**HEART DISEASE PREDICTION USING MACHINE LEARNING**

**DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF THE DIPLOMA OF**

IT ENABLE SERVICES & MANAGEMENT

**VIth Semester**

**Jan – May, 2021**

**Project Guide :- Submitted by :-**

**Mrs. Sunita Kr. Chaurasia Rahul kumar**

**1813111034**

[**RAJOKARI INSTITUTE OF TECHNOLOGY -(GOVT), RAJOKRI VILLAGE, RAJOKRI NEW DELHI, DELHI 110038**](https://www.bing.com/local?lid=YN4070x4833719631170764844&id=YN4070x4833719631170764844&q=Rajokari+Farms&name=Rajokari+Farms&cp=28.52436637878418%7e77.11663818359375&ppois=28.52436637878418_77.11663818359375_Rajokari+Farms)**,INDIA**

**Rajokari Institute of Technology**

**Department of Training and Technical Education, Delhi**

**DECLARATION**

I hereby declare that the project titled **Heart Disease Prediction Using Machine Learning** Submitted by me for Diploma in *Information Technology Enable Service Management System VIth* semester to *Rajokari Institute of Technology*, *Department of Training and Technical Education, Delhi,* comprises my own work and due acknowledgement has been made in text to all other material used.

Signature of Student……………………….

Name: **Rahul kumar**

Date: …………………………

**Rajokari Institute of Technology**

**Department of Training and Technical Education, Delhi**

**CERTIFICATE FROM GUIDE**

It is to certify that the project entitled “**\_\_\_\_Heart Disease Prediction Using Machine Learning submitted \_\_\_\_\_**”, submitted by **Mr. Rahul kumar** to the *Rajokari Institute of Technology, Department of Training and Technical Education, Delhi*, has been completed under my supervision and the work is carried out and presented in a manner required for its acceptance to Diploma in *Information Technology Enable Service Management System VIth semester.*

**Project Guide**

Signature: ………………………..

Name: **Mrs. Sunita Kr. Chaurasia**

Date: ………………………..

**Rajokari Institute of Technology**

**Department of Training and Technical Education, Delhi**

**EXAMINATION APPROVAL CERTIFICATE**

It is to certify that we have examined the project entitled “**Heart Disease Prediction Using Machine Learning submitted**”, submitted by **Mr. Rahul kumar** to the *Rajokari Institute of Technology, Department of Training and Technical Education, Delhi,* and hereby accord our approval of it as a study carried out and presented in a manner required for its acceptance *Diploma in Information Technology Enable Service Management VIth semester.*

**Internal Examiner External Examiner**

Signature**:** Signature**:**

Name**:** Name**:**

Date**:** Date**:**

**ACKNOWLEDGEMENT**

I have taken My efforts in this project. However, it would not have been possible without the kind support and help of many individuals. We would like to extend our sincere thanks to all of them.

We are highly indebted to Principal Sir. Mr. Porusu Ramanaiah And Mrs. Sunita kr. Chaurasia for her guidance and constant supervision as well as for providing necessary information regarding the project & also for her support in completing the project.

We would like to express our gratitude towards our families for their kind cooperation and encouragement which help me in completion of this project.

Our thanks and appreciations also go to people who have willingly helped us out with their abilities.

Table of Contents

[**Abstract** 8](#_Toc70289137)

[**LIST OF FIGURES** 9](#_Toc70289138)

[**LIST OF ABBREVIATIONS** 10](#_Toc70289139)

[**CHAPTER 1 : Introduction** 11](#_Toc70289140)

[**1.1 Problem definition** 11](#_Toc70289141)

[**1.2 Motivation** 11](#_Toc70289142)

[**1.3 Aim and Objectives** 12](#_Toc70289143)

[**1.4 Scope and Limitation.** 12](#_Toc70289144)

[**1.4.1 Scope.** 12](#_Toc70289145)

[**1.4.2 Limitations.** 12](#_Toc70289146)

[**1.5 Goal:** 12](#_Toc70289147)

[**1.6 Features & Predictor** 13](#_Toc70289148)

[**CHAPTER 2: RELATED WORKS** 14](#_Toc70289149)

[**CHAPTER 3** 15](#_Toc70289150)

[**3.1 DATASETS** 15](#_Toc70289151)

[**3.2 DATA WINDING** 15](#_Toc70289152)

[**CHAPATER 4: METHODS AND ALGORITHMS USED TOOLS** 18](#_Toc70289153)

[**4.1 Random Forest Classifier using Scikit-learn** 18](#_Toc70289154)

[4.1.1 Random Forest Classifier 18](#_Toc70289155)

[**Using Random Forest Classifier** 18](#_Toc70289156)

[**4.2 TOOLS** 19](#_Toc70289157)

[**4.3 SOFTWARE REQUIREMENTS** 19](#_Toc70289158)

[**HARDWARE REQUIREMENTS** 20](#_Toc70289159)

[**CHAPTER 5 : EXPERIMENTS** 21](#_Toc70289160)

