KNOWLEDGE & TECHNOLOGY

Bangladesh Army University of Engineering & Technology

Department of Computer Science and Engineering

A project on

Heart Disease Prediction using Machine Learning with Python

Submitted by

Sakib Haque Zisan
ID: 18204008
Muzakkir Hossain Minhaz
ID: 18204010
Md. Faisal Khan
ID: 18204034

Supervised by

Subrata Kumer Paul
Lecturer, Department of CSE

Rakhi Rani Paul
Lecturer, Department of CSE

BAUET

Department of Computer Science and Engineering

Bangladesh Army University of Engineering & Technology

October 2022

Bangladesh Army University of Engineering & Technology

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the project entitled "Heart Disease Prediction using Machine Learning with Python" by "Sakib Haque Zisan", ID: 18204008, "Muzakkir Hossain Minhaz", ID: 18204010, "Md. Faisal Khan", ID: 18204034, has been accepted as satisfactory in partial fulfillment of the requirement for the degree of Bachelor of Science in Computer Science and Engineering on October 2022.

Signature of Supervisor

(Subrata Kumer Paul)
(Lecturer)
Department of CSE, BAUET

Signature of Supervisor

(Rakhi Rani Paul)
(Lecturer)
Department of CSE, BAUET

KNOWLEDGE & TECHNOLOGY

Bangladesh Army University of Engineering & Technology

Department of Computer Science and Engineering

DECLARATION

I hereby declare that our project entitled "Heart Diseases Prediction using Machine Learning with Python" is the result of our work. We also ensure that it is not previously submitted or published elsewhere for awarding any degree or diploma.

The work has been accepted for a Bachelor of Science in Computer Science and Engineering at Bangladesh Army University of Engineering & Technology (BAUET).

Authors

Sakib Haque Zisan (18204008)

Muzakkir Hossain Minhaz (18204010)

Md. Faisal Khan (18204034)

ACKNOWLEDGMENT

We are highly indebted to our course teachers for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

Sakib Haque Zisan
ID: 18204008

Muzakkir Hossain Minhaz
ID: 18204010

Md. Faisal Khan
ID: 18204034

ABSTRACT

Day by day the cases of heart diseases are increasing at a rapid rate and it's very important and concerning to predict any such diseases beforehand. This diagnosis is a difficult task i.e. it should be performed precisely and efficiently. The research paper mainly focuses on which patient is more likely to have a heart disease based on various medical attributes. We prepared a heart disease prediction system to predict whether the patient is likely to be diagnosed with heart disease or not using the medical history of the patient. We used different algorithms of machine learning such as Logistic Regression and KNN to predict and classify the patient with heart disease. A quite Helpful approach was used to regulate how the model can be used to improve the accuracy of prediction of Heart Attacks in any individual. The strength of the proposed model was quite satisfying and was able to predict evidence of having heart disease in a particular individual by using KNN and Logistic Regression which showed good accuracy in comparison to the previously used classifier such as naive Bayes etc. So a quite significant amount of pressure has been lifted off by using the given model in finding the probability of the classifier to correctly and accurately identify the heart disease. The Given heart disease prediction system enhances medical care and reduces the cost. This project gives us significant knowledge that can help us predict patients with heart disease It is implemented in the .ipynb format.

List of Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

1.1 Introduction

Nowadays, heart disease prediction has been a major concept in the recent world that is impacting society towards health. The main concept is to identify the age group and heart rate using the Random forest algorithm. Our project tells how the heart rate and condition are estimated based on the inputs such as blood pressure and many more being provided by the user to a system. This is being much better way when it comes to others algorithms the implementation of RFA gives a better experience and provides an accurate result. This helps in the early prediction of the disease and is used in many ways, whereas it is being provided with the input, in order to find the heart rate based on the health condition.

1.2 Background

Millions of people are getting some sort of heart disease every year and heart disease is the biggest killer of both men and women in the United States and around the world. The World Health Organization (WHO) analyzed that twelve million deaths occur worldwide due to Heart diseases. In almost every 34 seconds heart disease kills one person in the world.

