ISSN No.: 2454- 2024 (online)

International Journal of Technical Research & Science

EVALUATE AND PROPOSE CLASSIFICATION TECHNIQUE FOR HEART DISEASE PREDICTION IN DATA MINING

Ria Bhatt[1], Jitendra Singh Chouhan[2]
E-Mail Id: jitendrasinghchauhan1984@gmail.com

Research Scholar[1], Associate Professor[2], Aravali Institute of Technical Studies, Udaipur (Rajasthan), India

Abstract-Data mining is a method in which the valuable data is mined from the rough data. The futuristic outcomes are forecasted using recent information in the prediction analysis. This research work deals with the prediction of the heart disease. There are several steps that are included in the heart disease prediction. The pre-processing, feature extraction and classification are some of these steps. The random forest and logistic regression based the hybrid scheme is introduced. The features are abstracted using RF(Random Forest). The implementation of LR (Logistic Regression) is done for classification. The analysis of performance of the recommended model for acquiring accuracy, precision and recall is completed in this research. The accuracy has obtained in predicting the heart disease from this model is evaluated 95%.

Keywords: Cardio-Vascular Disease, Data Mining, Random Forest, K-Means, Random Forest Classifier.

1. INTRODUCTION

Abrupt increase in data and databases generates essential requirement for novel methods and schemes. These tools need to be able to logically and mechanically convert the processed data into meaningful knowledge. Data mining has one more synonym called KDD. This refers to a process of important retrieval of implied, formerly unidentified and possibly meaningful knowledge from gathered data [1]. The KDD (knowledge discovery in databases) process can be applied to extract motivating info, uniformities or high-quality knowledge from the collected datasets.

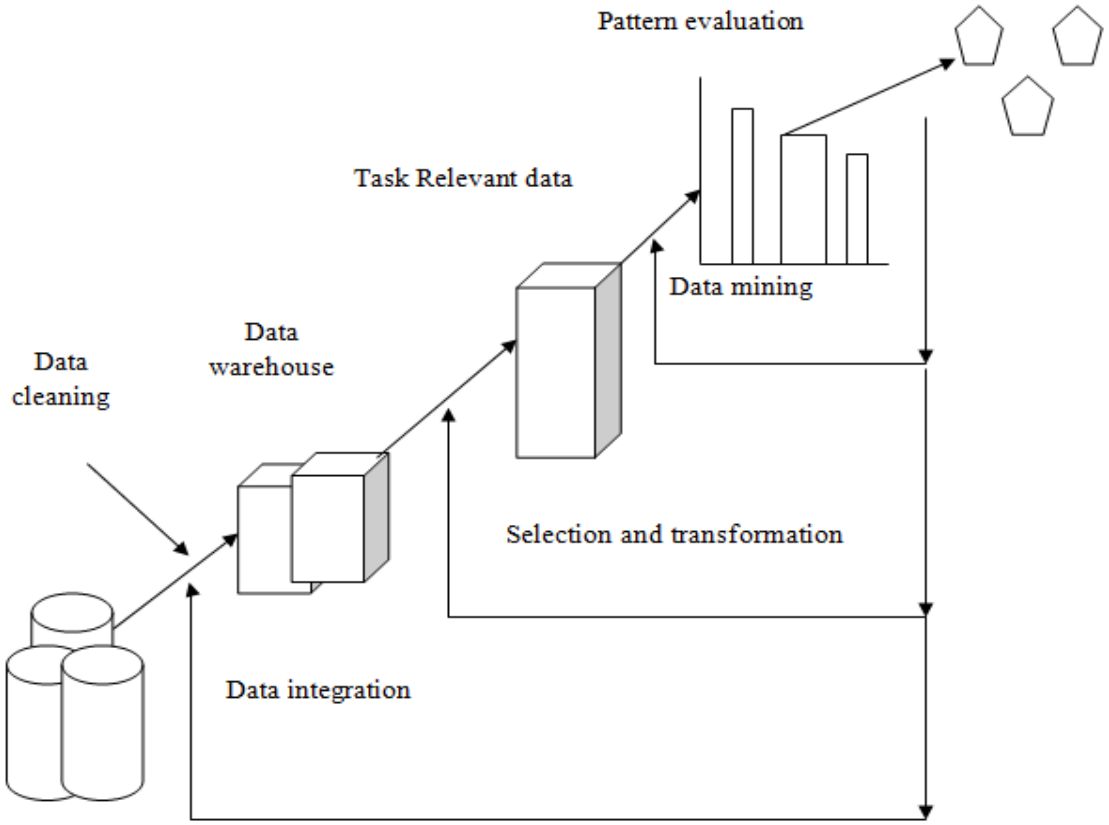

Fig. 1.1 Knowledge Discoveries in Database (KDD)

Various task of KDD. These steps range from collecting the rough data to some sort of novel knowledge. Following are the steps included in KDD:

1.2 Data Cleaning

In this step, noisy and redundant data is eradicated from the gathered data [2]. In this step we also check about outliers if any.