[**5.1 Exploratory Data Analysis** 21](#_Toc70289161)

[**Chapter 6 : Machine Learning + Predictive Analytics** 25](#_Toc70289162)

[**6.1 Prepare Data for Modeling** 25](#_Toc70289163)

[**6.2 Modeling /Training** 25](#_Toc70289164)

[**Model 1: Logistic Regression** 25](#_Toc70289165)

[**Model 2: K-NN (K-Nearest Neighbors)** 26](#_Toc70289166)

[**Model 3: SVM (Support Vector Machine)** 26](#_Toc70289167)

[**Model 4: Decision Trees** 27](#_Toc70289168)

[**Model 5: Random Forest**🏆 28](#_Toc70289169)

[**6.3 Making the Confusion Matrix** 29](#_Toc70289170)

[**6.4 MODEL SAVE** 29](#_Toc70289171)

[**CHAPTER 7 : CONNECT TO WEB PAGE USING A PYTHON** 30](#_Toc70289172)

[**CHAPTER 8 : RESULT** 31](#_Toc70289173)

[**9.1 Libraries used:** 32](#_Toc70289174)

[**CHAPTER 10 : CONCLUSIONS** 33](#_Toc70289175)

[**REFERENCES** 33](#_Toc70289176)

# **Abstract**

This report represents the mini-project assigned to seventh semester students for the partialfulfillment of COMP 484, Machine Learning, given by the department of computer science and engineering, KU. Cardiovascular diseases are the most common cause of death worldwide over the last few decades in the developed as well as underdeveloped and developing countries. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the mortality rate. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. In this project, we have developed and researched about models for heart disease prediction through the various heart attributes of patient and detect impending heart disease using Machine learning techniques like backward elimination algorithm, Random Forest Classifier Andric on the dataset available publicly in Kaggle Website, further evaluating the results using confusion matrix and cross validation. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine.

Keywords: Machine Learning, Random Forest Classifier, Cross-Validation, Backward Elimination, REFCV, Cardiovascular Diseases.

# **LIST OF FIGURES**

[Figure 1 Original Dataset Snapshot 15](#_Toc70288772)

[Figure 2 Import libaray & read csv file 15](#_Toc70288773)

[Figure 3 Describe data & dtypes 16](#_Toc70288774)

[Figure 4 summarizes the count, mean, standard deviation, min, and max for numeric variables. 16](#_Toc70288775)

[Figure 5 Checking null values 17](#_Toc70288776)

[Figure 6 It appears we have a good balance between the two binary outputs. 17](#_Toc70288777)

[Figure 7 Model import 19](#_Toc70288778)

[Figure 8 Random Forest Classification 19](#_Toc70288779)

[Figure 9 Within seconds, you can see whether something is positively or negatively correlated with our predictor (target) 21](#_Toc70288780)

[Figure 10 Chose to make a smaller pairplot with only the continuous variables, to dive deeper into the relationships. Also a great way to see if theirs a positive or negative correlation! 23](#_Toc70288781)

[Figure 11 accuracy 86 28](#_Toc70288782)

[Figure 12 Saving model 29](#_Toc70288783)

[Figure 13 Connecting to web page 30](#_Toc70288784)

[Figure 14 ENTER YOUR DITAILS 31](#_Toc70288785)

[Figure 15 It is showing predition report 31](#_Toc70288786)

# **LIST OF ABBREVIATIONS**

# 

1. IDE: Integrated Development Environment

2. REFCV: Recursive Feature Elimination using Cross-Validation

3. CV: Cross Validation

4. RFE: Recursive Feature Eliminatio

# **CHAPTER 1 : Introduction**

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This project aims to predict future Heart Disease by analysing data of patients which classifies whether they have heart disease or not using machine-learning algorithms.

## **1.1 Problem definition**

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today’s world, we can use various machine learning algorithms to analyse the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

## **1.2 Motivation**

Machine learning techniques have been around us and has been compared and used for analysis for many kinds of data science applications. The major motivation behind this research-based project was to explore the feature selection methods, data preparation and processing behind the training models in the machine learning. With first hand models and libraries, the challenge we face today is data where beside their abundance, and our cooked models, the accuracy we see during training, testing and actual validation has a higher variance. Hence this project is carried out with the motivation to explore behind the models, and further implement Random Forest Classifier model to train the obtained data. Furthermore, as the whole machine learning is motivated to develop an appropriate computer-based system and decision support that can aid to early detection of heart disease, in this project we have developed a model which classifies if patient will have heart disease in ten years or not based on various features (i.e., potential risk factors that can cause heart disease) using Random Forest Classifier. Hence, the early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which ca n be a great milestone in the field of medicine.

## **1.3 Aim and Objectives**

The main objective of developing this project are :

1. To develop machine learning model to predict future possibility of heart disease by implementing RandomForestClassifier.

2. To determine significant risk factors based on medical dataset which may lead to heart disease.

3. To analyse feature selection methods and understand their working principle.

4. The main objective of this research is to develop a heart prediction system. The system can discover and extract hidden knowledge associated with diseases from a historical heart data set

5. Heart disease prediction system aims to exploit data mining techniques on medical data set to assist in the prediction of the heart diseases.

