**MINOR-1 PROJECT**

**SYNOPSIS MID Report**

**Heart Disease Detection Using Machine Learning**

Submitted By :

|  |  |  |
| --- | --- | --- |
| **Name** | **Roll No** | **Branch** |
| Tanisha Goyal | R110219144 | CCVT |
| Vishal Gupta | R110219153 | CCVT |
| Vishakha Sharma | R110219151 | CCVT |
|  |  |  |

**Under the guidance of**

**Mr. Abhirup Khanna**

**Associate Professor**

**Department of Computer Science**



School of Computer Science

**UNIVERSITY OF PETROLEUM AND ENERGY STUDIES**

**B. Tech CCVT, 5th Sem,  Batch : 2019-23**

**Approved By**

(Mr. Abhirup Khanna) (Dr. Neelu J. Ahuja)

**Project Guide Cluster Head**



**School of Computer Science**

University of Petroleum & Energy Studies, Dehradun

**Synopsis Report (2021-22)**

1. **Project Title**

Heart Disease Detection

**2**  **Abstract**

Heart is the next major organ comparing to brain which has more priority in human body. Heart disease (also called as Cardiovascular disease) is one of the most critical human diseases in the world and affects human life very badly. With the rampant increase in the heart stroke rates at juvenile ages, we need to put a system in place to be able to detect the symptoms of a heart disease at an early stage and thus prevent it. In Heart disease, the heart is unable to push the required amount of blood to other parts of the body. Prediction the heart disease is a critical challenge in the area of clinical data analysis. The amount of data in the health care industry is huge. Data mining turns the large collection of raw healthcare data into information that can help to make informed decision and prediction on time. In this project, we aim at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of heart disease. It is crucial to select the correct combination of significant features that can improve the performance of the detection models. Detection models were developed using different combination of features, and seven classification techniques : Decision tree, Naive Bayes (NB), Logistics Regression(LR), k-NN, Support Vector Machine(SVM) , Neural Network and Vote ( a hybrid technique with Naive Bayes and Logistics Regression). Our main objective is to detect the heart disease with higher accuracy rate through the detection model for heart disease with the hybrid algorithm

**Keywords** - Detection, heart disease, hybrid algorithm,  Support Vector machine (SVM), Naive Bayes(NB), Symptoms.

**3. Introduction**

The heart is one of the main organs of the human body. It pumps blood through the blood vessels of the circulatory system. The circulatory system is extremely important because it transports blood, oxygen and other materials to the different organs of the body. Heart plays the most crucial role in circulatory system. If the heart does not function properly then it will lead to serious health conditions including death. Heart Disease (also known as cardiovascular disease) is one of the prevalent disease that can lead to reduce the lifespan of human beings nowadays. Heart Disease is the number one killer according to World Health Organization(WHO) statistics. Millions of people die every year because of heart disease and large population of people suffers from heart disease. There is no death of records regarding medical symptoms of patients suffering heart strokes. Heart Disease are the number one cause of death globally. Life is dependant on component functionality of heart, because heart is necessary part of our body. As estimate of a person’s risk for coronary heart disease is important for many aspects of health promotion and clinical medicine. Due to digital technologies are rapidly growing, healthcare centres store huge amount of data in there database that is very complex and challenging to analysis. The techniques and algorithms can be directly used on a dataset for creating some models or to draw vital conclusions, and inferences from the dataset. Common attributes used for heart disease are Age, Sex, Fasting Blood Pressure, Chest Pain type, Resting ECG(test that measures the electrical activity of the heart), Number of major vessels coloured by fluoroscopy, Threst Blood Pressure (high blood pressure), Serum Cholesterol(determine the risk for developing heart disease), Thala Ch (maximum heart rate achieved), ST depression (finding on an electrocardiogram, trace in the ST segment is abnormally low below the baseline), pain lock (chest pain location (substernal=1, otherwise=0)), Fasting blood sugar, Exang(exercise included angina), smoke, hypertension, food habits, weight, height and obesity. Table below summarizes the most common types of heart diseases as follows:-

