TITLE - Heart Disease Prediction

Group Members -

Mr. Vaibhav Pawar 2127052

Mr. Avdhoot Kumbhar 2127037

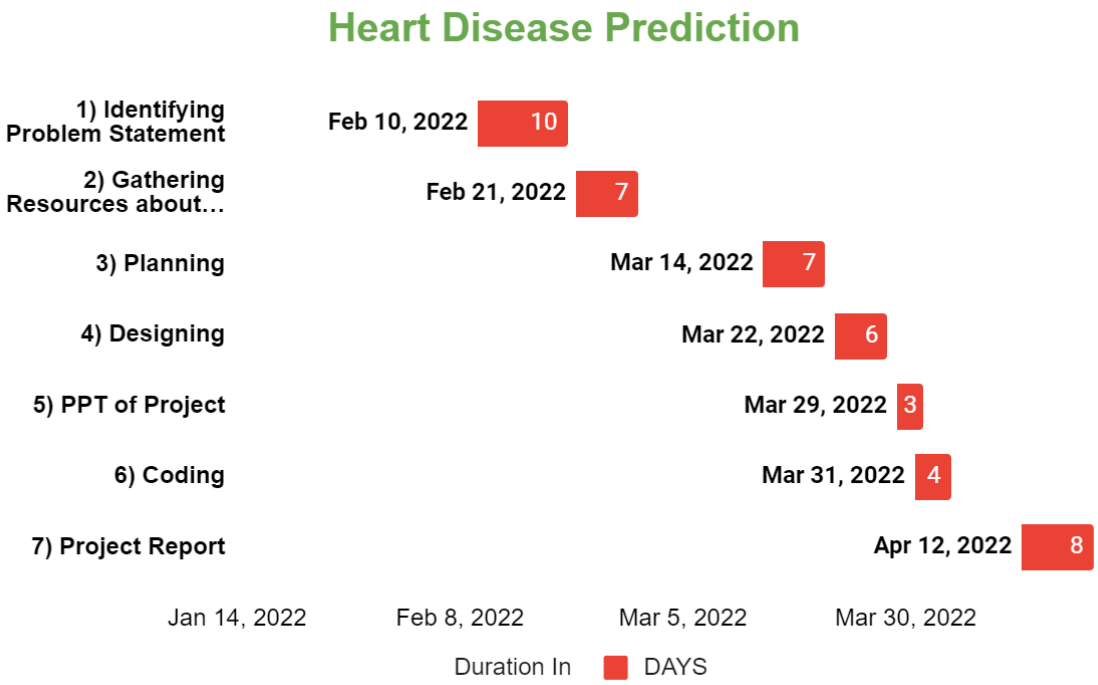
Mr. Yashraj Devrat 2127011

Mr. Shubham Keskar 2127029

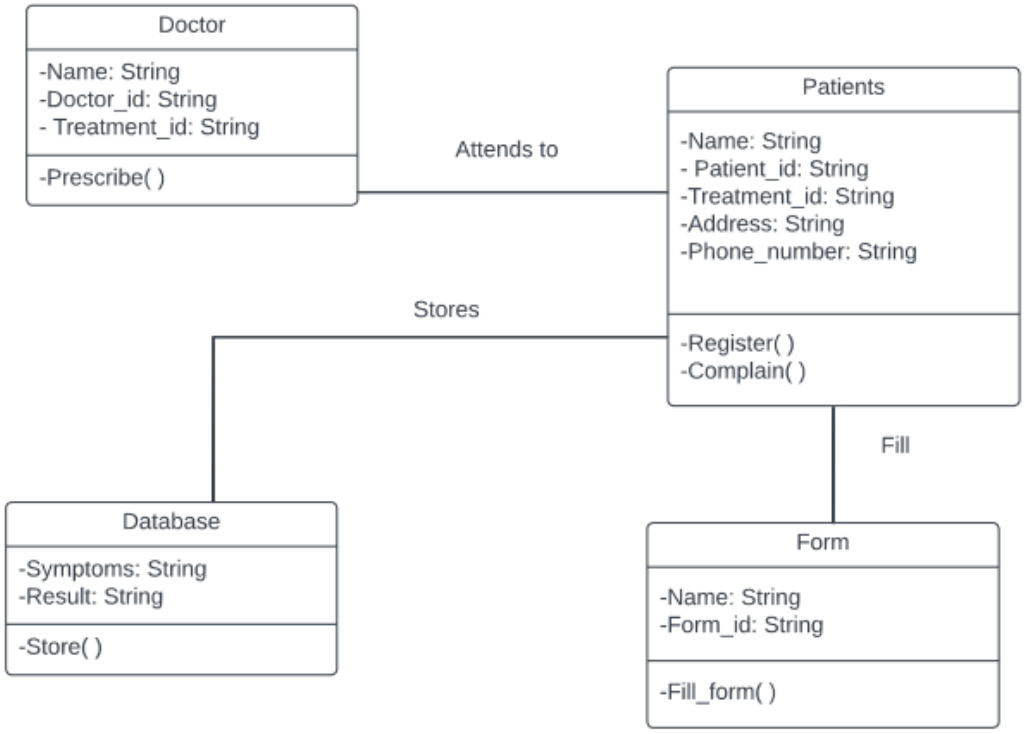
Mr. Tushar Mamadge 2127040

Project Plan –

TASK	START DATE	Duration
1) Identifying Problem Statement	2/10/2022	10
2) Gathering Resources about Project	2/21/2022	7
3) Planning	3/14/2022	7
4) Designing	3/22/2022	6
5) PPT of Project	3/29/2022	3
6) Coding	3/31/2022	4
7) Project Report	4/12/2022	8



Class Diagram –



Implementation Process –

Explanation – The Objective of this project is to create a model that can predict the patient's heart disease Status.Another Objective is to explore the data we have been given and find key insights into Heart Disease that could be helpful for the Medical community going forward. Dataset for this project is taken from Kaggle website https://www.kaggle.com/datasets/priyanka841/heart-disease-prediction-uci .Here we use Logistic Regression model.In this Dataset there are 303 Rows and 14 Columns, means 302 Persons Data out of which 241 are use for Training the Data and 61 use for Testing the Data.

```
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
```

Data Collection and Processing

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

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

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

   ca  thal  target
0   0     1       1
1   0     2       1
2   0     2       1
3   0     2       1
4   0     2       1
```

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

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

```
     slope  ca  thal  target
298      1   0     3       0
299      1   0     3       0