Medical diagnosis plays a vital role and is yet a complicated task that needs to be executed efficiently and accurately. To reduce the cost of achieving clinical tests appropriate computer-based information and decision support should be aided. Data mining is the use of software techniques for finding patterns and consistency in sets of data. Also, with the advent of data mining in the last two decades, there is a big opportunity to allow computers to directly construct and classify the different attributes or classes.

Learning the risk components connected with heart disease helps medicinal services experts to recognize patients at high risk of having Heart disease. Statistical analysis has identified risk factors associated with heart disease to be age, blood pressure, total cholesterol, diabetes, hypertension, family history of heart disease, obesity and lack of physical exercise, fasting blood sugar, etc [1].

Researchers have been applying different data mining Techniques to help medicinal services experts with progressed exactness in the judgment of heart disease. Neural networks, Naive Bayes, Decision Trees, etc. are some techniques used in the diagnosis of heart disease.

Applying Decision Tree techniques has shown useful accuracy in the diagnosis of heart disease. But assisting healthcare professionals in the diagnosis of the world's biggest killer demands higher accuracy. Our research seeks to improve diagnosis accuracy to improve health outcomes.

A Decision Tree is one of the data mining techniques that cannot handle continuous variables directly so the continuous attributes must be converted to discrete attributes. A couple of Decision Trees use binary discretization for continuous-valued features. Another important accuracy improvement is applying reduced error pruning to the Decision Tree in the diagnosis of heart disease patients. Intuitively, more complex models might be expected to produce more accurate results, but which technique is best? Seeking to thoroughly investigate options for accuracy improvements in heart disease diagnosis this paper systematically investigates comparing multiple classifier decision tree techniques.

This research uses Waikato Environment for Knowledge Analysis (WEKA).The information of the UCI repository is regularly introduced in a database or spreadsheet. In order to use this data for the WEKA tool, the data sets need to be in the ARFF format (attribute-relation file format). WEKA tool is used to pre-process the dataset. After reviewing all these 76 different attributes, the unimportant attributes are dropped and only the important attributes (i.e. 14 attributes in this case) are considered for analysis to yield more accurate and better results. The 14th one is basically a predicted attribute, which is referred to as Class. A thorough comparison between different decision tree algorithms within the WEKA tool and deriving the decisions out of it would help the system to predict the likely presence of heart disease in the patient and will definitely help to diagnose heart disease well in advance and able to cure it in right time

1.3 Objectives

The main objective of our project is to predict heart disease (healthy or affected) by taking user input. Some others objectives of our project is below,

- To choose the best possible algorithm among many prediction algorithms.
- To build an efficient model to predict heart disease with help of a machine learning algorithm.
- Compare the accuracy and precision of this algorithm with various other algorithms.
- Try to improve the model using various techniques.
- Convert the model into a user-friendly website or android app.

1.4 Conclusion

In this project, a machine learning technique is used to predict the heart disease of a person. Machine learning is where computers can learn to do something without the need to explicitly program them for the task. It uses data and produces a program to perform a task such as classification. Compared to knowledge engineering, machine learning techniques require messages that have been successfully pre-classified. The pre-classified messages make the training dataset which will be used to fit the learning algorithm to the model in the machine learning studio.

Chapter 2

PLANNING

2.1 Introduction

Planning is the process of thinking regarding the activities required to achieve the desired goal. Planning is based on foresight, the fundamental capacity for mental time travel. The evolution of forethought, the capacity to think ahead, is considered to have been a prime mover in human evolution. Planning is a fundamental property of intelligent behavior. It involves the use of logic and imagination to visualize not only the desired end result but the steps necessary to achieve that result. An important aspect of planning is its relationship to forecasting. Forecasting aims to predict what the future will look like, while planning imagines what the future could look like. Planning according to established principles is a core part of many professional occupations, particularly in fields such as management and business. Once a plan has been developed it is possible to measure and assess progress, efficiency, and effectiveness. As circumstances change, plans may need to be modified or even abandoned. [2]