1.3 Data Integration

The second step of data integration generally combines cleaned at a common place. The combined data is generally having homogenous nature.

1.4 Data Selection

Data appropriate for analysis is determined and retrieved from the collected data in this task.

1.5 Data Transformation

This step is also termed as data consolidation. In this step, the conversion of chosen data is carried out into different formats suitable for mining process.

1.6 Data mining

This is a very important step. This step implements efficient methods so that the highly meaningful patterns can be extracted.

1.7 Heart Disease

Heart disease defines any illness related to heart. The problems using the blood vessels, cardiovascular system and the heart are defined as the cardiovascular disease [6]. On the other hand, the problems and deformities in the heart itself are known as heart disorder. The Congenital heart disease, Arrhythmia, CAD, Myocardial infarction, Heart failure, Hypertrophic cardiomyopathy (this is related to heart muscle a type of chronic disease.), Pulmonary stenosis (narrowing or restriction of a blood vessel) etc. are the major types of heart diseases. The Arrhythmia is a kind of the heart rhythm abnormality.

1.4 Prediction of Heart Diseases Using Data Mining

The performance of data mining differs from technique to technique being adopted and chosen attributes. In general, the clinical databanks in the medical domain are useless and unpredictable. Therefore, there is the need of prior and appropriate preparations for implementing data mining approaches.

