Prediction On Heart Disease Using Data Mining Algorithms

[1] Shivani Aggarwal , [2] Arvind Kumar, [3] Sonam Kashyap,
[1]Ambedkar Institute of Advanced Communication
Technologies & Research
Geeta colony, New Delhi,India

Abstract : Heart disease is one of the major health problem in this modern era. In a survey by Public Health Foundation Of India in 2016 about 5,681 per 1,00,000 population in India suffered from heart disease. This paper gives a brief description about prediction of heart disease using data mining algorithms. The major reason to use data mining technologies is increasing amount of data in healthcare sector which is " not minded". Data mining is the exploration of hidden and previously unknown pattern, relationship and knowledge which is hard to detect using traditional tools. Data mining techniques help researchers to develop an intelligent heart disease prediction system. In this paper, we use the data mining methods for the risk prediction of heart disease like Naïve Bayes, Decision Tree, Random Forest and Logistic Regression. This paper also gives a comparative study of these algorithms using a web application Jupyter Notebook. The performance of these algorithms are compared according to their accuracies.

***IndexTerms* - Heart disease, Data mining, Naïve Bayes, Decision Tree, Random Forest, Logistic Regression.**

I. INTRODUCTION

Heart Disease Prediction Model supports medical executives in predicting heart disease chances established on the clinical data of patients. Heart is the most important part of human body. If heart is not able to work, a number of death occurs that's why heart disease is one of the cause of increase in number of death. According to world health organization report, one in three people have the problem of high blood pressure which is one of the symptom of heart disease. There are a number of conditions that affect the heart and these are coronary artery disease, cardiovascular disease, angina pectoris, cardiac arrest. In today's time of machines, clinical test results are given under the insight of experienced specialists that extract data from a expansive databases which leads to an expensive data evaluation. Data mining overcome this problem by making data extraction easier and less costly as compared to other methods. The main aim of data mining is to find out the premature unrevealed patterns and modes in databases which are provided and then use that information to built predictive model[7]. Data mining techniques are classified as classification, prediction, association rule, neural network. Data mining techniques used to analyze the group of data using different views to derive knowledgeable information. The knowledge which is discovered by the medical researchers are used to improve the quality of services provided by healthcare executives. Quality services involves diagnosing patients precisely and use the treatments that are effective for patients[8].. It is essential for medical people to find out the algorithm which are best fit and have precise accuracy, less cost, gives speed in the evaluation and has good memory utilization for classification and prediction of heart disease[2].

II. DATA MINING ALGORITHMS

Data Mining is a method of insignificantly extract implicit, previously unknown and potential useful information about the data[1]. Data mining techniques used in many research works like mathematics, cybernetics, genetics, marketing and predictions. Diagnosis is a complex task and needed a highly experience doctors and it also demand to executed accurately. Data mining prediction model help doctors to predict a large amount of patient data in few minutes instead manually it takes a long time to predict. By using some data mining algorithms we can predict heart disease patients record. In this paper we discuss about data mining classification algorithms and these algorithms are representated in fig.1.

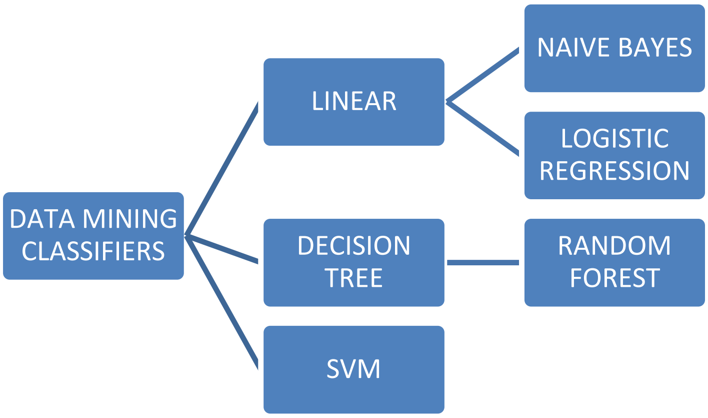


Fig.1. Representation of data mining classifier

A. Naïve Bayes:

Naïve bayes is one of the classification method of data mining which is construct on bayes theorem with the expectation that predictors are unrelated to each other.Naïve Bayes classifiers are trained in a regular way in supervised learning technique. Naïve bayes model is beneficial for extremely big datasets.