Some steps that we have planned to do before building our project:

i. Import the required libraries
ii. Load the dataset
iii. Preprocess the data
iv. Splitting the features and target
v. Splitting the data into training and testing
vi. Build the model
vii. Compare the model with other Machine Learning algorithms
viii. Improve the model

Data preprocessing in Machine Learning refers to the technique of preparing (cleaning and organizing) the raw data to make it suitable for building and training Machine Learning models.

The model building process involves setting up ways of collecting data, understanding and paying attention to what is important in the data to answer the questions you are asking, finding a statistical, mathematical, or simulation model to gain understanding and make predictions.

Accuracy is vital to any machine learning model and is the most often talked about. Without accurate predictions, there is no purpose for deploying the algorithm – so strive for the best accuracy you can within reasonable limitations. But in our case we will focus on precision more.

2.2 Conclusion:

We plan to build this system as much as helpful to all aspects in all the available field it could be used. Our policy and the output of the service will reflect the fact of usefulness in the world of technology. Thus we built a heart disease prediction system using the best possible algorithm. Our further plan is to develop this system such as we can use it all over the web and android os.

Chapter 3

ANALYSIS

3.1 Introduction

In Machine Learning, Data Analysis is the process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information by informing conclusions and supporting decision-making.

An Organized Dataset of individuals had been selected Keeping in mind their history of heart problems and in accordance with other medical conditions [3]. Heart disease is the diverse conditions by which the heart is affected. According to World Health Organization (WHO), the greatest number of deaths in middle-aged people are due to Cardiovascular diseases. We take a data source that is comprised of the medical history of 304 different patients of different age groups. This dataset gives us the much-needed information i.e. the medical attributes such as age, resting blood pressure, fasting sugar level, etc. of the patient that helps us in detecting whether the patient is diagnosed with any heart disease or not. This dataset contains 13 medical attributes of 304 patients that help us detect if the patient is at risk of getting a heart disease or not and it helps us classify patients that are at risk of having heart disease and that who are not at risk. This Heart Disease dataset is taken from the UCI repository. According to this dataset, the pattern which leads to the detection of patients prone to getting heart disease is extracted. These records are split into two parts: Training and Testing. This dataset contains 303 rows and 14 columns, where each row corresponds to a single record. All attributes are listed in *'Table 1'*.

Table 1: Various Attributes used are dataset

S. No	Observation	Description	Values
1.	Age	Age in Years	Continuous
2.	Sex	Sex of Subject	Male/Female
3.	CP	Chest Pain	Four Types
4.	Trestbps	Resting Blood Pressure	Continuous
5.	Chol	Serum Cholesterol	Continuous
6.	FBS	Fasting Blood Sugar	<, or > 120 mg/dl
7.	Restecg	Resting Electrocardiograph	Five Values
8.	Thalach	Maximum Heart Rate Achieved	Continuous
9.	Exang	Exercise Induced Angina	Yes/No

10.	Oldpeak	ST Depression when Workout compared to the Amount of Rest Taken	Continuous
11.	Slope	Slope of Peak Exercise ST segments	Up/Flat/Down
12.	Ca	Gives the number of Major Vessels Coloured by Fluoroscopy	0-3
13.	Thal	Defect Type	Reversible/Fixed/ Normal
14.	Num (Disorder)	Heart Disease	Not Present/Present

3.2 Conclusion

In this chapter we have discussed the dataset we used. We train our model based on some conditions of a heart. Mainly we used 1-13 attributes to train our system and check the validation by using the 14 attributes from the dataset. The above table (Table 1) describes our dataset attributes. here we see the description and values of the corresponding attributes.

Chapter 4

DESIGN

4.1 Introduction

The project design includes everything from who is responsible for completing the project to a description of the project, its goals, outcomes, and objectives. It describes when these goals, outcomes, and objectives will be reached, and the major deliverables, products, or features that will be completed.