**Table 1: Different types of heart disease**

|  |  |
| --- | --- |
| Arrhythmia | The heart beat is improper whether it may irregular, too slow or too fast. |
| Cardiac arrest | An unexpected loss of heart function, consciousness and breathing occur suddenly. |
| Congestive heart failure | The heart does not pump blood as well as it should, it is the condition of chronic. |
| Congenital heart disease | The heart’s abnormality which develops before birth. |
| Coronary artery disease | The heart’s major blood vessels can damage or any disease occurs in the blood  vessels. |
| High Blood Pressure | It has a condition that the force of the blood against the artery walls is too high. |
| Peripheral artery disease | The narrowed blood vessels which reduce flow of blood in the limbs, is the  circulatory condition. |
| Stroke | Interruption of blood supply occur damage to the brain. |

Figure 1 depicts the parts of human heart such as Left atrium , Right atrium, Right Ventricle, Left Ventricle, Aorta, Pulmonary vein, Pulmonary valve, Pericardium, Pulmonary artery, Tricuspid valve, Aortic valve, Mitral valve, Superior vena cava and Interior vena cava.

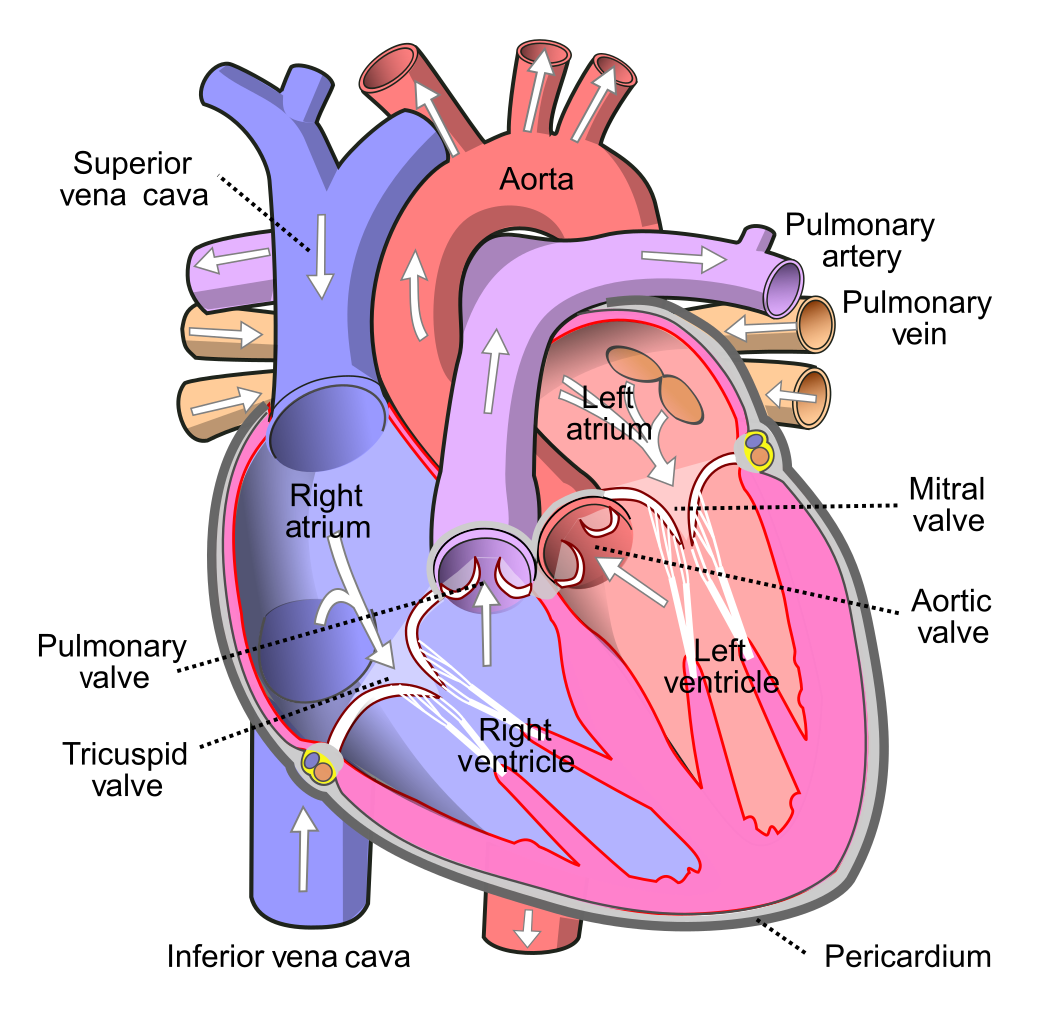


Figure 1 Human Heart

However the potential they have to help us foretell similar possibilities in seemingly health adults ae going unnoticed. For instance: As per Indian Heart Association(IHA), 50% of heart strokes occur under 50 years of age and 25% of all heart strokes occur under 40 years of age in Indians. Urban population is thrice as vulnerable to heart attacks as rural population. Also, it is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Prediction of heart disease early plays a crucial role for the treatment. If heart disease could be predicted before, lots of patient deaths would be prevented and also a more accurate and efficient treatment way could be provided. A need to develop such a medical diagnosis system arises day by day. The important key points of such medical diagnosis systems are reducing cost and obtaining more accurate rate efficiently. Developing a medical diagnosis system based on machine learning for prediction of heart disease provides more accurate diagnosis than traditional way and reduces cost of treatment. Heart attack is a common cause of death worldwide. Table below shows the symptoms of heart attack are as follows:

**Table 2 : Symptoms of Heart Attack**

|  |  |
| --- | --- |
| Chest pain | It is the most common symptom of heart attack. If someone has a blocked artery or is having a heart attack, he may feel pain, tightness or pressure in the chest. |
| Nausea, Indigestion, Heartburn and Stomach Pain | These are some of the often over looked symptoms of heart attack. Women tend to show these symptoms more than men. |
| Pain in the Arms | The pain often starts in the chest and then moves towards the arms, especially in the left side. |