300      1   2     3       0
301      1   1     3       0
302      1   1     2       0
```

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

```
(303, 14)
```

```
# getting some info about the data
heart_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       303 non-null    int64  
 1   sex       303 non-null    int64  
 2   cp        303 non-null    int64  
 3   trestbps  303 non-null    int64  
 4   chol      303 non-null    int64  
 5   fbs       303 non-null    int64  
 6   restecg   303 non-null    int64  
 7   thalach   303 non-null    int64  
 8   exang     303 non-null    int64  
 9   oldpeak   303 non-null    float64
 10  slope     303 non-null    int64  
 11  ca        303 non-null    int64  
 12  thal      303 non-null    int64  
 13  target    303 non-null    int64  
dtypes: float64(1), int64(13)
memory usage: 33.3 KB
```

```
# checking for missing values
heart_data.isnull().sum()
```

```
age         0
sex         0
cp          0
trestbps    0
chol        0
fbs         0
restecg     0
thalach     0
exang       0
oldpeak     0
slope       0
ca          0
thal        0
target      0
dtype: int64
```

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

```
              age         sex          cp    trestbps        chol
fbs  \
count  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
std      9.082101    0.466011    1.032052   17.538143   51.830751
0.356198
min     29.000000    0.000000    0.000000   94.000000  126.000000
0.000000
25%     47.500000    0.000000    0.000000  120.000000  211.000000
0.000000
50%     55.000000    1.000000    1.000000  130.000000  240.000000
0.000000
75%     61.000000    1.000000    2.000000  140.000000  274.500000
0.000000
max     77.000000    1.000000    3.000000  200.000000  564.000000
1.000000