4.2 Flowchart

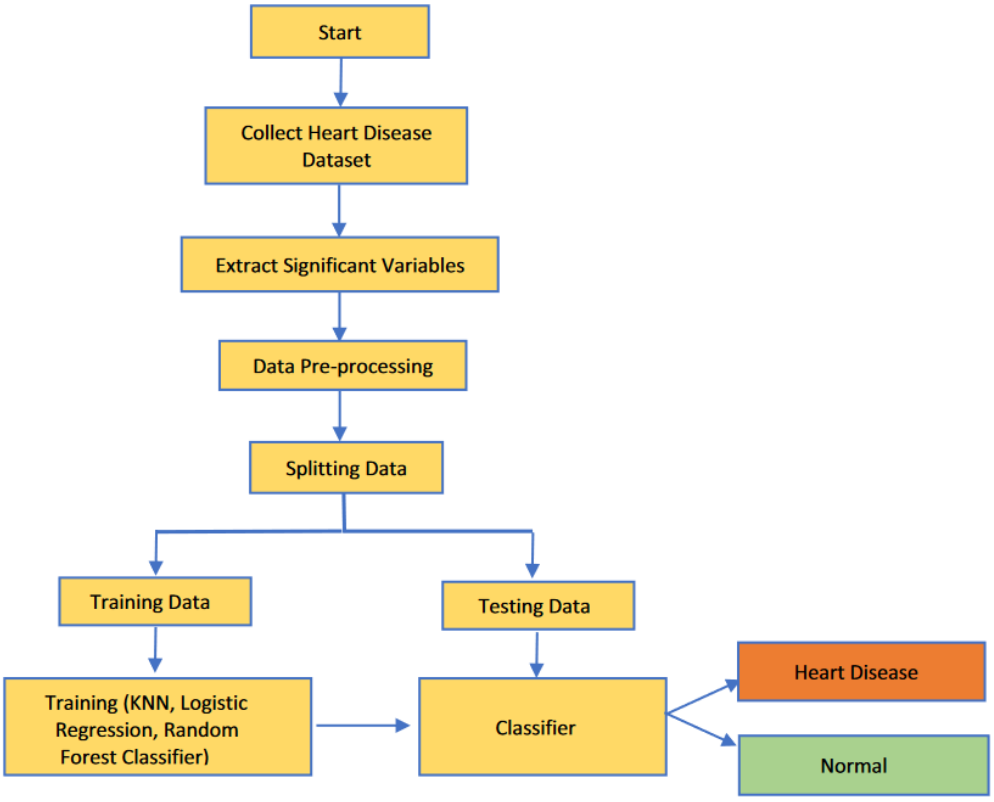

Fig 4.2.1: Flowchart of Heart Disease Prediction system using Machine Learning algorithm.

This is the flowchart of our project. At first, we have to select the dataset that we want to use in our model. After that, we have to extract significant variables from the dataset and pre-process the dataset. After that we split the data into two parts, these are training data and testing data. From the training data, we train by using the Logistic Regression algorithm, then classify the data and test data used to test the classification. Finally, we will give input in the trained model and will see the output as Heart Disease (affected) or Normal (healthy).