| Feeling Dizzy and Light Headed | Things that lead to the loss of balance. |
| Fatigue | Simple chores which begin to set a feeling of tiredness should not be ignored. |
| Sweating | Some other cardiovascular diseases which are quite common are stroke, heart failure, hypertensive heart disease, rheumatic heart disease, Cardiomyopathy, Cardiac arrhythmia, Congenital heart disease, Valvular heart disease, Aortic aneurysms, Peripheral artery disease and Venous thrombosis. Heart diseases may develop due to certain abnormalities in the functioning of the circulatory system or may be aggravated by certain lifestyle choices like smoking, certain eating habits, sedentary life and others. If the heart diseases are detected earlier then it can be treated properly and kept under control. Here, early detection is the main key. Being well informed about the whys and wherefores of heart disease will help in prevention summarily. |



Figure 2 Heart Disease

The application of data mining brings a new dimension to heart disease detection. Various techniques in data mining and neural networks have been employed to find the severity of heart disease among humans. The severity of the disease is classified based on various methods like K- Nearest Neighbour Algorithm(KNN), Decision Tree(DT), Genetic Algorithm(GT), and Naive Bayes (NB). The nature of heart disease is complex and hence, the disease must be handled carefully. Not doing so may affect the heart or cause premature death. It is harder to use data mining techniques without prior and appropriate preparations. Data Redundancy and inconsistency in a raw dataset affect the predicted outcome of the algorithm so to apply the machine learning algorithms to its full potential, an effective preparation is needed to pre-process the datasets. Furthermore, unwanted features can reduce the performance of the data mining techniques as well. Thus, along with data preparation, a proper feature selection method is needed to achieve high accuracy in heart disease detection using significant features and data mining techniques. Although it has been quite clear that feature selection is as important as the selection of a suitable techniques. The heart disease datasets were collected from the data source. Kaggle datasets was selected because it is a commonly used database by machine learning researches with records that are most complete. Seven classification techniques (k-NN, Decision Tree, Naïve bayes, Logistics Regression, Vote, Support Vector Machine and neural networks) were applied to create prediction models for this project to prepare dataset. In the end , our main goal is to detect the heart disease earlier if we detect any of the symptoms as soon as possible and to prevent it.

**4. Literature Review**

In the past, lot of work has been done related to heart disease Prediction using data mining techniques. A brief literature review is presented here.