          restecg     thalach       exang     oldpeak       slope
ca  \
count  303.000000  303.000000  303.000000  303.000000  303.000000
303.000000
mean     0.528053  149.646865    0.326733    1.039604    1.399340
0.729373
std      0.525860   22.905161    0.469794    1.161075    0.616226
1.022606
min      0.000000   71.000000    0.000000    0.000000    0.000000
0.000000
25%      0.000000  133.500000    0.000000    0.000000    1.000000
0.000000
50%      1.000000  153.000000    0.000000    0.800000    1.000000
0.000000
75%      1.000000  166.000000    1.000000    1.600000    2.000000
1.000000
max      2.000000  202.000000    1.000000    6.200000    2.000000
4.000000

             thal      target
count  303.000000  303.000000
mean     2.313531    0.544554
std      0.612277    0.498835
min      0.000000    0.000000
25%      2.000000    0.000000
50%      2.000000    1.000000
75%      3.000000    1.000000
max      3.000000    1.000000
```

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

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

Splitting the Features and Target

```
X = heart_data.drop(columns='target', axis=1)
Y = heart_data['target']

print(X)
```

```
     age  sex  cp  trestbps  chol  fbs  restecg  thalach  exang  oldpeak
\
0     63    1   3       145   233    1        0      150      0      2.3
1     37    1   2       130   250    0        1      187      0      3.5
2     41    0   1       130   204    0        0      172      0      1.4
3     56    1   1       120   236    0        1      178      0      0.8
4     57    0   0       120   354    0        1      163      1      0.6
..   ...  ...  ..       ...   ...  ...      ...      ...    ...      ...
298   57    0   0       140   241    0        1      123      1      0.2
299   45    1   3       110   264    0        1      132      0      1.2
300   68    1   0       144   193    1        1      141      0      3.4
301   57    1   0       130   131    0        1      115      1      1.2
302   57    0   1       130   236    0        0      174      0      0.0

     slope  ca  thal
0        0   0     1
1        0   0     2
2        2   0     2
3        2   0     2
4        2   0     2
..     ...  ..   ...
298      1   0     3
299      1   0     3
300      1   2     3
301      1   1     3
302      1   1     2

[303 rows x 13 columns]
```

```
print(Y)
```

```
0      1
1      1
2      1
3      1
4      1
      ..
298    0
299    0
300    0
301    0
302    0
Name: target, Length: 303, dtype: int64
```

Splitting the Data into Training data & Test Data

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

```
print(X.shape, X_train.shape, X_test.shape)
```

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

Model Training

Logistic Regression

```
model = LogisticRegression()
```

```
# training the LogisticRegression model with Training data
model.fit(X_train, Y_train)
```

```
/usr/local/lib/python3.7/dist-
packages/sklearn/linear_model/_logistic.py:818: ConvergenceWarning: lbfgs
failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown
in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  extra_warning_msg=_LOGISTIC_SOLVER_CONVERGENCE_MSG,
```

```
LogisticRegression()
```

Model Evaluation

Accuracy Score

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

```
print('Accuracy on Training data : ', training_data_accuracy)
```

```
Accuracy on Training data :  0.8512396694214877
```

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