4.3 Use-case Diagram:

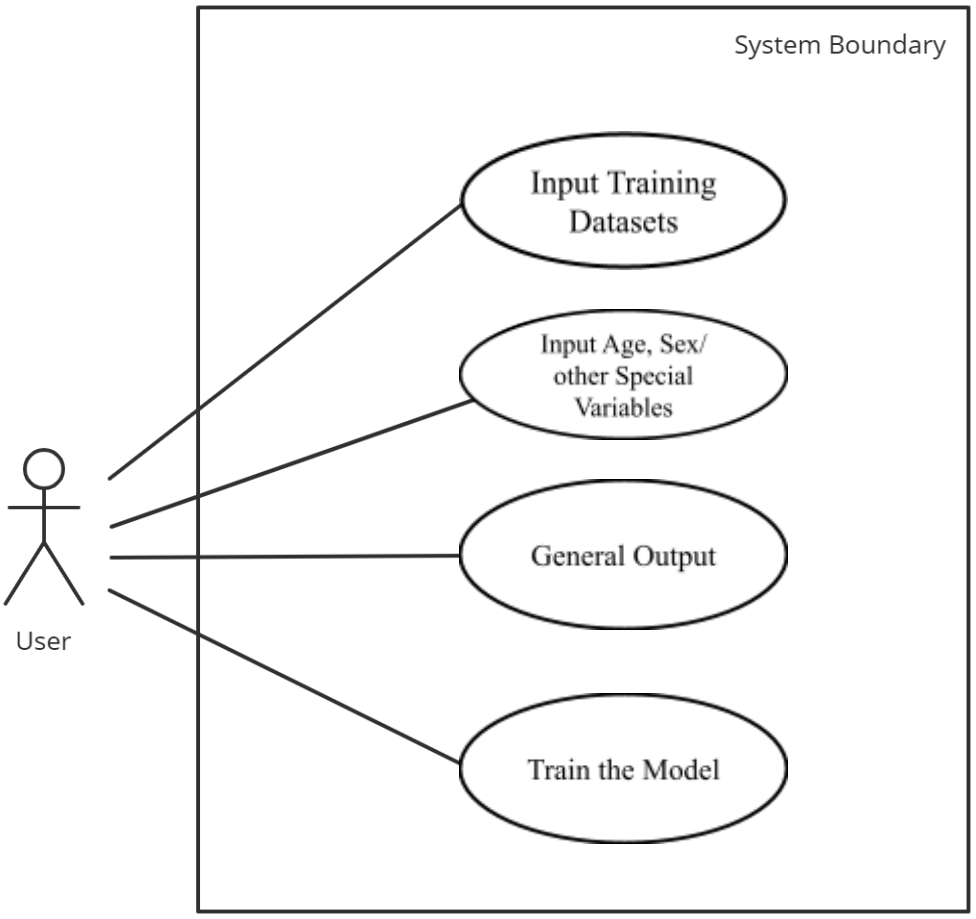

Fig 4.3.1: Use Case diagram of Heart Disease Prediction system using Machine Learning algorithm.

This is the use case diagram of our project. At first, input the training datasets, then input age/sex/cp or other parameters and show the general output which is to predict whether the heart is affected or not. The model gives 100% precision and 97% accuracy while predicting whether the heart is affected or not by given input.

4.4 Conclusion

In this chapter we have discussed the design process we have followed to build our system. We have given the flow chart and use case diagram of our system and describe them here. A flowchart is a picture of the separate steps of a process in sequential order. Use-case diagrams describe the high-level functions and scope of a system. These diagrams also identify the interactions between the system and its actors. The use cases and actors in use-case diagrams describe what the system does and how the actors use it, but not how the system operates internally. Though we have built it using any prior knowledge about the system we have managed to build it in a way that it uses the best possible algorithm and gives the best result.

Chapter 5

IMPLEMENTATION

5.1 Introduction

The implementation phase involves putting the project plan into action. It's here that the project manager will coordinate and direct project resources to meet the objectives of the project plan. As the project unfolds, it's the project manager's job to direct and manage each activity, every step of the way. Our project is implemented as per the requirements. All the options proposed in the design are implemented here. [4].