6.Provides new approach to concealed patterns in the data.

7.Helps avoid human biasness.

8.Reduce the cost of medical tests.

## **1.4 Scope and Limitation.**

### **1.4.1 Scope.**

Here the scope of the project is 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 modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions

### **1.4.2 Limitations.**

Medical diagnosis is considered as a significant yet intricate task that needs to be carried out precisely and efficiently. The automation of the same would be highly beneficial. Clinical decisions are often made based on doctor’s intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

## **1.5 Goal:**

* Predict whether a patient should be diagnosed with Heart Disease. This is a binary outcome.
* Positive (+) = 1, patient diagnosed with Heart Disease
* Negative (-) = 0, patient not diagnosed with Heart Disease
* Experiment with various Classification Models & see which yields greatest accuracy.
* Examine trends & correlations within our data
* Determine which features are most important to Positive/Negative Heart Disease diagnosis

## **1.6 Features & Predictor**

Our Predictor (Y, Positive or Negative diagnosis of Heart Disease) is determined by 13 features (X):

1. age (#)

2. sex : 1= Male, 0= Female (Binary)

3. (cp)chest pain type (4 values -Ordinal):Value 1: typical angina ,Value 2: atypical angina, Value 3: non-anginal pain , Value 4: asymptomatic

4. (trestbps) resting blood pressure (#)

5. (chol) serum cholesterol in mg/dl (#)

6. (fbs)fasting blood sugar > 120 mg/dl(Binary)(1 = true; 0 = false)

7. (restecg) resting electrocardiography results(values 0,1,2)

8. (thalach) maximum heart rate achieved (#)

9. (exang) exercise induced angina (binary) (1 = yes; 0 = no)

10. (oldpeak) = ST depression induced by exercise relative to rest (#)

11. (slope) of the peak exercise ST segment (Ordinal) (Value 1: up sloping , Value 2: flat , Value 3: down sloping )

12. (ca) number of major vessels (0–3, Ordinal) colored by fluoroscopy

13. (thal) maximum heart rate achieved — (Ordinal): 3 = normal; 6 = fixed defect; 7 = reversible defect

# **CHAPTER 2: RELATED WORKS**

Now we've got our data split into training and test sets, it's time to build a machine learning model.

We'll train it (find the patterns) on the training set.

And we'll test it (use the patterns) on the test set.

We're going to try 3 different machine learning models:

1. Logistic Regression
2. K-Nearest Neighbours Classifier
3. Support Vector machine
4. Decision Tree Classifier
5. Random Forest Classifier

|  |  |  |  |
| --- | --- | --- | --- |
| S.NO | Model | Training Accuracy % | Testing Accuracy % |
| 1 | Logistic Regression | 74 | 74 |
| 2 | K-Nearest Neighbours Classifier | 75 | 75 |
| 3 | Support Vector machine | 75 | 75 |
| 4 | Decision Tree Classifier | 69 | 69 |
| 5 | Random Forest Classifier | 86 | 86 |

# **CHAPTER 3**

## **3.1 DATASETS**

he dataset is publicly available on the Kaggle Website at [4] which is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. It provides patient information which includes over 4000 records and 14 attributes. The attributes include: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting, sugar blood, resting electrocardiographic results, maximum heart rate, exercise induced angina, ST depression induced by exercise, slope of the peak exercise, number of major vessels, and target ranging from 0 to 2, where 0 is absence of heart disease. The data set is in csv (Comma Separated Value) format which is further prepared to data frame as supported by pandas library in python.

Table

Description automatically generated

Figure 1 Original Dataset Snapshot

The education data is irrelevant to the heart disease of an individual, so it is dropped. Further with this dataset pre-processing and experiments are then carried out

