the symptoms they went through.

The quality of theoretical models depends on the method used, the set of data used, the number of symbols and information in the sample, the study methods, and the title offered to the model. We believe that data with adequate measurements and validity tests can be used to create a model that is more accurate in predicting complete heart disease. It is a very critical aspect of data planning that will be used by the learning algorithm system and get good performance, the information details will be improved accordingly. The appropriate algorithm can also be used when creating a prediction model.

Also, we can try other Algorithm Models. We may be seeing some development algorithms in the future as the years go by. Indeed, the use of a reading machine to diagnose heart disease is an important field because it will benefit hospital physicians and individuals. It is indeed a growing industry, yet not all of it is published despite the availability of (big data) of patient information in research facilities or hospitals.

CHAPTER 10

ACKNOWLEDGMENT

We would like to express our deep gratitude to all our guides and reviewers for their patient guidance, enthusiastic encouragement, and useful critiques for this project work. Also to their advice and assistance in keeping our progress on schedule during previous reviews.

Finally, we wish to thank our parents for their continuous support and encouragement throughout our study.

CHAPTER 11

REFERENCES

[1] Avinash Golande, Pavan Kumar T, ”Heart Disease Prediction Using Effective Machine Learning Techniques”, International Journal of Recent Technology and Engineering, Vol 8, pp.944-950,2019.

[2] T. Nagamani, S.Logeswari, B.Gomathy, ” Heart Disease Prediction using Data Mining with MapReduce Algorithm”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3, January 2019.

[3] Fahd Saleh Alotaibi,” Implementation of Machine Learning Model to Predict Heart Failure Disease”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.

[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, ”Design And Implementation Heart Disease Prediction Using Natives Bayesian”, International Conference on Trends in Electronics and Information(ICOEI 2019.

[5] P. V. Ankur Makwana, “Identify the patients at high risk of re-admission hospital in the next year”. International Journal of Science and Research, vol. 4pp. 2431-2434, 2015.

[6] The data can be collected from https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[7] Obasi, T.; Shafiq, M.O. Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attacks and Diseases. In Proceedings of the 2019 IEEE International Conference on Big Data, Big Data 2019, Los Angeles, CA, USA, 9–12 December 2019; pp. 2393–2402.

[8] Ramalingam, V.V.; Dandapath, A.; Raja, M.K. Heart disease prediction using machine learning techniques: A survey. Int. J. Eng. Technol. **2018**, 7, 684–687.

[9] Chen, M.; Hao, Y.; Hwang, K.; Wang, L.; Wang, L. Disease Prediction by Machine Learning Over Big Data from Healthcare
Communities. IEEE Access **2017**, 5, 8869–8879.

[10] Aljanabi, M.; Qutqut, M.; Hijjawi, M. Machine Learning Classification Techniques for Heart Disease Prediction: A Review. Int. J.
Eng. Technol. **2018**, 7, 5373–5379.

[11] Pasha, S.J.; Mohamed, E.S. Novel Feature Reduction (NFR) Model with Machine Learning and Data Mining Algorithms for Effective Disease Risk Prediction. IEEE Access **2020**, 8, 184087–184108.

[12] Swain, D.; Pani, S.K.; Swain, D. A Metaphoric Investigation on Prediction of Heart Disease using Machine Learning. In Proceedings of the 2018 International Conference on Advanced Computation and Telecommunication, ICACAT, Bhopal, India, 28–29 December 2018; pp. 1–6.

[13] Weng, S.F.; Reps, J.M.; Kai, J.; Garibaldi, J.M.; Qureshi, N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE **2017**, 12, e0174944.

[14] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," 2018 Second International Conference on Electronics, Communication, and Aerospace Technology (ICECA), Coimbatore, 2018, pp. 1275-1278.

[15] P. S. Kohli and S. Arora, "Application of Machine Learning in Disease Prediction," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-4.

[16] C. T. and A. Choudhary, "Heart Disease Diagnosis using a Machine Learning Algorithm," 2019 Innovations in Power and Advanced Computing Technologies (i-PACT*)*, Vellore, India, 2019, pp. 1-4.

[17] 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.

[18] Perumal, R. Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques. Int. J. Adv.
Sci. Technol. **2020**, 29, 4225–4234.