```
print('Accuracy on Test data : ', test_data_accuracy)
```

```
Accuracy on Test data :  0.819672131147541
```

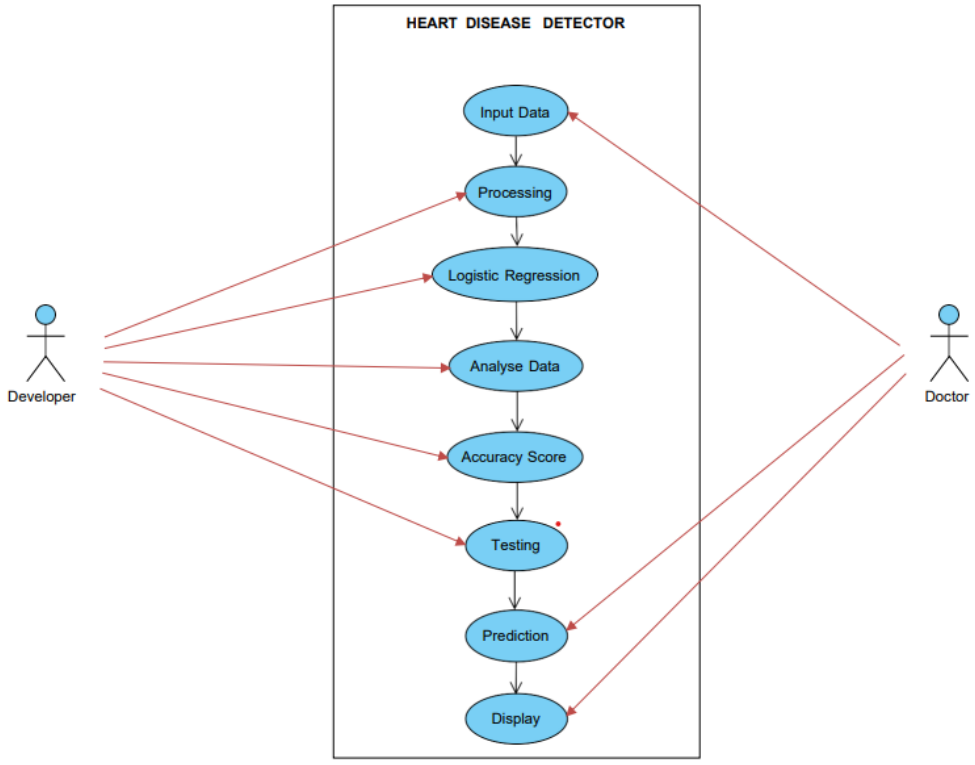
Building a Predictive System

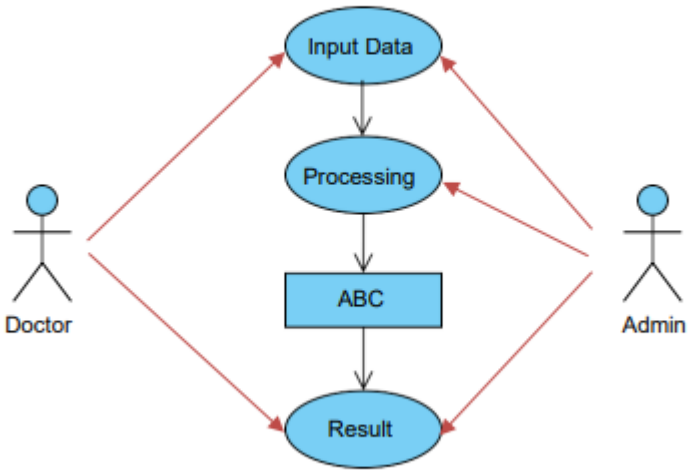
```
input_data = (62,0,0,140,268,0,0,160,0,3.6,0,2,2)

# change the input data to a numpy array
```

```
input_data_as_numpy_array= np.asarray(input_data)

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

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

if (prediction[0]== 0):
  print('The Person does not have a Heart Disease')
else:
  print('The Person has Heart Disease')
```