Bayes Rule:

$$P(O|N) = (P(N|O).P(O))/P(N)$$

Where,
$P(O|N)$ = Posterior Probability
$P(N|O)$ = Likelihood
$P(O)$ = Class Prior Probability
$P(N)$ = Predictor Prior Probability

Naïve bayes defines determinant and minor relationship between two randomly generated events. It helps to assess posterior possibilities that holds observations. Naive Bayes compute the probability which is showing the coorect or incorrect result in diagnosis.

B. Decision Tree*:*

Decision Trees (DTs) are a supervised learning method which is non parametric in nature used for data mining classification techniques[3]. Decision trees are appear as trees. Decision tree hass tree structured classifier and are of two types of nodes: Decision nodes and Leaf nodes. Decision node decide which direction we have to go for taking a decision. Leaf node indicates the claasification of the example. Decision tree can be used both for classification and regression. The advantages of decision tree are easy understanding of datasets and can handle both numerical and unambiguous data. The limitation of decision tree is it works only for small amount of training data if the data get complex accuracy of decision tree may get affected.

C. Random Forest:

Random forest is a grouped classifier that consist of many decision trees. It is one of the most accurate algorithm in machine learning algorithms. It gave good performance on a number of problems because it is non sensitive to noise in the dataset and it is not affected by overfitting. Random forest is built by merging the prediction of several trees, each of which are trained in isolation.

D. Logistic Regression:

Logistic regression a term of regression which is utilize to estimate binary or multi- category variable and also the response variable is separate. It's employed to classify low dimensional knowledge having non-linear boundaries. It offer changes within the share of variables and provides the rank to variable individually as per their importance. The main thought of supplying regression is to work out the results of every variable properly.

III. LITERATURE SURVEY

Singh Navdeep, Jindal Sonika (2018)[1] style a perceptive model for heart disease acknowledgment using data mining strategies that are fit for enhancing the consistency of heart infections conclusion In this paper the researches used data sets of three hundred three records and fourteen elements that are composed from the net dataset depository of Archive.ics.edu/ml/datasetFor the execution of the project, the platform used is Python 3.6. They give a model that use two different classification data mining algorithms i.e. genetic rule and Naive Bayes.

B.Venkatalakshmi, M.V Shivsankar (2014) [3] examines that after conducting many experiments on the same dataset they conclude that naïve bayes algorithm gives better performance than decision tree. This paper also infers that genetic algorithm can lessen the size of data to give the minimal subgroup of elements that are used for heart disease prediction in future.

Nidhi, B and Kiran, J (2012) [7] inspect data mining classifiers and explained them in their work that are appear in recent years for predicting cardiovascular disease designation. This paper concludes that decision tree with the assistance of genetic algorithm program shows better accuracy and performance and gives set choices. In their analysis, neural network gave the highest accuracy of 100%. The target of their effort supplies a brief study of different data processing methods.

Aqueel Ahmed, Shaikh Abdul Hannan, (September 2012) [8] defines various classification algorithm and gives their performances. In this research algorithms like decision tree and SVM shows more accuracy than any other algorithms. They suggest may different measures are used to predict heart disease in patients that can help in future for medical participators in finding more accurate data.

.

G.Subbalakshmi, Mr. K. Ramesh, Mr. M. Chinna Rao, (2011) [10] describes decision support system for heart disease prediction and to define it use naïve bayes classifier. They extract information using naïve bayes technique from a traditional data depository. In this complicatd and intercoonected queries can also be answered using naïve bayes algorithm.

IV. METHODOLOGY

In this paper, we have to analysis various data mining techniques which is very helpful for providing accurate data for heart disease diagnosis. For analyzing the work in the field of data mining and heart disease prediction the main method they were used by inspecting information from many publications, reviews and journals.

For implementation of our project, we used different data mining algorithms and the platform used is Python.

