Heart Disease Prediction Using Machine Learning Algorithms

A Project Report Submitted in the
Partial Fulfillment of the Requirements
for the Award of the Degree of

BACHELOR OF TECHNOLOGY

IN

INFORMATION TECHNOLOGY

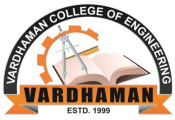
Submitted by

T.S.Jyothirmaye Chowdary	17881A1244
S.Rakesh	17881A1227
S.Hanish Rao	17881A1209

SUPERVISOR
E.Ravi kumar
Assistant Professor

Department of Information Technology
VARDHAMAN COLLEGE OF ENGINEERING, HYDERABAD
An Autonomous Institute, Affiliated to JNTUH

May, 2021



VARDHAMAN COLLEGE OF ENGINEERING, HYDERABAD

An Autonomous Institute, Affiliated to JNTUH

Department of Information Technology

CERTIFICATE

This is to certify that the project titled **Heart Disease Prediction Using Machine Learning Algorithms** is carried out by

T.S.Jyothirmaye Chowdary	**17881A1244**
S.Rakesh	**17881A1227**
S.Hanish Rao	**17881A1209**

in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Information Technology** during the year 2020-21.

Signature of the Supervisor	**Signature of the HOD**
E.Ravi Kumar	**Dr.Muni Sekhar Velpuru**
Assistant Professor	**Associate Professor and HOD**

Acknowledgement

The satisfaction that accompanies the successful completion of the task would be put incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crown all the efforts with success.

We wish to express our deep sense of gratitude to **E.Ravi Kumar**, Assistant Professor and Project Supervisor, Department of Information Technology, Vardhaman College of Engineering, for his able guidance and useful suggestions, which helped us in completing the project in time.

We are particularly thankful to **Dr.Muni Sekhar Velpuru**, the Head of the Department, Department of Information Technology, his guidance, intense support and encouragement, which helped us to mould our project into a successful one.

We show gratitude to our honorable Principal **Dr. J.V.R. Ravindra**, for providing all facilities and support.

We avail this opportunity to express our deep sense of gratitude and heartful thanks to **Dr. Teegala Vijender Reddy**, Chairman and **Sri Teegala Upender Reddy**, Secretary of VCE, for providing a congenial atmosphere to complete this project successfully.

We also thank all the staff members of Information Technology department for their valuable support and generous advice. Finally thanks to all our friends and family members for their continuous support and enthusiastic help.

T.S.Jyothirmaye Chowdary (17881A1244)

S.Rakesh (17881A1227)

S.Hanish (17881A1209)

Abstract

Coronary Illness is a major cause of death worldwide. The medical services field has a large amount of data, arranging that information using certain methods,data mining is one of the procedures used. This framework predicts the emerging prospects for coronary heart disease, namely, The results of this framework provide an opportunity for heart disease up to YES / NO. The databases used were collected up to the clinical boundaries. This framework examines those boundaries using mine data planning techniques. Datasets are managed by a python system using the basic Machine Learning Algorithm, Support Vector Classifier, K Neighbors Algorithm, Random Forest Algorithm showing excellent calculation between the above strategies to the appropriate level of coronary disease.

Table of Contents

List of Figures

Abbreviations

Abbreviation	Description
CVD	Cardiovascular disease
WHO	World health organization
KNN	K nearest Neighbours
SVM	Support Vector Machine

CHAPTER 1

Introduction

1.1 Introduction

Over past decade heart disease is the main reason for deaths. A person dies due to this every minute.A effective detection technique has to be introduced to lower the number of deaths. The algorithms used in this project are Decision tree,K-nearest Neighbour(KNN)s, Support Vector Classifier(SVC),Random Forest.

One of the significant organs in a human body is heart. Heart siphons blood through circulatory framework. The blood and oxygen is flowed by circulatory framework and if heart doesn't work as expected, the whole blood framework gets fallen. In this way, if heart doesn't work as expected , it prompts genuine medical issue, it can even prompt passings.

CVD has been expanding ordinarily on the planet. As indicated by WHO, an assessment of 17 million individuals bite the dust every year because of CVD, strokes and incomplete cardiovascular failures. Subsequently, it is important to record the significant indications and propensities that add to CVD. There are different tests that are performed earlier the determination of CVD, including ECG,Cholesterol,auscultation,blood pressing factor and glucose. These tests take quite a while and surprisingly long if the patient's condition perhaps basic and they begin taking prescription promptly, so it is vital to focus on the tests. There are numerous wellbeing propensities that add to CVD. Because of the expansion in the measure of information, ML has become the most arising field . ML makes it conceivable to separate information from enormous information. Moreover,the distinctive ML calculations has been looked at utilizing enhancement calculations. In light of grouping, 70 level of information is prepared and the rest 30 rate is tried. [1]

1.2 Types of Heart Disease:

CVD aka the heart disease includes blood, heart off body. Myocardial infarctions is the part of CVD. Coronary Heart Disease (CHD) is another type of a Heart disease, in this, plaque is developed in coronary arteries. This plaque blocks the vessels completely through course of times.

The Symptoms of Heart Attacks are:

1. Chest Pain: Most common symptom is the chest pain. The cause happens because, a blockage in a coronary vessel.
2. Low in oxygen:cause for plaque, the degrees of oxygen decreases in body and causes tipsiness and equilibrium misfortune.
3. Arms pain: This sort of agony regularly begins in chest and moves towards basically left arm.
4. Excessive Swearing: Sweating is also a common symptom.
5. Tiredness: This causes to fatigue,ehich means chores becomes harder.
6. Bradycardia: The patients heartbeat becomes slower to 60 bpm.
7. Diabetics: The patient have a rate of 100 bpm and ocasionally has heart rate 130 bpm.
8. Cerebrovascular Disease: Patient will have high pulse than ordinary, its typically 200bpm or higher than that which causes a coronary episode.

Some different reasons can be the propensities for way of life like smoking and other eating habits.It is assessed that more than 17.5 million passings are reason for CVD. India contains a more than 30M CVD cases at the present time. More prominent than two lakhs open a medical procedure each year are finished.

CHAPTER 2

Literature Survey

2.1 Existing System

A research which was conducted in 2000 by Shusaku Tsumoto, says that human beings were not able of arranging huge size data, so it was suggested to use data mining techniques which are available for finding various kinds of patterns which is available from huge database and then can again be used for clinical research purposes and to perform various operations on it.

A few tests were led on the clinical informational indexes utilizing different characterizes and highlight determination methods. There is an examination on coronary illness information set.Many of those contain great precision.

Medical services has came long way for treatment of patients with various types of diseases.

Finding of patients accurately and managing viable medicines have become a significant test.

Determination of the condition exclusively relies on In the current system,practical utilization of different gathered information is tedious. [2]

2.2 Proposed System

In proposed framework, We created project, in which the framework predicts the potential outcomes of event of coronary illness as far as percentage's, It additionally figures out which calculation is best for the forecast of heart disease.It is truly appropriate for coronary illness expectation continuously.

The primary errand of the information expectation is finished utilizing four strategies.

Main Method used for the prediction is Decision tree,KNN algorithm,SVM and Random forest. The systems use 13 medical attributes or parameters

as inputs, the datasets go through mining and displays the accuracy of each algortithm. By implementation of computerized system, limitations in the proposed system will be reduced. [3]

Advantages:

1)Lower man power.

2)Higher efficiency.

3)High Accuracy.

4)Less consumption of time.

CHAPTER 3

Analysis

3.1 Software requirement specifications

3.1.1 Software Requirements:

The technical requirements in order to develop algorithms for the prediction of heart disease are:

-Programming language: Python
-Programming Environment: Google colab or Jupyter Notebook
-Operating System: Windows 7 or above
-Browser: Google Chrorme or Mozilla Fire-fox or IE.

3.1.2 Hardware Requirements:

This requirements includes device that have the access to any of the latest browser viz, IE9,Mozilla,Chrome,etc.,which supports above techniques.Memory requirements are same as the browser requirements.

-CPU type: Intel Corei5
-Clock speeed: 1.8GHz
-RAM size: 4GB
-Hard disk capacity: 1TB
-Keyboard type: Basic English Keyboards
-Monitor type: 15 inch colour monitor

3.1.3 Functional Requirements:

These are the techniques that are being used in this project, they are:

-Decision Tree

-KNN Algorithm

-SVC

-Random Forest.

3.2 Architecture Design or Flow Chart:

Figure 3.1: Architecture Design

CHAPTER 4

Design

4.1 UML/ER diagrams:

4.1.1 Use Case Diagram

This below outline shows relationship among entertainers and usecase, it is a bunch of situations which portrays the connections between a client and framework. There are two main components, they are:

1)Actors

2)User.

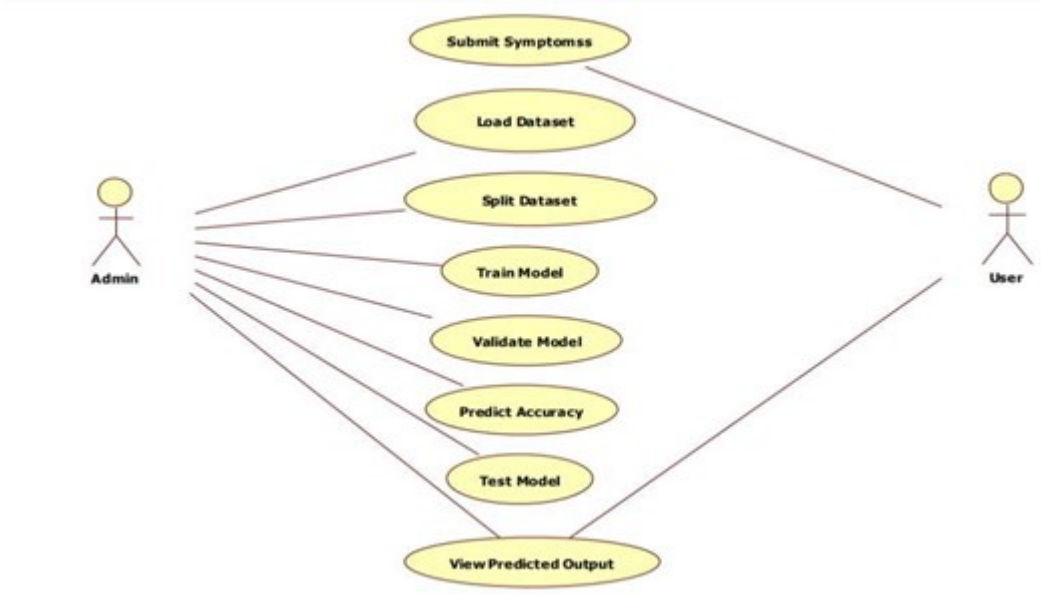


Figure 4.1: Usecase Diagram

4.1.2 Class Diagram

The below figure gives gives static views. These are built with structural thing like Interfaces, classes and Collaboration between them.

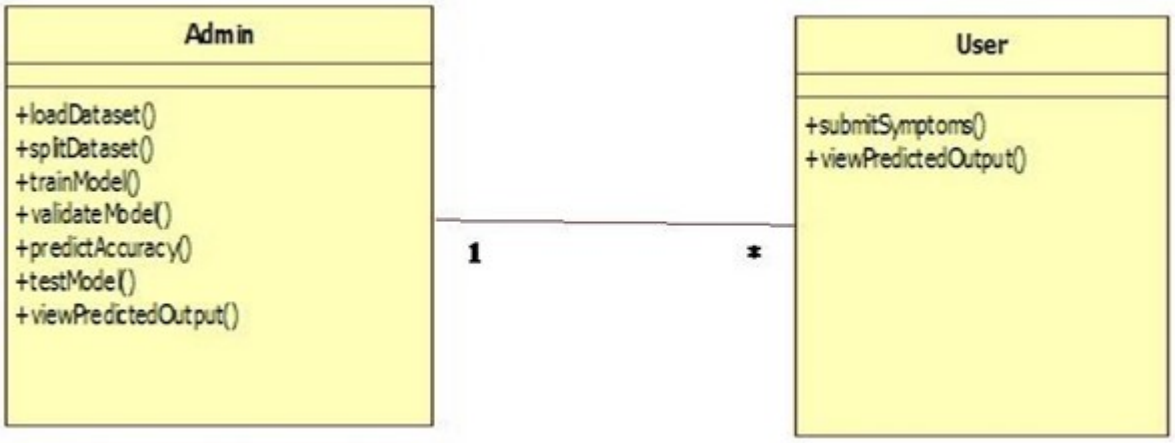


Figure 4.2: Class Diagram

4.1.3 Sequence Diagram

This chart shows time grouping of articles taking part in connections. This graph likewise shows as equal and vertical lines, different cycles and articles that lives simultaneously,and additionally as flat bolts and messages will be traded in the middle them,in request to their event.