* In 2015 research done by Nguyen Cong Long et al. on disease prediction using firefly algorithm. The classifier is trained by using rough set theory. The results are compared with other classification techniques such as Naïve Bayes and SVM. Limitation of this study is that rough set attribute is unmanageable when there is large number of attributes.
* R. Sharmila et al, proposed to use non- linear classification algorithm for heart disease prediction. It is proposed to use bigdata tools such as Hadoop Distributed File System (HDFS), Mapreduce along with SVM for prediction of heart disease with optimized attribute set. This work made an investigation on the use of different data mining techniques for predicting heart diseases. It suggests to use HDFS for storing large data in different nodes and executing the prediction algorithm using SVM in more than one node simultaneously using SVM. SVM is used in parallel fashion which yielded better computation time than sequential SVM.
* P.Sai Chandrasekhar Reddy et al, proposed Heart disease prediction using ANN algorithm in data mining. Due to increasing expenses of heart disease diagnosis disease, there was a need to develop new system which can predict heart disease. Prediction model is used to predict the condition of the patient after evaluation on the basis of various parameters like heart beat rate, blood pressure, cholesterol etc. The accuracy of the system is proved in java.
* Ashwini shetty recommended developing the prediction system which will diagnosis the heart disease from patient’s medical dataset. 13 risk factors of input attributes have taken into account to build the system. After analysis of the data from the dataset, data cleaning and data integration was performed.
* S.Prabhavathi proposed Decision tree based Neural Fuzzy System (DNFS) technique to analyse and predict of various heart disease. This paper reviews the research on heart disease diagnosis. DNFS stand for Decision tree based Neural Fuzzy System. This research is to create an intelligent and cost effective system, and also to improve the performance of the existing system. Specifically in this paper, data mining techniques are used to enhance heart disease prediction. The result of this research shows that the SVM and neural networks results highly positive manner to predict heart disease. Still the data mining techniques are not encouraging for heart disease prediction.
* Syed Umar Amin, Dr. Rizwan Beg and Kavita Agarwal have proposed a hybrid system using Genetic Algorithms and Artificial Neural Networks for prediction of heart disease based on risk factors. The neural network was trained with Backpropagation Algorithm. It was pointed out that Backpropagation Algorithm has two major disadvantages. First problem is that finding out initial weights which are globally optimized is almost impossible. Second problem is slowness of Backpropagation Algorithm in convergence. The problems were solved by using Genetic Algorithm for optimization connection weights of Artificial Neural Network so as to obtain better performance from the network. The neural network used in the study had 12 input, 10 hidden and 2 output nodes. The results show that training accuracy is 96.2% and obtained validation accuracy is 89%.
* Saba Bashir, M.Younus Javed and Usman Qamar have another study to predict heart disease. The proposed method uses Decision Tree, Support Vector Machine and Naive Bayes as a hybrid model. Majority voting scheme was obtained by these three classifiers. There were two steps in the proposed approach. First one was producing every three classifiers’ decision. Second one was combining the decisions in order to acquire new model based on majority voting scheme. The results show that the accuracy rate obtained from the study is much higher than the others. 74% sensitivity, 82% accuracy and 93% specificity were obtained from the study to predict the heart disease.

**5. Problem Statement**

Heart disease may sedentary life and others. If the heart diseases are detected earlier then it can be treated properly and can be kept under control. In this project, earlier detection is the main key. The use of machine learning algorithms can be used to detect the heart disease at earlier stage :-

1. Support Vector Algorithm (SVM)
2. Naïve Bayes (NB)
3. Hybrid Algorithms
4. Kaggle datasets

**6. Objectives**

* Our objective behind this project is to develop a heart disease detection system with accurate few tests and attributes present in heart disease.
* Our goal is to predict the risk of having heart disease with few attributes and faster efficiency in form of datasets.
* Also, the system will help to provide effective treatment to future patients and can also avoid severe consequences.

**7. Methodology**

In our proposed project, it is a research conducted where asset of variables are kept constant while other set of variables are being measured as the dataset. This is more practical as we need to see weather the dataset of a person matches with all the details presented and cross checked with the previous data.

