ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of this project would be incomplete without mentioning the people who made it possible, without whose constant guidance and encouragement would have made efforts go in vain. I consider myself privileged to express gratitude and respect towards all those who have guided through the completion of projects.

We convey thanks to our project guide **Ms Monali Gandhi**, Information Technology department, CGPIT for providing encouragement, constant support and guidance which was of a great help to complete this project work successfully.

We are grateful to **Ms Palak Desai**, Head of the Department, Information Technology, CGPIT for giving us the support and encouragement that was necessary for the completion of this project.

We would also like to express our gratitude to **Dr Rajkumar V. Patil**, Director, Chhotubhai Gopalbhai Patel Institute of Technology for providing us a congenial environment to work in.

We would like to thank all the faculty members for their patience, understanding and guidance that gave us strength and will power to work through the long tedious hours for developing a project and preparing the report.

Last but not least, We would also like to thank our colleagues, friends, who have co-operated during the preparation of our report and without them, this project has not been possible. Their ideas helped me a lot to improve our project report.

Hemish Radadiya(20103100810012)

Kanaiyalal Bhadani(201703100810058)

Milan Vaghani(2017031008100063)

ABSTRACT

Heart disease is one of the most significant causes of mortality in today's world. Heart disease proves to be the leading cause of death for both men and women. This affects the human life very badly. The diagnosis of heart disease in most cases depends on a complex combination and huge volume of clinical and pathological data. Machine learning has been shown to be effective assisting in making decisions and predictions from the large quantity of data produced by the health care industry. In this paper, various traditional machine learning algorithms that aims in improving the accuracy of heart disease prediction has been applied. In heart diseases, accurate diagnosis is primary. But, the traditional approaches are inadequate for accurate prediction and diagnosis. In order to apply deep learning technique very large datasets are required which are not available in medical and clinical research. To address this issue, surrogate data is generated from Cleveland dataset. The generated synthetic dataset is utilized with traditional machine learning algorithms as well as with deep learning model. The predicted results show that there is an improvement in classification accuracy. The generated synthetic dataset plays a vital role to improve the classification prediction particularly when dealing with sensitive data.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1 Introduction

1.1 Background

Heart is an important organ of the human body. It pumps blood to every part of our anatomy. If it fails to function correctly, then the brain and various other organs will stop working, and within few minutes, the person will die.Heart diseases have emerged as one of the most prominent cause of death all around the world. According to World Health Organisation, heart related diseases are responsible for the taking 17.7 million lives every year, 31% of all global deaths. Medical organisations, all around the world, collect data on various health related issues. These data can be exploited using various machine learning techniques to gain useful insights. But the data collected is very massive and, many a times, this data can be very noisy.Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart related diseases accurately

1.2 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 heartdisease in human. Early detection of cardiac diseases can decrease the mortality rate and overallcomplications. However, it is not possible to monitor patients every day in all cases accurately andconsultation 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 analyze the data for hidden patterns. The hidden patterns can beused for health diagnosis in medicinal data.

.

1.3 Motivation

The main motivation for us to go for this project was the slow and inefficient traditional manual heart disease prediction system. This made us think why not make it automated fast and much efficient. Also,disease prediction techniques are in use by department like helthcare where they use patient's data to prediction of occurance of heart disease.

1.4 Objective

The main objective of developing this project are:

1. To develop machine learning model to predict future possibility of heart disease byimplementing Logistic Regression.

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

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

1.5 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, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions

1.6 Applications

- Quick medication available
- User friendly
- Doctors can dicison more accurate

Chapter 2 System Planning

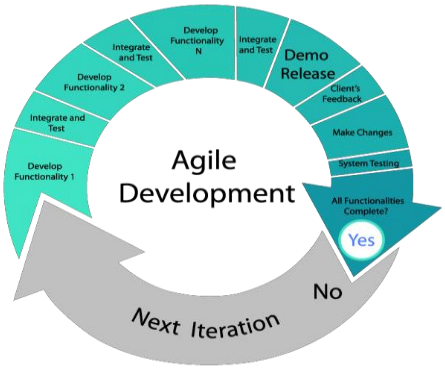
2.1 Project Development Approach

Each project needs to be developed with software model which makes the project with high quality, reliable and cost-effective so for our project we have selected Iterative model.

- The advantage of this model is that there is a working model of the system at a very early stage of development, which makes it easier to find functional or design flaws. Finding issues at an early stage of development enables to take corrective measures in a limited budget and whenever any problem occurs we can resume work at that point no need to backtrack the whole process.
- Advantages of the iterative model are some working functionality can be developed quickly and early in the life cycle and we need not go from the first if any problem occurs we can continue from that point only.so, this model is reliable too.

Agile model:

Agile SDLC model is a combination of iterative and incremental process models with a focus on process adaptability and customer satisfaction by rapid delivery of working software product. Agile Methods break the product into small incremental builds. These builds are provided in iterations. Each iteration typically lasts from about one to three weeks. Iteration involves cross-functional teams working simultaneously on various areas like planning, requirements analysis, design, coding, unit testing, and acceptance testing. At the end of the iteration, a working product is displayed to the customer and important stakeholders. Here is a graphical illustration of the Agile Model:



2.2 System Modules

2.2.1 Gathering Patient's data

- Admin will gather user's data for building a model.

2.2.2 Training patient's data

- Model will be trained on Jupyter or Colab Notebook for process a UCI dataset.
- There is an algorithm for trained or build a classifier.
- The weight generated by the algorithm is predict the accuracy of heart disease.

2.2.3 Analysis heart disease

- The system will generate the accuracy of heart disease using a different machine learning algorithem.

2.3 Functional Requirements

- In this system have two functional requirements are there admin and user.

ID	**Title & Description**
FR 1	*Title: Admin* *Desc: The admin has to login using his username and password. After admin login they can a upload a dataset, they can train the model using machine learning approach.*
FR 2	*Title: User* *Desc: User has to login using user id and password, Then user has to input patient data and then based on training model user input data will check and it will give the output.*

2.4 Non Functional Requirements

Portability: Discuss the hardware/software portability according to system it supports.

Security: One or more requirements about protection of your system and its data. Do not discuss solutions (e.g. passwords) in a requirements document.

Performance: The framework will be utilized by numerous representatives all the while. It should allow accessibility to each and every piece of its customers.

Reliability: Details about error rate and accuracy of your system.

Reusability: discuss about how your system will react with the new extension.