```
[0]
The Person does not have a Heart Disease

/usr/local/lib/python3.7/dist-packages/sklearn/base.py:451: UserWarning: X
does not have valid feature names, but LogisticRegression was fitted with
feature names
  "X does not have valid feature names, but"
```

Use Case Diagram –





Project Report –

Heart Disease Prediction

A

Project Report

Submitted

by

Mr. Vaibhav Pawar	**2127052**
Mr. Avdhoot Kumbhar	**2127037**
Mr. Yashraj Devrat	**2127011**
Mr. Shubham Keskar	**2127029**
Mr. Tushar Mamadge	**2127040**

In partial fulfillment for the requirement of Project Based Learning-II

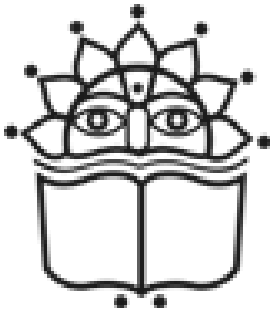
of

Second Year of Artificial Intelligence and Data Science

Under the guidance of

Prof R.V.Panchal

(Assistant Professor)

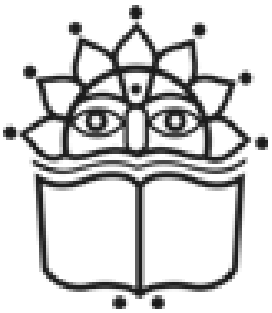


Department of Artificial Intelligence and Data Science

Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering and Technology

Bhigwan Road, Vidyanagari

Baramati-413133



Vidya Pratishthan's
Kamalnayan Bajaj Institute of Engineering and Technology, Baramati
Department of Artificial Intelligence and Data Science

Certificate

This is to certify that following students

Mr. Vaibhav Pawar	**2127052**
Mr. Avdhoot Kumbhar	**2127037**
Mr. Yashraj Devrat	**2127011**
Mr. Shubham Keskar	**2127029**
Mr. Tushar Mamadge	**2127040**

have successfully completed their project work on

Heart Disease Prediction

during the academic year **2021-2022** in the partial fulfillment towards the completion of **Project Based Learning-II** in **Artificial Intelligence and Data Science**

Project Guide
(Name of Guide/Supervisor)

Head, Deptt. of AI & DS
(Digambar Padulkar)

Principal
(**Dr. R. S. Bichkar)**

Internal Examiner **External Examiner**

Acknowledgments

It gives us great pleasure in presenting the project report on ' Heart Disease Prediction'. We would like to take this opportunity to thank our internal supervisor Mr. R. V. Panchal for giving us all the help and supervision that we needed. We are really grateful to them for their kind support. Their valuable suggestions were very helpful. We are also grateful to Mr. Digambar Padulkar, Head of Artificial Intelligence and Data Science Engineering Department, VPKBIET for his indispensable support, suggestions.

Mr. Vaibhav Pawar
Mr. Avdhoot Kumbhar
Mr. Yashraj Devrat
Mr. Shubham Keskar
Mr. Tushar Mamadge

Contents

1

Synopsis

1.1 Project Title

Heart Disease Prediction

1.2 Technical Keyword

Classification,Machine Learning,Heart Disease Prediction,Training and Testing.

1.3 Problem Statement

Heart disease can be managed effectively with a combination of lifestyle changes, medicine and, in some cases, surgery. With the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved. The predicted results can be used to prevent and thus reduce cost for surgical treatment and other expensive. The overall objective of my work will be to predict accurately with few tests and attributes the presence of heart disease. Attributes considered form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with few attributes and faster efficiency the risk of having heart disease. Decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the data set and databases. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs.

1.4 Abstract

In the medical field, the diagnosis of heart disease is the most difficult task. The diagnosis of heart disease is difficult as a decision relied on grouping of large clinical and pathological

data. Due to this complication, the interest increased in a significant amount between the researchers and clinical professionals about the efficient and accurate heart disease prediction. In case of heart disease, the correct diagnosis in early stage is important as time is the very important factor. Heart disease is the principal source of deaths widespread, and the prediction of Heart Disease is significant at an untimely phase. Machine learning in recent years has been the evolving, reliable and supporting tools in medical domain and has provided the greatest support for predicting disease with correct case of training and testing. The main idea behind this work is to study diverse prediction models for the heart disease and selecting important heart disease feature using Random Forests algorithm. Random Forests is the Supervised Machine Learning algorithm which has the high accuracy compared to other Supervised Machine Learning algorithms such as logistic regression etc. By using Random Forests algorithm we are going to predict if a person has heart disease or not.

1.5 Goals and Objective

The Goals and Objective of this project is to create a model that can predict the patient's heart disease status.Another Objective is to explore the data we have been given and find key insights into Heart Disease that could be helpful for the Medical community going forward.

2

Technical Keywords

2.1 Area of Project

Project Consists Of Machine Learning Field.Different Machine Learning Techniques are used to analyze the Heart Disease Dateset.

2.2 Technical Keywords

1]Classification- Different Classification Technique are Applied to Classify the Dataset.

2]Machine Learning - Different Machine Learning Techniques are involved in Project.for example Classification of Heart Disease Dataset With Different Parameters.

3]Heart Disease Prediction - Using Different classifier doing a Heart Disease Prediction.

4]Training And Testing - Training and Testing Technique are used to find Accuracy.

3

Introduction

3.1 Motivation of Project