**Data Collection and Pre-processing** :- The data set for the research was taken from the Kaggle data repository. 10 data accessed from the Kaggle Machine Learning Repository is freely available. In particular, the Cleveland and Hungarian databases have been used by many researchers and found to be suitable for developing a mining model, because of lesser missing values and outliers. The data is cleaned and pre processed before it is submitted to the proposed algorithm for training and testing.

**Pattern Matching and Prediction:-** The overall objective of our work is to predict more accurately the presence of heart disease. In this project, Kaggle repository dataset are used to get more accurate results. Data mining classification technique which was applied is Naive Bayes.

**Accuracy Calculation:-** The goal field refers to the presence of heart disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with the Kaggle database have concentrated on simply attempting to distinguish presence (values 1,2,3,4) from absence (value 0). Attributes with categorical values were converted to numerical values since most machine learning algorithms require integer values.

**MACHINE LEARNING ALGORITHMS:-**

**Naïve Bayes (NB):** They are the collection of classification algorithms based on Bayes’s Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle ,i.e. every pair of features being classified is independent of each other.

**Support Vector Machine (SVM):** The main purpose of this algorithm is to use its wings to improve its accuracy of test results. Attributes such as age, blood pressure, thickness of the artery, etc., are fed into the support vector machine (SVM) which is used to predict risk of heart attack chances in a person.

**Hybrid Algorithm (HB):** It is an algorithm that combines two or more algorithms that solve the same problem. It is usually done to combine desired features of each dataset, so that overall algorithm is better than the individual algorithm.