5.2 Source Code and Output:

We have implemented our project on Google Colaboratory [5]. Here we added the screenshots of our code by section wise:

```
[ ]  import numpy as np
     import pandas as pd
     from sklearn.model_selection import train_test_split
     from sklearn.linear_model import LogisticRegression
     from sklearn.metrics import accuracy_score
```

Fig 5.2.1: Imported libraries or dependencies

```
# loading the csv data to a Pandas DataFrame
heart_data = pd.read_csv('/content/heart_disease_data.csv')
```

```
[4] # print first 5 rows of the dataset
    heart_data.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Fig 5.2.2: Data collection and processing - I

```
[5] # print last 5 rows of the dataset
    heart_data.tail()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
298	57	0	0	140	241	0	1	123	1	0.2	1	0	3	0
299	45	1	3	110	264	0	1	132	0	1.2	1	0	3	0
300	68	1	0	144	193	1	1	141	0	3.4	1	2	3	0
301	57	1	0	130	131	0	1	115	1	1.2	1	1	3	0
302	57	0	1	130	236	0	0	174	0	0.0	1	1	2	0

```
[6] # number of rows and columns in the dataset
    heart_data.shape

    (303, 14)
```

Fig 5.2.3: Data collection and processing - II

```
[10] # statistical measures about the data
     heart_data.describe()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
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	303.000000
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	0.544554
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	0.498835
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	0.000000
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	0.000000
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	1.000000
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	1.000000
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	1.000000

```
[8] # checking the distribution of Target Variable
    heart_data['target'].value_counts()

    1    165
    0    138
    Name: target, dtype: int64
```

Fig 5.2.4: Data collection and processing - III

```
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, stratify=Y, random_state=2)

[15] print(X.shape, X_train.shape, X_test.shape)

     (303, 13) (242, 13) (61, 13)

[16] print(Y.shape, Y_train.shape, Y_test.shape)

     (303,) (242,) (61,)
```

Fig 5.2.5: Splitting the data into Training and Testing Data

```
[17] model = LogisticRegression()

[18] # train the LogisticRegression model with Training data
     model.fit(X_train, Y_train)
```

Fig 5.2.6: Train the Model with Logistic Regression

```
[19] # accuracy on training data
     X_train_prediction = model.predict(X_train)
     training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

[20] print('Accuracy on Traing Data: ', training_data_accuracy)

     Accuracy on Traing Data:  0.8512396694214877

[21] # accuracy on test data
     X_test_prediction = model.predict(X_test)
     test_data_accuracy = accuracy_score(X_test_prediction, Y_test)

[22] print('Accuracy on Test Data: ', test_data_accuracy)

     Accuracy on Test Data:  0.819672131147541
```

Fig 5.2.7: Accuracy Score

```
[23] input_data = (60,1,0,145,282,0,0,142,1,2.8,1,2,3)

     # change the input data to a numpy array
     input_data_as_np = np.asarray(input_data)

     # reshape the numpy array ass we are predicting for only on instance
     input_data_reshaped = input_data_as_np.reshape(1, -1)

     prediction = model.predict(input_data_reshaped)
     print(prediction)

     if prediction[0] == 0:
       print("Healthy Heart")
     else:
       print("Affected Heart")