2.5 Hardware and Software Requirements

Hardware Requirement

- *Processor: Any processor (Ex. Intel).*
- *RAM: 1024MB.*
- *Space on disk: minimum 100mb.*
- *Device: Any device that can access the internet.*

Software Requirement

- *Operating system: Any OS with clients to access the internet.*
- *Network: WI-FI internet or cellular network.*
- *Pycharm: Crate a python file.*
- *Google Chrome: Medium to find reference to do system testing, display.*

2.6 Timeline Chart

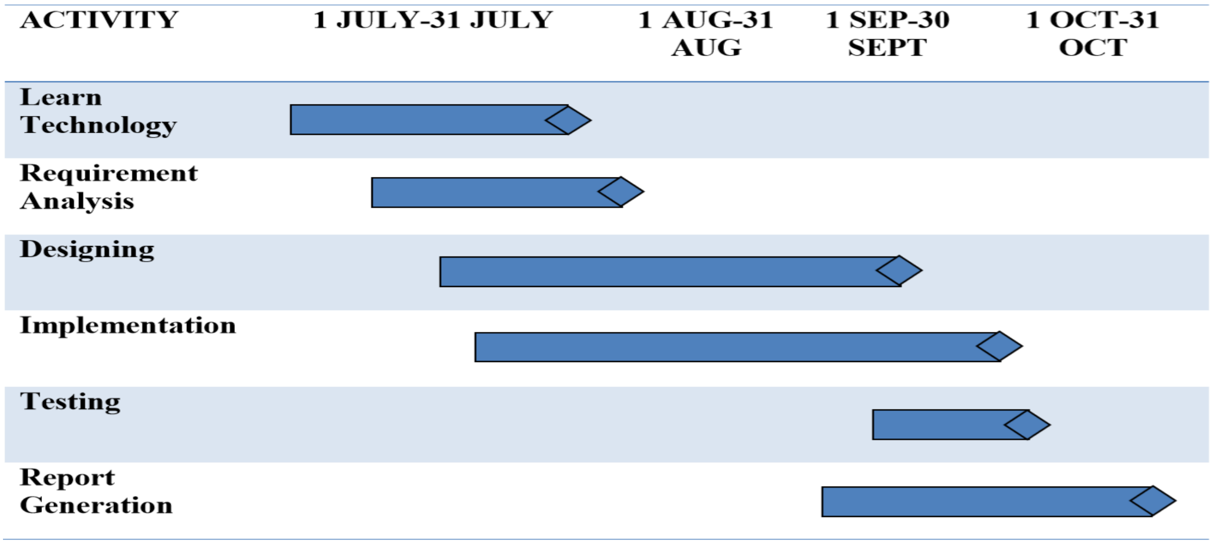
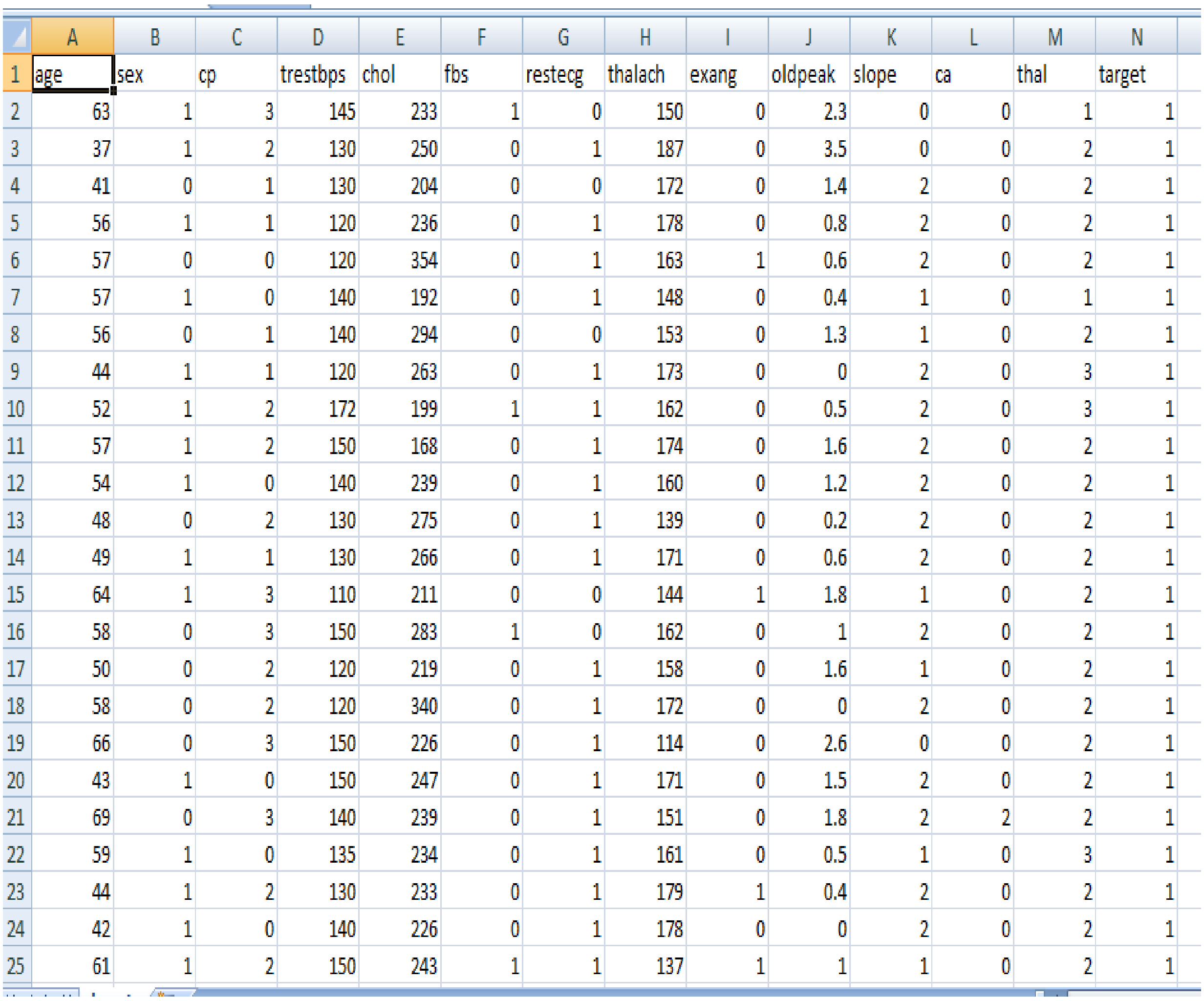


Figure 2.6: Timeline Chart

Chapter 3 System Design

3.1 Database Schema



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

Figure 3.1 Dataset Diagram

3.2 Use Case Diagram

A use case diagram at its simplest is a representation of a user's interaction with the system that shows the relationship between the user and the different use cases in which the user is involved.