## **3.2 DATA WINDING**

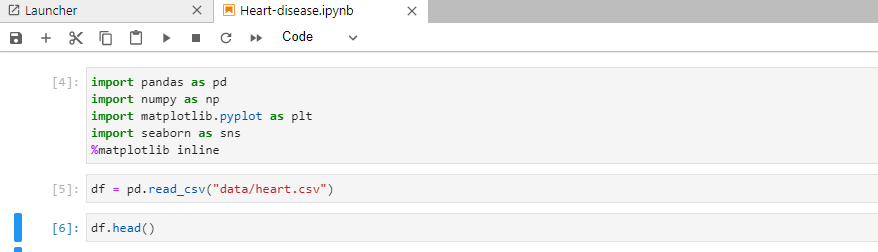
****

Figure 2 Import libaray & read csv file

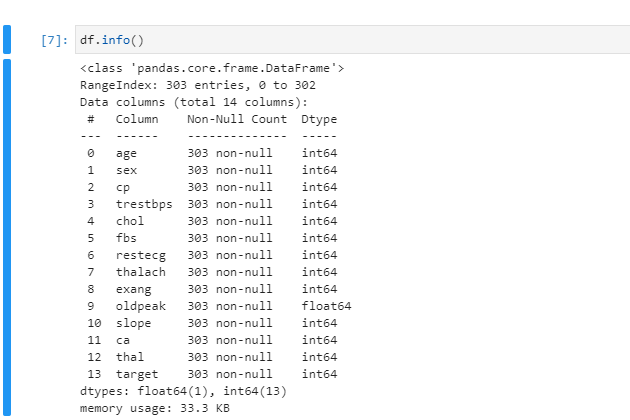
****

Figure 3 Describe data & dtypes

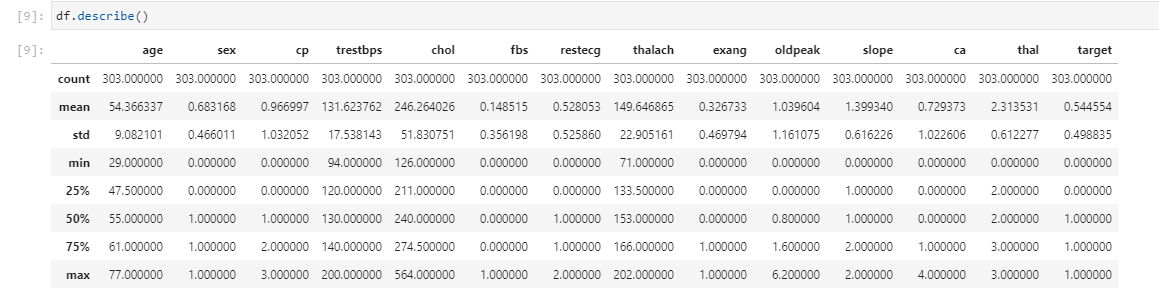


Figure 4 summarizes the count, mean, standard deviation, min, and max for numeric variables.

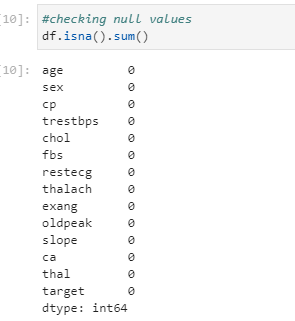


Figure 5 Checking null values

Lets see if theirs a good proportion between our positive & negative **binary predictor.**



Figure 6 It appears we have a good balance between the two binary outputs.

# **CHAPATER 4: METHODS AND ALGORITHMS USED TOOLS**

## **4.1 Random Forest Classifier using Scikit-learn**

Random Forest Classifier is ensemble algorithm. In next one or two posts we shall explore such algorithms. Ensembled algorithms are those which combines more than one algorithms of same or different kind for classifying objects. For example, running prediction over Naive Bayes, SVM and Decision Tree and then taking vote for final consideration of class for test object.

### 4.1.1 Random Forest Classifier

Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

In Laymen’s term,

Suppose training set is given as : [X1, X2, X3, X4] with corresponding labels as [L1, L2, L3, L4], random forest may create three decision trees taking input of subset for example,

1. [X1, X2, X3]
2. [X1, X2, X4]
3. [X2, X3, X4]

### **Using Random Forest Classifier**

The code for using Random Forest Classifier is similar to previous classifiers.

1. Import library
2. Create model
3. Train
4. Predict

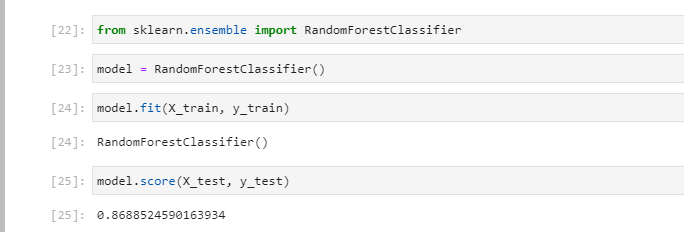
****

Figure 7 Model import

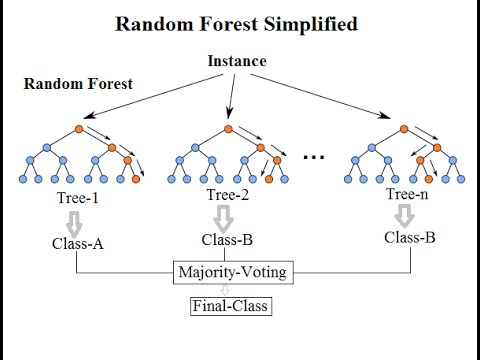


Figure 8 Random Forest Classification

## **4.2 TOOLS**

For application development, the following Software Requirements are:

Operating System: Windows 7 or any Linux Debian Distro.

Language: R and Shiny Tools: RStudio IDE, Microsoft Excel (Optional).