The specification of the simple run time scenario is in a graphical manner.

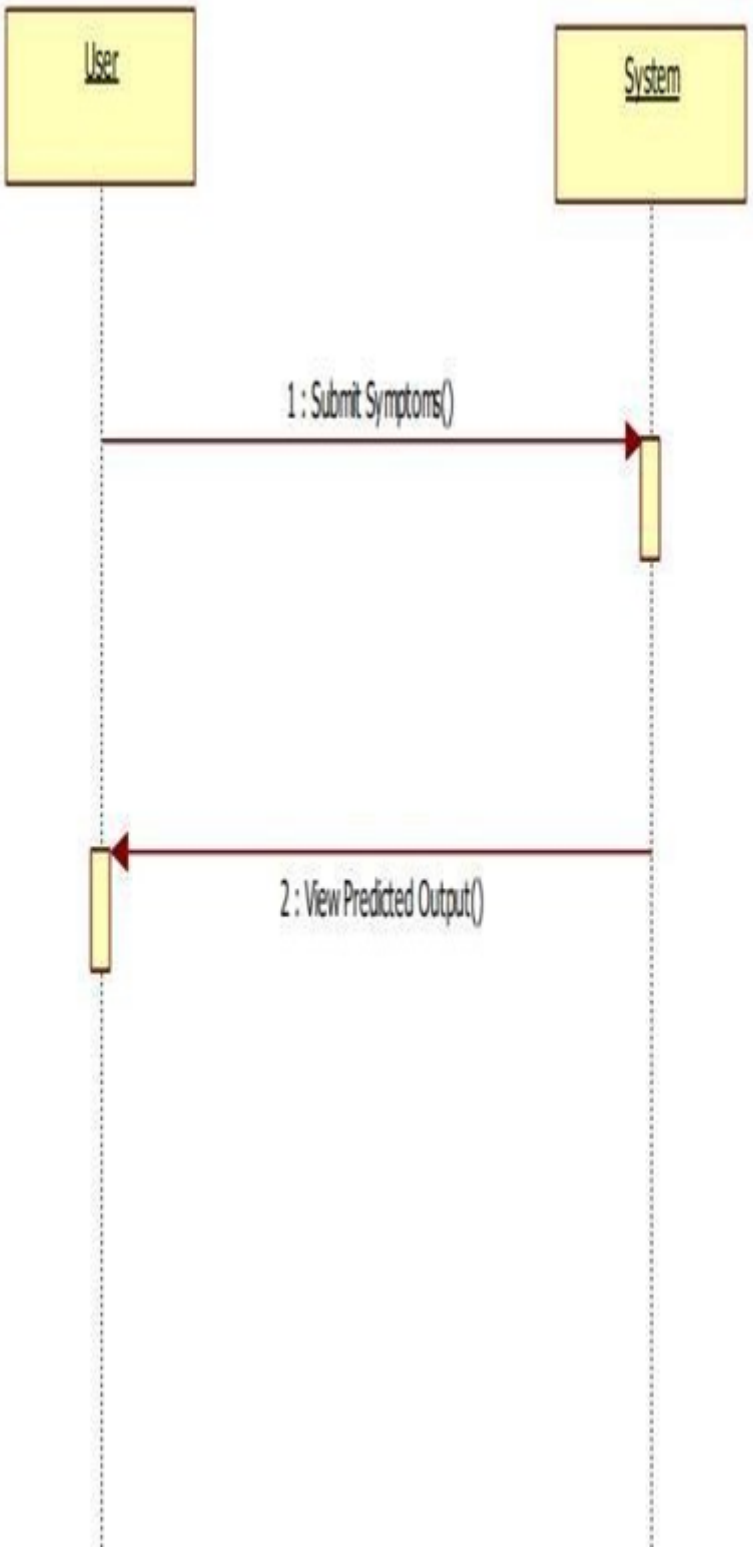


Figure 4.3: Sequence diagram of User

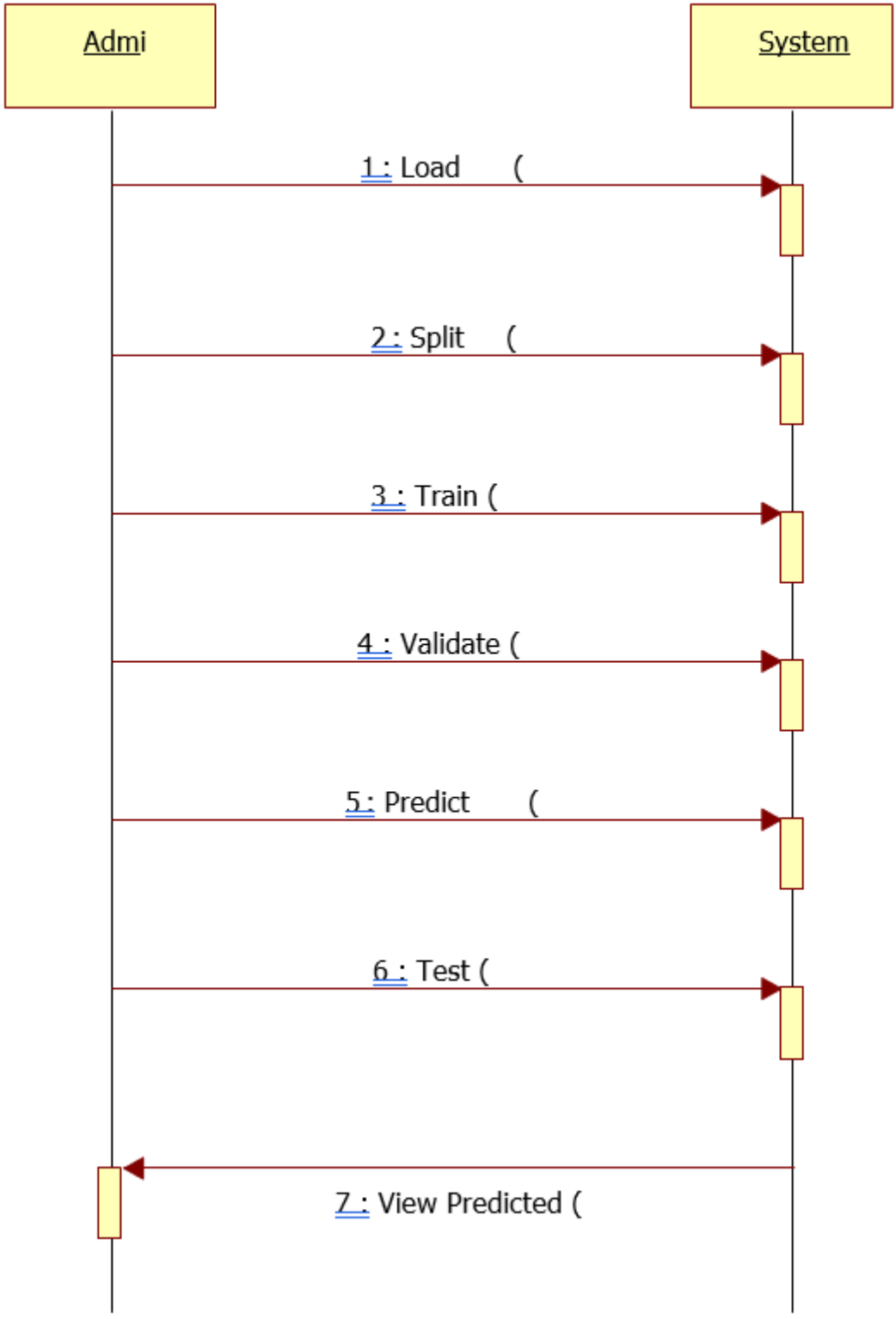


Figure 4.4: Sequence diagram of Admin

4.1.4 Activity Diagram

This chart shows the condition of graph where the greater part of the states are in real life states and a large portion of them are set off by the

fruition of activities in source states.This mostly centers around the stream driven by the interior preparing.

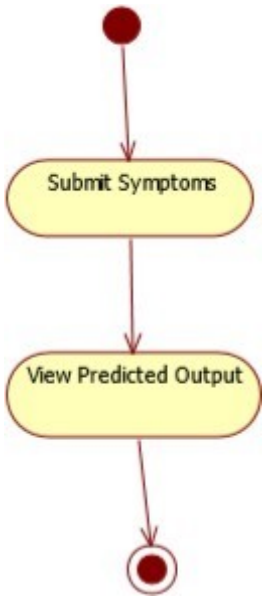


Figure 4.5: Activity diagram of User

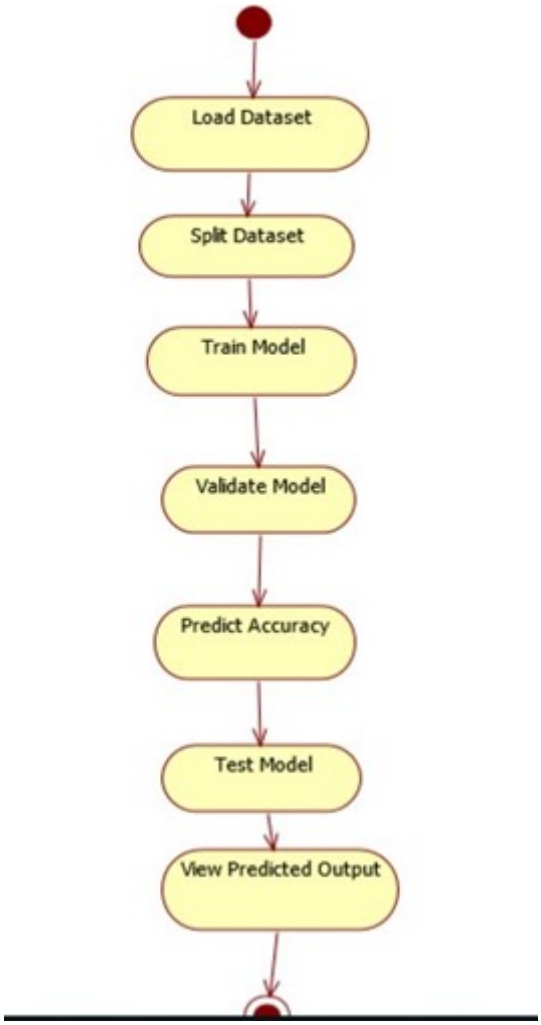


Figure 4.6: Activity diagram of Admin

4.1.5 ER

ER diagram represnts the Entity and relationships.