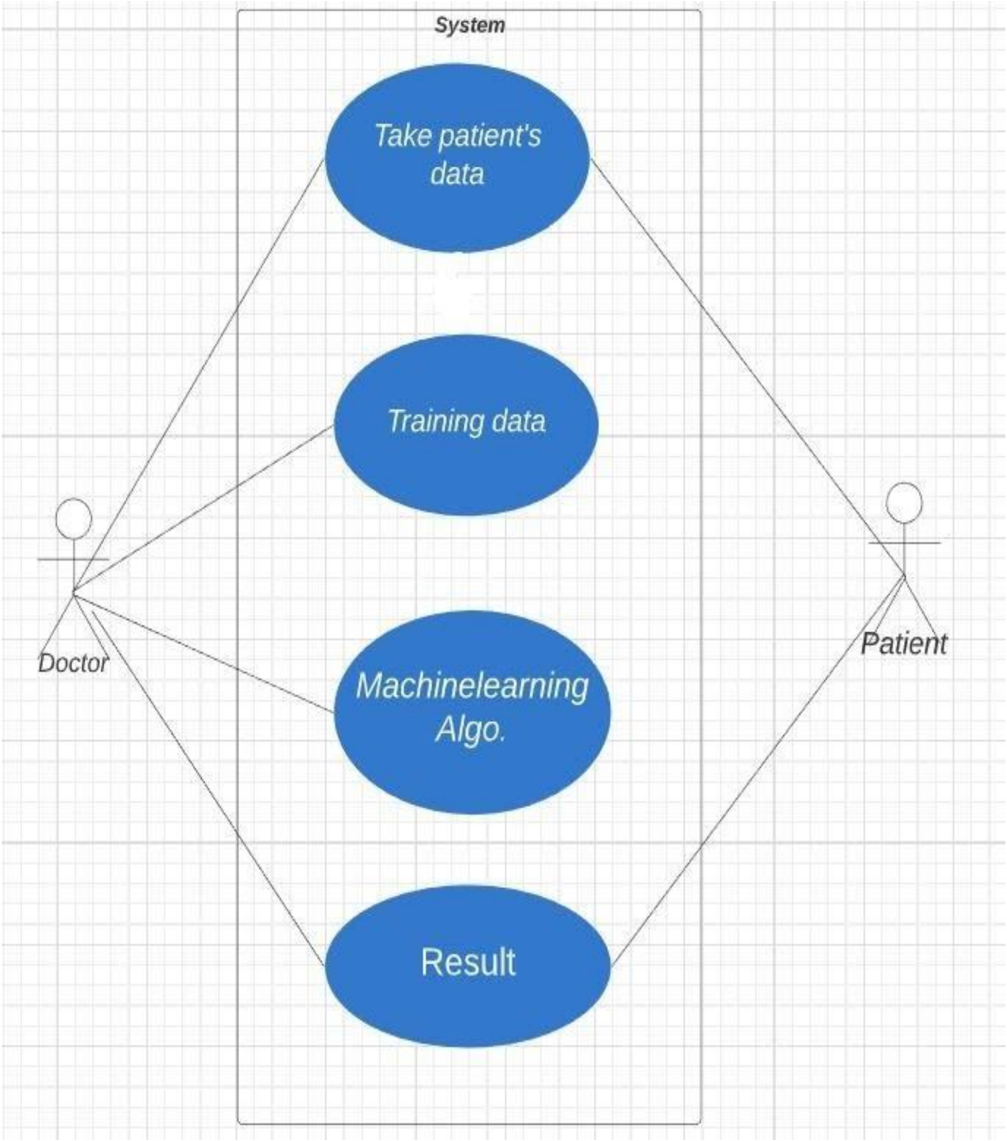


Figure 3.2 Usecase Diagram

3.3 Sequence Diagram

A sequence diagram shows object interactions arranged in time sequence. It depicts the objects and classes involved in the scenario and the sequence of messages exchanged between the objects needed to carry out the functionality of the scenario.

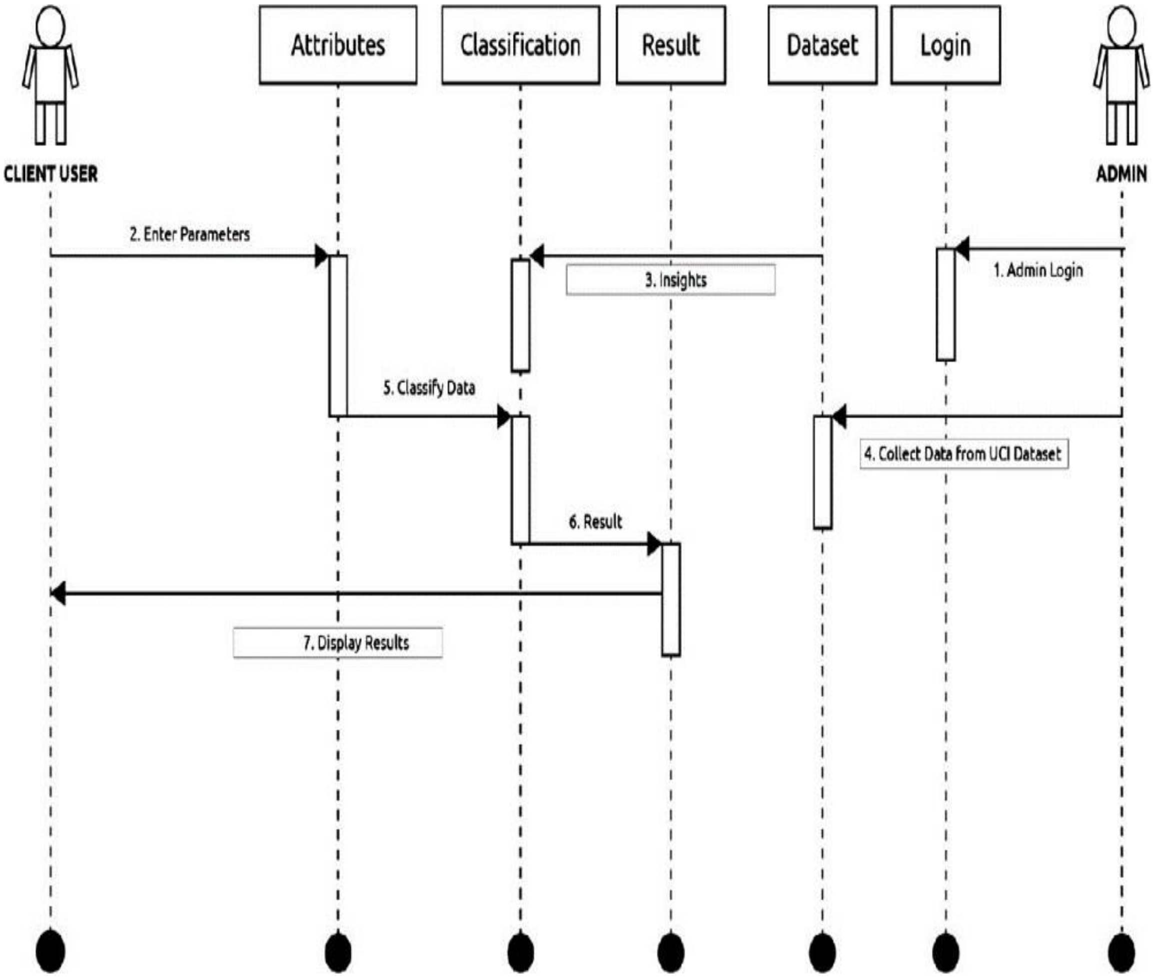


Figure 3.3: Sequence Diagram

3.4 Activity Diagram

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency.