[19] Latha, C.B.C.; Jeeva, S.C. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques.
Inform. Med. Unlocked **2019**, 16, 100203.

[20] Anancy-Obiri, D.; Sarku, E. Predicting the Presence of Heart Diseases using Comparative Data Mining and Machine Learning
Algorithms. Int. J. Comput. Appl. **2020**, 176, 17–21.

[21] Kumar, N.K.; Sindhu, G.; Prashanthi, D.; Sulthana, A. Analysis and Prediction of Cardio Vascular Disease using Machine Learning
Classifiers. In Proceedings of the 2020 6th International Conference on Advanced Computing and Communication Systems
(ICACCS), Coimbatore, India, 6–7 March 2020; pp. 15–21.

[22] Gupta, A.; Kumar, R.; Arora, H.S.; Raman, B. MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis. IEEE Access
2019, 8, 14659–14674.

[23] Sultana, M.; Haider, A.; Uddin, M.S. Analysis of data mining techniques for heart disease prediction. In Proceedings of 2016
3rd International Conference on Electrical Engineering and Information and Communication Technology, IEEE ICT 2016, Dhaka,
Bangladesh, 22–24 September 2016; pp. 1–5.

[24] Mohan, S.; Thirumalai, C.; Srivastava, G. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. IEEE
Access **2019**, 7, 81542–81554.

[25] Kodi, S.; Vivekanandam, R. Analysis of Heart Disease using in Data Mining Tools Orange andWeka Sri Satya Sai University Analysis of Heart Disease using in Data Mining Tools Orange and Weka. Glob. J. Comput. Sci. Technol. **2018**, 18.

[26] Soni, Jyoti, et al. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." *International Journal of Computer Applications* 17.8 (2011): 43-48.

[27] Chitra, R., and V. Seenivasagam. "Review of heart disease prediction system using data mining and hybrid intelligent techniques."
ICTACT Journal on soft computing 3.04 (2013): 605-609.

[28] Masethe, Hlaudi Daniel, and Mosima Anna Masethe. "Prediction of heart disease using classification algorithms." *Proceedings of the world Congress on Engineering and Computer science*. Vol. 2. 2014.

[29] Purusothaman, G., and P. Krishnakumari. "A survey of data mining techniques on risk prediction: Heart disease." *Indian Journal of Science and Technology* 8.12 (2015): 1.

[30] Take, Hidayet. "Improvement of heart attack prediction by the feature selection methods." *Turkish Journal of Electrical Engineering & Computer Sci.*

[31] Chicco, D.; Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection
fraction alone. BMC Med. Inform. Decis. Mak. **2020**, 20, 1–16.

[32]Karthick, D.; Priyadharshini, B. Predicting the chances of occurrence of Cardio Vascular Disease (CVD) in people using classification techniques within fifty years of age. In Proceedings of the 2nd International Conference on Inventive Systems and Control,
ICISC 2018, Coimbatore, India, 19–20 January 2018; pp. 1182–1186

[33] Obasi, T.; Shafiq, M.O. Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack
and Diseases. In Proceedings of the 2019 IEEE International Conference on Big Data, Big Data 2019, Los Angeles, CA, USA, 9–12 December 2019; pp. 2393–2402.

[34] B. Dun, E. Wang, and S. Majumder, "Heart disease diagnosis on medical data using ensemble learning," 2016.

[35] R. S. Singh, B. S. Saini, and R. K. Sunkaria, "Detection of coronary artery disease by reduced features and extreme learning machine," *Medicine and Pharmacy Reports*, vol. 91, no. 2, pp. 166–175, 2018.

[36] F. Yaghouby, F. Yaghouby, A. Ayatollahi, and R. Soleimani, "Classification of cardiac abnormalities using reduced features of heart rate variability signal," *World Applied Sciences Journal*, vol. 6, no. 11, pp. 1547–1554, 2009.

[37] B. M. Asl, S. K. Setarehdan, and M. Mohebbi, "Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal," *Artificial Intelligence in Medicine*.

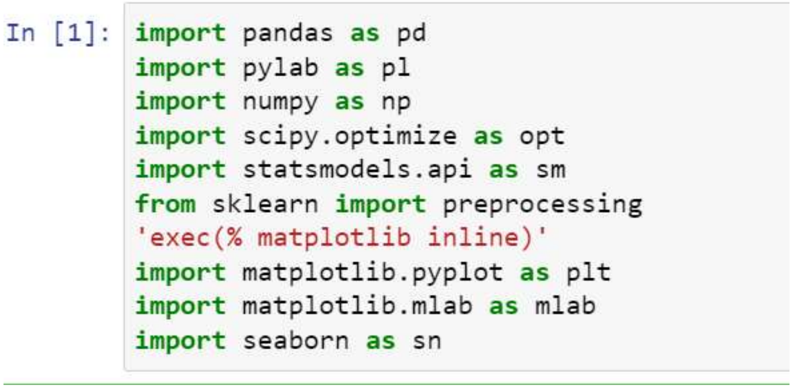
[38] Amin, M.S., Telematics and Informatics, https://doi.org/10.1016/j.tele.2018.11.007

[39] Seyedamin Pouriyeh, Sara Vahid, Giovanna Sannino, Giuseppe De Pietro, Hamid Arabnia, Juan Gutierrez. (2017) A Comprehensive Investigation and Comparison of Machine Learning Techniques in The Domain of Heart Disease‘, Published in Computers and Communications (ISCC), 2017 IEEE Symposium on 3-6 July 2017, DOI: 10.1109/ISCC.2017.8024530.

[40] G.Parthiban, S.K.Srivatsa.(2012) Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients‘, International Journal of Applied Information Systems (IJAIS), ISSN: 2249-0868, Volume 3– No.7.

SOURCE CODE

Code: Loading the libraries.

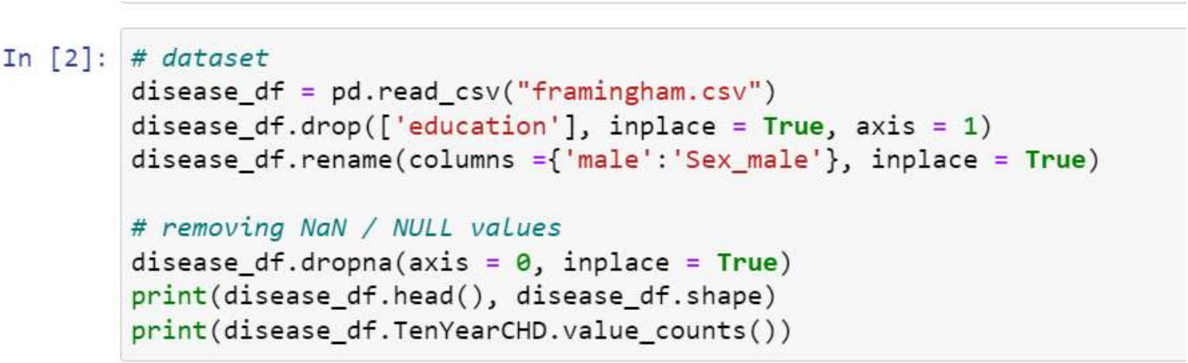


```
In [1]: import pandas as pd
        import pylab as pl
        import numpy as np
        import scipy.optimize as opt
        import statsmodels.api as sm
        from sklearn import preprocessing
        'exec(% matplotlib inline)'
        import matplotlib.pyplot as plt
        import matplotlib.mlab as mlab
        import seaborn as sn
```

Data Preparation :

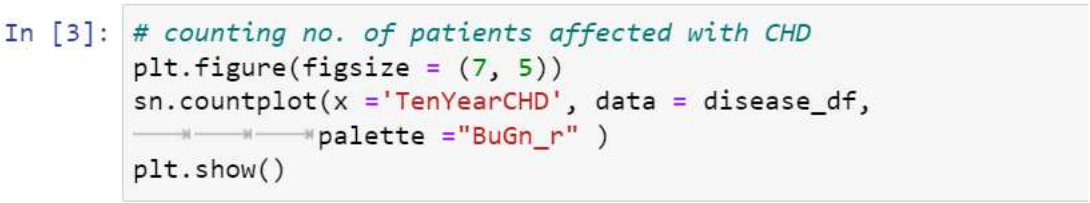
The dataset is publicly available on the Kaggle website and comes from ongoing cardiovascular research for residents of the town of Framingham.