Technologies used: R, Unix, Shiny

## **4.3 SOFTWARE REQUIREMENTS**

Operating System Any OS with clients to access the internet Network Wi-Fi Internet or cellular Network

Visio Studio Create and design Data Flow and Context Diagram

Github Versioning Control

Google Chrome Medium to find reference to do system testing,

## **HARDWARE REQUIREMENTS**

For application development, the following Software Requirements are:

Processor: Intel or high

RAM: 1024 MB

Space on disk: minimum 100mb For running the application:

Device: Any device that can access the internet Minimum space to execute: 20 MB

# **CHAPTER 5 : EXPERIMENTS**

## **5.1 Exploratory Data Analysis**

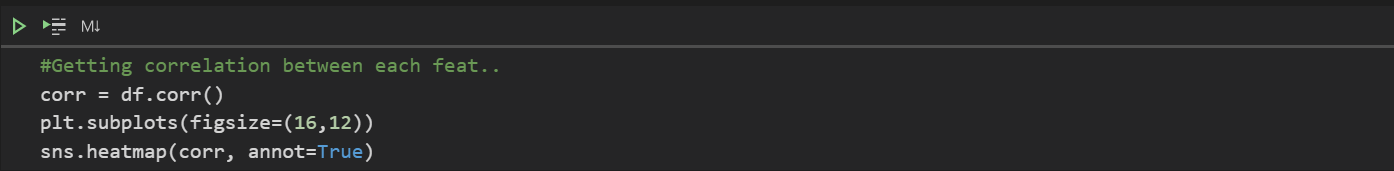
Correlation Matrix visualization Before Feature Selection shows It shows that there is no single feature that has a very high correlation with our target value. Also, some of the features have a negative correlation with the target value and some have positive. The data was also visualized through plots and bar graphs**



Figure 9 Within seconds, you can see whether something is positively or negatively correlated with our predictor (target)

Correlation Matrix- let’s you see correlations between all variables.

We can see there is a **positive correlation** between chest pain (cp) & target (our predictor). This makes sense since, the greater amount of chest pain results in a greater chance of having heart disease. Cp (chest pain), is a ordinal feature with 4 values:

Value 1: typical angina ,

Value 2: atypical angina,

Value 3: non-anginal pain ,

Value 4: asymptomatic.

In addition, we see a **negative correlation** between exercise induced angina (exang) & our predictor. This makes sense because when you excercise, your heart requires more blood*,* but narrowed arteries slow down blood flow*.*

Pairplots are also a great way to immediately see the correlations between all variables. But you will see me make it with only continuous columns from our data, because with so many features, it can be difficult to see each one. So instead I will make a pairplot with only our continuous features.