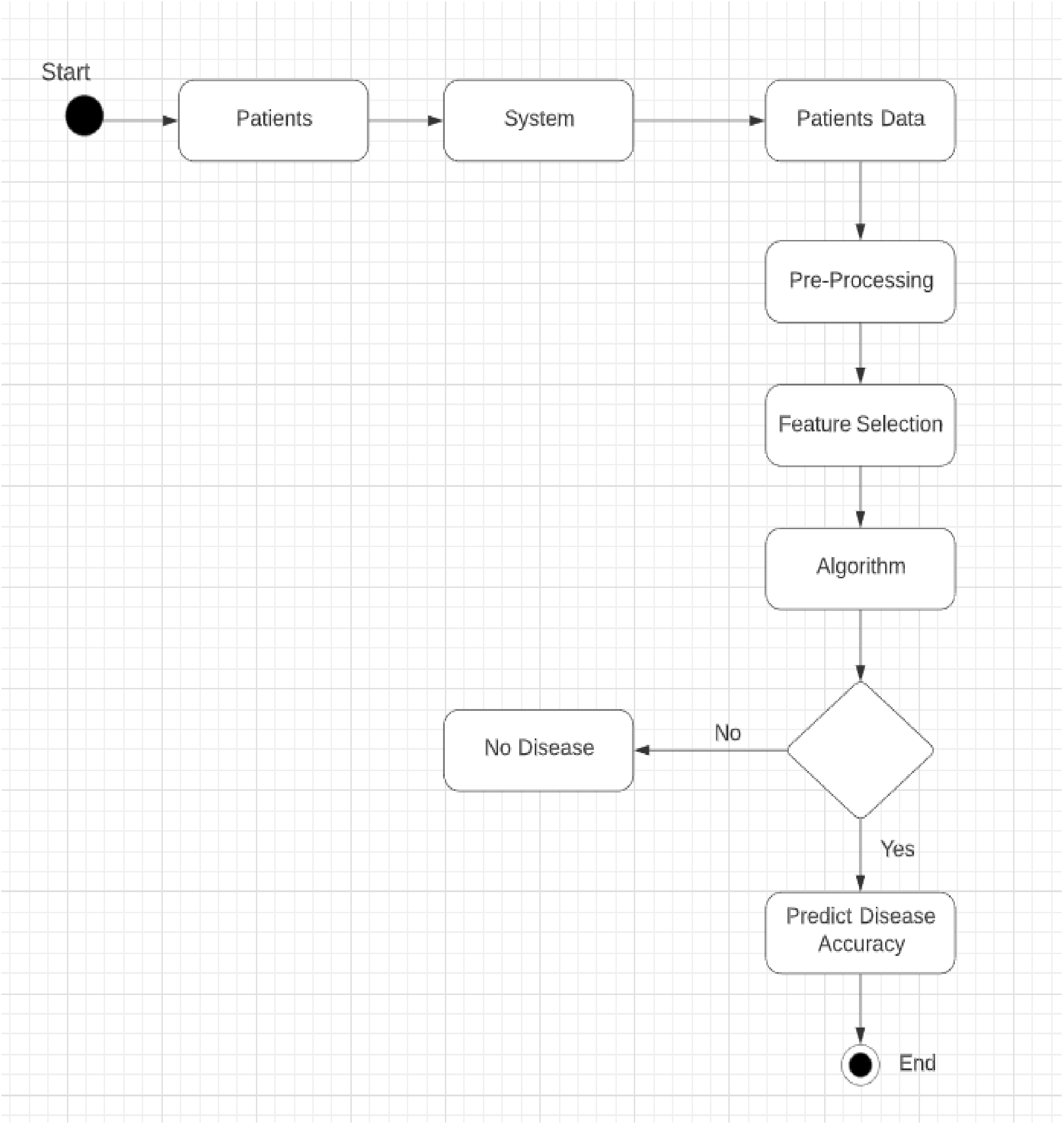


Figure 3.4: Activity Diagram

3.5 Data Flow Diagram

A Data Flow Diagram(DFD) is a graphical representation of the "flow" of data through an information system, modelling its process aspects. A DFD is often used as a preliminary step to create an overview of the system without going into great detail, which can later be elaborated.