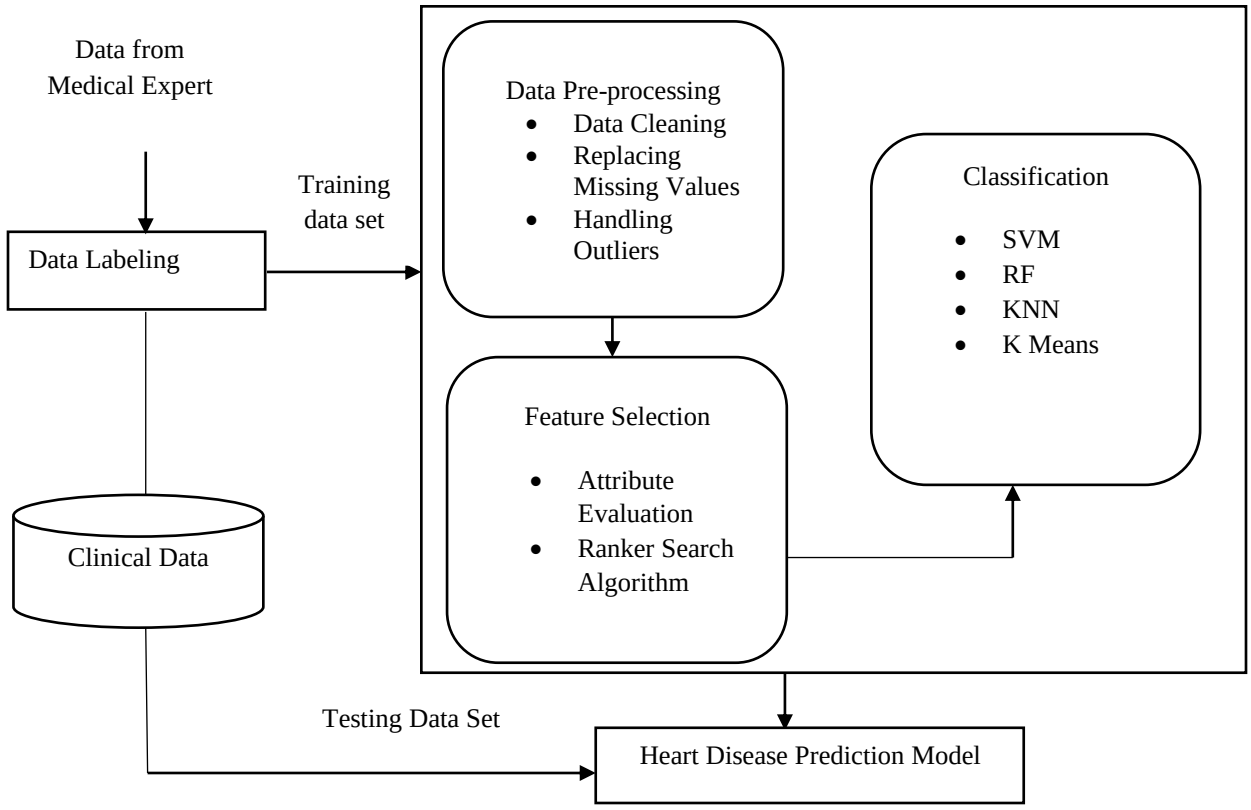

Fig. 1.2 Prediction Model for Heart Disease

1.5 Classification Algorithm

1.5.1 SVM (Support Vector Machine)

It as a training algorithm is employed to learn classification and regression rules from data. This algorithm follows the concept of statistical learning. This algorithm provides solution of the major issues indirectly without proving solution of more complicated issues Usually, this algorithm can be applied in two ways. The first way is the use of statistical programming and, the second way is use kernel functions. Whenever the training data is linearly separable, then a pair (w denotes weight, b denotes bias) exists such that

$$W^T X_i + b \geq 1 for\ all\ X_I \in P, \text{ and}$$
$$W^T X_i + b \leq -1\ for\ all\ X_I \in P$$

The prediction rule is represented by:

$$f = sign(W.X) + b)$$

1.5.2 Random Forest (RF)

The RF technique is simple, fast and adaptable machine learning algorithm. This algorithm combines several tree predictors. This algorithm is able to handle different data types like numerical, binary, and nominal. In this algorithm, a number of trees are generated and these trees are combined together so that the appropriate and accurate result can be generated. This algorithm can resolve both regression and classification issues. Classification is the most important function in machine learning [8].

1.5.3 K- Means Clustering

Clustering is a process used for grouping the similar objects. The grouping of objects is based on their features. It is an unsupervised learning method. The ordinary grouping of instances is derived for the unlabelled data. These clusters are homogeneous in nature. The objects of one cluster are different from another cluster in terms of features i.e every cluster is heterogeneous with other cluster. The intra-clustering is high amid the objects and the inter-clustering similarity is assessed low amidst the clusters in the clustering. Cluster analysis approaches include partitioning clustering, hierarchal clustering etc. K-means clustering is simpler and more user-friendly. Manhattan distance d_M between the tuples X_1 and $X_2$is also referred as 1st order Minkowski distance. This distance is represented by:

$$d_M(X_1, X_2) = \sum_{i=1}^{n} |x_{1i} - x_{2i}|$$

The Euclidean distance d_E between two tuples X_1 and X_2 is measured as:

$$d_E(X_1, X_2) = \sum_{i=1}^{n} (x_{1i} - x_{2i})^2$$

2. LITERATURE REVIEW

Emphasized on the detection of heart disease in which earlier data and information had considered [1-5]. For this purpose, the Navies Bayesian was utilized to develop the SHDP that was capable for risk factors prediction related to heart disease.The data mining techniques were able and proved as remedy in this situation[6-8]. Thus, the deployment of data mining techniques had done for this.

This paper proposed to provide details related to different approaches of info retrieval through data mining techniques. These techniques were carried to predict health related disorders in research. The medical data sets were employed for the analysis of various algorithms. All these algorithms had been analyzed and applied here. The random forest was implemented on Apache Spark in this prediction solution[10-11]. It was demonstrated that the accuracy obtained from this approach was evaluated 98%. The comparison among Naïve-Bayes classifier was also illustrated in this paper in which the random forest approach had performed efficiently in terms of considerable margin as compare to earlier used approach.

3. RESEARCH METHODOLOGY

Following are the various phases of heart disease prediction:

3.1 Data Acquisition

The data is collected from various clinical organizations to perform experiments.

3.2 Data Preprocessing

For applying machine learning techniques such that completeness can be introduced, and a meaningful analysis can be achieved on the data, the data preprocessing is performed.

3.3 Feature Selection

This step makes use of a subset comprising extremely unique features for diagnosing heart diseases. These selective features relate to existing class of features. In the proposed method, the random forest model is applied for the feature selection.