Data set:

In this research paper, we take three hundred three records and forteen elements collected from the net data depository of https://www.kaggle.com/. Kaggle is the world largest community of data scientist and offering a public data platform and a short form of AI education. The data parameters are listed below in Table 1:

Sr. No.	Attributes Name	Description
1	(age)	Patient Age
2	(sex)	Male/Female
3	(cp)	Chest pain type
4	(trestbps)	Resting blood pressure (in mm Hg on admission to the hospital)
5	(chol)	Serum cholestoral (mg/dl)
6	(fbs)	Fasting blood sugar
7	(restecg)	Resting ECG results
8	(thalach)	Maximum heart rate achieved
9	(expand)	Exercise induced angina
10	(old peak)	ST depression included by exercise relative to rest
11	(slope)	The slope of the peak exercise ST segment
12	(ca)	Number of major vessels (0-3) colored by flouroscopy
13	(thal)	3 = normal, 6 = fixed defect, 7 = reversible defect.
14	(target)	Outcome (1= test is positive, 0= test is negative)

Table 1: Data Parameters

V. TOOL USED

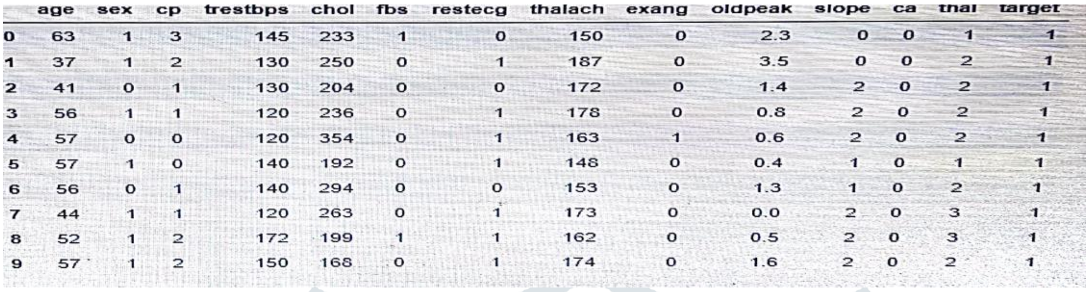
Now a days the most efficient application is Jupyter Notebook for heart disease diagnosis. Jupyter Notebook allows creating and sharing documents and it is open source web application in nature. There uses include data transformation, machine learning. Jupyter Notebook also improve web browser technologies. Jupyter runs code in many programming languages but Python is essential for installing Jupyter Notebook.

It requires 32 or 64 bit computer, 32MB available, Linux, OS X or Windows and also requires 300 MB to download Anaconda plus another 300 to install it. Anaconda is freemium open sorce distribution of Python programming language for large scale data processing and predictive analysis.

VI. EXPERIMENTAL WORK

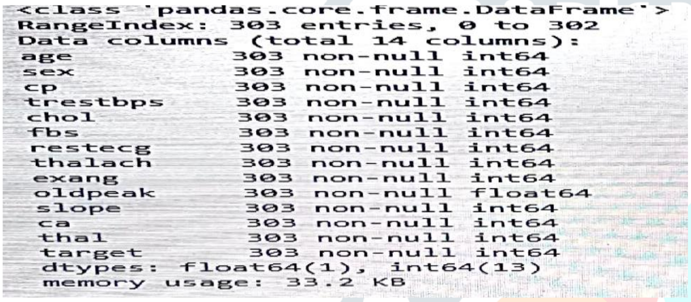
A. The Data:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
% matplotlib inline
import math
df = pd.read_csv(‘heart.csv’)
df.head(10)
```



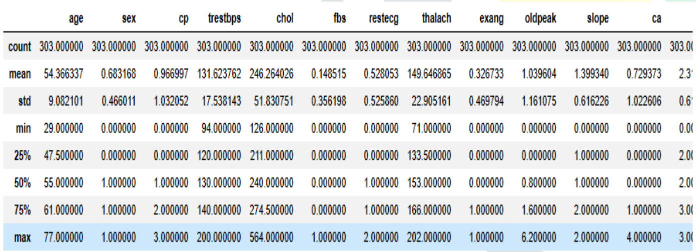
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

df.info()



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 303 entries, 0 to 302
Data columns (total 14 columns):
age         303 non-null int64
sex         303 non-null int64
cp          303 non-null int64
trestbps    303 non-null int64
chol        303 non-null int64
fbs         303 non-null int64
restecg     303 non-null int64
thalach     303 non-null int64
exang       303 non-null int64
oldpeak     303 non-null float64
slope       303 non-null int64
ca          303 non-null int64
thal        303 non-null int64
target      303 non-null int64
dtypes: float64(1), int64(13)
memory usage: 33.2 KB
```

df.describe()