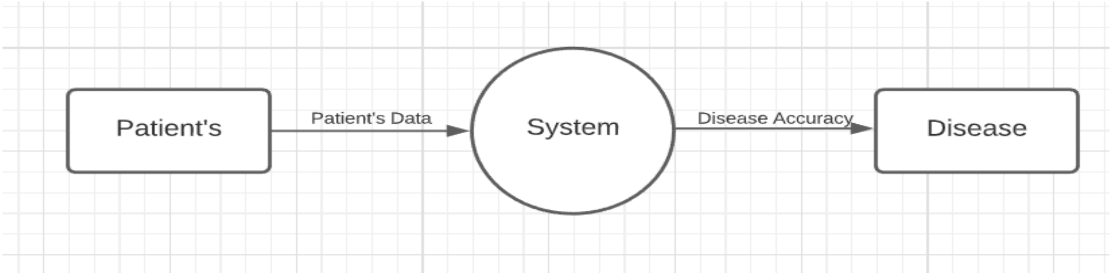


Figure 3.5.1 Level-0 DFD Diagram

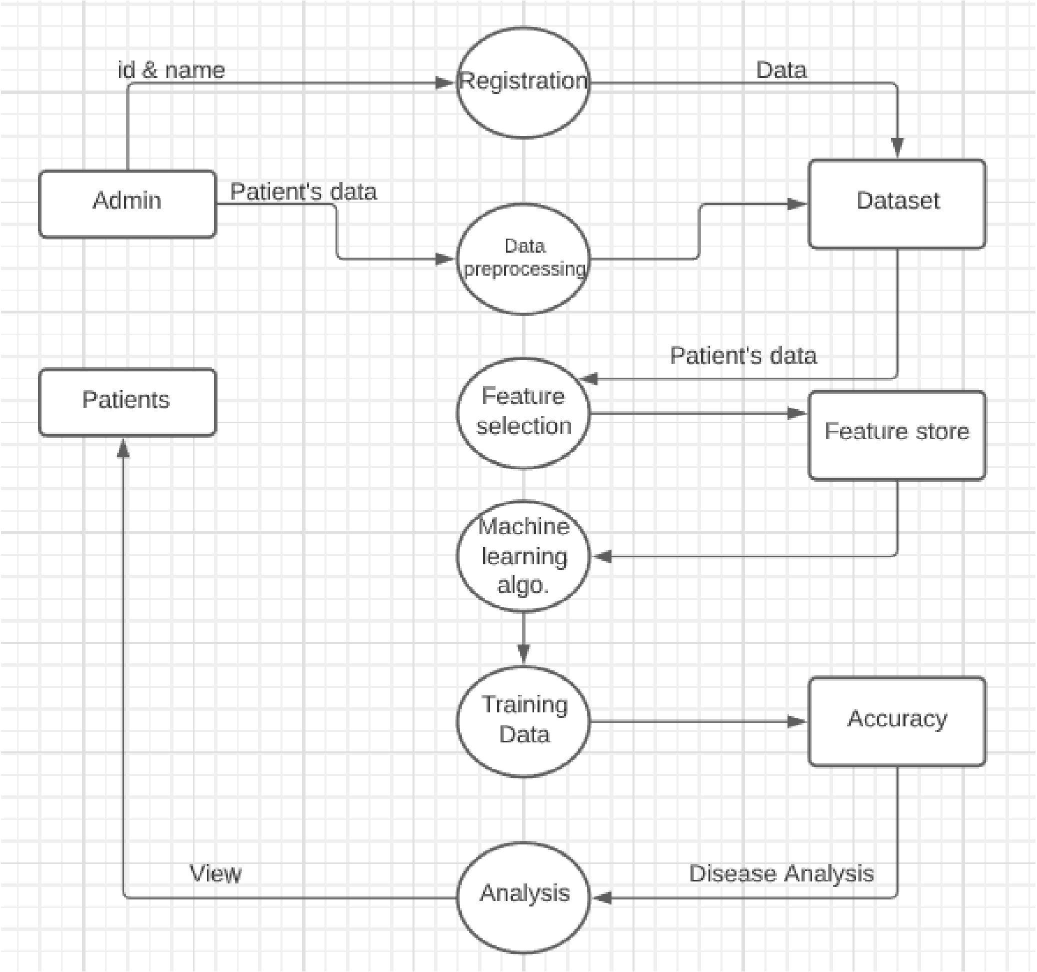
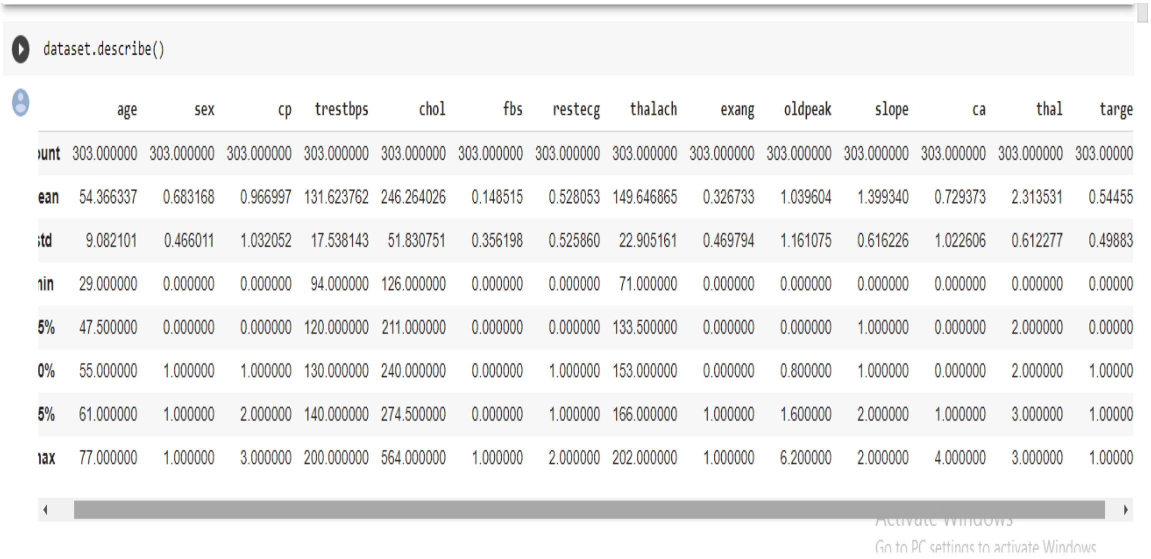


Figure 3.5.2 : Level-1 DFD

Chapter 4 Implementation and Testing

4.1 Snapshots

- **Dataset Describe**



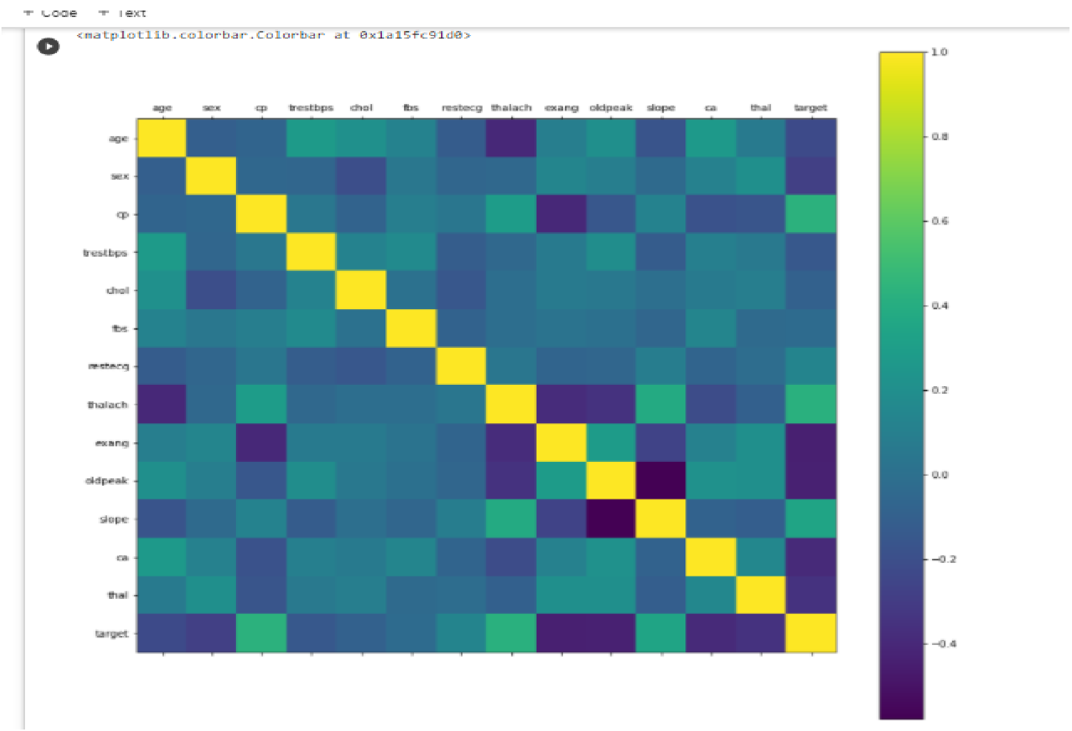
```
dataset.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	targe
unt	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	303.00000
ean	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	0.54455
td	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	0.49883