The main motivation of doing this research is to present a heart disease prediction model for the prediction of occurrence of heart disease. Further, this research work is aimed towards identifying the best classification algorithm for identifying the possibility of heart disease in a patient. This work is justified by performing a comparative study and analysis using classification algorithms namely logistic Regression and Random Forest are used at different levels of evaluations. Although these are commonly used machine learning algorithms, the heart disease prediction is a vital task involving highest possible accuracy. Hence, the algorithms are evaluated at numerous levels and types of evaluation strategies. This will provide researchers and medical practitioners to establish a better understanding and help them identify a solution to identify the best method for predicting the heart diseases.

3.2 Literature Survey

According to Ordonez [1] the heart disease can be predicted with some basic attributes taken from the patient and in their work have introduced a system that includes the characteristics of an individual human being based on totally 13 basic attributes like sex, blood pressure, cholesterol and others to predict the likelihood of a patient getting affected by heart disease. They have added two more attributes i.e. fat and smoking behaviour and extended the research dataset. The data mining classification algorithms such as Decision Tree, Naive Bayes, and Neural Network are utilized to make predictions and the results are analysed on Heart disease database. Yılmaz, [2] have proposed a method that uses least squares support vector machine (LS-SVM) utilizing a binary decision tree for classification of cardiotocogram to find out the patient condition. Duff, et al. [3] have done a research work involving five hundred and thirty-three patients who had suffered from cardiac arrest and they were integrated in the analysis of heart disease probabilities. They performed classical statistical analysis and data mining analysis using mostly Bayesian

networks. Frawley, et al. [4] have performed a work on prediction of survival of Coronary heart disease (CHD) which is a challenging research problem for medical society. They also used 10-fold cross-validation methods to determine the impartial estimate of the three prediction models for performance comparison purposes. Lee, et al. proposed a novel methodology to expand and study the multi-parametric feature along with linear and nonlinear features of Heart Rate Variability diagnosing cardiovascular disease. They have carried out various experiments on linear and non-linear features to estimate several classifiers, e.g., Bayesian classifiers, CMAR, C4.5 and SVM. Based on their experiments, SVM outperformed the other classifiers.

4

Problem Definition and Scope

4.1 Problem Statement

Heart disease can be managed effectively with a combination of lifestyle changes, medicine and, in some cases, surgery. With the right treatment, the symptoms of heart disease can be reduced and the functioning of the heart improved. The predicted results can be used to prevent and thus reduce cost for surgical treatment and other expensive. The overall objective of my work will be to predict accurately with few tests and attributes the presence of heart disease. Attributes considered form the primary basis for tests and give accurate results more or less. Many more input attributes can be taken but our goal is to predict with few attributes and faster efficiency the risk of having heart disease. Decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the data set and databases. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Data mining holds great potential for the healthcare industry to enable health systems to systematically use data and analytics to identify inefficiencies and best practices that improve care and reduce costs.

4.2 Goals and Objective

4.2.1 Statement of 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, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions

4.2.2 Methodology of problem Solving

There are four phases that involve in the spiral model: 1) Planning phase - Phase where the requirement are collected and risk is assessed. This phase where the title of the project has been discussed with project supervisor. From that discussion, Heart Prediction System has been proposed. The requirement and risk was assessed after doing study on existing system and do literature review about another existing research.

2) Risk analysis Phase - Phase where the risk and alternative solution are identified. A prototype are created at the end this phase. If there is any risk during this phase, there will be suggestion about alternate solution.

3) Engineering phase - At this phase, a software are created and testing are done at the end this phase.

4) Evaluation phase - At this phase, the user do evaluation toward the software. It will be done after the system are presented and the user do test whether the system meet with their expectation and requirement or not. If there is any error, user can tell the problem about system

4.2.3 Outcome

1] It is very helpful to medical Community to Predict Accurate results.

2]By using classification models user can analyze the Dataset.

3]Then Software will display overall feature wise user opinion in the form of graphs.

4.2.4 Application

1] In Medical Community.

2] Analysis of Heart Disease.

4.2.5 Constraints

Data Mining techniques does not help to provide effective decision making.

4.2.6 S/W Resources

1]Platform : Operating System: Windows 7 or above,Ubuntu 12 or above

2]IDE: Jupyter Notebook,Google Colab Notebook

3] Programming Language : Python