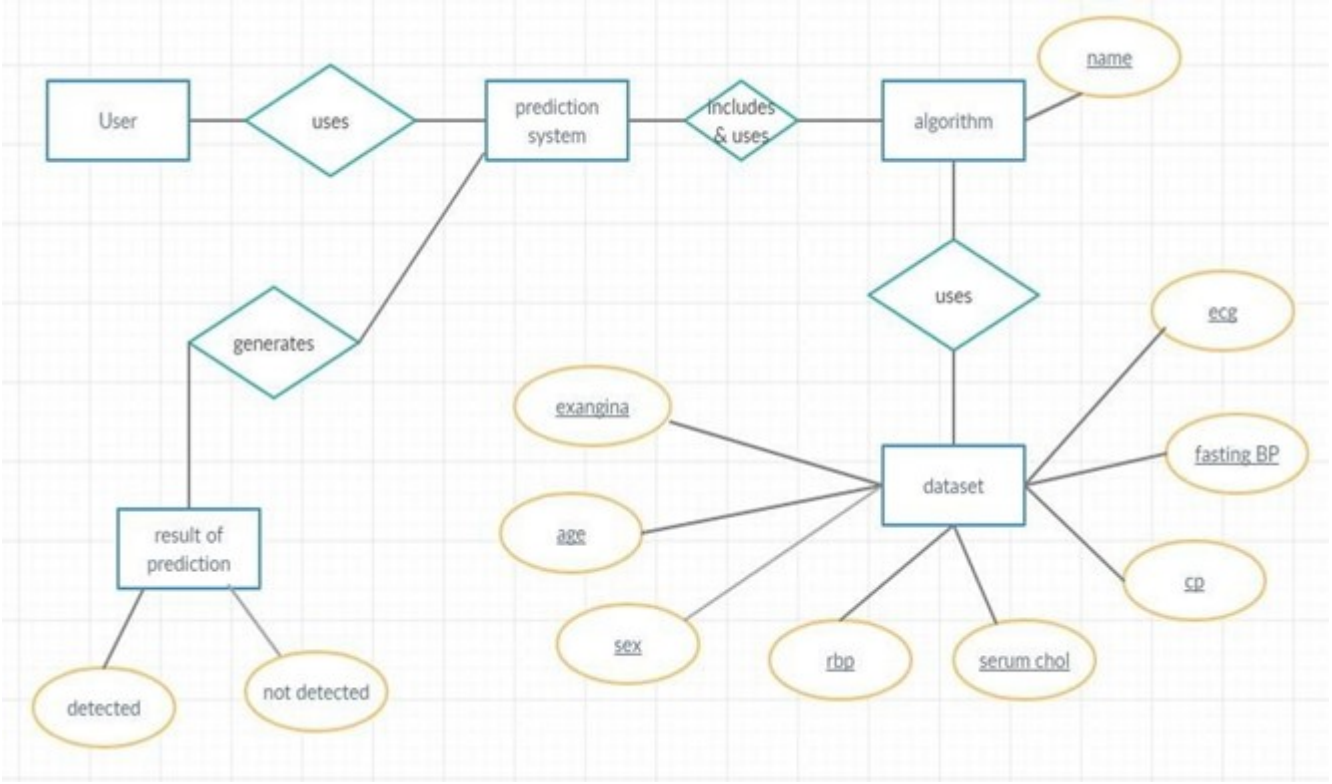


Figure 4.7: ER Diagram

4.2 Dataset Design

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	0
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	0
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	0
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	0
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	0
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	0
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
27	71	0	1	160	302	0	1	162	0	0.4	2	2	2	1
28	59	1	2	150	212	1	1	157	0	1.6	2	0	2	1
29	51	1	2	110	175	0	1	123	0	0.6	2	0	2	1

Figure 4.8: Dataset

CHAPTER 5

Implementation and Result Analysis

5.1 Modules

DATA SOURCE:

The data used to anticipate coronary artery disease was taken from the UCI ML database. Various databases used to create ML techniques. The database is a real dataset and contains about 300 examples of information covering 14 clinical boundaries. Database parameters are related to tests that are considered to indicate coronary heart disease such as bp level, type of chest pain and electro-cardiographic effect and so on.

This data set comprises of 76 atrributes,all the distributed investigations allude to the utilization of a subset of 14 of them. In Particular, Cleveland data set is the solitary information base that was utilized by the ML analysts to date. In this project, apart from Cleveland, we also used other database like Switzerland and Budapest.

The "target" issue shows the presence of coronary coronary heart disorder in a patient,Its either int value 0 (Not present) or 1 (presence).[4]

The attributes that are used in this system are:

1. age - The age in years
2. sex - type of sex (1 = male; 0 = female)
3. cp - type of chest pain – Value as 1: typical angina – Value as 2: atypical angina – Value as 3: non-anginal pain – Value as 4: asymptomatic
4. trestbps - resting blood pressure (in mm Hg on admission to the hospital)
5. chol - The serum cholestoral in mg/dl
6. fbs - fating blood sugar > 120 mg/dl (1 = true; 0 = false)

7. restecg - resting electro-cardio-graphic results – Value as 0: normal – Value as 1: having ST-T wave abnormality (The t wave inversions and/or S-T elevation or depression of 0.05 mV or greater)
   – Value as 2: showing probable or definite left ventricular hypertrophy by Estes' criteria

8. thalach - The maximum heart rate that's achieved

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

10. oldpeak - The ST depression induced by the exercise relative to rest

11. slope - slope of the peak exercise in ST segment – Value as 1: upsloping – Value as 2: flat – Value as 3: downsloping

12. ca - number of the major vessels (0-3) colored by/with flourosopy

13. thal - If 3 = normal; 6 = fixed defect; 7 = it is a reversible defect

14. target (the predicted attribute) - The diagnosis of the heart disease (angiographic disease status)
   – Value 0 if: $< 50\%$ diameter narrowing
   – Value 1 if: $> 50\%$ diameter narrowing

5.1.1 Sample datasets

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

Figure 5.1: Sample Dataset

5.1.2 Description of the datasets

1.CLEVELAND:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	tar
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000

Figure 5.2: Cleveland dataset

2.Budapest:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	tar
count	297.000000	297.000000	297.000000	297.000000	297.000000	297.000000	297.000000	297.000000	297.000000	297.000000	297.000000	297.000000	297.000000	297.000
mean	54.542088	0.676768	3.158249	131.693603	247.350168	0.144781	0.996633	149.599327	0.326599	1.055556	1.602694	0.676768	4.730640	0.461
std	9.049736	0.468500	0.964859	17.762806	51.997583	0.352474	0.994914	22.941562	0.469761	1.166123	0.618187	0.938965	1.938629	0.499
min	29.000000	0.000000	1.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	1.000000	0.000000	3.000000	0.000
25%	48.000000	0.000000	3.000000	120.000000	211.000000	0.000000	0.000000	133.000000	0.000000	0.000000	1.000000	0.000000	3.000000	0.000
50%	56.000000	1.000000	3.000000	130.000000	243.000000	0.000000	1.000000	153.000000	0.000000	0.800000	2.000000	0.000000	3.000000	0.000
75%	61.000000	1.000000	4.000000	140.000000	276.000000	0.000000	2.000000	166.000000	1.000000	1.600000	2.000000	1.000000	7.000000	1.000
max	77.000000	1.000000	4.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	3.000000	3.000000	7.000000	1.000

Figure 5.3: Budapest dataset

3.Switzerland:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	tar
count	270.000000	270.000000	270.000000	270.000000	270.000000	270.000000	270.000000	270.000000	270.000000	270.00000	270.000000	270.000000	270.000000	270.000
mean	54.433333	0.677778	3.174074	131.344444	249.659259	0.148148	1.022222	149.677778	0.329630	1.05000	1.585185	0.670370	4.696296	0.444
std	9.109067	0.468195	0.950090	17.861608	51.686237	0.355906	0.997891	23.165717	0.470952	1.14521	0.614390	0.943896	1.940659	0.497
min	29.000000	0.000000	1.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.00000	1.000000	0.000000	3.000000	0.000
25%	48.000000	0.000000	3.000000	120.000000	213.000000	0.000000	0.000000	133.000000	0.000000	0.00000	1.000000	0.000000	3.000000	0.000
50%	55.000000	1.000000	3.000000	130.000000	245.000000	0.000000	2.000000	153.500000	0.000000	0.80000	2.000000	0.000000	3.000000	0.000
75%	61.000000	1.000000	4.000000	140.000000	280.000000	0.000000	2.000000	166.000000	1.000000	1.60000	2.000000	1.000000	7.000000	1.000
max	77.000000	1.000000	4.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.20000	3.000000	3.000000	7.000000	1.000

Figure 5.4: Switzerland dataset

5.2 Understanding the dataset

- The visualizations are used for better understanding of the data and to look at any processing that we might want to do.

- Confusion matrix: It is a network or a table that is utilized to quantify the exhibition of a ML calculation, it as a rule is a regulated learning..

Each column of disarray grid addresses occurrences of a genuine class and every col addresses examples of anticipated class.It mirrors a reality that its simple for us see what sort of disarrays happens in characterization calculations.

For instance, calculations ought to need to anticipate an example as cici cause real class is cici, yet calculation came out as cjcj.In the instance of mislabeling component cm[i,j]cm[i,j] and it will be augmented by one, when disarray grid is built.

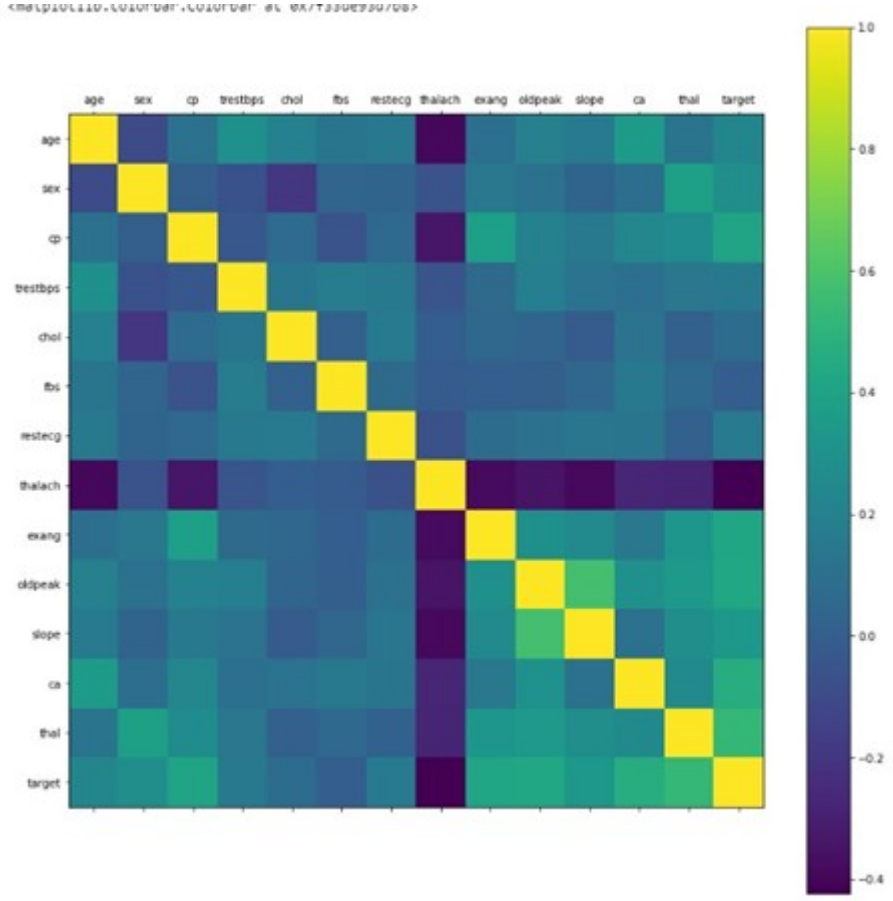


Figure 5.5: Confusion matrix

By taking a gander at the network, it's not difficult to see a couple of highlights that have negative relationship with target variable while some of them have positive.

Now, we will take a look at all the histpgrams for each and every variable.

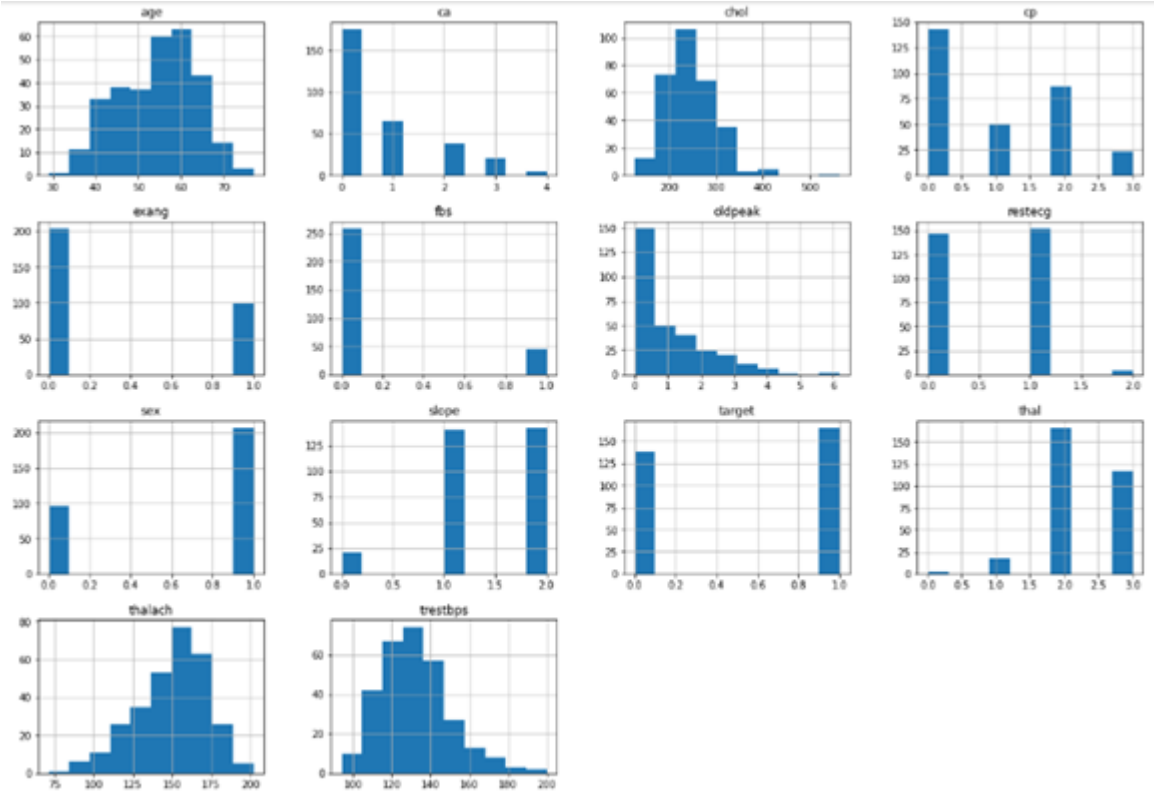


Figure 5.6: Histogram

-The above histograms shows that each variable has different range of distributions.

-So, using the scaling before predictions is of great use. And also the categorical features do stand out.

-It is always good to work with dataset where target variable are of approx equal size.

5.3 Processing the data

In the wake of investigating dataset, It can be seen that it is expected to change over some straight out factors into sham factors and those are to be scaled prior to preparing the ML models.

At First, we will utilize the get.dummies strategy for making faker segments for the clear cut factors.

For handling information, we need to import a couple of libraries, to part the dataset for testing and preparing, we have utilized the train.test.split technique. To scale the highlights, we need to utilize StandardScaler.

5.4 Splitting the data set

- We then have imported train.test.split for splitting the data set into training and test datasets. Then, we have imported all ML models that we will be using for training and testing the data.

- ML algorithms works in two stages,the test stage and train stage.

- The preparation dataset is our underlying dataset that is utilized to prepare calculation to comprehend and how to apply the advances like neural organizations, and to learn and create the mind boggling results.

- It likewise incorporates both the information and furthermore the comparing anticipated yield. The principle motivation behind the train dataset is for furnishing a calculation with "ground truth" information.

- The test informational index, however,it is utilized to evaluate how well the calculation was prepared with the preparation set. You can't just re-utilize the train set in the testing stage cause calculation will be as of now "knowing" anticipated yield, which losses reason for test the calculations.

- The data is not ready for our ML applications.

1. DECISION TREE:

- we consider the preparation set as the root. - Highlight esteem are liked to be as an unmitigated. In the event that qualities are continous, they must be discretized before the structure the model. - Based on credits, records must be conveyed recursively. - Utilization of measurable techniques for the requesting ascribes, as root or as the inward hub. - IIn Decision Tree, the fundamental test is the distinguishing proof of quality for root hub in the each level. The interaction is known as trait choice. We likewise have 2 famous properties choice measures., They are:

1)IG - Information Gain

2)Giini Index

Decision Tree Code:

```
    dt_scores = []
for i in range(1, len(X1.columns) + 1):
   dt_classifier = DecisionTreeClassifier(max_features = i, random_state = 0)
    dt_classifier.fit(X_train1, y_train1)
    dt_scores1.append(dt_classifier.score(X_test1, y_test1))

   plt.plot([i for i in range(1, len(X1.columns) + 1)], dt_scores1,
   color = 'green')
for i in range(1, len(X1.columns) + 1):
    plt.text(i, dt_scores1[i-1], (i, dt_scores1[i-1]))
plt.xticks([i for i in range(1, len(X1.columns) + 1)])
plt.xlabel('Max features')
plt.ylabel('Scores')
plt.title('Decision Tree Classifier scores for different number of
maximum features')
```

2. K Nearest Neighbours :

–Algorithm:

-Consider m as the quantity of preparing information tests. Furthermore, p being an obscure point.
- First Store preparing tests in cluster of information focuses arr[]. This implies all component of the exhibit addresses a tuple (x, y).

for i== 0 too next m:
Calculate distance using euclidean formula for d(arr[i], p).
Also, Make a set specifically S of K littlest distance got before. Every one of these distances relates to effectively ordered information point.
Return most of the mark among S.

KNN Code:

```
    knn_scores = []
for k in range(1,31):
    knn_classifier = KNeighborsClassifier(n_neighbors = k)
    knn_classifier.fit(X_train1, y_train1)
    knn_scores1.append(knn_classifier.score(X_test1, y_test1))

plt.plot([k for k in range(1, 31)], knn_scores1, color = 'red')
for i in range(1,31):
    plt.text(i, knn_scores1[i-1], (i, knn_scores1[i-1]))
plt.xticks([i for i in range(1, 31)])
plt.xlabel('Number of Neighbors (K)')
plt.ylabel('Scores')
plt.title('K Neighbors Classifier scores for different K values')
```

3. SVM :

SVM Code:

```
    svc_scores = []
for i in range(len(kernels)):
    svc_classifier = SVC(kernel = kernels[i])
    svc_classifier.fit(X_train1, y_train1)
    svc_scores1.append(svc_classifier.score(X_test1, y_test1))

    colors = rainbow(np.linspace(0, 1, len(kernels)))
plt.bar(kernels, svc_scores1, color = colors)
for i in range(len(kernels)):
    plt.text(i, svc_scores1[i], svc_scores1[i])
plt.xlabel('Kernels')
plt.ylabel('Scores')
plt.title('Support Vector Classifier scores for different kernels)')
```

4. RF (Random Forest):

We can comprehend the working of the Random Forest calculation with assistance of the accompanying advances:

Stage 1 :Right off the bat, start with choice of arbitrary example from the given dataset.

Stage 2 : Then, the calculation will build the choice tree for every single example. At that point it will get forecast result from each choice tree.

Stage 3 : In this progression, casting a ballot will be performed for every single anticipated outcome.

Stage 4 : In last advance, select the most casted a ballot foreseeing result as the last expectation result.

Random Forest Code:

```
    rf_scores = []
for i in estimators:
    rf_classifier = RandomForestClassifier(n_estimators = i,
    random_state = 0)
    rf_classifier.fit(X_train1, y_train1)
    rf_scores1.append(rf_classifier.score(X_test1, y_test1))

    colors = rainbow(np.linspace(0, 1, len(estimators)))
plt.bar([i for i in range(len(estimators))], rf_scores1,
color = colors,width = 0.8)
for i in range(len(estimators)):
    plt.text(i, rf_scores1[i], rf_scores1[i])
plt.xticks(ticks = [i for i in range(len(estimators))],
labels = [str(estimator) for estimator in estimators])
plt.xlabel('Number of estimators')
plt.ylabel('Scores')
plt.title('Random Forest Classifier scores for different number of estimators')
```

These calculations are applied to the equivalent dataset to break down the best calculation as far as exactness.

The KNN model has predicted for an accuracy level of 87Decision Tree classifier has predicted an accuracy level of a 79Random Forest classifier has predicted an accuracy level of a 84SVC has predicted an accuracy level of a 83In this way, the best calculation for the forecast of coronary illness is the KNN model. In this undertaking, Four Machine learning calculations were applied on the dataset to anticipate the best calculation for Heart Disease and a GUI is assembled utilizing the best calculation i.e KNN. [5]

5.4.1 GUI Implementation

Pseudo code

1. import Libraries

   import tkinter

2. import mat plotlib.pyplot as plt

3. from mat plotlib.cm import rainbow from sklearn.tree import DecisionTree Classifier from sklearn.ensemble import RandomForestClassifier

4. Importing Datasets heart= pd.readcsv ("Dataset.csv")

5. Extracting Independent and Dependent variables

   y = heart['target']

   X = heart.drop (['target'], axis = 1)

6. Parting the dataset into preparing and test set.

7. Fitting Random Forest classifier to the training set

   classifier= RandomForestClassifier (nestimators=
   classifier.fit (Xtrain, Ytrain)

8. Predicting the test set result finalResult=classifier.predict([inputValues])
   print (finalResult)

9. Taking User Input deftakeInput(): inputValues=[]
   age1 = ((int (age.get ())- 29) / (77-29))
   inputValues.append (age1)

10. Creating the Result Prediction Window substituteWindow= tkinter.Tk
    () substituteWindow.title ("RESULT PREDICTION")
    substituteWindow.columnconfigure (0, weight=2) substituteWindow.rowconfigure
    (5, weight=1) if finalResult[0] == 1:
    label1=tkinter.Label(substituteWindow, text="HEART DISEASE DETECTED",
    font=('Impact', -35), fg='0080ff')
    label1.grid (row=0, column=1, columnspan=6)
    else:

    label1 = tkinter.Label (substituteWindow, text="NO DETECTIOIN OF
    HEART DISEASES",
    font= ('Impact', -35)) label1.grid (row=2, column=1, columnspan=6)
    label2 = tkinter.Label (substituteWindow, text="Do not forget to exer-
    cise daily. ",
    font=('Impact',-20), fg='green') label2.grid (row=3, column=1,columnspan=6)
    substituteWindow.mainloop ()

11. Creation of Main Window mainWindow= tkinter.Tk () mainWindow.title
    ("HEART DISEASE PREDICTION")

    label1 = tkinter.Label(mainWindow, text="HEART DISEASE PREDIC-
    TION MODEL",
    font=('Impact', -35), bg='ff8000') label1.grid (row=0, column=0, columnspan=6)
    label2 = tkinter.Label(mainWindow, text="Enter the details carefully",
    font=('Impact', -20) , fg='white', bg='ff00bf') label2.grid (row=1, col-
    umn=0, columnspan=6)
    analyseButton=tkinter.Button(mainWindow,text=". ANALYZE/PREDICT",
    font=('Impact', -15), bg = 'red', command=takeInput)

mainWindow.mainloop ()

12. Frame for the feature inputs ageFrame= tkinter.LabelFrame (mainWindow, text="Age(yrs)")
ageFrame.grid (row=2, column=0) ageFrame.config (font= ("Courier", -15)) age= tkinter.Entry (ageFrame) age.grid (row=2, column=2, sticky='nw')

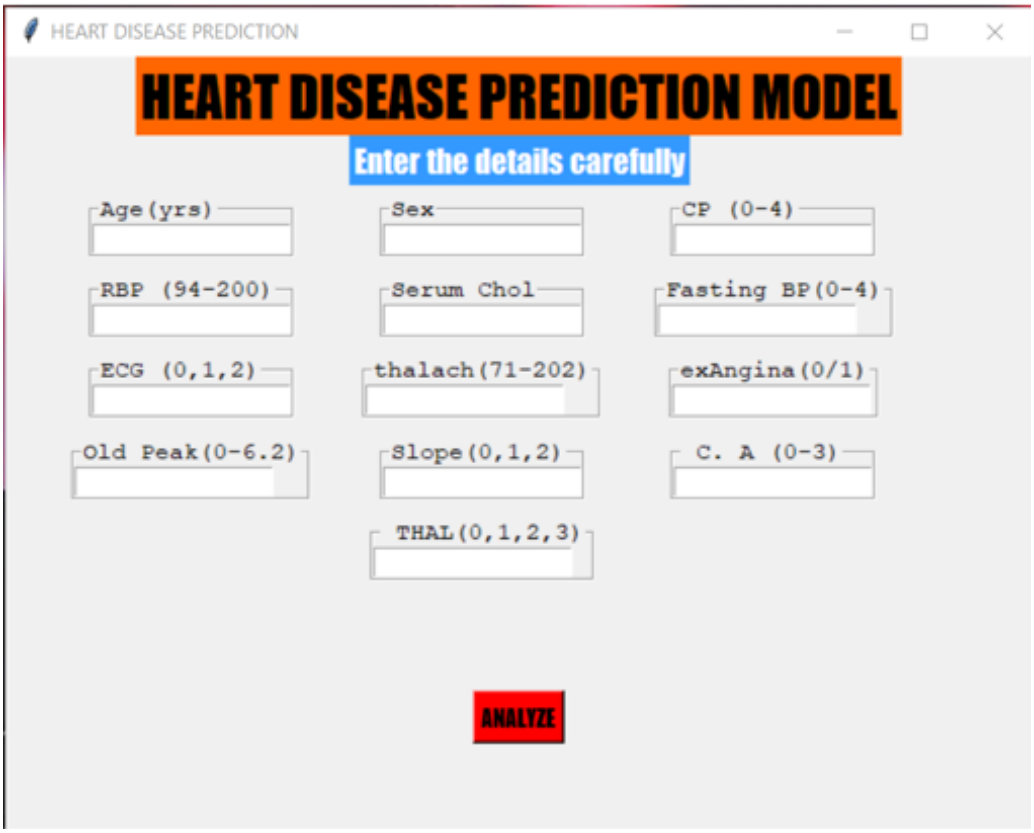


Figure 5.7: GUI

Now, Give the input values to the fields to predict the occurrence of the Heart Disease.

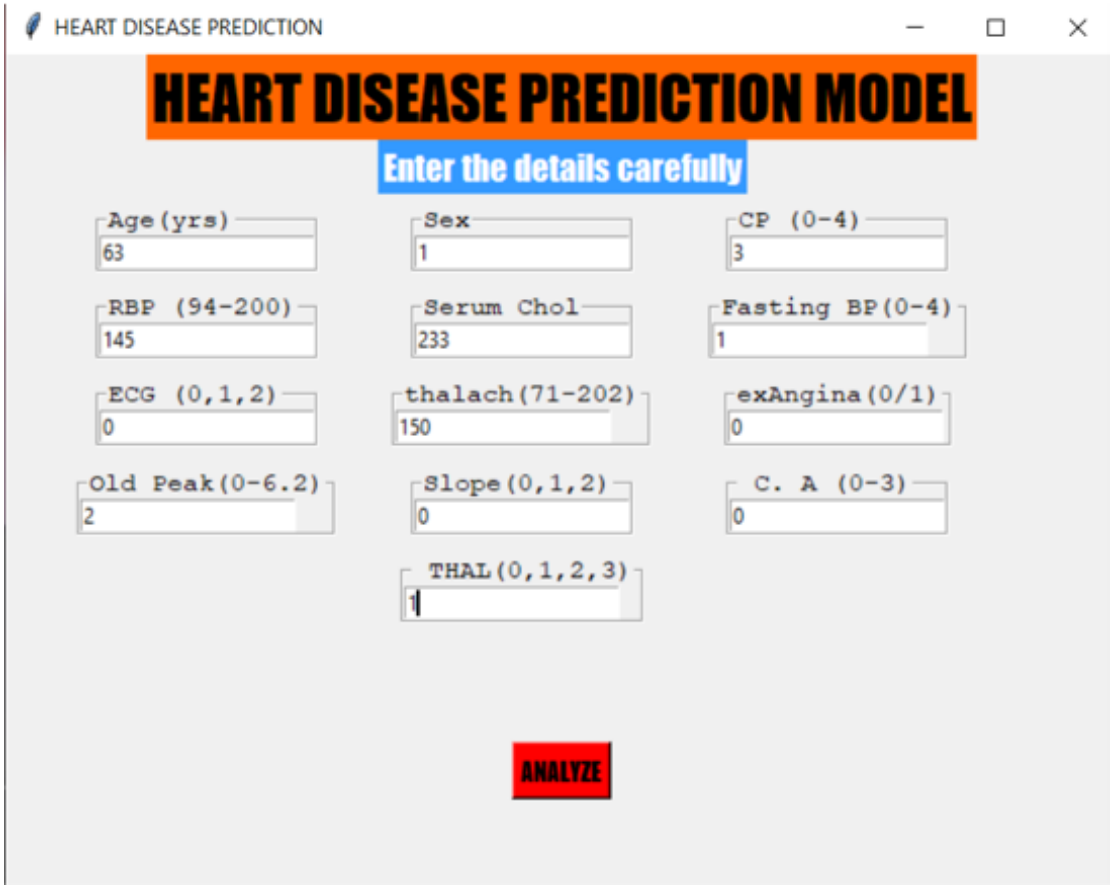


Figure 5.8: Inputting the values into the GUI

After giving the input values, the system predicts the occurrence of the Heart Disease as seen in the following picture.

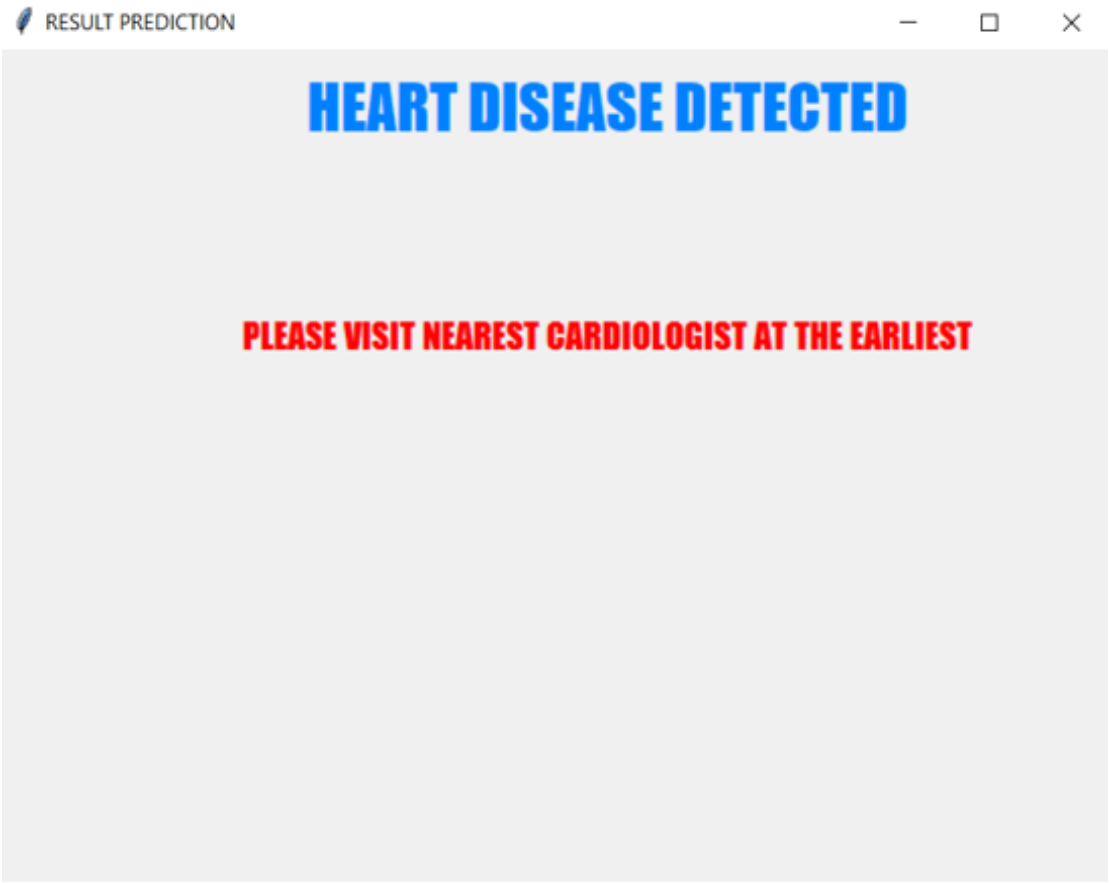


Figure 5.9: Final Output

CHAPTER 6

Testing and Result

This project has been tested upon various datasets like Cleveland,Budapest,Switzerland, which contains 14 attributes in total whose output screens are as follows:

6.1 Understanding the data using confusion matrix

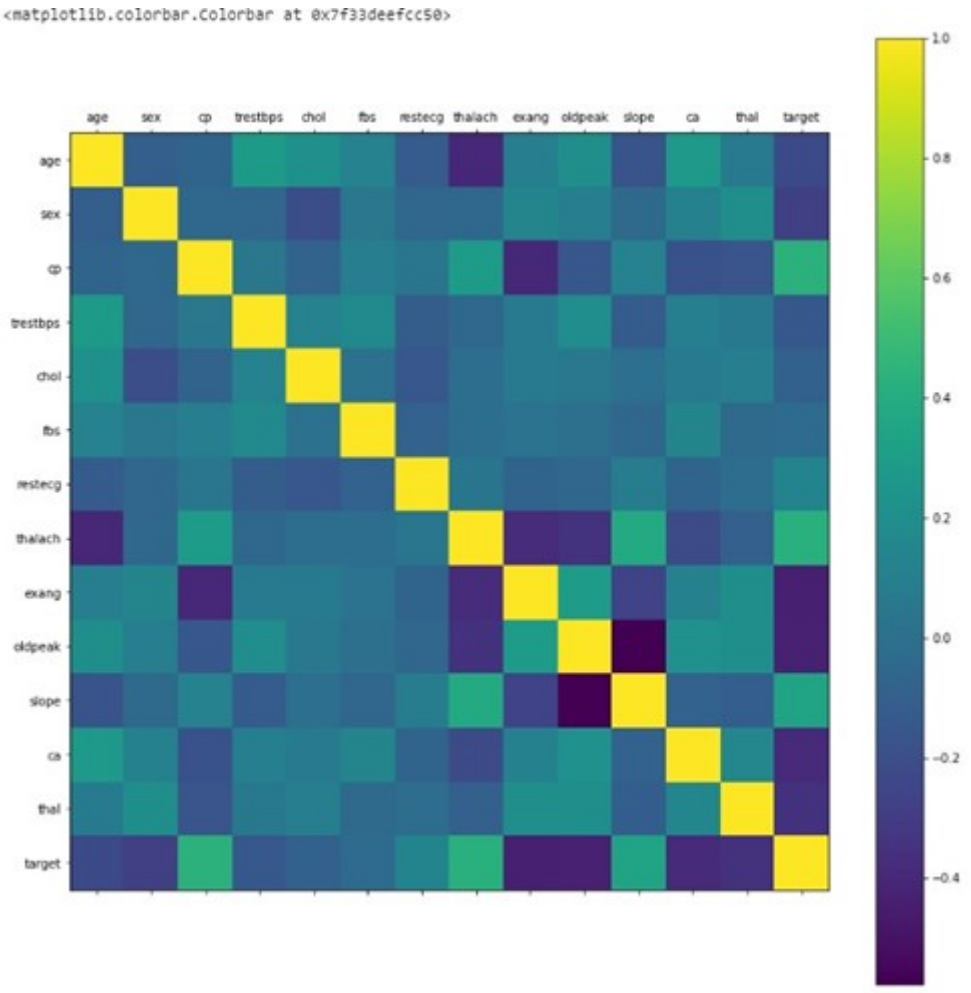


Figure 6.1: Confusion matrix of Cleveland

Confusion matrix helps us in the better understanding the relationship

between the attributes, We can easily identify the correlation between target variable and the other attributes, which helps us in better understanding of the attributes and the relation needed to built the model. We have found the confusion matrix for all datasets

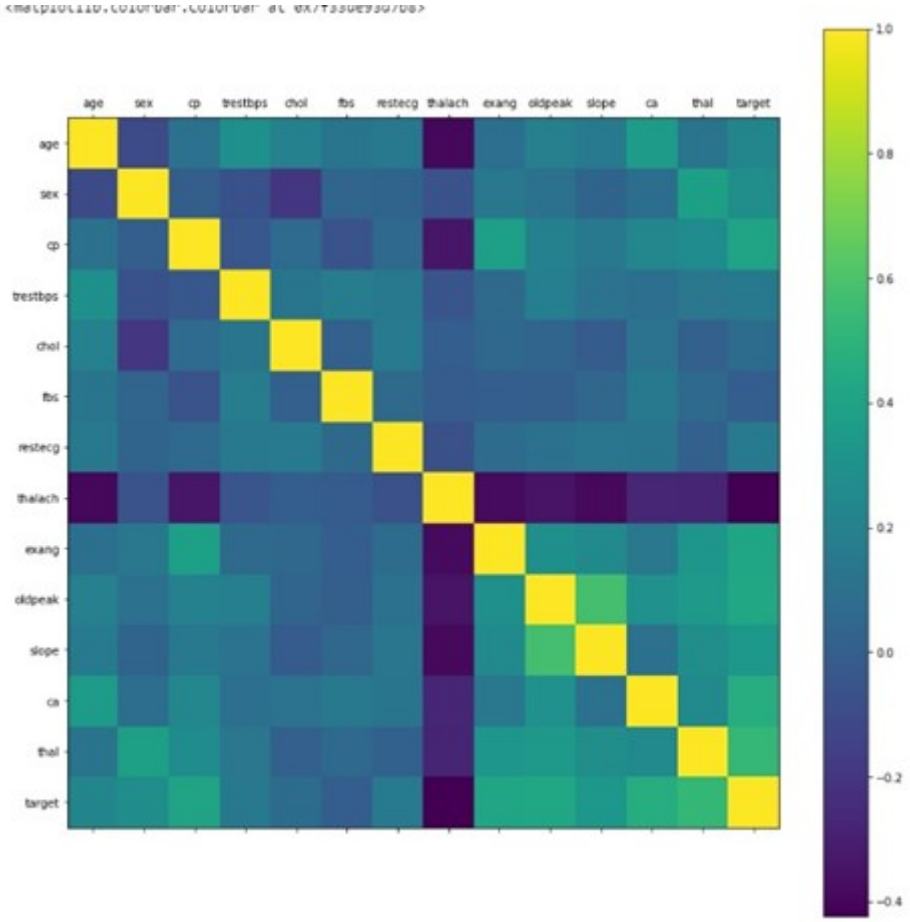


Figure 6.2: Confusion matrix of Budapest

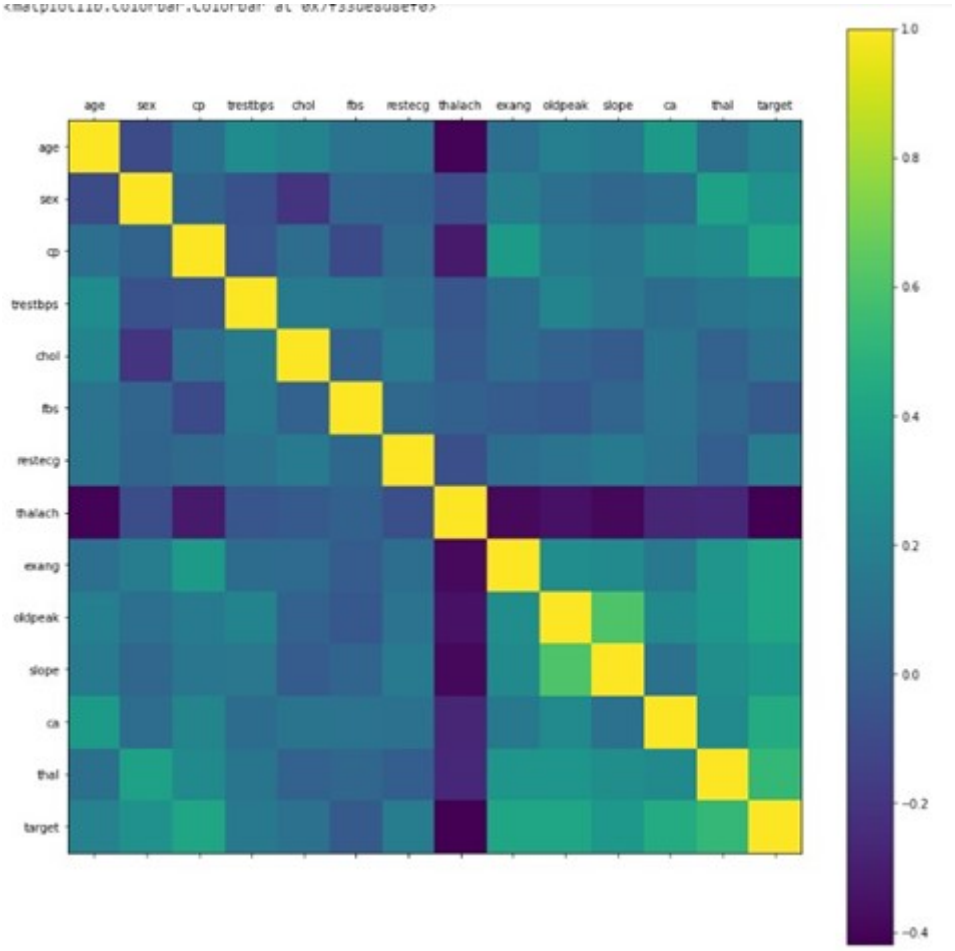


Figure 6.3: Confusion matrix of Switzerland

6.2 Decision Tree

We have applied Decision tree algorithm to find the final scores of each dataset which helps us in finding the accuracy, based on the accuracy we will be creating a GUI with the best algorithm.

The following are the results for all the three datasets

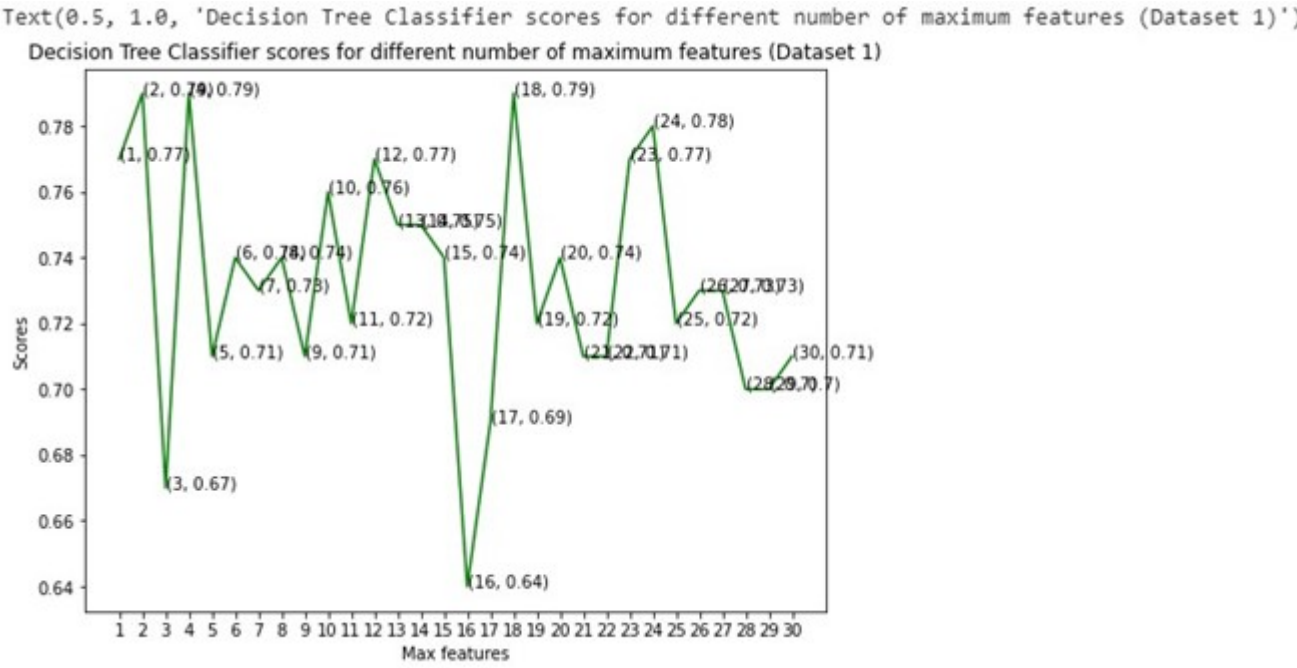


Figure 6.4: Result of Cleveland (Decision Tree)

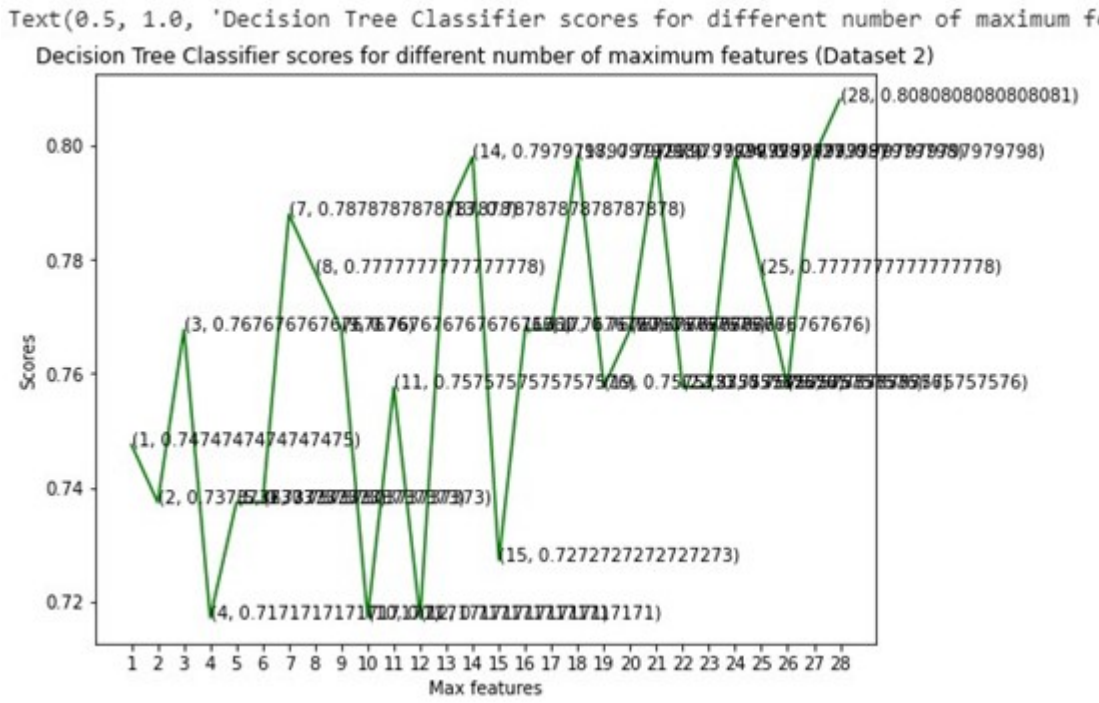


Figure 6.5: Result of Budapest (Decision Tree)

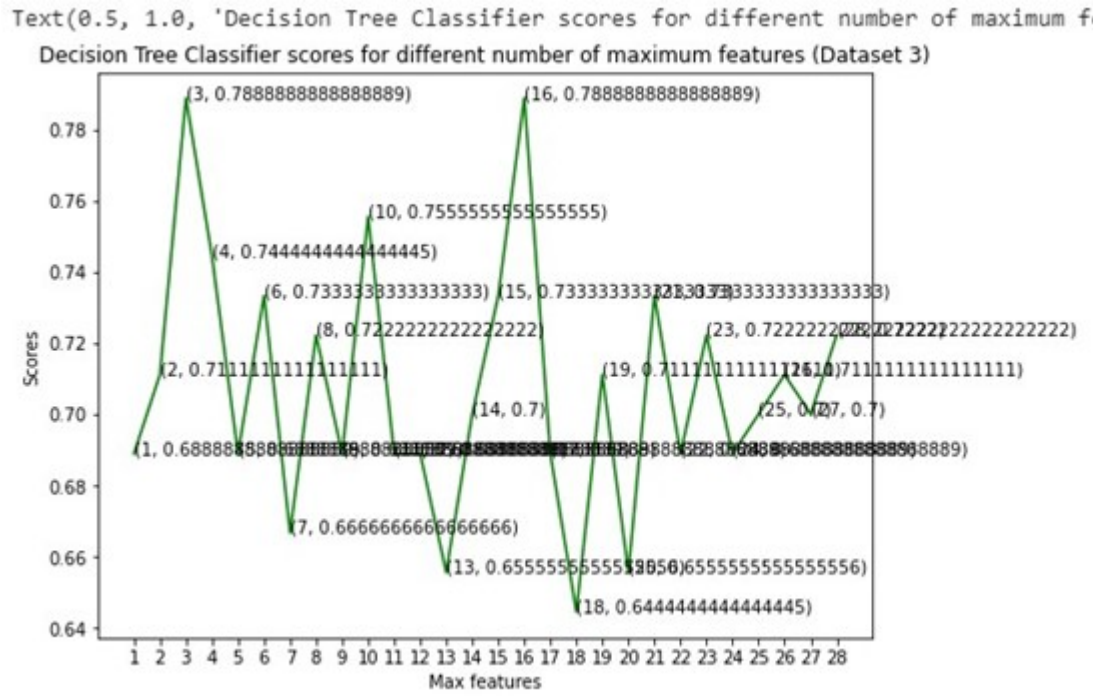


Figure 6.6: Result of Switzerland(Decision Tree)

6.3 KNN

We have applied KNN algorithm to find the final scores of each dataset which helps us in finding the accuracy, based on the accuracy we will be creating a GUI with the best algorithm.

The following are the results for all the three datasets

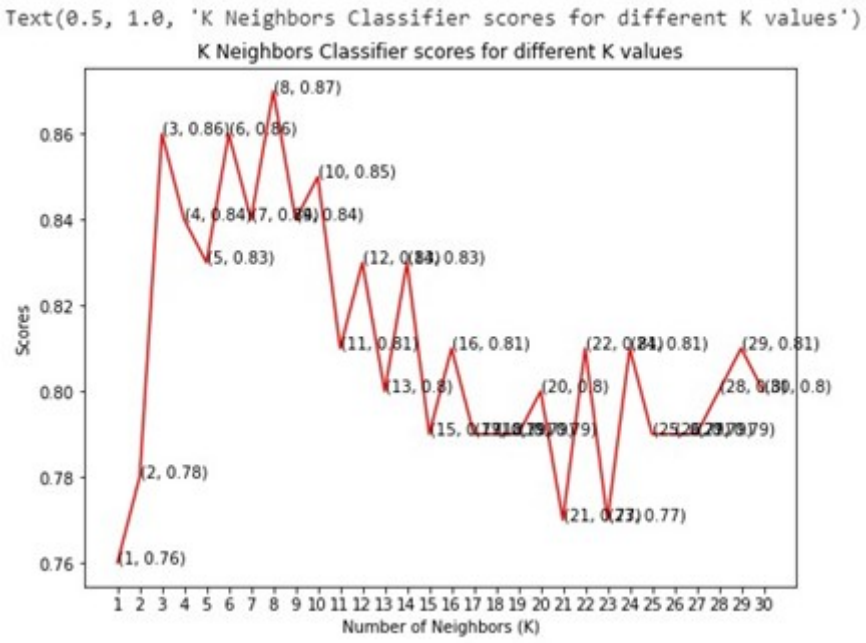


Figure 6.7: Result of Cleveland (KNN)

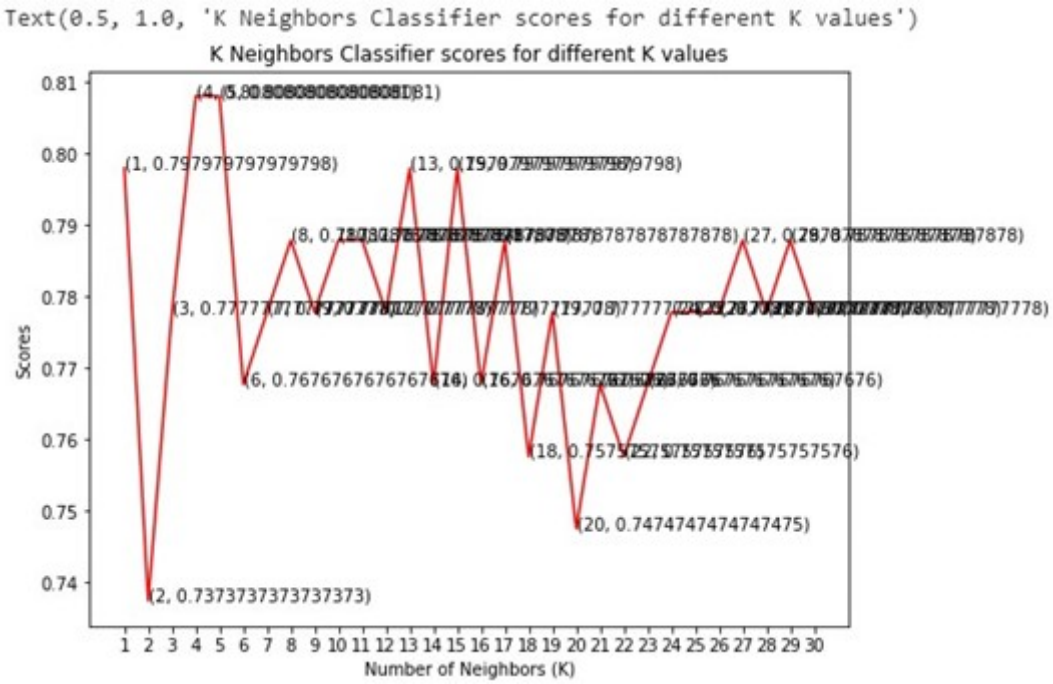


Figure 6.8: Result of Budapest (KNN)

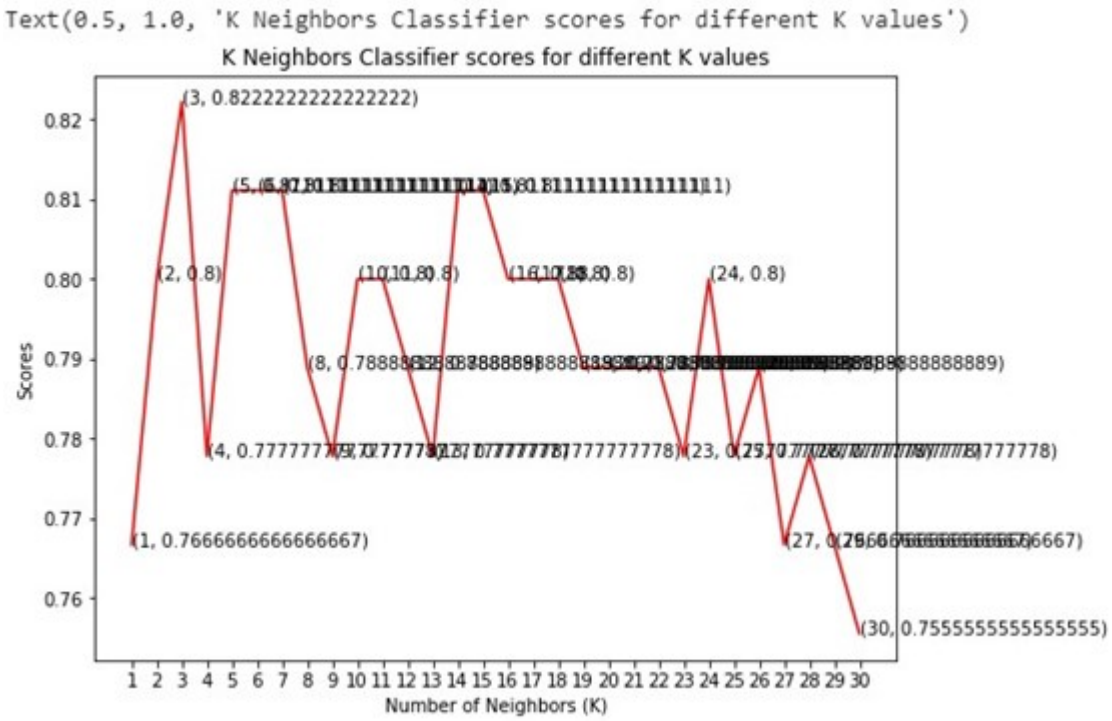


Figure 6.9: Result of Switzerland (KNN)

6.4 SVM

We have applied SVM algorithm to find the final scores of each dataset which helps us in finding the accuracy, based on the accuracy we will be creating a GUI with the best algorithm.

The following are the results for all the three datasets

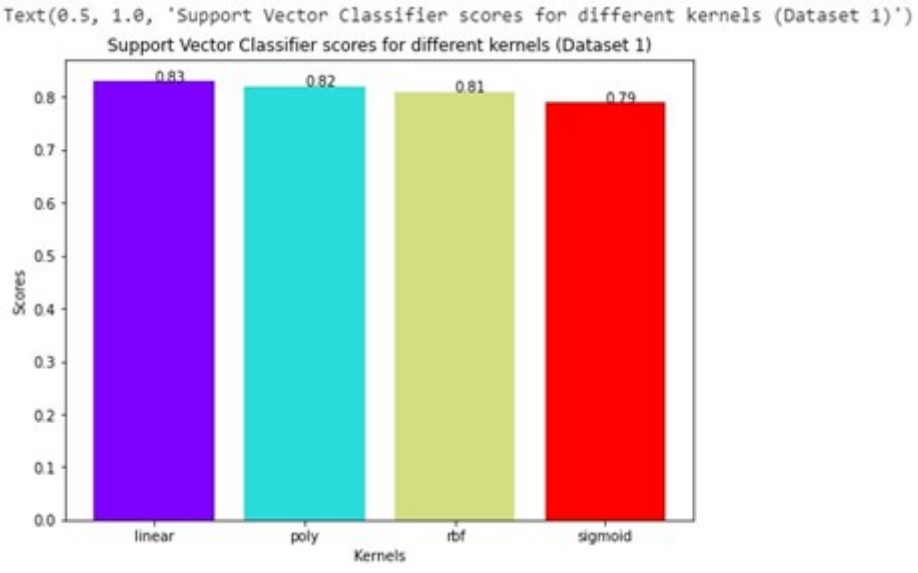


Figure 6.10: Result of Cleveland (SVM)

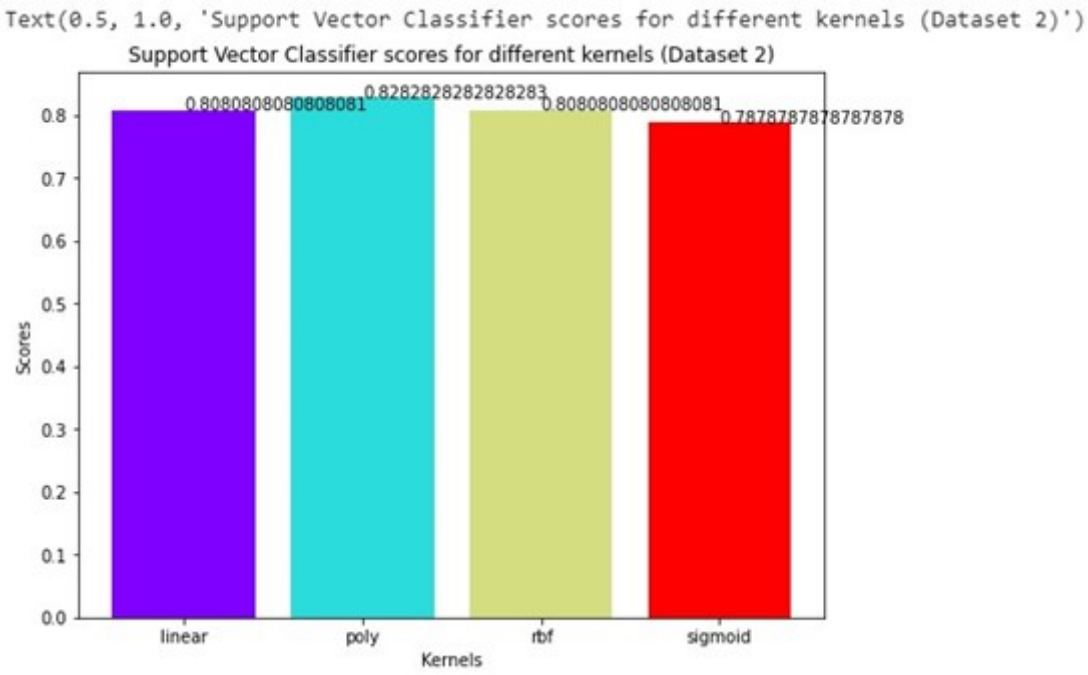


Figure 6.11: Result of Budapest (SVM)

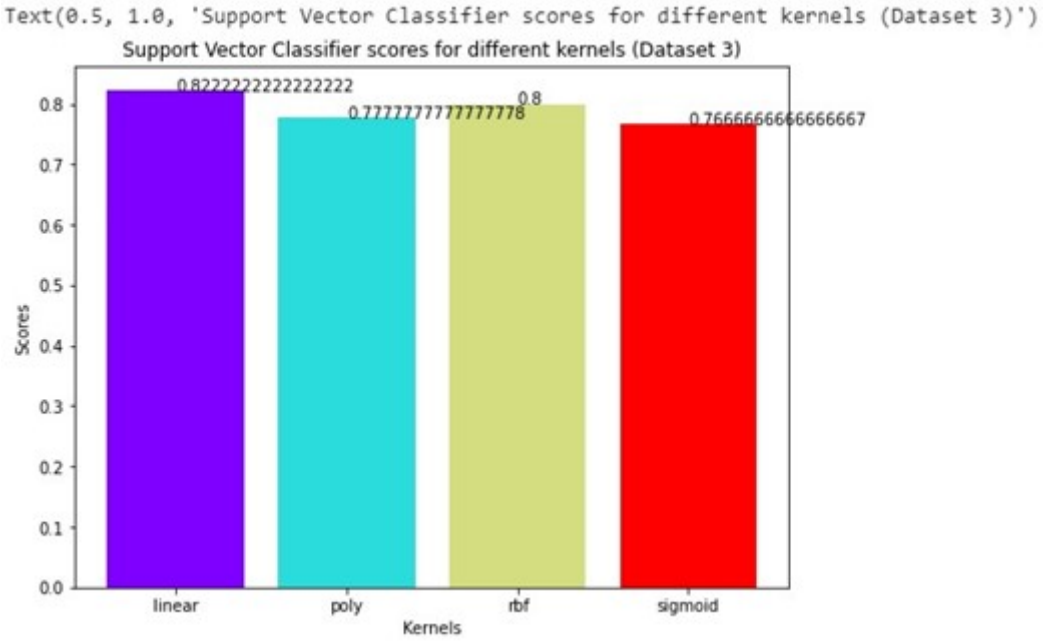


Figure 6.12: Result of Switzerland (SVM)

6.5 Random Forest

To find the final scores of each dataset which helps us in finding the accuracy, based on the accuracy we will be creating a GUI with the best algorithm.

The following are the results for all the three datasets

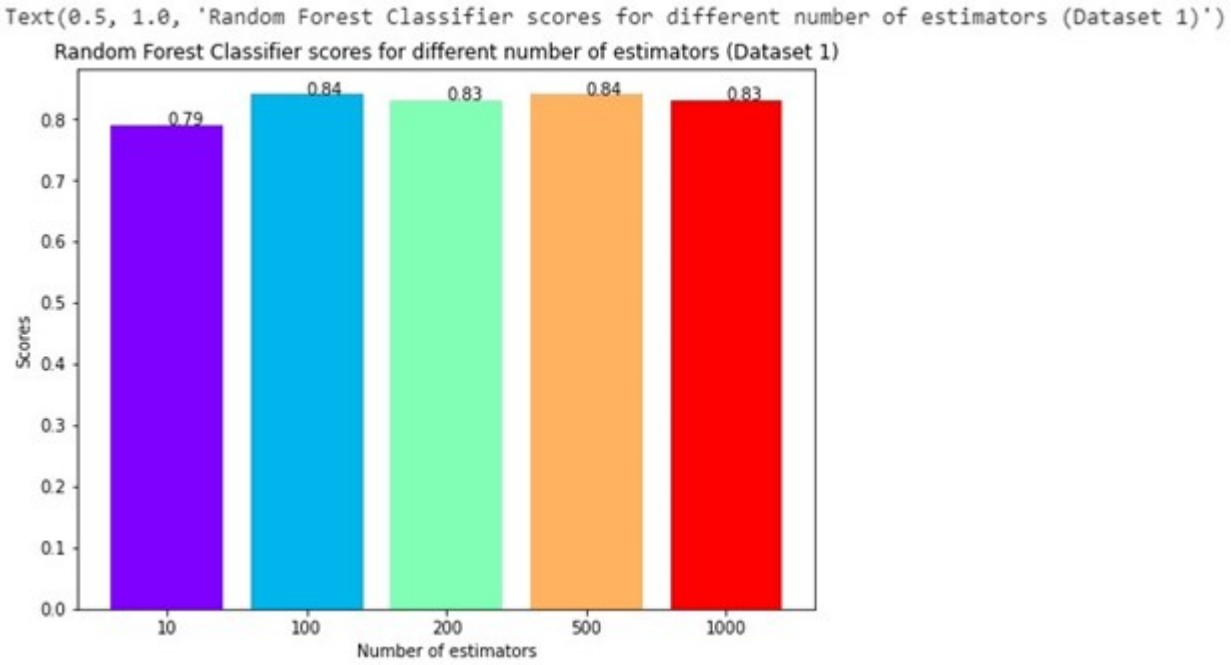


Figure 6.13: Result of Cleveland (Random Forest)

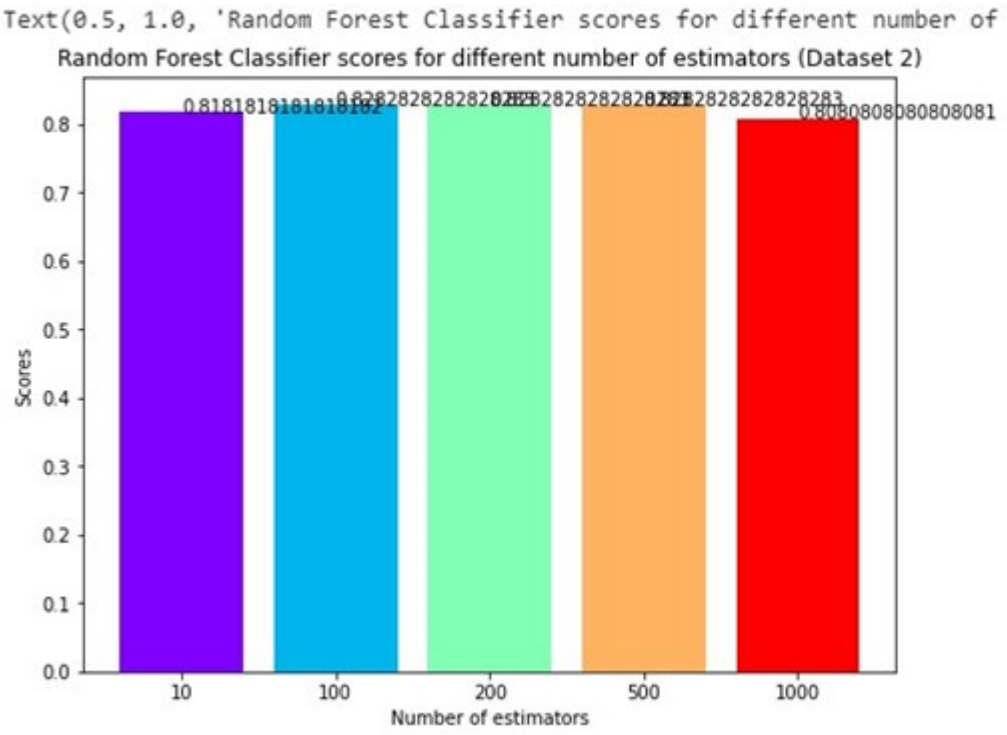


Figure 6.14: Result of Budapest (Random Forest)

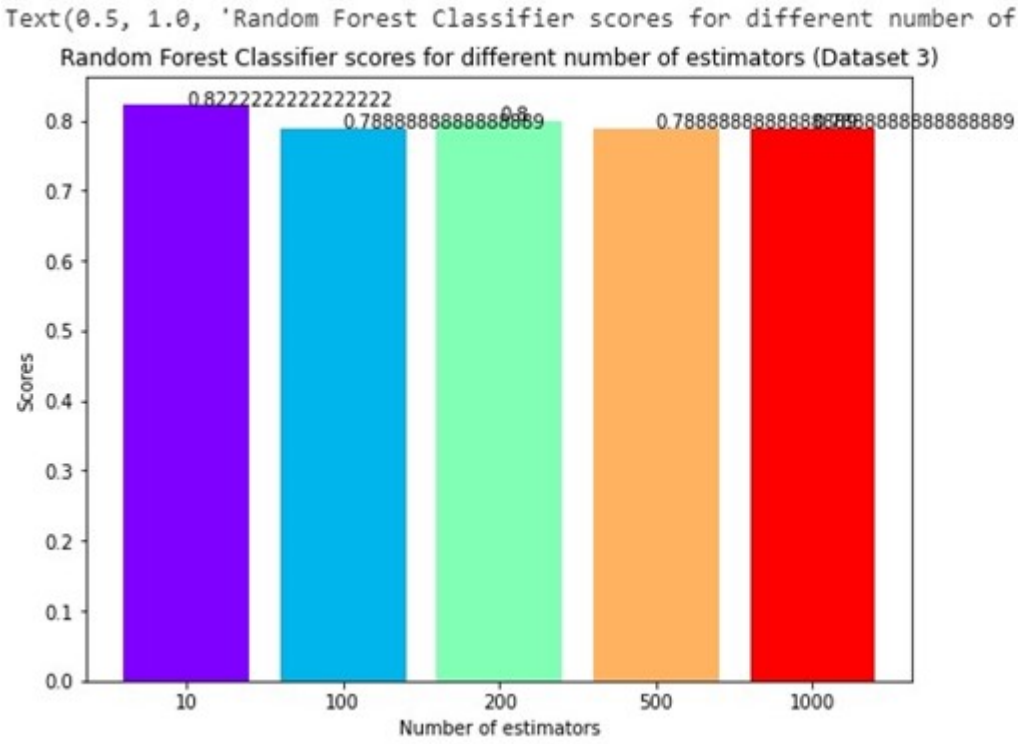


Figure 6.15: Result of Switzerland (Random Forest)

The below plot is the final comparison of accuracy for all four algorithms.

As we can see from the below plot that KNN has the highest accuracy, So KNN algorithm is used to build the GUI to predict the occurrence of heart

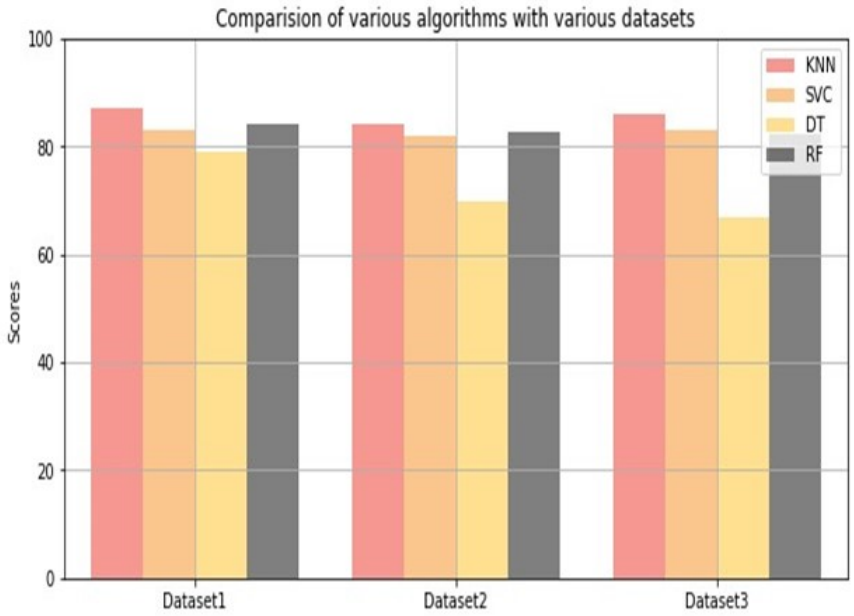


Figure 6.16: Final Comparison of the Algorithms

disease.

CHAPTER 7

Conclusion

The point of the venture is to know whether the patient has a coronary illness or not. The records gave in the informational indexes are isolated into train set and test sets. After pre-preparing information, the information mining and arrangement strategies specifically choice trees, KNN Algorithm, irregular timberlands are applied. This is finished utilizing Python Programming language. Results are produced for both the preparation datasets and furthermore test information sets.The Supervised information mining calculations were applied on dataset to foresee potential outcomes of having a coronary illness of the patient, they were broke down with the order model named KNN, Random Forest and Decision Tree. Later on, this framework with the pre-owned ML arrangement calculation can likewise be utilized to foresee or to analyze different illnesses. This work can likewise be broadened or improved for robotization of heart infections investigation additionally including some other ML calculations. While utilizing various sorts of information mining and ML methods to foresee event of coronary illness have been summed up. Decide forecast execution of every calculation and furthermore apply on the proposed framework for regions it required. Utilize more applicable component determinations strategy, to improve the precision of calculations. There are likewise a few medicines for patient, on the off chance that they determined to have a specific kind of coronary illness once. Information mining can likewise be of entirely learned structure for such appropriate dataset.

All in all, as its distinguished through the writing overview, it is accepted just a minimal achievement has been accomplished in production of prescient model for the coronary illness patients and subsequently there is a requirement for the mixes and more unpredictable model to build the precision of forecast of it in the right off the bat set of coronary illness. With more measure

of information being taken care of into data set and the framework will be exceptionally astute.

REFERENCES

[1] Santhana Krishnan and S Geetha. "Prediction of Heart Disease Using Machine Learning Algorithms." In: *2019 1st international conference on innovations in information and communication technology (ICIICT)*. IEEE. 2019, pp. 1–5.

[2] Austin H Chen, Shu-Yi Huang, Pei-Shan Hong, Chieh-Hao Cheng, and En-Ju Lin. "HDPS: Heart disease prediction system". In: *2011 computing in cardiology*. IEEE. 2011, pp. 557–560.

[3] K Aravinthan. "Heart Attack Prediction Using Data Mining Techniques". In: *International Journal of Pure and Applied Mathematics* 119.12 (2018), pp. 16119–16123.

[4] Sharyu U Kamble, Vaishnavi S Jawanjal, Pooja P Velapure, Priya K Jadhav, and Sanjivani S Kadam. "Heart Disease Prediction using Machine Learning Techniques". In: *IJETT* 6.1 (2019).

[5] NN Khanom, F Nihar, SS Hassan, and L Islam. "Performance Analysis of Algorithms on Different Types of Health Related Datasets". In: *Journal of Physics: Conference Series*. Vol. 1577. 1. IOP Publishing. 2020, p. 012051.