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	
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	303.0
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	2.3
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	0.6
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	0.0
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	2.0
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	2.0
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	3.0
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	3.0

```
x = df.drop(‘target’, axis = 1)
y = df[‘target’]
X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.2,random_state=0)
```

B. Prediction of heart disease using machine learning algorithm :

B.1.Naive Bayes:

Main Source Code:

```
from sklearn.naive_bayes import GaussianNB
nb = GaussianNB()
nb.fit(X_train, y_train)
print(“training score”, nb.score(X_train, y_train))
y_pred_nb = nb.predict(X_test)
print(“testing score’, nb.score(X_test, y_test))
```

Output:

```
Training score = 0.86
Testing score = 0.85
```

Attributes Importance In Naïve Bayes:

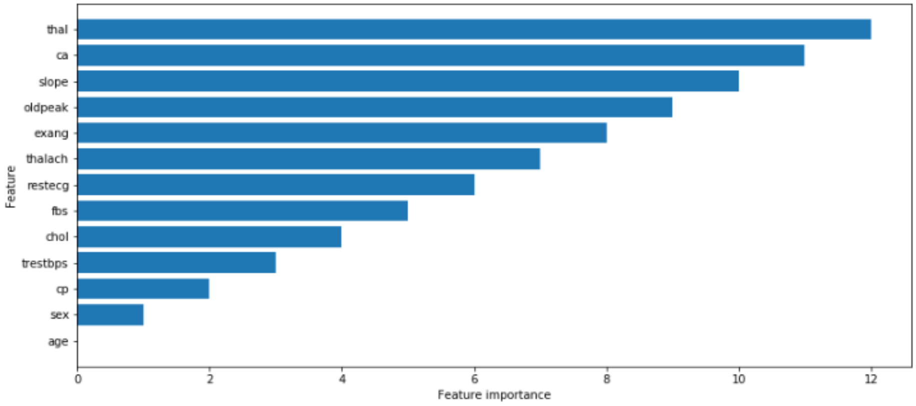


Fig 1. Attributes Importance

B.2. Decision Tree:

Main Source Code:

```
from sklearn .tree import DecisionTreeClassifier
Tree = DecisionTreeClassifier(random_state = 0)
Tree.fit (X_train, y_train)
print(“training score”, Tree.score(X_train, y_train))
y_pred = Tree.predict(X_test)
print(“testing score”, Tree.score(X_test, y_test))
```

Output:

```
Training score = 1.0
Testing score = 0.78
```

Attribute Importance In Decision Tree:

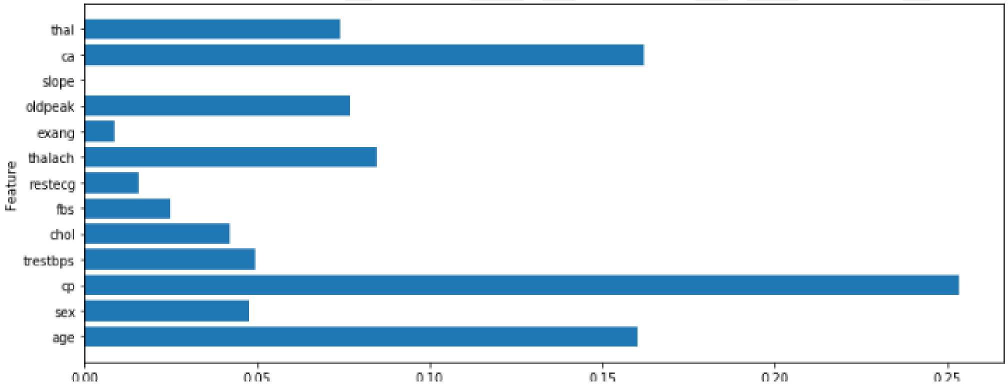


Fig 2. Attributes Importance

B.3. Random Forest:

Main Source Code:

```
from sklearn.ensemble_import RandomForestClassifier
random_forest =
RandomForestClassifier(n_estimators=100)
random_forest.fit(X_train, y_train)
print(“training score”, random_forest.score(X_train, y_train))
```

```
y_pred = random_forest.predict(X_test)
print("testing score", random_forest.score(X_test, y_test))
```

Output:

```
  Testing score = 1.0
 Training score = 0.868
```

Attribute Importance In Random Forest:

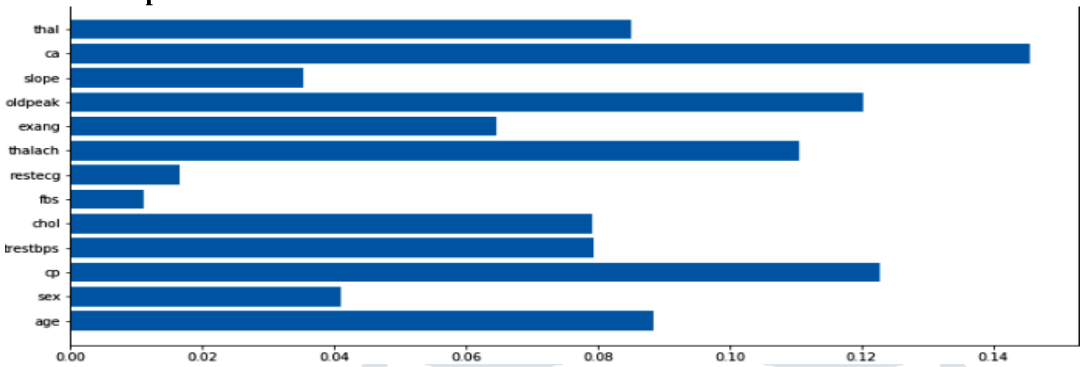


Fig.3. Attributes Importance

B.4. Logistic Regression:

Main Source Code:

```
from skelearn.linear_model  import LogisticRegression
logreg = LogisticRegression()
logreg.fit(X_train, y_train)
print("Training score,",logreg.score(X_train, y_train))
y_pred _logreg = logreg.predict(X_test)
print("Testing score", logreg.score(X_test, y_test))
```

Output:

```
Testing score = 0.867
Training score = 0.852
```

Attribute Importance In Logistic Regression:

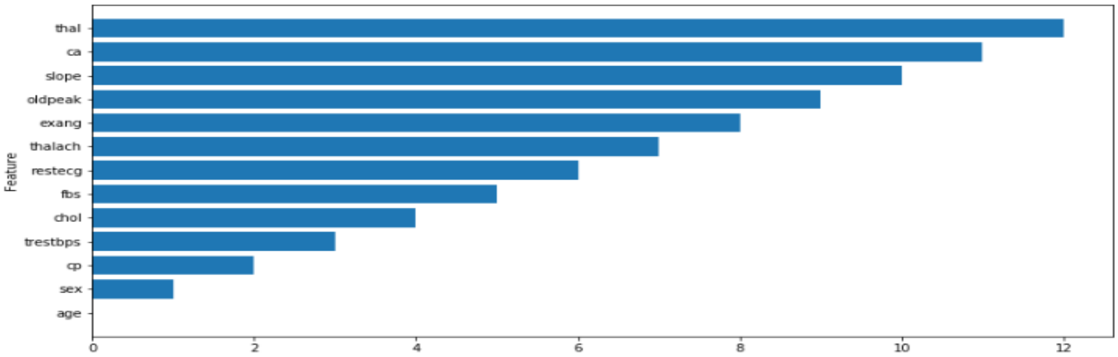


Fig.4. Attributes Importance

VII. COMPARATIVE ANALYSIS

In our experimental work, we use four machine learning algorithms which are naïve bayes, decision tree, random forest and logistic regression in which we check the accuracy of data heart disease patients data. The comparison of these algorithms accuracy are shown in Table 2 and graphical representation are shown in Fig 5.