in	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	0.00000
5%	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	0.00000
0%	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	1.00000
5%	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	1.00000
ax	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	1.00000

Figure 4.1: Dataset Describe

The scale of each feature column is different and quite varied as well. While the maximum for age reaches 77, the maximum of chol (serum cholestoral) is 564.

- ➢ **Corelation matrix for all attribute**

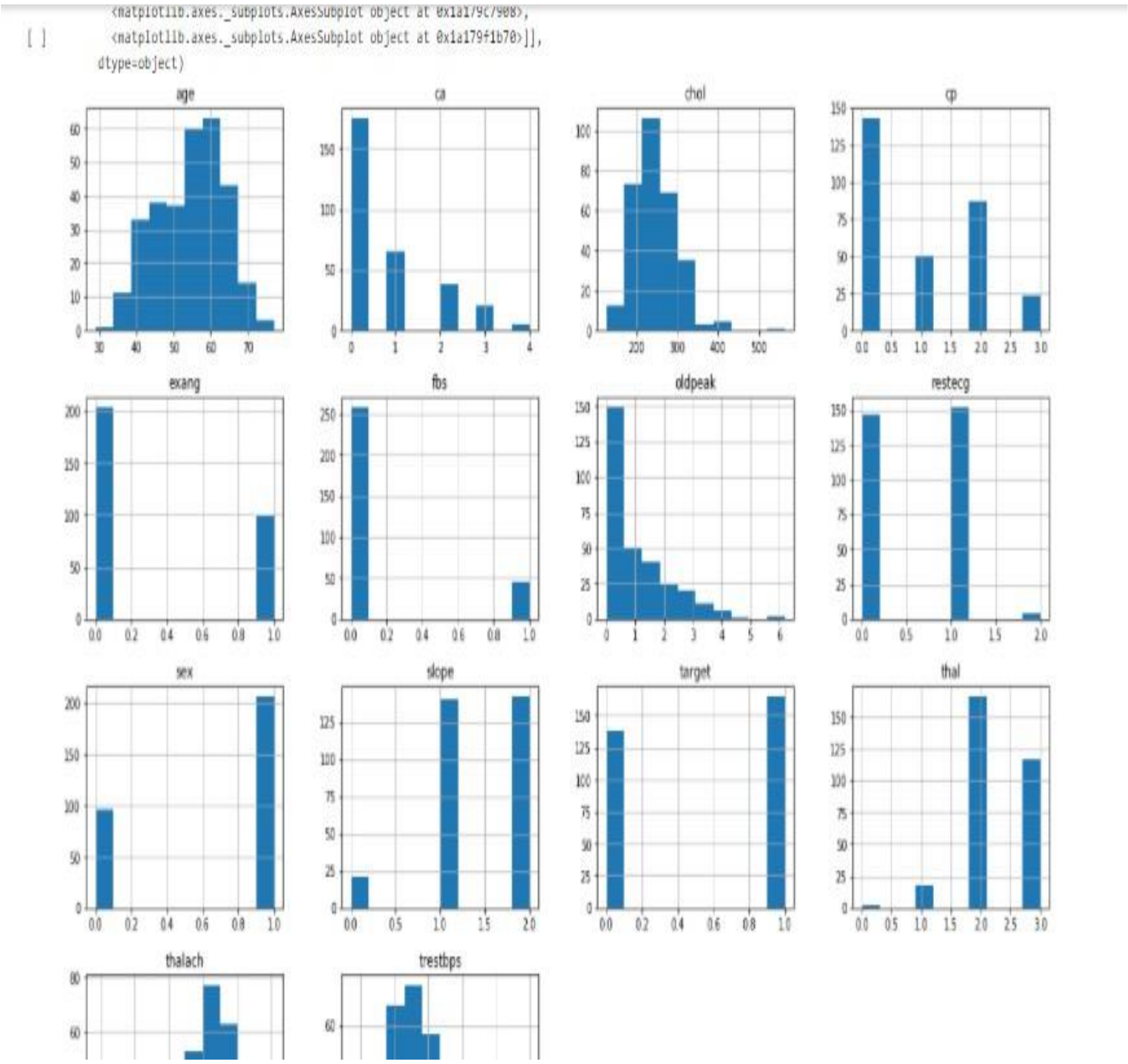