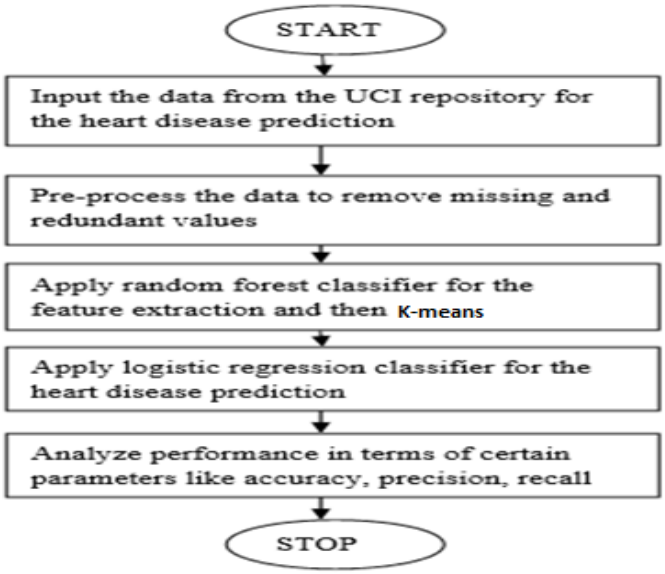

Figure 4.1 Proposed Methodologies

4. PROPOSED ALGORITHM

'T' represents Training Dataset which used as input
Algorithm:

1. Read the Training Dataset T;
2. Extract features of the Training dataset T;
3. Apply cross validation for the data division;
4. Divide Training set into training and testing data;
5. Apply K-mean algorithm
    5.1. Read the training dataset T
    5.2. Calculate the mean and standard deviation of predictor variable in each class
    5.3. Repeat
        5.3.1 Calculate the probability of f(i) using gauss density equation in each class
        5.3.2. Until the probability of all predictors variable (f1,f2 upto fn) has been calculated
6. Calculate the likelihood for each class
7. Get the greatest likelihood
8. Apply Random Forest Classifier
    8.1. Read the training dataset T
    8.2. for each slave agent do
        8.2.1. for Ij=1 to m do
            If D(SVj<-SVi)<hyperplane
                SVj<-SVi
                Update SVj
            End if
        End for
    End for

4.1 Performance Metrics

4.1.1 Confusion Matrix

It is used to find accuracy and correctness of classification model. It is used where the output is of two or more categories.

Table-4.1 Confusion Metrics

		Prdicted Values	
		1	0
Actual Values	1	True (+)	False(-)
	0	False(+)	True(-)

4.1.2 True Positive (TP)

In this actual value and predicted value both are true. This means that the outcome is true as desired.

4.1.3 False Negative (FN)

In this case actual value is true but the predicted value is false. Outcome is not as desired.

4.1.4 False Positive (FP)

In this case actual value is false but the predicated value is true.

4.1.5 True Negative (TN)

In this case actual and the predicted value both are false.

5. RESULT

This work uses Cleveland dataset which is most commonly used for predicting heart diseases. This dataset has 14 attributes. In this research work, the implementation and comparison of several models is performed for predicting the heart disease. The DT, Multilayer perception, NB, Ensemble classification method in which random forest, naïve bayes models are combined, proposed models are compared with regard to certain performance parameters.

Table-5.1 Accuracy Analysis

Models	Accuracy
Decision Tree	75.41 percent
Naïve Bayes	83.61 percent
Multilayer perceptron	83.61 percent
Ensemble Method	85.25 percent
Proposed Method	95.08 percent

The figure 5.2 illustrates that a variety of models including DT, NB, multilayer perceptron, ensemble and proposed models are compared in respect to accuracy. The result which derived from the analysis is that proposed model achieves the highest accuracy that of almost 95%. Thus performing better than other classifiaction algorithms for predicting cardiovascular disorder.