Loading the Dataset.



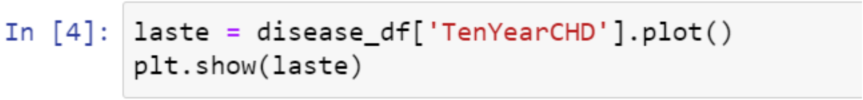
```
In [2]: # dataset
        disease_df = pd.read_csv("framingham.csv")
        disease_df.drop(['education'], inplace = True, axis = 1)
        disease_df.rename(columns ={'male':'Sex_male'}, inplace = True)

        # removing NaN / NULL values
        disease_df.dropna(axis = 0, inplace = True)
        print(disease_df.head(), disease_df.shape)
        print(disease_df.TenYearCHD.value_counts())
```

Code: Ten Year’s CHD Record of all the patients available in the dataset :



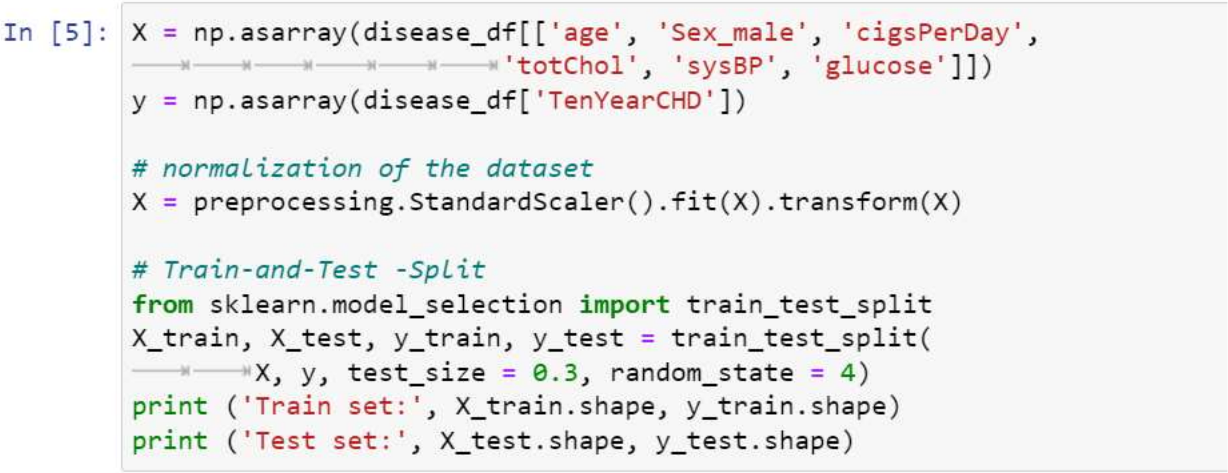
```
In [3]: # counting no. of patients affected with CHD
        plt.figure(figsize = (7, 5))
        sn.countplot(x ='TenYearCHD', data = disease_df,
                    palette ="BuGn_r" )
        plt.show()
```

Code: Counting number of patients affected by CHD where (0= Not Affected; 1= Affected) :



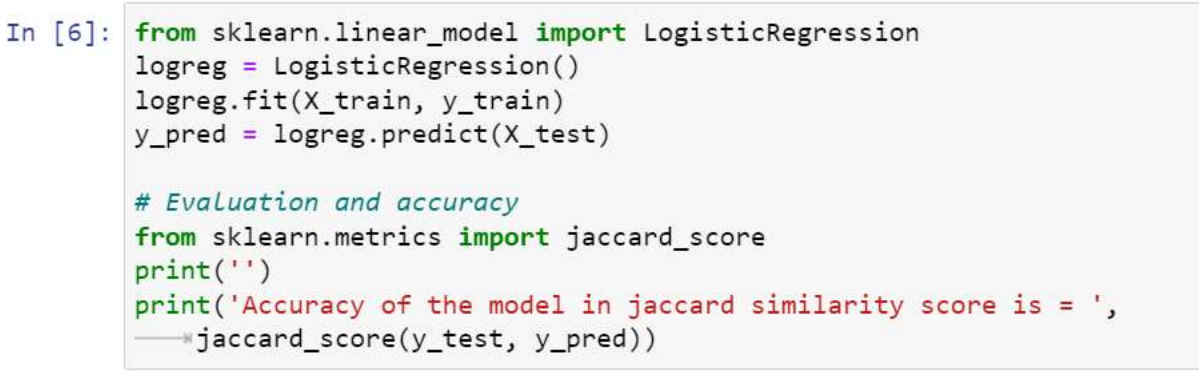
```
In [4]: laste = disease_df['TenYearCHD'].plot()
        plt.show(laste)
```