```
<matplotlib.colorbar.Colorbar at 0x1a15fc91d0>
```

Figure 4.2: Corelation matrix between attribute

Taking a look at the correlation matrix above, it's easy to see that a few features have negative correlation with the target value while some have positive.

- **Histogram for all attribute**



```
        <matplotlib.axes._subplots.AxesSubplot object at 0x1a179c7908>,
[ ]     <matplotlib.axes._subplots.AxesSubplot object at 0x1a179f1b70>]],
      dtype=object)
```

Figure 4.3: Histogram for each attribute has different range of distribution

Taking a look at the histograms above, we can see that each feature has a different range of distribution. Thus, using scaling before our predictions should be of great use. Also, the categorical features do stand out.

- **Target of class**

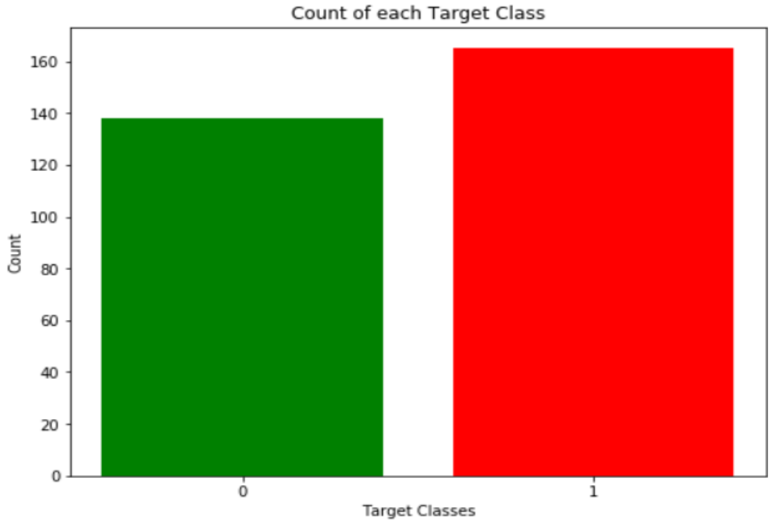


Figure 4.4: Count of each Target Class

The two classes are not exactly 50% each but the ratio is good enough to continue without dropping/increasing our data.

- **K-Nearest Neighbors algo.**

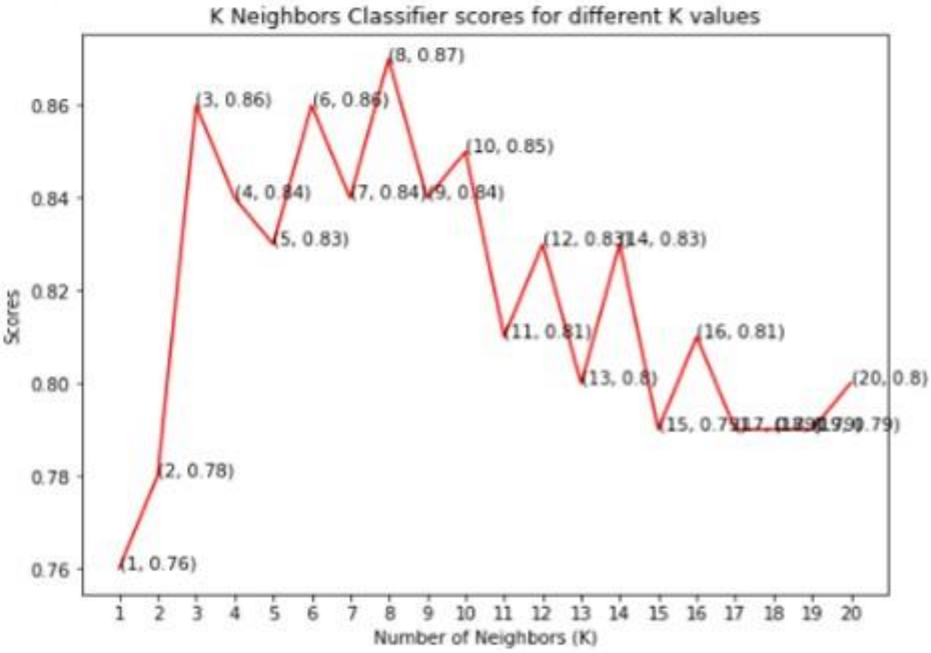


Figure 4.5: Accuracy using KNN algo.

From the plot above, it is clear that the maximum score achieved was 0.87 for the 8 neighbors.

- **Support Vector Clasifier**

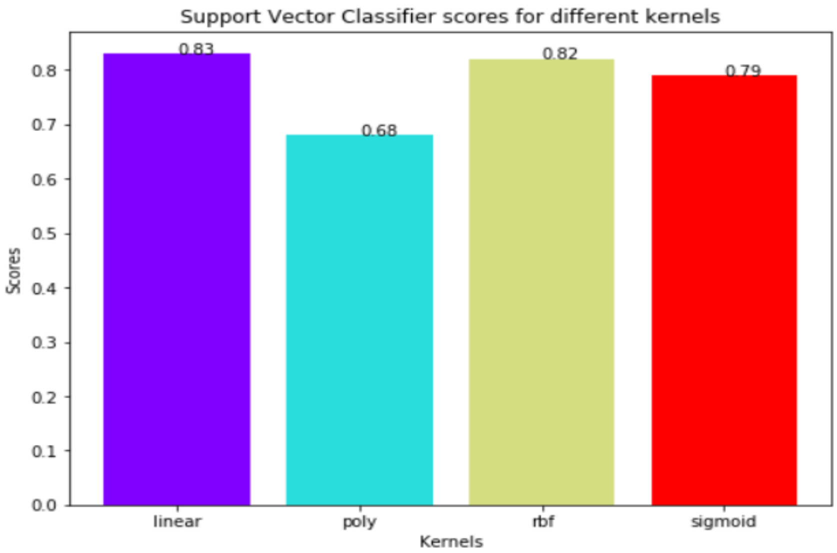


Figure 4.6: Accuracy using SVC algo.

The linear kernel performed the best, being slightly better than rbf kernel, The score for Support Vector Classifier is 83.0% with linear kernel.

- **Decision Tree Classifer**

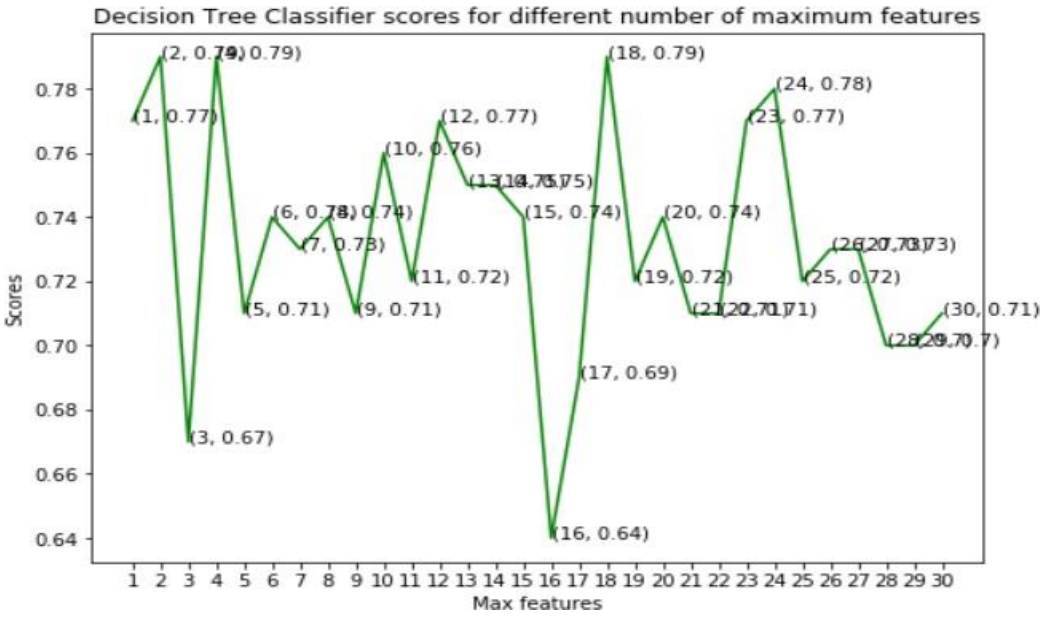


Figure 4.7: Accuracy using Decision Tree Classifer algo.

The model achieved the best accuracy at three values of maximum features, 2, 4 and 18. The score for Decision Tree Classifier is 79.0% with [2, 4, 18] maximum features.

- **Random Forest Classifer**

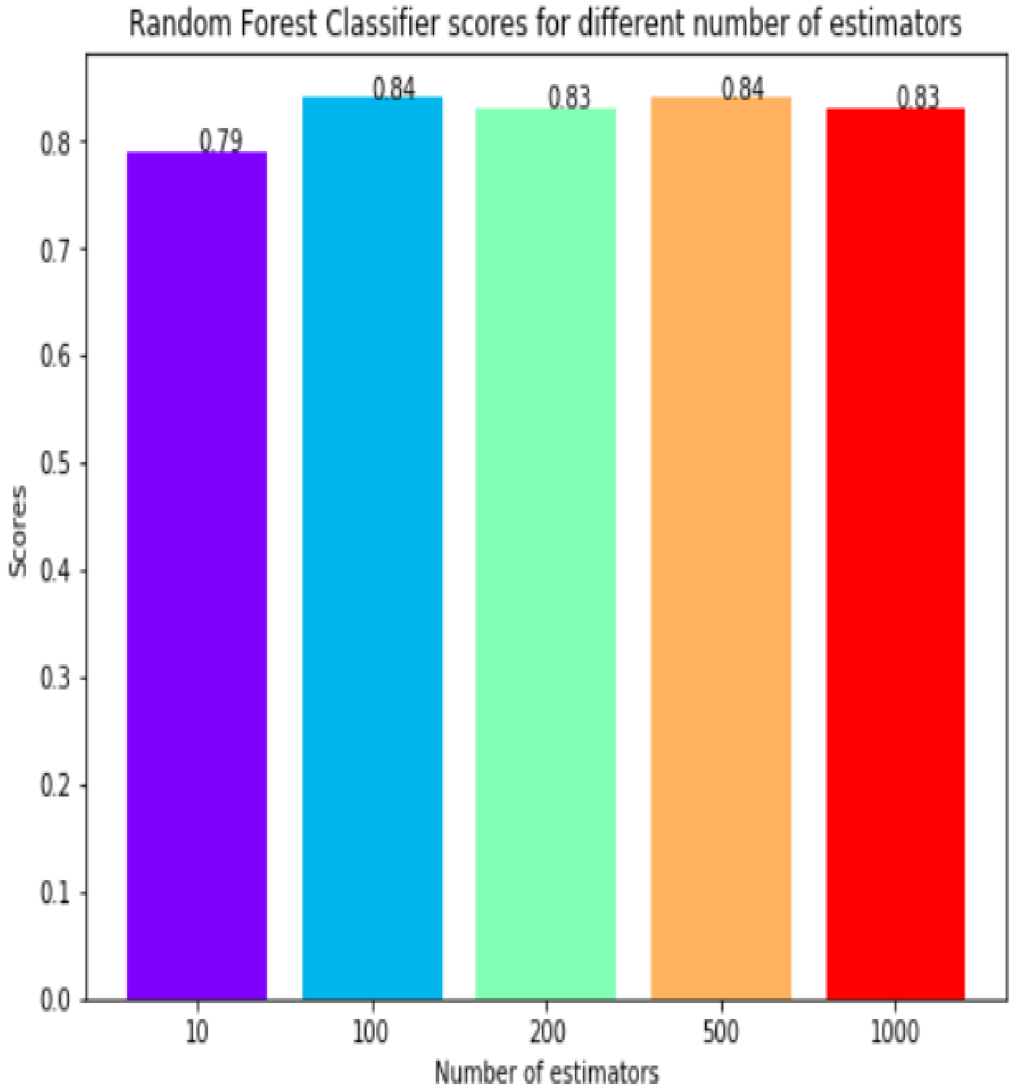


Figure 4.8: Accuracy using Random Forest Classifer algo.

The maximum score is achieved when the total estimators are 100 or 500. The score for Random Forest Classifier is 84.0% with [100, 500] estimators.

➢ Comparing Different ML Algo.

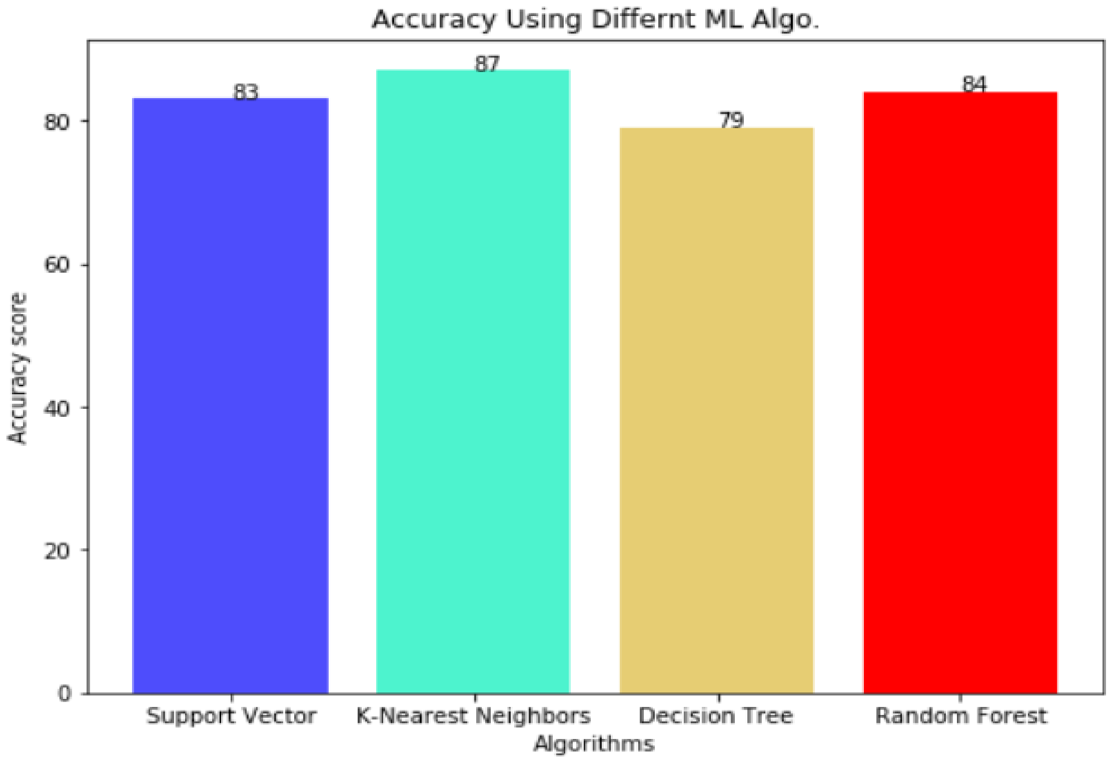


Figure 4.9: Accuracy using comparing different algo.