**ALGORITHM**

Step 1: Start

Step 2: Add for each attribute A

Step 3: Traverse attribute list for A at examined Node

Step 4: Calculate Probability

Step 5: Set value of A in a class

Step 6: Update class for A

Step 7: Add all values in A

Step 8: IF no values of A are not found in A then go to step 4

Step 9: IF There is any attribute

Step 10: Next Attribute and go to step 3 else

Step 11: Stop

**8. Experimental Setup**

1. SOFTWARE REQUIREMENTS

* **Operating System :** Microsoft Windows
* Programming Language : JAVA

2. HARDWARE REQUIREMENTS

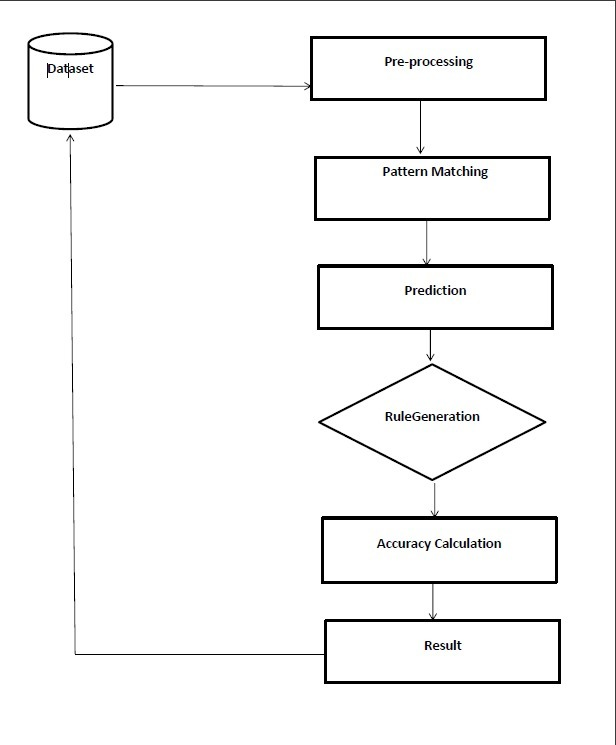
* **For Program**

Processor : 11th Gen Intel® Core(TM) i7-1165G7 @ 2.80GHz

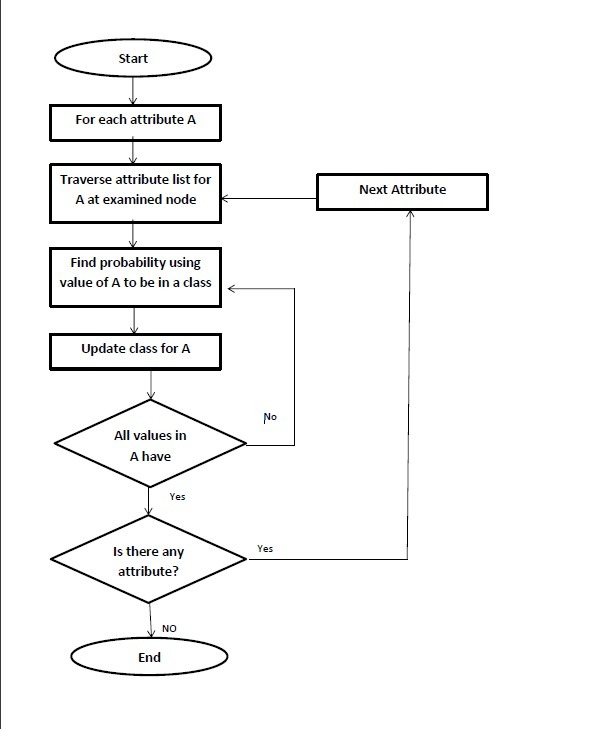
Disk Drive : Hard Drive

RAM: 8GB or higher

**9(a). DATA FLOW DIAGRAM**

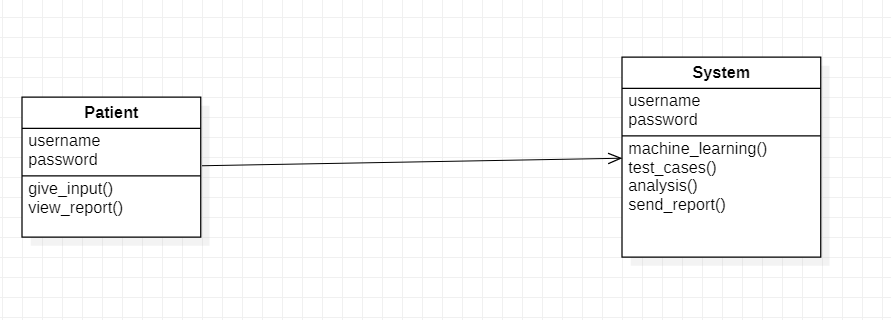


**9(b). FLOWCHART OF NAÏVE BAYES ALGORITHM**

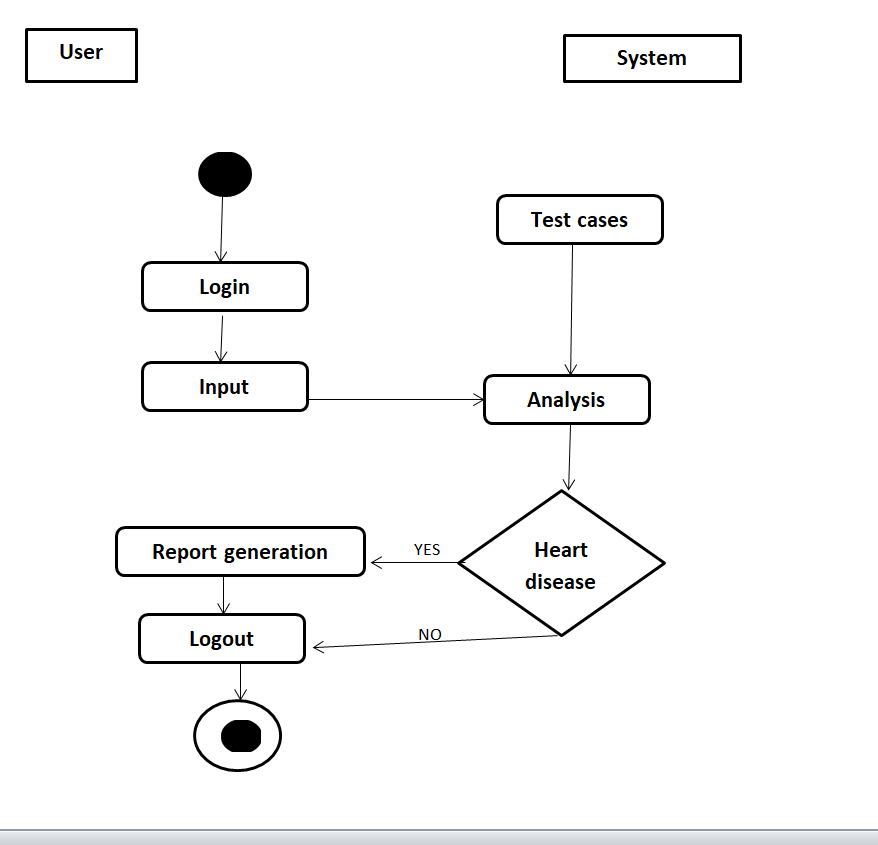


**9(c ). UML DIAGRAM:-**

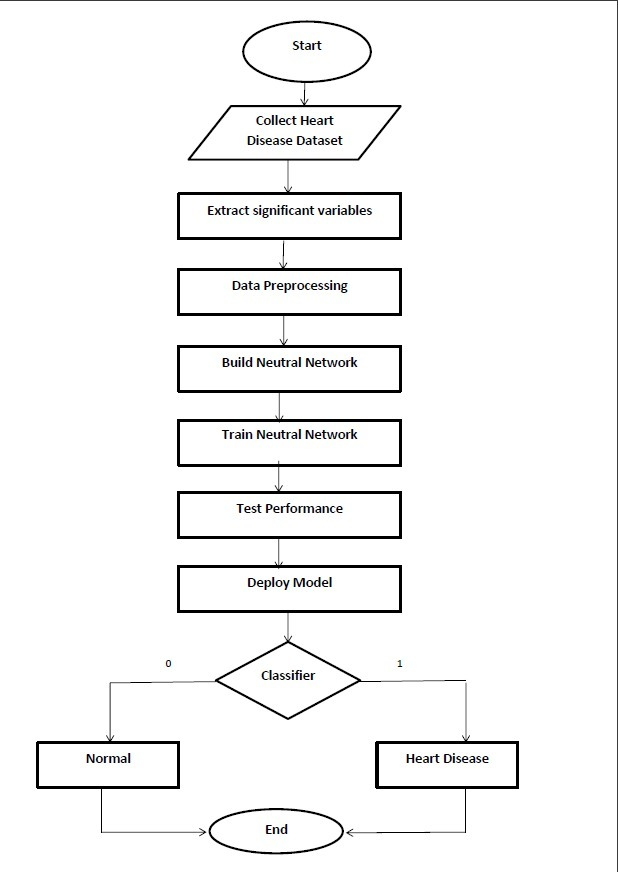
**CLASS DIAGRAM**

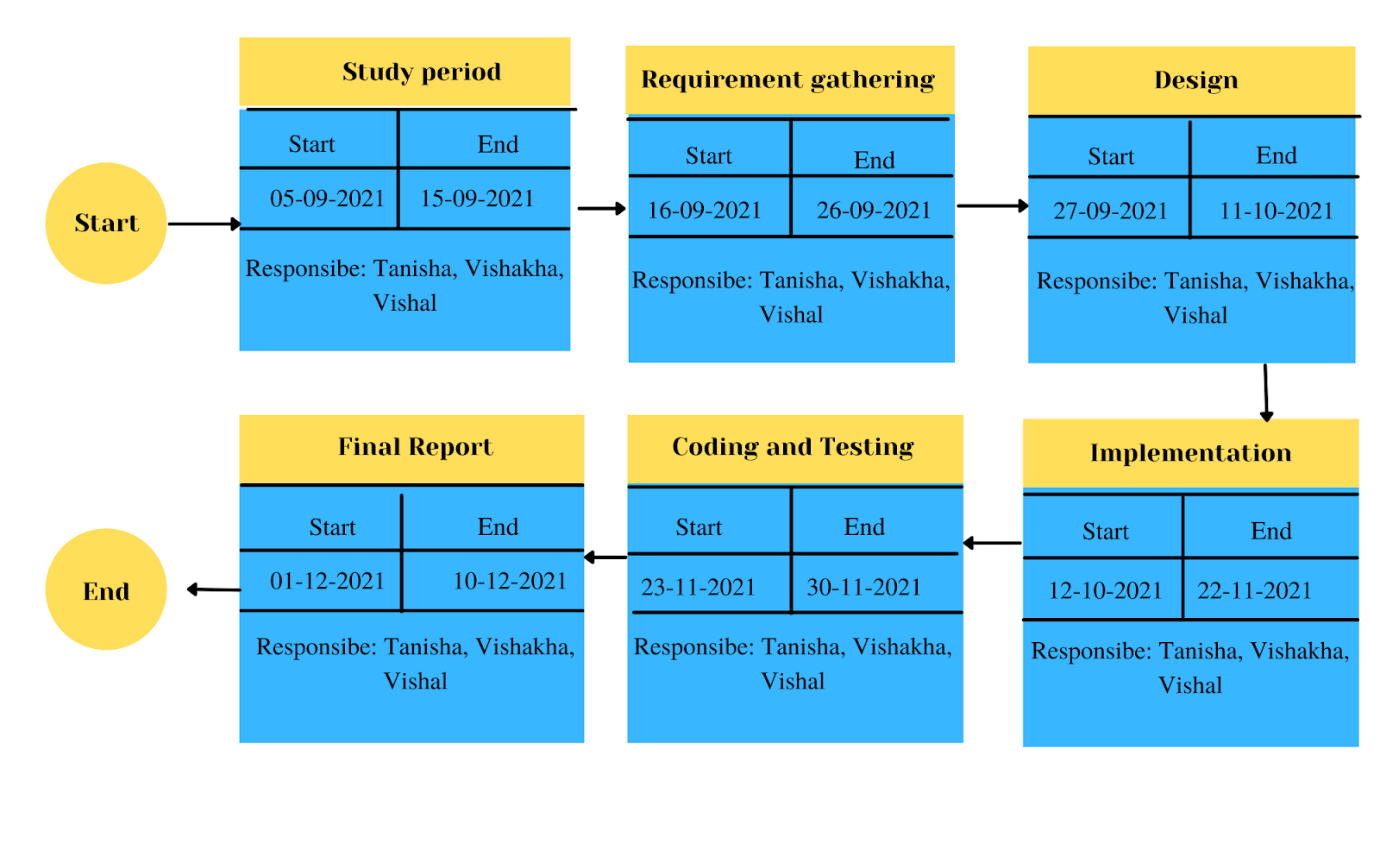
****

**ACTIVITY DIAGRAM:-**

****

**9(d). PROJECT FLOWCHART**



**10. Schedule (PERT CHART)**

**References & GIT Link**

1. Brownlee, J. (2016). Naive Bayes for Machine Learning. Retrieved March 4, 2019, from https://machinelearningmastery.com/naive-bayes-for-machine-learning/
2. Animesh Hazra, Arkomita Mukherjee, Amit Gupta,  Asmita Mukherjee, “Heart Disease Diagnosis and  Prediction Using Machine Learning and Data Mining  Techniques: A Review”, Research Gate Publications,  July 2017, pp.2137-2159.
3. T.Mythili, Dev Mukherji, Nikita Padaila and Abhiram  Naidu, “A Heart Disease Prediction Model using SVM Decision Trees- Logistic Regression (SDL)”,  International Journal of Computer Applications, vol. 68,  16 April 2013.
4. PREDICTION OF HEART DISEASE  USING K-MEANS and ARTIFICIAL  NEURAL NETWORK as HYBRID  APPROACH to IMPROVE ACCURACY  Amita Malav#1, Kalyani Kadam#2, Pooja Kamat #3 #Computer Science, Symbiosis International University, Pune, India
5. Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using k-nearest neighbor and genetic algorithm. Procedia  Technology, 10,85-94.
6. Dua, D., Karra Taniskidou, E., 2017. UCI Machine Learning Repository. Irvine, CA: University Dua, D., Karra Taniskidou, E., 2017. UCI Machine Learning Repository. Irvine, CA: University
7. A, A. S., & Naik, C. (2016). Different Data Mining Approaches for Predicting Heart Disease, 277– 281. https://doi.org/10.15680/IJIRSET.2016.0505545