Code: Training and Test Sets: Splitting Data | Normalization of the Dataset



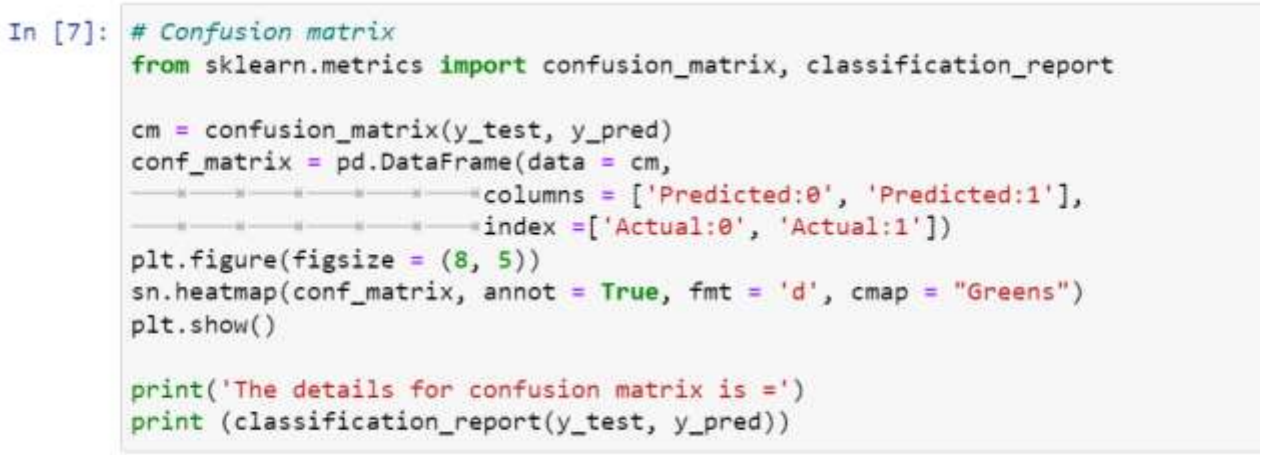
```
In [5]: X = np.asarray(disease_df[['age', 'Sex_male', 'cigsPerDay',
                                   'totChol', 'sysBP', 'glucose']])
        y = np.asarray(disease_df['TenYearCHD'])

        # normalization of the dataset
        X = preprocessing.StandardScaler().fit(X).transform(X)

        # Train-and-Test -Split
        from sklearn.model_selection import train_test_split
        X_train, X_test, y_train, y_test = train_test_split(
                X, y, test_size = 0.3, random_state = 4)
        print ('Train set:', X_train.shape, y_train.shape)
        print ('Test set:', X_test.shape, y_test.shape)
```

Code: Modeling of the Dataset | Evaluation, and Accuracy :



```
In [6]: from sklearn.linear_model import LogisticRegression
        logreg = LogisticRegression()
        logreg.fit(X_train, y_train)
        y_pred = logreg.predict(X_test)

        # Evaluation and accuracy
        from sklearn.metrics import jaccard_score
        print('')
        print('Accuracy of the model in jaccard similarity score is = ',
            jaccard_score(y_test, y_pred))
```

Code: Using Confusion Matrix to find the Accuracy of the model :



```
In [7]: # Confusion matrix
        from sklearn.metrics import confusion_matrix, classification_report

        cm = confusion_matrix(y_test, y_pred)
        conf_matrix = pd.DataFrame(data = cm,
                                   columns = ['Predicted:0', 'Predicted:1'],
                                   index =['Actual:0', 'Actual:1'])
        plt.figure(figsize = (8, 5))
        sn.heatmap(conf_matrix, annot = True, fmt = 'd', cmap = "Greens")
        plt.show()

        print('The details for confusion matrix is =')
        print (classification_report(y_test, y_pred))
```