In this project, We used Machine Learning to predict whether a person is suffering from a heart disease. After importing the data, We analysed it using plots. Then, We did generated dummy variables for categorical features and scaled other features. We then applied four Machine Learning algorithms, K Neighbors Classifier, Support Vector Classifier, Decision Tree Classifier and Random Forest Classifier. We varied parameters across each model to improve their scores. In the end, K Neighbors Classifier achieved the highest score of 87% with 8 nearest neighbors.

4.2 Test Cases

Sample test cases are given as below:

Test ID	Case	Test Data	Expected Result	Actual Result	Pass/Fail
1	Gathering data	- Patient’s data	The admin will enter the patient’s record then it will store in dataset folder.	The patient’s record store to a particular dataset.	Pass
2	Processing data	- Patient’s data	The model should process data and clear unnecessary patient’s data to dataset.	Model processing patient’s data using proper algorithm.	Pass
3	Training data	- Patient’s data	The model should train the patient’s data.	Model trains patient’s data using proper ML algorithm and generating a accuracy of disease.	Pass
4	Analysis of data	-	The model should generate the accuracy and analysis of disease.	Model should generate the analysis of disease.	Pass

Conclusion and Future Scope

It can be concluded from that a reliable, secure, fast and an efficient system has been developed replacing a manual and unreliable system. This system can be implemented for better results regarding the heart disease prediction. The advantage of this model is that provide digital solution for heart disease prediction systerm. The limitation of project is to require an physical equipment and processing power for better output.This model will be useful in identifying the possible patient who may suffer form heart disease.

Future work could also include adding several well-structured heart disease prediction system for all people. When a patient is predicted as positive for heart disease ,then the medical data for the patient can be closely analysed by the doctors.

References

Paper references

1. J. S. Rumsfeld, K. E. Joynt, and T. M. Maddox, “Big data analytics to improve cardiovascular care: promise and challenges”, Nature Reviews Cardiology, Vol.13, No.6, pp.350, 2016,Date accessed:[03/10/2020].

2. M. Lichman, “UCI Machine Learning Repositry”, [Online]. https://archive.ics.uci.edu/, 2013,Date accessed:[11/09/2020].

Web references

1. https://ieeexplore.ieee.org/document/8474922,Date accessed:[19/08/2020].

2. UCI, Heart Disease Data Set.[Online]. Available (Accessed on Aug. 22 2020): https://www.kaggle.com/ronitf/heart-disease-uci.