Table 5.2: Precision Analysis

Models	**Precision**
Decision Tree	75 percent
Naïve Bayes	84 percent
Multilayer perceptron	85 percent
Ensemble Method	86 percent
Proposed Method	95 percent

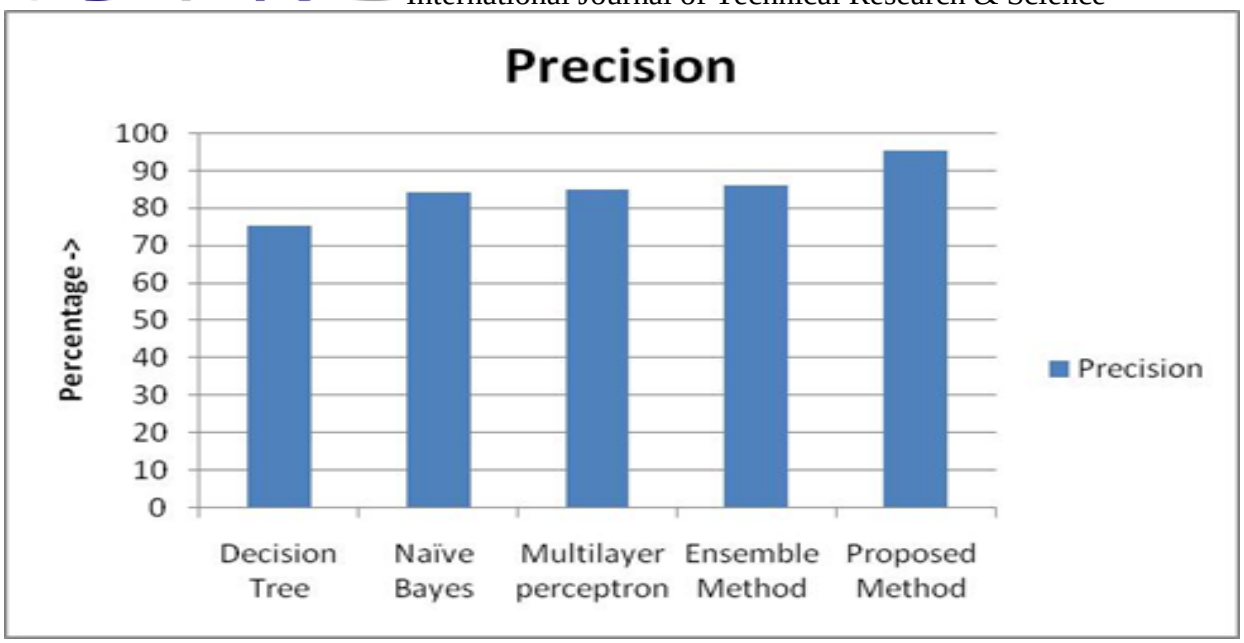

Fig. 5.3 Precision analysis

As shown in fig. 5.3 the various models of including DT, NB, MLP, ensemble and proposed models are compared with respect to precision. From the result of analysis it is determined that proposed model has highest precision rate almost 95%, thus performing better than other classification algorithm for cardiovascular disorders.

Table 5.3: Recall Analysis

Models	Precision
Decision Tree	75 percent
Naïve Bayes	84 percent
Multilayer perceptron	84 percent
Ensemble Method	85 percent
Proposed Method	95 percent