     [0]
     Healthy Heart
```

Fig 5.2.8: Building the predictive system, and showing the output

5.3 Conclusion:

In this chapter, we have discussed the implementation part of our project. As we know Logistic Regression algorithm is a supervised learning algorithm and supervised ML is able to separate messages and classified the correct categories efficiently. It is also able to score the model and weigh them successfully. For instance, based on user input of the parameters it predicts whether the heart is affected or not.

Chapter 6

TESTING AND INTEGRATION

6.1 Introduction

Testing is the process of evaluating a system or its component(s) with the intent to find whether it satisfies the specified requirements or not. In simple words, testing is executing a system in order to identify any gaps, errors, or missing requirements contrary to the actual requirements.

Project integration management is the coordination of all elements of a project. This includes coordinating tasks, resources, stakeholders, and any other project elements, in addition to managing conflicts between different aspects of a project, making trade-offs between competing requests, and evaluating resources. [6]

We have tested different parts of our project and integrated them to get the best performance possible.

6.2 Conclusion

This project “Heart Disease Prediction using Machine Learning with Python” is a Data Mining project. In this chapter, we have discussed our project testing and integration. Project integration management is the coordination of all aspects of a project, including its processes and related systems, to ensure that a project is well executed. In doing so, project managers can better balance the demands of stakeholders, the expectations of clients, and the tasks during a project.

Chapter 7

MAINTENANCE

7.1 Introduction:

Project maintenance means any usual action, activity, expense, replacement, adjustment, or repair taken to retain a project or grant item in a serviceable, operational, or normal condition, or the routine efforts and expenses necessary to restore it to serviceable or normal condition, including the routine recurring work.

The importance of project management in organizations can't be overstated. When it's done right, it helps every part of the business run more smoothly. It allows your team to focus on the work that matters, free from the distractions caused by tasks going off track or budgets spinning out of control. [7]

So for these reasons, we have also decided to maintain our project.

7.2 Conclusion:

This chapter discusses the maintenance of the system. So, all of the modules in this application are maintained to make sure that the model works properly in near future. This is a crucial part of the project development.

An integrated project management and control system is used at all stages of software development project. Its main principle is– to gather as much information on the project and process of its development, as possible. Such approach helps integrated mangers handle their projects more effective.

Chapter 8

CONCLUSION

8.1 Introduction:

A cardiovascular disease detection model has been developed using three ML classification modeling techniques. This project predicts people with cardiovascular disease by extracting the patient medical history that leads to fatal heart disease from a dataset that includes patients' medical history such as chest pain, sugar level, blood pressure, etc. This Heart Disease detection system assists a patient based on his/her clinical information of them been diagnosed with a previous heart disease. The algorithms used in building the given model are Logistic regression, Random Forest Classifier and KNN [8]. The accuracy of our model is 85.1%. Use of more training data ensures the higher chances of the model to accurately predict whether the given person has a heart disease or not [9]. By using these, computer aided techniques we can predict the patient fast and better and the cost can be reduced very much. There are a number of medicaldatabases that we can work on as these Machine learning techniques are better and they can predict better than a human being which helps the patient as well as the doctors. Therefore, in conclusion this project helps us predict the patients who are diagnosed with heart diseases by cleaning the dataset and applying logistic regression and KNN to get an accuracy of an average of 87.5% on our model which is better than the previous models having an accuracy of 85%. Also, it is concluded that accuracy of KNN is the highest among the three algorithms that we have used i.e. 88.52%. 'Figure 6' shows 44% of people that are listed in the dataset are suffering from Heart Disease.

8.2 Advantages

1. Increased accuracy for effective heart disease diagnosis.
2. Handles the roughest(enormous) amount of data using random forest algorithm and feature selection.
3. Reduce the time complexity of doctors.
4. Cost-effective for patients.

8.3 Disadvantages

1. The prediction of cardiovascular disease results is not accurate.
2. Data mining techniques do not help to provide effective decision-making.
3. Cannot handle enormous datasets for patient records.

8.4 Conclusion:

From this project, it can be concluded that the machine learning algorithm is one of the important parts in order to create the Heart Disease Prediction application. To make it more efficient, improvements need to be implemented in the future.

REFERENCES

[1] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, *"Hybrid intelligent modeling schemes for heart disease classification," Applied Soft Computing,* vol. 14, pp. 47–52, 2014.

[2] https://medium.com/

[3] Dangare C S & Apte S S (2012). *Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications,* 47(10), 44-8.

[4] https://www.wikipedia.org/

[5] https://colab.research.google.com/

[6] https://www.w3schools.com/

[7] https://www.geeksforgeeks.org/

[8] Piller L B, Davis B R, Cutler J A, Cushman W C, Wright J T, Williamson J D & Haywood L J (2002). *Validation of heart failure events in the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) participants assigned to doxazosin and chlorthalidone. Current controlled trials in cardiovascular medicine,* 3(1), 10.

[9] Dangare Chaitrali S and Sulabha S Apte. *"Improved study of heart disease prediction system using data mining classification techniques." International Journal of Computer Applications* 47.10 (2012): 44-8.