S. No	Algorithm Used	Training Accuracy (%)	Testing Accuracy (%)

1.	Naïve Bayes	86%	85%
2.	Decision Tree	100%	78%
3.	Random Forest	100%	86.8%
4.	Logistic Regression	86.7%	85.2%

Table 2 . Comparitive analysis of machine learning algorithms

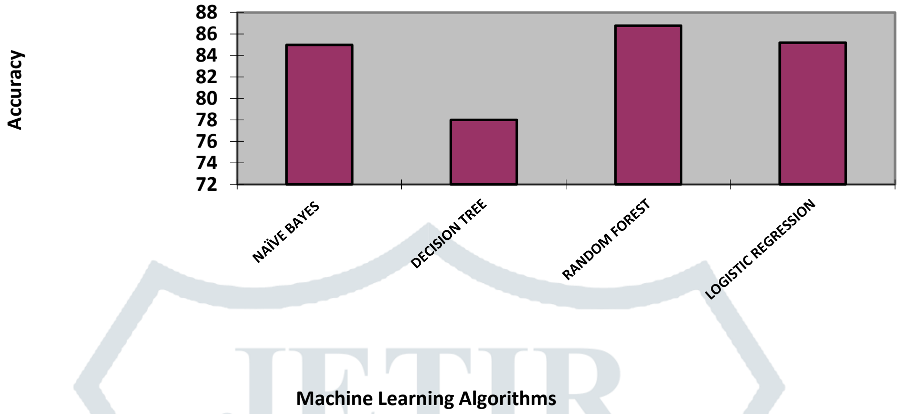


Fig.5. Performance level of machine learning algorithms

VIII. FUTURE WORK AND CONCLUSION

In this paper, different machine learning algorithms are conducted to find best accuracy given algorithm for predicting the patients of heart disease. The analysis at jupyter book shows that Random Forest with 14 attributes has shown the highest accuracy on testing data i.e., 86.8% so far. On the other hand, it also shows that Naïve Bayes has also a good accuracy as compare to existing researchers who find accuracy of other techniques. As future work, we have to considering more data for expanding this study and also useful for improving the accuracy of decision tree.

References

[1] Singh Navdeep, Jindal Sonika:" Heart disease prediction system using hybrid technique of data mining algorithms", International Journal of Advanced Research,Ideas and Innovations in Technology, Vol. 4, Issue 2, 2018.

[2] G. Purusothaman and P. Krishnakumari;" A Survey of Data Mining Techniques on Risk Prediction: Heart Disease", Indian Journal of Science and Technology, Vol.8(12), ISSN 0974-5645, June 2015.

[3] B.Venkatalkshmi, M.V Shivshankar; " Heart Disease Diagnosis Using Predictive Data mining",International Journal of Innovative Research in Science,Engineering and Technology, Vol.3, Issue 3, March 2014.

[4] Abhishek Taneja; "Heart Disease Prediction System Using Data Mining Technique", Oriental Journal Of Computer Science And Technology ,Vol.6(4),pp:457-466, December 2013.

[5] R.Chitra and V.Seenivasagam; "Review of Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Technique", ICTACT Journal On Soft Computing,Vol.3,Issue.4, July,2013.

[6] Vikas Chaurasia, et al; "Early Prediction of Heart Diseases Using Data Mining Techniques", Carib.j.SciTech,Vol.1,pp:208-217,2013.

[7] Nidhi Bhatia, Kiran Jyoti; "An Analysis of Heart Disease Prediction using Different Data Mining Techniques",

[8] Aqueel Ahmed, Shaikh Abdul Hannan; " Data Mining Techniques to Find Out Heart Diseases:An Overview", International Journal of Innovative Technology and Exploring Engineering, Vol.1,Issue 4, September 2012.

[9] Mai Shouman, Tim Turner, Rob Stocker; "Using Data Mining Techniques in Heart Disease Diagnosis and Treatment", IEEE, Vol.12, 2012.

[10] G.Subbalakshmi et al; " Decision Support in Heart Disease Prediction System using Naïve Bayes", Indian Journal of Computer Science and Engineering, Vol.2,Issue 2, May 2011.

[11] Jyoti Soni, Ujma Ansari; " Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications, Vol.17,Issue 8, March 2011.

[12] Sellappan Palaniappan, Rafiah Awang; "Intelligent Heart Disease Prediction System Using Data Mining Techniques", IEEE, Vol.8,2008.

Book References

[13] Jiawei Han and Micheline Kamber – Data Mining – Concept and Techniques; Second Edition; Elsevier Inc, 2006.