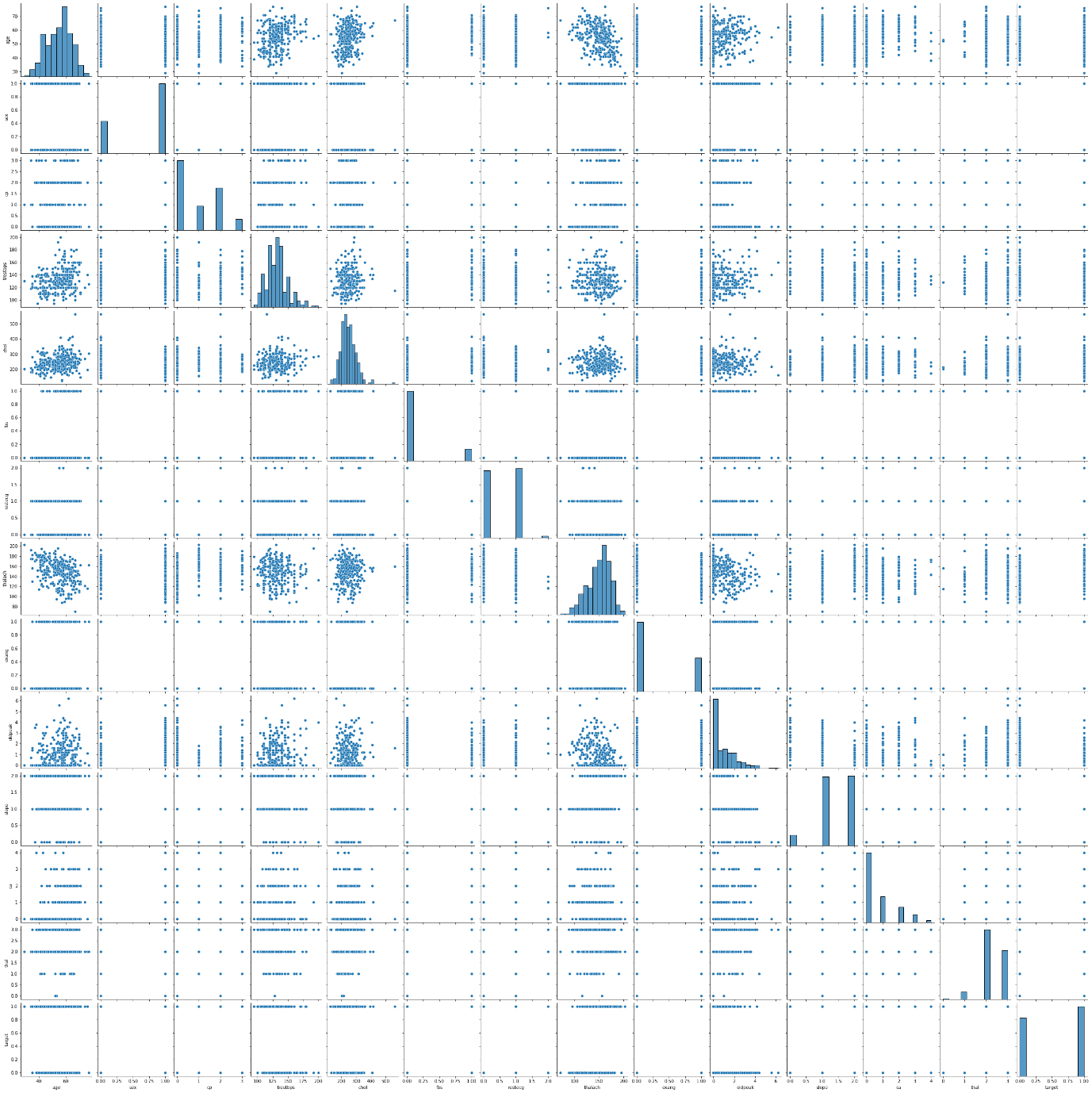


Figure 10 Chose to make a smaller pairplot with only the continuous variables, to dive deeper into the relationships. Also a great way to see if theirs a positive or negative correlation!



ST segment depression occurs because when the ventricle is at rest and therefore repolarized. If the trace in the ST segment is abnormally low below the baseline, this can lead to this Heart Disease. This is **supports** the plot above because low ST Depression yields people at greater risk for heart disease. While a high ST depression is considered normal & healthy. The “slope” hue, refers to the peak exercise ST segment, with values: 0: upsloping , 1: flat , 2: downsloping). Both positive & negative heart disease patients exhibit **equal distributions**of the 3 slope categories.

# **Chapter 6 : Machine Learning + Predictive Analytics**

## **6.1 Prepare Data for Modeling**

To prepare data for modeling, just remember ASN (Assign,Split, Normalize).

Assign the 13 features to X, & the last column to our classification predictor, y

X = data.iloc[:, :-1].values  
y = data.iloc[:, -1].values

Split: the data set into the Training set and Test set

from sklearn.model\_selection import train\_test\_split  
x\_train, x\_test, y\_train, y\_test = train\_test\_split(X,y,test\_size = 0.2, random\_state = 1)

Normalize: Standardizing the data will transform the data so that its distribution will have a mean of 0 and a standard deviation of 1.

from sklearn.preprocessing import StandardScaler  
sc = StandardScaler()  
x\_train = sc.fit\_transform(x\_train)  
x\_test = sc.transform(x\_test)

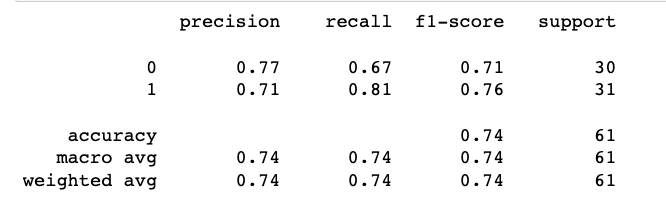
## **6.2 Modeling /Training**

Now we’ll Train various Classification Models on the Training set & see which yields the highest accuracy. We will compare the accuracy of Logistic Regression, K-NN (k-Nearest Neighbours), SVM (Support Vector Machine), Decision Trees, Random Forest.

Note: these are all supervised learning models.

**Model 1: Logistic Regression**

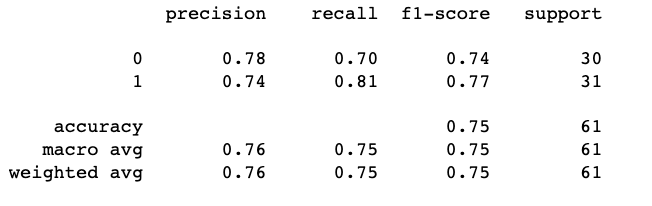
**from** **sklearn.metrics** **import** classification\_report   
**from** **sklearn.linear\_model** **import** LogisticRegression  
  
**model1 = LogisticRegression(random\_state=1)** # get instance of model  
**model1.fit(x\_train, y\_train)** # Train/Fit model   
  
**y\_pred1 = model1.predict(x\_test)** # get y predictions  
**print(classification\_report(y\_test, y\_pred1))** # output accuracy



**Accuracy 74%**

**Model 2: K-NN (K-Nearest Neighbors)**

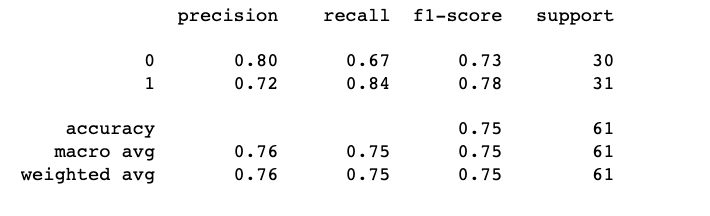
**from** **sklearn.metrics** **import** classification\_report   
**from** **sklearn.neighbors** **import** KNeighborsClassifier  
  