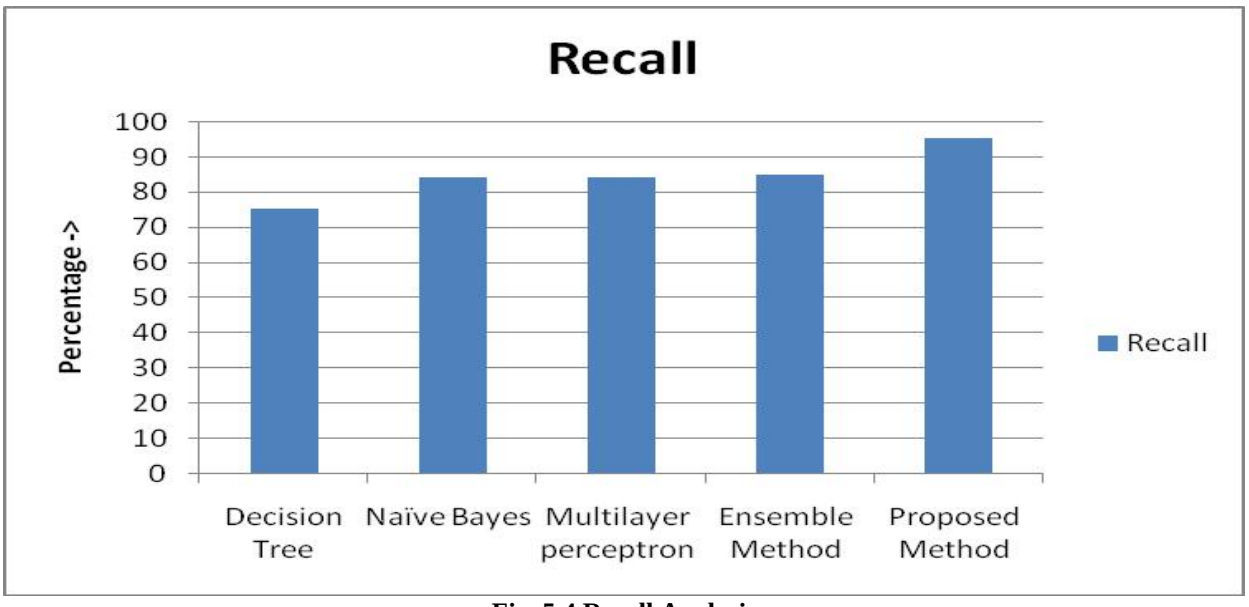

Fig. 5.4 Recall Analysis

As shown in figure 5.4, the various models like DT, NB, multilayer perceptron, and ensemble are compared with proposed model in respect of recall. From the analysis result drawn is that the prediction of cardiovascular disorder from the proposed model is approx 95%.

CONCLUSION

The term heart disease refers to a heart related disorder. The problems using the blood vessels, circulatory system and the heart are defined as the cardiovascular disease [9]. On the other hand, the problems and deformities in the heart itself are known as cardiovascular disease. One of the major reasons of deaths death in the UK, US, Canada, and Australia is heart disease as stated by the CDC. The heart disease may lead to every one out of deaths in Union States. The heart disease consists of numerous kinds of diseases due to which various parts of the organ are infected. To conclude, it is analyzed in this work that heart disease prediction is very challenging as the large number of features included in it. The various models are tested for the heart disease prediction like decision tree, naïve bayes, multilayer perceptron, ensemble classifier. The novel model in which the random forest and logistic regression are integrated is introduced to predict heart disorders. The extraction of features is generated using RF and the logistic regression is carried out to perform the classification. The recall, accuracy and precision obtained from the proposed model is computed as 95 percent.

REFERENCES

[1] K.Gomathi, Dr.Shanmugapriyaa, “Heart Disease Prediction Using Data Mining Classification”, 2016, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 4 Issue II.

[2] H. Benjamin Fredrick David and S. Antony Belcy, “Heart Disease Prediction using Data Mining Techniques”, 2018, ICTACT Journal on Soft Computing, Volume: 09, Issue: 01.

[3] G. Purusothaman and P. Krishnakumari, “A Survey of Data Mining Techniques on Risk Prediction: Heart Disease”, 2015, Indian Journal of Science and Technology, vol. 8, no. 12.

[4] K. Srinivas, G. RaghavendraRao and A. Govardhan, “Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques”, 2010, Proceedings of 5th International Conference on Computer Science & Education, China, pp. 24–27.

[5] J Peter and K. Somasundaram, “An Empirical Study on Prediction of Heart Disease Using Classification Data Mining Techniques”, 2012, Proceedings of IEEE International Conference on Advances in Engineering, Science and Management (ICAESM), pp. 514-518.

[6] H. D. Masethe and M. A. Masethe, “Prediction of Heart Disease using Classification Algorithms”, 2014, Proceedings of the World Congress on Engineering and Computer Science (WCECS), San Francisco, USA.

[7] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction”, 2011, International Journal of Computer Applications (0975 – 8887) Volume 17, No.8.

[8] Syed Immamul Ansarullah1, Pradeep Kumar Sharma2, Abdul Wahid3, Mudasir M Kirmani, “Heart Disease Prediction System using Data Mining Techniques: A study”, 2016, International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 08.

[9] V. Manikandan and S. Latha, “Predicting the Analysis of Heart Disease Symptoms Using Medical Data Mining Methods”, 2013, International Journal of Advanced Computer Theory and Engineering, Vol. 2, Issue 2.

[10] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, “Design and Implementing Heart Disease Prediction Using Naives Bayesian”, 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI).

[11] Monika Gandhi, Shailendra Narayan Singh, “Predictions in heart disease using techniques of data mining”, 2015, International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE).

[12] Rashmi G Saboji, “A scalable solution for heart disease prediction using classification mining technique”, 2017, International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS).