Health Monitoring System

Submitted by

Rahul Kumar 13000114066

Rahul Saha 13000114067

Ravi Kumar 13000114073

Sahasradal Kishor Ghara 13000114083

Under the guidance of

**Mr. Mrinal Kanti Nath**

Submitted for the partial fulfillment for the degree of Bachelor of Technology in Computer Science and Engineering

Techno India,

EM 4/1, SaltLake, Sector V, Kolkata – 700 091

## **CERTIFICATE**

This is to certify that the project entitled “Health Monitoring System”, prepared by Rahul Kumar (13000114066), Rahul Saha (13000114067), Ravi Kumar (13000114073) and Sahasradal Ghara (13000114083) of B.Tech (Computer Science & Engineering), Final Year, has been done according to the regulations of the Degree of Bachelor of Technology in Computer Science & Engineering. The candidates have fulfilled the requirements for the submission of the project report.

It is to be understood that, the undersigned does not necessarily endorse any statement made, opinion expressed or conclusion drawn thereof, but approves the report only for the purpose for which it has been submitted.

----------------------------------------------- -----------------------------------------------

**Mr. Mrinal Kanti Nath** **Ms. Poulami Dutta**

Asst. Professor, Dept. of CSE In-charge, Dept. of CSE

-----------------------------------------------

(Signature of the External Examiner with Designation and Institute)

--------------------------------------------------------------------------------------------

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

--------------------------------------------------------------------------------------------

**ACKNOWLEDGEMENT**

We would like to express our sincere gratitude to **Mr. Mrinal Kanti Nath** of the department of Computer Science and Engineering, whose role as project guide was invaluable for the project. We are extremely thankful for the keen interest he / she took in advising us, for the books and reference materials provided for the moral support extended to us.

Last but not the least we convey our gratitude to all the teachers for providing us the technical skill that will always remain as our asset and to all non-teaching staffs for the gracious hospitality they offered us.

Place: Techno India, Salt Lake

Date:

………………………………

………………………………

………………………………

………………………………

**CONTENTS**

|  |  |  |
| --- | --- | --- |
|  | **Table of Contents** | **Page Number** |
| **1** | **Introduction** | 1 |
| 1.1 | Briefing | 1 |
| 1.2 | Problem Domain | 1 |
| 1.3 | Related Studies | 2 |
| 1.4 | Glossary | 5 |
| **2** | **Problem Definition** | 7 |
| 2.1 | Scope | 7 |
| 2.2 | Exclusion | 7 |
| 2.3 | Assumptions | 7 |
| **3** | **Project Planning** | 8 |
| 3.1 | Software Lifecycle Model | 8 |
| 3.2 | Scheduling | 9 |
| 3.3 | Cost Analysis | 11 |
| **4** | **Requirement Analysis** | 12 |
| 4.1 | Requirement Matrix | 12 |
| 4.2 | Requirement Elaboration | 13 |
| **5** | **Design** | 17 |
| 5.1 | Technical Environment | 17 |
| 5.2 | Hierarchy of Modules | 18 |
| 5.3 | Detailed Design | 18 |
| 5.4 | Test Planning | 19 |
| 6 | **Conclusion** | 20 |
| 6.1 | Project Benefits | 20 |
| 6.2 | Future Scope for Improvement | 20 |
| 7 | **References** | 21 |

**Tables and Figures**

|  |  |  |
| --- | --- | --- |
|  | **Table Name** | **Page Number** |
| 1 | Confusion Matrix | 4 |
| 2 | Glossary | 6 |
| 3 | Constant for Cost Analysis | 11 |
| 4 | Features of the Heart Disease dataset | 13 |
| 5 | Descriptive statistics of Heart Disease dataset | 14 |
| 6 | Corelation matrix between numeric attributes | 15 |
| 7 | Technical Environment | 17 |
| 8 | Test Planning | 19 |

|  |  |  |
| --- | --- | --- |
|  | **Figure Name** | **Page Number** |
| 1 | Random Forest Schema | 2 |
| 2 | SVM Schema | 3 |
| 3 | Iterative waterfall model | 8 |
| 4 | Project Plan | 9 |
| 5 | Gantt Chart 1 | 10 |
| 6 | Gantt Chart 2 | 10 |
| 7 | Gantt Chart 3 | 11 |
| 8 | Traceability Matrix | 12 |
| 9 | Statistics | 12 |
| 10 | Heart Disease Diagram | 15 |
| 11 | Hierarchy of Modules | 18 |
| 12 | DFD | 19 |

1. **INTRODUCTION**
   1. **BRIEFING**

Heart disease is the leading cause of death in the United States. More than  
600,000 Americans die from it each year. One out of every four deaths every  
year is from heart disease. Heart disease is the build up of plaque in a person’s  
arteries. As the plaque builds up, the arteries narrow reducing, blood flow to  
the heart. A doctor can perform several tests to diagnose heart disease, including chest X-rays, coronary angiograms, electrocardiograms and exercise stress tests. Fortunately, heart disease is treatable. Eating a healthy diet, exercising regularly, maintaining a healthy weight and taking medications are four ways to reduce the risk of developing heart disease.

The main objective of this study was to compare the algorithms which can classify the  
heart disease correctly based on different performance metrics. There are 13 dependent  
variables in the data set and 1 independent variable to be predicted. The original data set contains predicted variables from 0 to 4 representing a healthy heart starting from 0 to severely unhealthy heart at 4. For this study, 0 to 4 class labels were changed to 0 and 1. The predicted class can be either 0 or 1, meaning the heart is either 0 (“Healthy”) or 1 (“Unhealthy”). Techniques such as feature selection, grid search and probability calibration were used to get the optimal results. In this study, algorithms such as Support Vector Machine, Naïve Bayes, Adaboost, Random Forest are used. It can be concluded that Support Vector Machine best the algorithms for this data set and possibly other heart disease data sets. For the proper conclusion for this study to be applied clinically, it needs to be further elaborated with the help of experts in both heart and machine leaning domains.

**1.2 PROBLEM DOMAIN**

* **Business Domain:** HealthCare
* **Technical Domain:** ML
* **Software Requirements:** Python 3
* **Operating System**: Windows, Linux, Mac
* **Hardware requirements**: 50GB SATA HDD, 4GB RAM, Intel i3 Processor
  1. **RELATED STUDIES**

Data mining (DM) [2] is the extraction of useful information from large data sets that results in predicting or describing the data using techniques such as classification, clustering, association, etc. Data mining has found extensive applicability  
in the healthcare industry such as in classifying optimum treatment methods, predicting disease risk factors, and finding efficient cost structures of patient care. Research using data mining models have been applied to diseases such as diabetes, asthma, cardiovascular diseases, AIDS, etc. Various techniques of data mining such as naïve Bayesian classification, support vector machines, decision trees, logistic regression, etc. have been used to develop models in healthcare research.

**Random Forest**

A random forest is a classifier consisting of a collection of decision trees, where  
each tree is constructed by applying an algorithm A on the training set S and  
an additional random vector, θ where θ is sampled i.i.d.(independently and identically distributed) from some distribution. The prediction of the random forest is obtained by a majority vote over the predictions of the individual trees.

The Random Forest algorithm works in following steps:

1. Picks random K data points from the training data.  
2. Builds a decision tree for these K data points.  
3. Chooses the Ntree subset from the trees and performs step 1 and step 2  
4. Decides the category or result on the basis of the majority of votes.

To understand Random Forest more intuitively it is better to understand decision trees and they can be better understood with the help of a diagram.

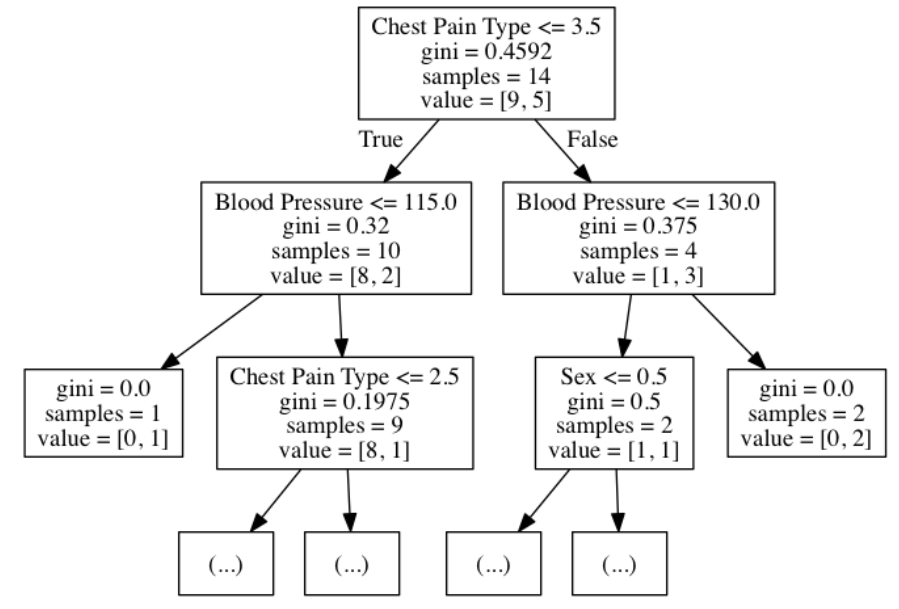
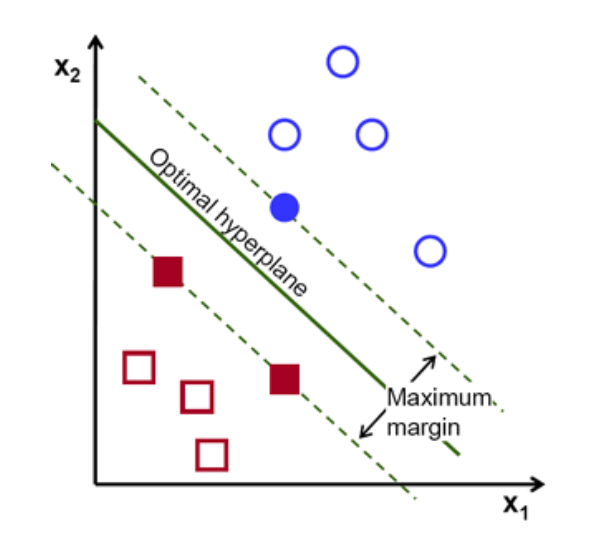


Figure 1 Random Forest Schematic.

Referring to Figure 1, it illustrates how decision trees work. If decision trees are used to predict whether there is a heart disease or not, then it will decide according to the above mapping. Gini is the coefficient of the spread of the data, samples are the number of data taken for classifying the node, value is the array of samples which are classified as true or false.

**Support Vector Machine (SVM)**

A Support Vector Machine (SVM) is a classifier which distinct the various classes of data by the use of a hyper-plane. SVM is modelled with the training data and it outputs the hyper-plane in the test data. The SVM model tries to find the space in the matrix of data where different classes of data can be widely differentiated and draws a hyper-plane.

  
 Figure 2 Support Vector Machine. Reprinted from Introduction to SVM

In Figure 2, Red and Blue are the classes of labelled training data points. To classify  
them linearly a hyper-plane can be drawn but the question is: There is more than one way to draw a hyper-plane so which one is optimal? An optimal hyper-plane is chosen which maximizes the margin between the classes. Hyper-plane need not always be 10 linear. A hyper-plane in SVM can also work as a non-linear classifier using technique known as kernel-trick.

**Naïve Bayes**

Naïve Bayes or Naïve Bayes classifier in a machine learning context is a classifier  
which uses the Bayes theorem to classify the data and it assumes that the probability  
of certain feature X is totally independent of another feature Y. Bayes theorem can be easily explained with the following example. Probability of spanners produced by machine A is 0.6, machine B is 0.4. A defect in spanners in the whole of production is 1 percent and the probability of defected spanners produced by machine A is 50 percent and machine B is 50 percent. In this scenario Bayes theorem can be used to answer what is the probability of a defected machine produced by machine B.

**Ada Boost**

Ada Boost is the one of the robust techniques proposed by Freund and Schapire and  
most practical and widely used boosting algorithm. Boosting is an ensemble technique in machine learning used for creating highly accurate predictions or strong classifier from relatively weak and inaccurate classifiers. Boosting is an iterative process  
where a model is trained in data and finds weak learners. The second model is trained  
in data which learns from the mistakes of previous training and fixes the errors and the  
process continues until the training data is correctly predicted.

**Performance Metrics:**

In a machine learning context, performance metrics is the measurement of algorithm  
on how well algorithm performs based on different criteria such as accuracy, precision,recall etc. Different performance metrics are discussed below.  
  
**Confusion Matrix**

Confusion matrix is a technique to show how the classifier is confused while predicting. In a binary classification problem, the confusion matrix is often illustrated as below.

Table 1. Confusion Matrix

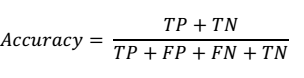
|  |  |  |
| --- | --- | --- |
| Correct classification | Classified as | |
|  | + (1) | - (0) |
| + (1) | TP (1,1) | FN (1,0) |
| - (0) | FP(0, 1) | TN(0,0) |

In Table 1, TP is the true positive value which means the positive value is truly classified, FP is the false positive value which means positive value which is falsely classified, FN is the false negative value which means negative value which is falsely classified and TN is the true negative which means negative value which is truly classified.

From the confusion matrix, different performance metrics can be calculated. Using Table 1 as an example, accuracy, precision, recall and F1-score are explained in more detail below.

**Accuracy**

Accuracy or predictive accuracy is the measure of the proportion of instances that are correctly classified. It shows how close the predicted value is to the true value or the theoretical value. Formula for calculating the accuracy is given below.



Accuracy is the measure of how close or near the predicted value is to the actual or  
theoretical value. For example, if the actual value of a person’s height is 6 feet and  
measured value or predicted value is 5.9 then it is quite an accurate measurement.

**Precision**

Precision is defined as the proportion of true positive instances which are classified as positive. It shows how close predicted values are to each other.

Precision is the measure of how close or near the predicted values are to each other.  
For example, if actual value of person’s height is 6 feet and measured value is or predicted value is 5.5 and every time measurement is taken height is 5.4 or 5.6 then prediction is quite precise but not accurate.

**Recall**

Recall is defined as the proportion of positive instances are that correctly classified as positive. Recall is also often called sensitivity.

Recall is simply the measure of how many true samples are predicted from the all the samples.

**F1\_Score**

F1 score is defined as the measure which combines both precision and recall and tries to convey the balance between them.

**Tools Used**

Different tools are used for this study. All of them are free and open source.  
1. Python 3.6  
2. NumPy 1.11.3  
3. Matplotlib 1.5.3  
4. Pandas 0.19.1  
5. Seaborn 0.7.1  
6. SciPy and Scikit-learn 0.18.1

* 1. **GLOSSARY**

Table 2. Glossary

|  |  |
| --- | --- |
| **ML** | **Machine Learning** |
| **Mac** | **Macintosh** |

1. **PROBLEM DOMAIN:**

**2.1 SCOPE:**

The first step of this study was to use an available heart disease data set and compare  
different machine learning algorithms to understand their performance based on different performance metrics. In the second step, scope for the study and its purpose was defined. The scope of the study was defined as per study purpose, resources and  
schedule. The studies purpose was to understand the machine learning algorithms  
behavior in particular heart disease data sets and to try to infer the results. The resources include a computer used for the analysis, this is crucial in defining the scope of the study because computer used defines how much and how fast data can be analyzed. Computer used in this study was enough to perform major tasks but not enough for in-depth and optimal results. In the third step, the basic knowledge to get started with the research was defined. For that several online resources were used and fundamentals for the study were understood. In fourth step, data was collected online from machine learning repository UCI, Irvine. In the fifth step, descriptive and exploratory data analysis as well as data cleaning was done as per the knowledge gained from step 3. It helped to understand the data more such as features included, correlation between features, missing values and outliers. In the sixth phase, different literature was studied related to this heart disease data set. The literature was available online and the findings of the literatures were reviewed. In the seventh phase, different machine learning algorithms were trained and the results were obtained as per different performance metrics. In the eighth phase, results were interpreted and compared with other existing literature on the subject. In ninth and final phase, conclusions were derived based on the results obtained.

**2.2 EXCLUSIONS:**

The article encompasses various types of studies, for example, randomized controlled trials, non-randomized controlled trials, longitudinal studies, cohort or case-control studies and descriptive and qualitative studies.

**2.3 ASSUMPTIONS:**

It is assumed that problems such as missing data, inconsistent data, and duplicate data have all been resolved.

**3. PROJECT PLANNING:**

**3.1 Software Life Cycle Model:**

The first SDLC to be developed and implemented was the classic Waterfall Model. In "The Waterfall" approach, the whole process of software development is divided into separate phases. In Waterfall model, typically, the outcome of one phase acts as the input for the next phase sequentially. In this model phases do not overlap. The next phase is started only after the defined set of goals is achieved for previous phase and it is signed off.

It does not allow for much reflection or revision. Once an application is in the testing stage, it is very difficult to go back and change something that was not well-documented or thought upon in the concept stage.

In this Project we have used a modified and more comprehensive form of the Waterfall Model called the Iterative Waterfall Model.

Iterative process starts with a simple implementation of a subset of the software requirements and iteratively enhances the evolving versions until the full system is implemented. At each iteration, design modifications are made and new functional capabilities are added. During each iteration, the development module goes through the requirements, design, implementation and testing phases. Each subsequent release of the module adds function to the previous release. Parallel development can be planned. Testing and debugging during smaller iteration is easy.

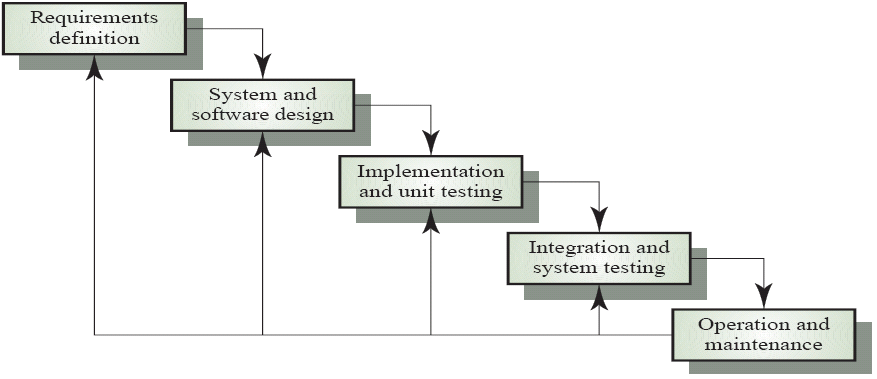
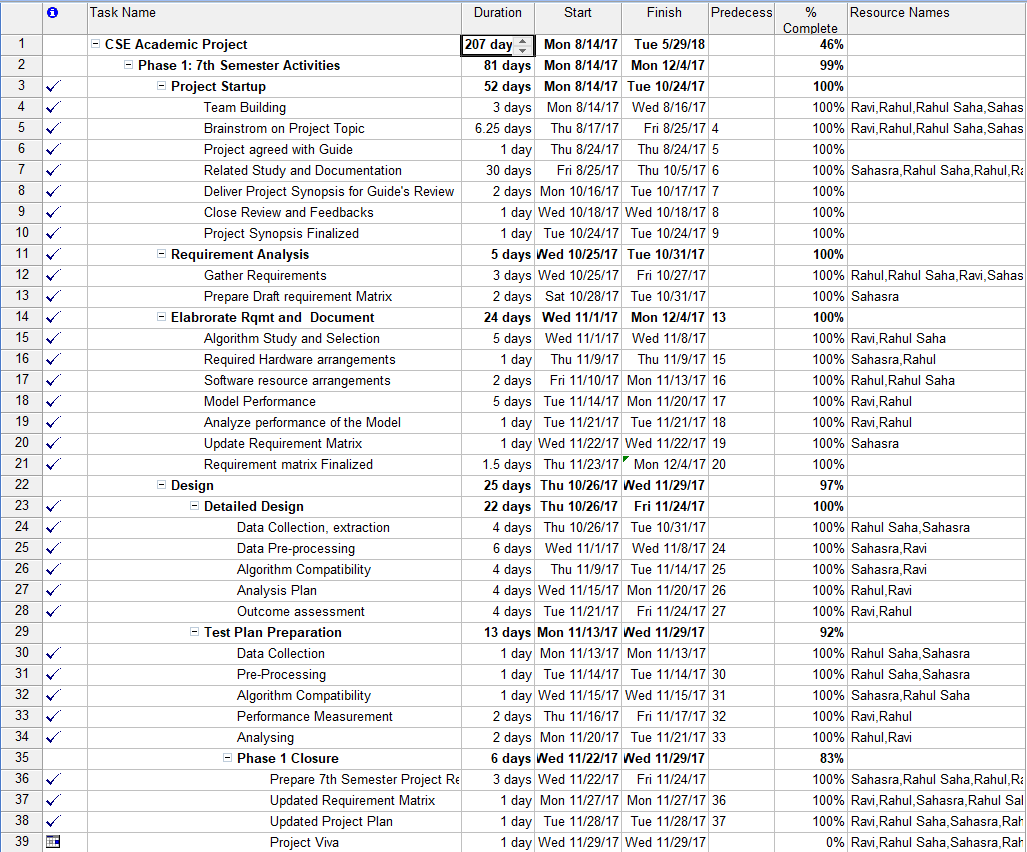


Figure 3: Iterative Waterfall Model

**3.2 SCHEDULING**

The project plan and the Gantt chart of our project work is shown below :

****

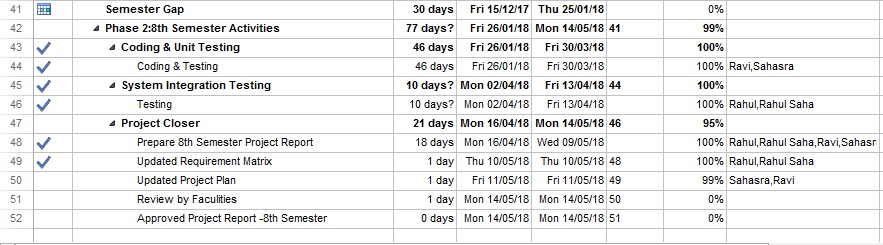
****

Figure 4: Project plan

**Scheduling Diagram:**

**Gantt Chart:**

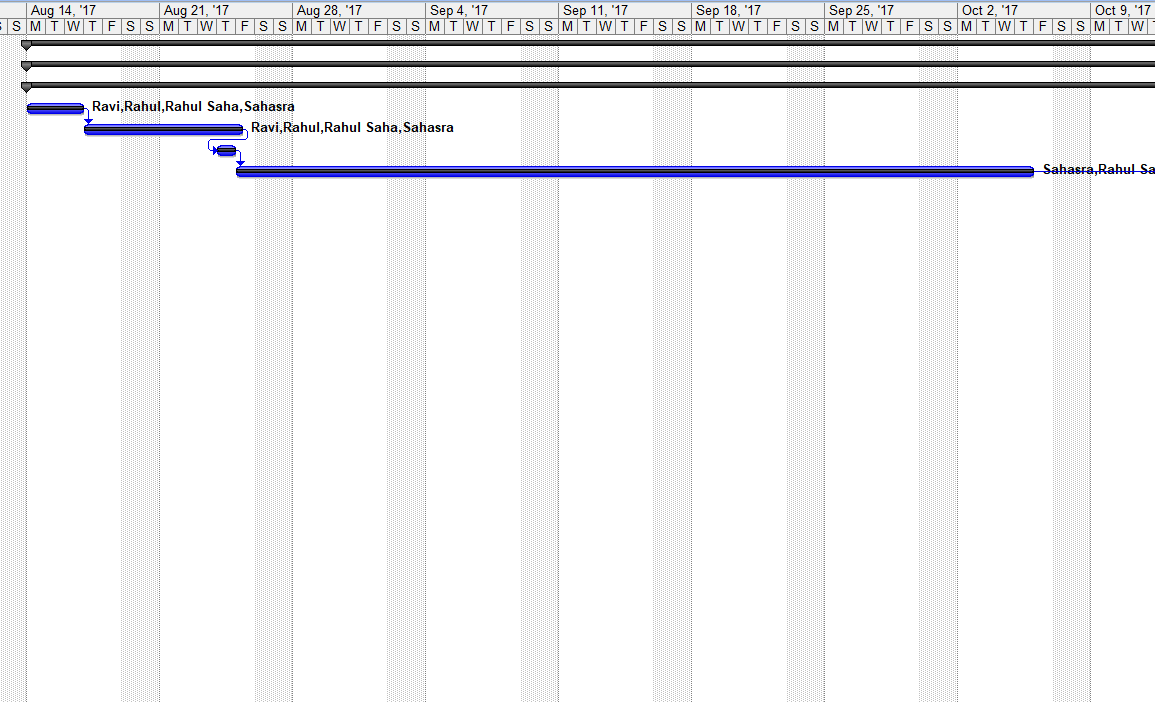
****

Figure 5: Gantt Chart 1

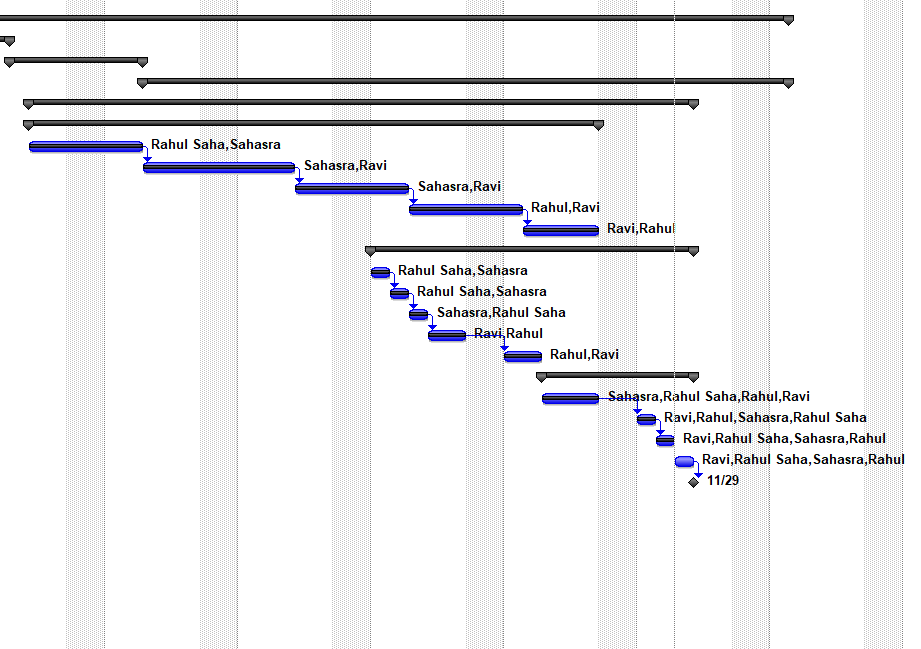
****

Figure 6: Gantt Chart 2

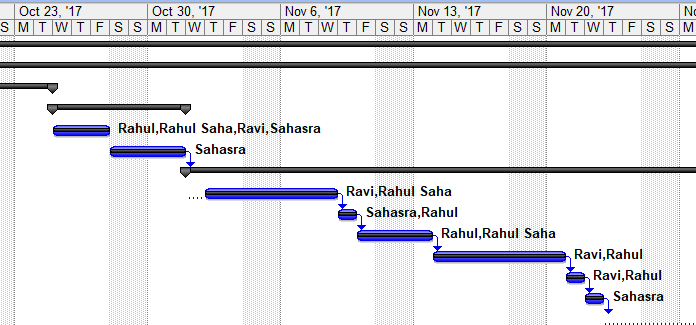
****

Figure 7: Gantt Chart 3

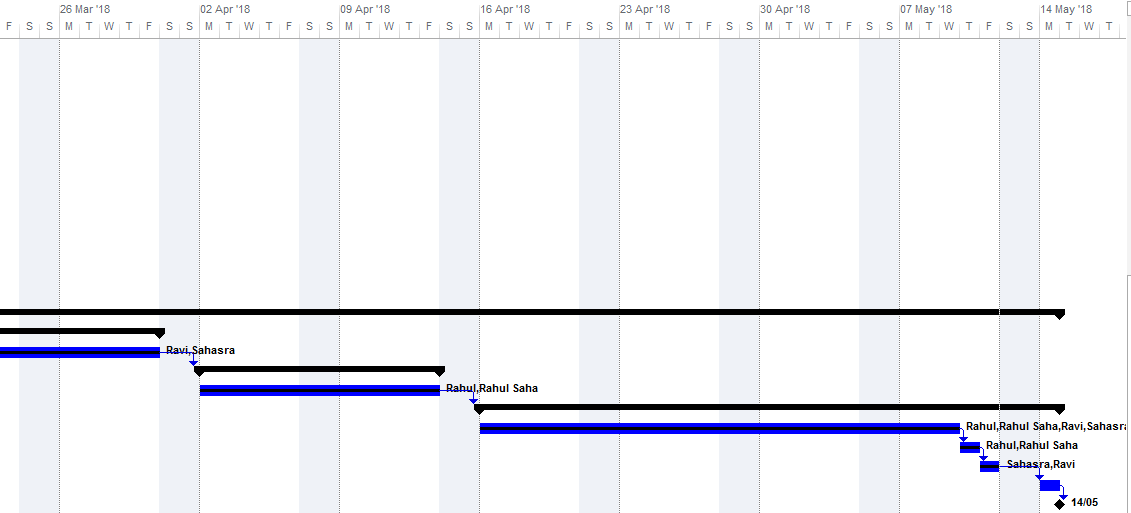
****

Figure 7: Gantt Chart 4

**3.3 Cost Analysis**

A project is feasible if it is possible to finish it within fixed budget constraints. Cost Analysis, sometimes also referred to as Cost/Benefit Analysis (CBA), is a systematic approach used to calculate and compare the benefits and costs of a project. For the purpose of this project, we are using a heuristic estimation technique called Constructive Cost Estimation Model (COCOMO).

The basic **COCOMO** gives an approximate estimate of the project parameters. This estimation model is given by the following expressions:

Effort Applied (E) = ab×(KLOC)bb PM

Development Time (D)=cb×(Effort Applied)db months

Where:

1. KLOC is the estimated size of the software product expressed in Kilo Lines of Code.
2. Effort Applied is the total effort required to develop the software product, expressed in person-months (PMs).
3. Development Time is the estimated time to develop the software, expressed in months.
4. ab, bb, cb and db are constants as given in the following table:

***Table 3 : Constants for Cost Analysis***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Software Project** |  |  |  |  |
| Organic | 2.4 | 1.05 | 2.5 | 0.38 |
| Semi-detached | 3.0 | 1.12 | 2.5 | 0.35 |
| Embedded | 3.6 | 1.20 | 2.5 | 0.32 |

**3.3.1. Estimation of effort, development time and productivity**

Lines of code = 5,000 = 5 KLOC (estimated)

Effort = 2.4 × (5)1.05 = 13 man-months (estimated)

Development Time = 2.5 × (13)0.38 = 6.62 months

Productivity = KLOC/Effort = 5/13 KLOC

1. **REQUIREMENT ANALYSIS:**

**4.1 REQUIREMENT MATRIX:**

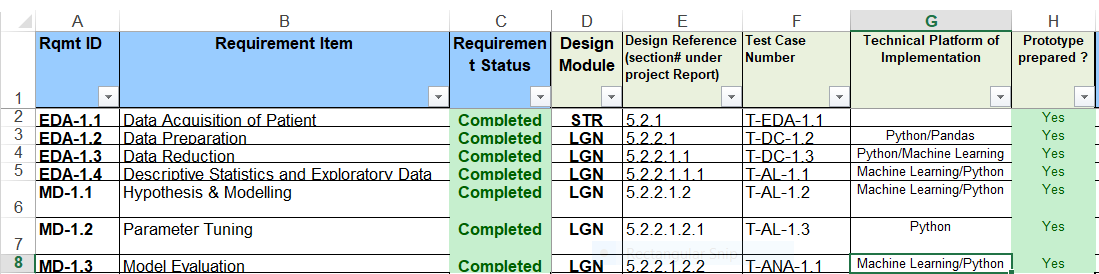


Figure 8: Traceability Matrix

**Statistics**

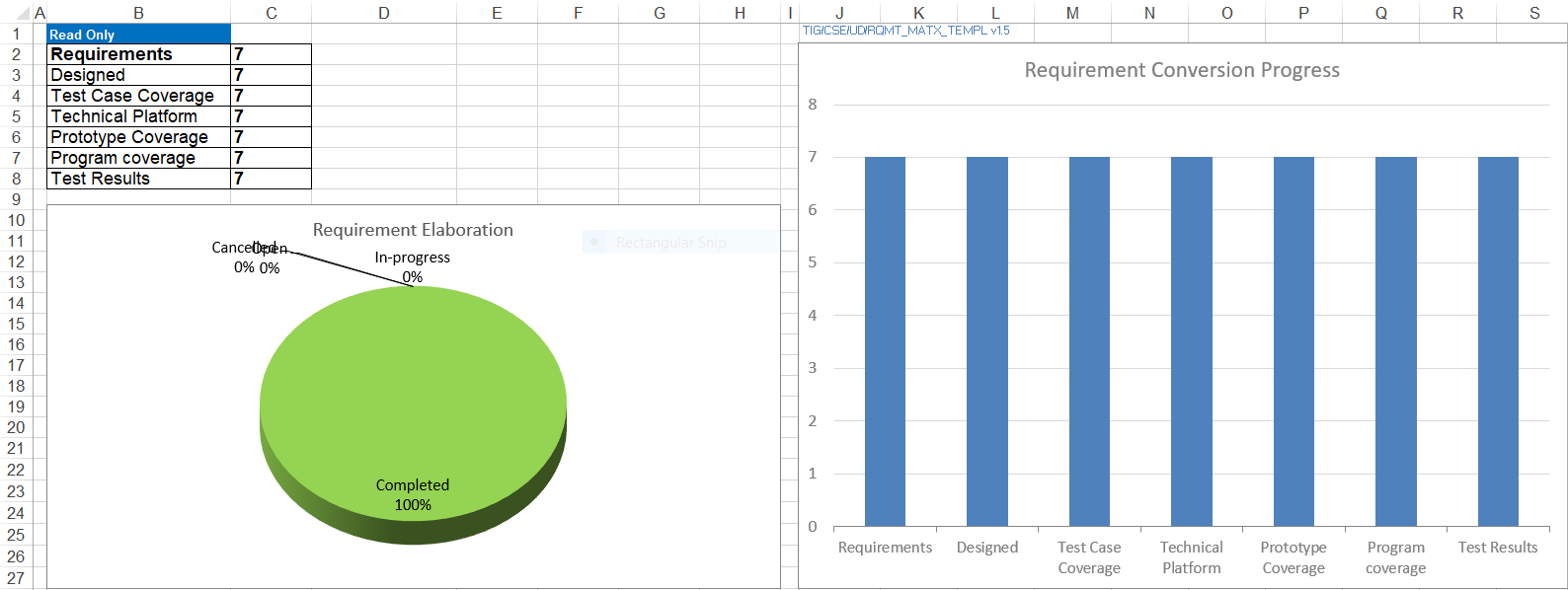


Figure 9: Statistics

**4.2 REQUIREMENT ELABORATION**

**4.2.1 Data Acquisition of Patient**

We used the heart disease data set [1] available from the UCI Machine Learning Repository. This data set dates from 1988 and consists of four databases: Cleveland (303 instances), Hungary (294), Switzerland (123), and Long Beach VA (200). Each database provides 76 attributes, including the predicted attribute. There are many missing attribute values. In addition, the Cleveland data set became corrupted after the loss of a computing node, and the surviving data set contains only 14 attributes per instance. Counting only instances with non-missing values for these 14 attributes, the total for all four databases is 299 instances (297 from Cleveland alone). This is the data set I will be using, and for simplicity I will be referring to it as the Cleveland data set.

Table 4. Features of the Heart Disease dataset

|  |  |
| --- | --- |
| **Features** | **Description** |
| Age | Age in years |
| Sex | Gender instance (0 = Female, 1 = Male) |
| Cp | Chest pain type (1: typical angina, 2: atypical angina, 3: non anginal pain, 4: asymptomatic) |
| Trestbps | Resting blood pressure in mm Hg |
| Chol | Serum cholesterol in mg/dl |
| Fbs | Fasting blood sugar > 120 mg/dl (0 = False, 1= True) |
| Restecg | Resting ECG results (0: normal, 1: ST-T wave abnormality, 2: LV hypertrophy) |
| Thalach | Maximum heart rate achieved |
| Exang | Exercise induced angina (0: No, 1: Yes) |
| Oldpeak | ST depression induced by exercise relative to rest |
| Slope | Slope of the peak exercise ST segment (1: up-sloping, 2: flat, 3: down-sloping) |
| Ca | Number of major vessels coloured by fluoroscopy (values 0 - 3) |
| Thal | Defect types: value 3: normal, 6: fixed defect, 7: irreversible defect |
| Num | Diagnosis of heart disease (0: Healthy, 1: Unhealthy) |

There are actually 76 attributes in the dataset but only 14 attributes are used for this  
study, these 14 attributes are in Table 3.

* + 1. **Data Preparation:**

The Cleveland dataset used for this study contains 303 rows, 5 numerical and 9 categorical attributes. The data set has 6 missing values. Since the missing values are from categorical attributes it was replaced with mode value of the respective columns. There are 43 outliers in the dataset but it is not a noise or wrong input. Data falling outside the 3 standard deviation are considered outliers. Patients who are diagnosed with heart disease have these outliers, meaning their cholesterol and blood pressure data is above the normal scale. So, in real clinical scenario it is not outlier and therefore it is decided to keep these outliers as they are. Standardization was used to rescale the data. That helped the machine learning models to perform better.

* + 1. **Data Reduction:**

For the Cleveland dataset, Sequential Backward Elimination is used. Any features in the data set whose p-value is less than 0.05 is significant and is considered essential in this data set. The logistic model was used as a classifier for sequential backward elimination.  
Out of 13 dependent attributes 9 attributes are selected by using sequential backward elimination. Data reduced features selected are listed below.  
1. Sex  
2. Chest pain type (Cp)  
3. Resting blood pressure (Trestbps)  
4. Maximum heart rate achieved (Thalach)  
5. ST depression induced by exercise (Oldpeak)  
6. Slope of the peak exercise ST segment (Slope)  
7. Number of major vessels coloured by fluoroscopy (Ca)

* + 1. **Descriptive Statistics and Exploratory Data Analysis:**

In this section, both descriptive statistics and exploratory data analysis are discussed.  
Out of 14 attributes, the data contains only five numeric attributes, out of which only four are tabulated below. Oldpeak attribute is excluded from the table because it needs deep understanding of heart physiology to interpret its results.

Table 5. Descriptive statistics of Heart Disease dataset.

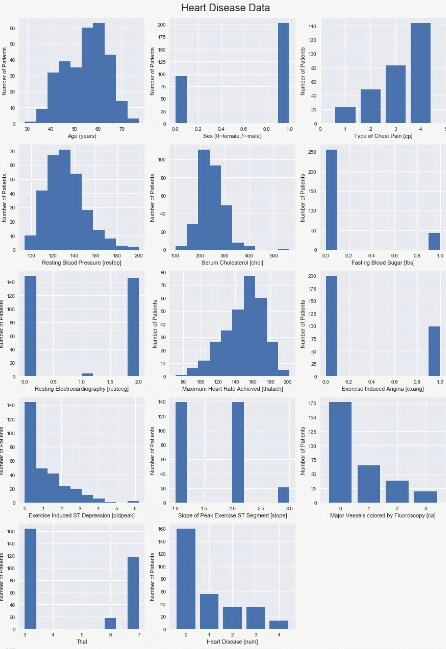
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Age** | | **Trestbps** | | | **Chol** | | **Thalach** |
| **mean** | | 54.43 | | 131.68 | 246.69 | | 149.60 | | |
| **std** | | 9.03 | | 17.59 | 51.77 | | 22.87 | | |
| **min** | | 29.00 | | 94.00 | 126.00 | | 71.00 | | |
| **median** | | 56.00 | | 130.00 | 241.00 | | 153.00 | | |
| **mode** | | 58.00 | | 120.00 | 197.00 | | 162.00 | | |
| **max** | | 77.00 | | 200.00 | 564.00 | | 202.00 | | |

In Table 4, the distribution of data is non-symmetrical and skewed. Population of data is 303, whose age range is 29 to 77. Age distribution of data is left skewed which shows population with low age are absent. Mean, median and mode of age column are 54.43, 56.00 and 58.00 respectively. Standard deviation is 9.03 which shows the sparseness of the age, ranging from 53.40 to 64.46, this age range visits the doctor most. Similarly, Trestbps, blood pressure column’s data is right skewed, which shows either population with normal blood pressure is present or there is lack of high blood pressure. Average blood pressure is 131.68 which is slightly above normal blood pressure, 120, this shows population is in pre-high blood pressure zone. Minimum blood pressure is 94 which is higher than low blood pressure 90, this shows population has not low blood pressure problems. In Chol, column mean value is 246.69 which mean population has high cholesterol level, cholesterol border line value is from 200 to 239. Mode value is 197.00 which shows most of the population has desirable cholesterol level, which is less than 200. Cholesterol mean value of 246.69 is abnormal in contrast to mode value of 197. In Thalach, maximum heart rate achieved, mean value is 149.60 with standard deviation of 22.87 with minimum value of 71.00 and maximum value of 202.00 which shows heart beat rate in average is less than normal of 200 beats per minute.

Table 6. Corelation matrix between numeric attributes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Age | Trestbps | Chol | Thalach |
| Age | 1.00 | 0.28 | 0.20 | -0.39 |
| Trestbps | 0.28 | 1.00 | 0.13 | -0.04 |
| Chol | 0.20 | 0.13 | 1.00 | -0.003 |
| Thalach | -0.39 | -0.04 | -0.003 | 1.00 |

From Table 6, it can be concluded that age and maximum heart rate achieved has  
some correlation. As, age increases, heart rate decreases except that none of other  
attributes shows any high correlation with each other.



**4.2.5 Hypothesis & Modelling:**

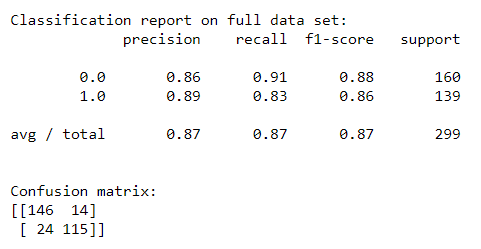
We used the full Cleveland data set to fit various classification algorithms like logistic regression, support vector machines, decision tree, random forest, Adaboost and xGBoost model. For the purpose of comparative analysis, six Machine Learning algorithms are discussed. The different Machine Learning (ML) classification algorithms like logistic regression, support vector machines, decision tree, random forest, adaboost and xGBoost model are used. The reason to choose these algorithms is based on their popularity.

* + 1. **Parameter Tuning**

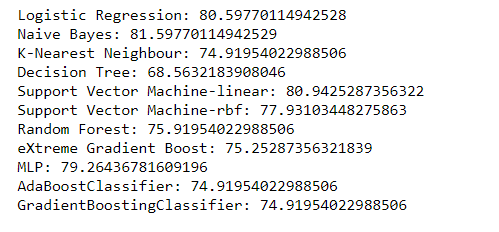
Scikit-learn library contains the class called GridSearchCV which performs the parameter tuning.

* + 1. **Model Evaluation**

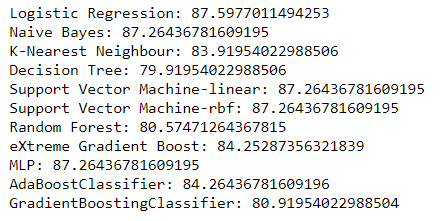
The test set is the portion of data where the model is tested, it is often the dependent  
variable of the data. In this study, 30 percent of the data was used for testing. When  
cross validated data is tested, it will perform better or worse depending on the model  
used. So, to ensure every model is functioning in its optimum, a technique named parameter tuning was used.



**Model After Applying PCA**



**Model After Applying LDA**



**5. Design:**

**5.1 Technical Environment:**

Table 7

|  |  |
| --- | --- |
| Activity | Tool/program |
| IDE | Jupyter Notebook/ Spyder |
| Documentation | Ms. Word 2016 |

Table 8

|  |  |
| --- | --- |
| Component | Specification |
| Processor | INTEL CORE i3 PROCESSOR |
| Operating System | Windows, Linux, Mac |
| Memory | 4 GB |
| Hard Disk | 500 GB |

**5.2 Hierarchy of Modules**

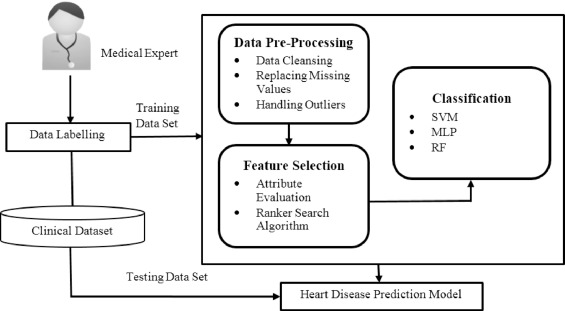


Figure 8: Hierarchy of Modules

**5.3 Detailed Design**

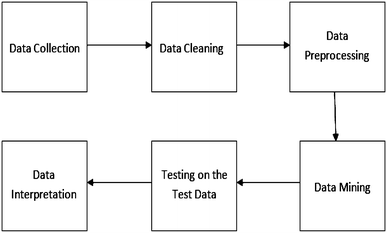


Figure 9: Detailed Design

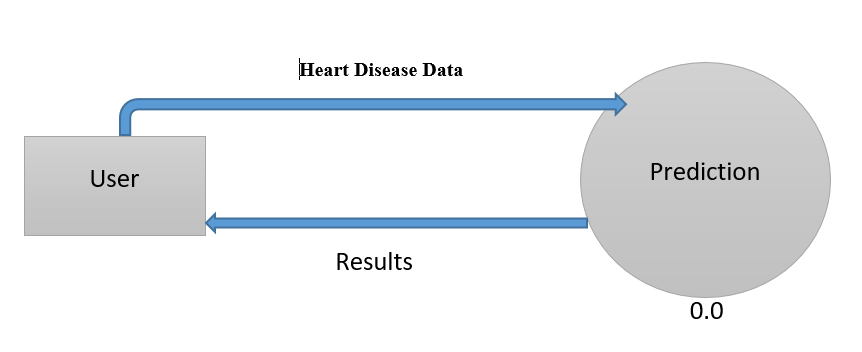


Figure 10: DFD Diagram

**5.4 Test Planning**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test no.** | **Test Case** | **Input** | **Expected result** | **Observed result** |
| 1 | T-MD-1.2 | Testing Set | Found good accuracy | Working |

Table 9: Test Planning

1. **CONCLUSION**

The goal of the project was to compare the algorithms with different performance metrics using machine learning. All data were pre-processed and used for the prediction on the test. Every algorithm performed better in some situation and worse in an other. SVC and Naïve Bayes are the likely models to work best in the dataset used in this study.  
To be able to diagnose the heart disease accurately using machine learning has many  
significances. Different devices can be manufactured which will monitor the heart related activities and diagnose the disease. These devices will prove to be helpful where heart disease experts are not available. With further research machine learning can also be used to diagnose the heart disease before the human experts can do.  
One of the main achievements of this project was that project helped to understand the  
algorithms better. When tested through various situations, the algorithms performed  
differently which helped me to understand the algorithm’s working mechanism. This  
thesis can be first learning step in heart disease diagnosis with machine learning and it  
can be extended further for future research. There are several limitations to this study  
primarily the knowledge base of the author, secondly, the tools used in this study such  
as processing power of the computer and thirdly the time limitation available for the  
study. This type of study requires state-of-art resources and expertise in respective  
domains.

* 1. **PROJECT BENEFITS:**

To be able to diagnose the heart disease accurately using machine learning has many significances. Different devices can be manufactured which will monitor the heart related activities and diagnose the disease. These devices will prove to be helpful where heart disease experts are not available. With further research machine learning can also be used to diagnose the heart disease before the human experts can do.

* Approach for early detection of heart disease by utilizing variety of feature
* Attain the most efficient model of the multiple rule based combination.
  1. **FUTURE SCOPE:**

Further work involves development of the system using the mentioned methodologies and thus training and testing the system. Future work may also involve the development of a tool to predict the risk of disease of a prospective patient. The framework can also be extended for use on other models such as neural networks

1. **References / Bibliography:**

[1] Blake, C.L., Mertz, C.J.: “UCI Machine Learning Databases”, http://mlearn.ics.uci.edu/databases/heart-disease/, 2004.

[2] Chapman, P., Clinton, J., Kerber, R. Khabeza, T., Reinartz, T., Shearer, C., Wirth, R.: “CRISP-DM 1.0: Step by step data mining guide”, SPSS, 1-78, 2000.

[3] Charly, K.: “Data Mining for the Enterprise”, 31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer, 7, 295-304, 1998.

[4] Fayyad, U: “Data Mining and Knowledge Discovery in Databases: Implications fro scientific databases”, Proc. of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington, USA, 2-11, 1997.

[5] Giudici, P.: “Applied Data Mining: Statistical Methods for Business and Industry”, New York: John Wiley, 2003.

[6] Han, J., Kamber, M.: “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers, 2006.

[7] Ho, T. J.: “Data Mining and Data Warehousing”, Prentice Hall, 2005.

[8] Kaur, H., Wasan, S. K.: “Empirical Study on Applications of Data Mining Techniques in Healthcare”, Journal of Computer Science 2(2), 194-200, 2006.