**model2 = KneighborsClassifier()** # get instance of model  
**model2.fit(x\_train, y\_train)** # Train/Fit model   
  
**y\_pred2 = model2.predict(x\_test)** # get y predictions  
**print(classification\_report(y\_test, y\_pred2))** # output accuracy



**Accuracy 75%**

**Model 3: SVM (Support Vector Machine)**

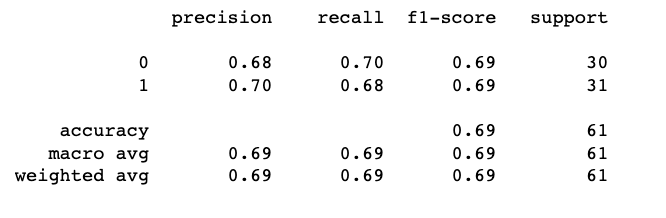
**from** **sklearn.metrics** **import** classification\_report   
**from** **sklearn.svm** **import** **SVC**  
  
**model3 = SVC(random\_state=1)** # get instance of model  
**model3.fit(x\_train, y\_train)** # Train/Fit model   
  
**y\_pred3 = model3.predict(x\_test)** # get y predictions  
**print(classification\_report(y\_test, y\_pred3))** # output accuracy



**Accuracy 75%**

**Model 4: Decision Trees**

**from** **sklearn.metrics** **import** classification\_report   
**from** **sklearn.tree** **import** DecisionTreeClassifier  
 **model5 = DecisionTreeClassifier(random\_state=1)** # get instance of model  
**model5.fit(x\_train, y\_train)** # Train/Fit model   
  
**y\_pred5 = model5.predict(x\_test)** # get y predictions  
**print(classification\_report(y\_test, y\_pred5))** # output accuracy



**Accuracy 69%**

**Model 5: Random Forest**🏆

**from** **sklearn.metrics** **import** classification\_report   
**from** **sklearn.ensemble** **import** RandomForestClassifier  
  
**model6 = RandomForestClassifier(random\_state=1)**# get instance of model  
**model6.fit(x\_train, y\_train)** # Train/Fit model   
  
**y\_pred6 = model6.predict(x\_test)** # get y predictions  
**print(classification\_report(y\_test, y\_pred6))** # output accuracy

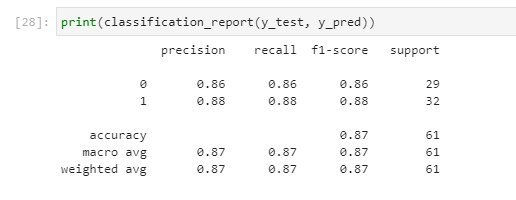


Figure 11 accuracy 86

Precision, Recall, F1-score and Support:

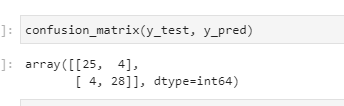
Precision : be “how many are correctly classified among that class”

Recall : “how many of this class you find over the whole number of element of this class”

F1-score : harmonic mean of precision and recall values.  
F1 score reaches its best value at 1 and worst value at 0.  
F1 Score = 2 x ((precision x recall) / (precision + recall))  
  
Support: # of samples of the true response that lie in that class.

## **6.3 Making the Confusion Matrix**

**from sklearn.metrics import confusion\_matrix, accuracy\_score  
cm = confusion\_matrix(y\_test, y\_pred6)  
print(cm)  
accuracy\_score(y\_test, y\_pred6)**



## **6.4 MODEL SAVE**

**Approach 1 : Pickle approach**

Following lines of code, the LR\_Model which we created in the previous step is saved to file, and then loaded as a new object called Pickled\_RL\_Model.

The loaded model is then used to calculate the accuracy score and predict outcomes on new unseen (test) data.

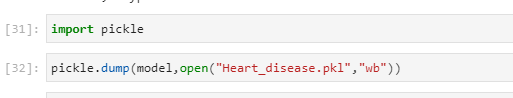


Figure 12 Saving model

# **CHAPTER 7 : CONNECT TO WEB PAGE USING A PYTHON**

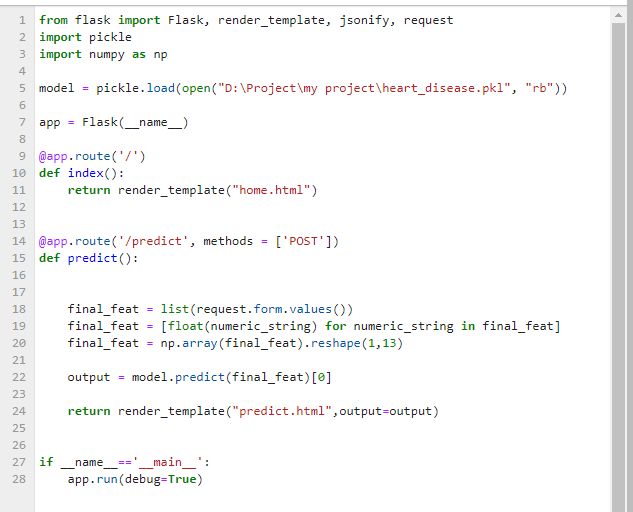
****

Figure 13 Connecting to web page

# **CHAPTER 8 : RESULT**

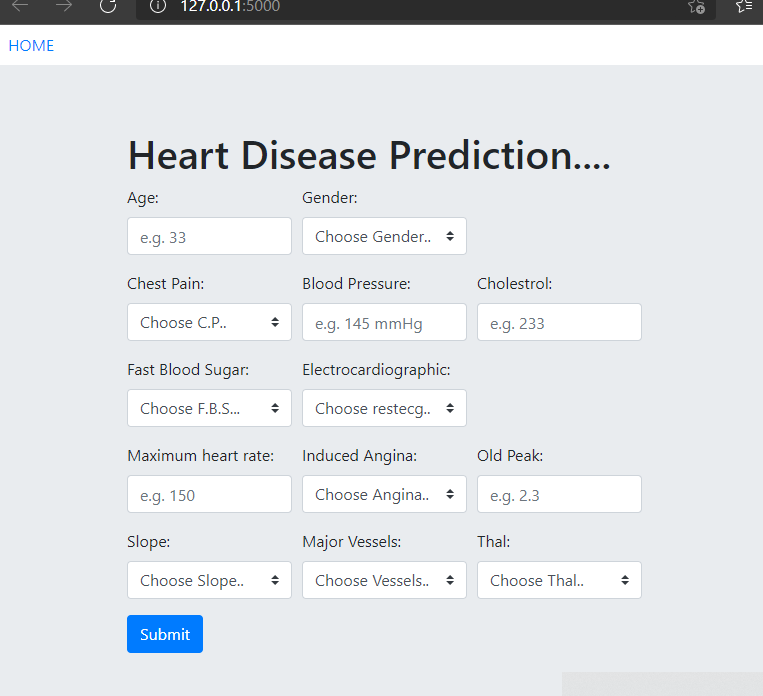


Figure 14 ENTER YOUR DITAILS

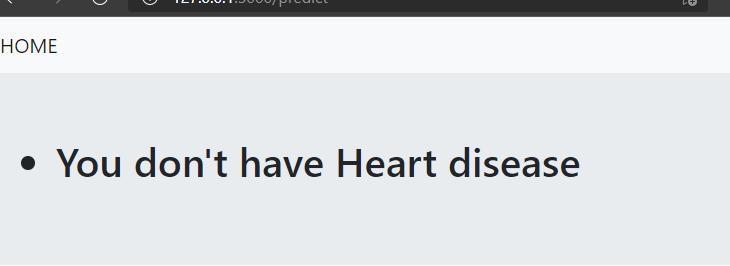


Figure 15 It is showing predition report

**CHAPTER 9 : CODE**

The coding portion were carried out to prepare the data, visualize it, pre-process it, building the model and then evaluating it. The code has been written in Python programming language using Jupyter Notebook as IDE. The experiments and all the models building are done based on python libraries. The code is available in the Git repository given in following link:

[**https://github.com/Rahul1-hy/Heart\_disease\_predition**](https://github.com/Rahul1-hy/Heart_disease_predition)

## **9.1 Libraries used:**

1. NumPy

2. Pandas

3. Matplotlib (pyplot, rcparams, matshow)

4. Pickle

5. flask

6.seabron

# **CHAPTER 10 : CONCLUSIONS**

1. Out of the 13 features we examined, the top 4 significant features that helped us classify between a positive & negative Diagnosis were chest pain type (cp), maximum heart rate achieved (thalach), number of major vessels (ca), and ST depression induced by exercise relative to rest (oldpeak).

2. Our machine learning algorithm can now classify patients with Heart Disease. Now we can properly diagnose patients, & get them the help they needs to recover. By diagnosing detecting these features early, we may prevent worse symptoms from arising later.

3. Our Random Forest algorithm yields the highest accuracy, 86%. Any accuracy above 70% is considered good, but be careful because if your accuracy is extremely high, it may be too good to be true (an example of Over fitting). Thus, 80% is the ideal accuracy!

# **REFERENCES**

[1] A. H. M. S. U. Marjia Sultana, "Analysis of Data Mining Techniques for Heart Disease Prediction," 2018.

[2] M. I. K. ,. A. I. ,. S. Musfiq Ali, "Heart Disease Prediction Using Machine Learning Algorithms".

[3] K. Bhanot, "towarddatascience.com," 13 Feb 2019. [Online]. Available: https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machinelearning-36f00f3edb2c.

[4] [Online]. Available: https://www.kaggle.com/ronitf/heart-disease-uci#heart.csv.

[5] M. A. K. S. H. K. M. a. V. P. M Marimuthu, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach".