*International Journal of Engineering & Technology*, 7 (3.1) (2018) 166-170

International Journal of Engineering & Technology

*Website: www.sciencepubco.com/index.php/IJET*

***Research paper***

# Predictive Analysis using Data Mining Techniques for Heart Disease Diagnosis

**Siddharth Joshi, Ashish Sasanapuri, Shreyash Anand, Saurav Nandi,Varsha Nemade***

*Department of Computer Engineering, MPSTME , NMIMS University, Shirpur, India*
**Corresponding author E-mail: Varsha.Nemade@nmims.edu*

**Abstract**

Due to technological advancements in the field of computer science and data warehousing techniques. The healthcare industry ranging from small clinics to large hospital campuses use Content management system which has made the storage and accessing of data a faster option. But these large amounts of data generated are regrettably not mined and the data remains unexploited. Through this research we aim to demonstrate the use of Data Mining algorithm by using python programming language in order to create a desktop-based application which will cater to our aim. This Paper will analyze the performance by comparing the metrics of data analysis like accuracy, precision and recall in order introducing our software solution which tries to be more accurate than the work previously done on Cleveland, VA Hungarian data sets taken from UCI repository [1].

***Keywords***: *Data mining; heart diseases; predictive Analysis.*

## 1. Introduction

Heart Disease is defined as the conglomeration of symptoms and diseases that affects the heart and blood vessel of the human body. Chronic diseases present in the heart muscle, circulatory diseases are major causes leading to a heart attack. Cardiovascular disease basically affects the heart and the blood streams in a manner dependent on the factor through which blood is pumped into the body. Contraction of coronary arteries results in reduced supply of blood and oxygen supply to the heart thus affecting it severely [2]. The research by Deeanna Kelley [3] states that although conventionally reducing the intake of dietary fatty acids was recommended to prevent any cardiovascular disease it is not advisable to do so because the effect of fatty acids (cholesterol) have different effect on lifestyles of different body. Early detection of disease will help curb such problem. The likelihood of being prone to heart attack can be determined by a person's habit and also the information about the essential vitals of body which are likely to affect his heart.

We used the data provided by UCI Repository [4] i.e. Cleveland, Hungarian and VA Long beach. The original data set contains about 76 attributes. We have chosen to Select fourteen attributes because [Table1.] and all the performance metric i.e. accuracy, precision in the previous research on these data sets were done on these 14 specific attributes. Our aim by introducing this solution is to provide the user i.e. patients and doctors the ease of an intuitive desktop GUI in which they can enter their details among the attribute for which the system is trained for and get a prediction of whether they have a heart disease or not. The data sets contain the information about whether a person has a heart disease or not, because of which we were able to train the system using Naïve Bayes Classifier and Logistic Regression.

| Name | Type | Description |
|---|---|---|
| Age | Continuous | Age in years |
| Sex | Discrete | Value 1: Male |
| | | Value 0: Female |
| Chest Pain | Discrete | Value 1: Typical type 1 angina |
| | | Value 2: Typical type 2 angina |
| | | Value 3: Non-angina pain |
| | | Value 4: Asymptomatic |
| Fasting Blood Sugar | Continuous | Value 1: >120 mg/dl |
| | | Value 0: <120 mg/dl |
| Restecg | Discrete | Resting electrographic results: |
| | | Value 0: Normal |
| | | Value 1: Having ST-T wave abnormality |
| | | Value 2: showing probable or definite left ventricular hypertrophy |
| Exang | Discrete | Exercise-induced angina |
| | | Value 1: Yes |
| | | Value 0: No |
| Slope | Discrete | Slope of the peak exercise segment |
| | | Value 1: Unsloping |
| | | Value 2: Flat |
| | | Value 3: Down sloping |
| CA | Discrete | Number of major vessels colored by fluoroscopy ranging in between value 0-3 |
| Thal | Discrete | Value 3: Normal |
| | | Value 6: Fixed defect |
| | | Value 7: Reversible defect |
| Trestbps | Continuous | Resting blood pressure in mm Hg |
| Chol | Continuous | Serum Cholesterol in mg/dl |
| Thalach | Continuous | Maximum heart rate achieved |
| Old peak ST | Continuous | Depression induced by exercise relative to rest |
| Diagnosis | Continuous | Value 0: No disease |
| | | Value 1: Mild disease |
| | | Value 2: Severe disease |

## 2. Related Work

Researchers have previously tried to forecast the presence of heart disease using features. Varma, Srivastava, and Negi [8] have created a model in which they diagnose coronary artery disease. They portray a hybrid system for this in which the authors used the dataset from Department of Cardiology at Indira Gandhi Medical college. The number of tuple contains 335 records and has 26 attributes. In their implementation the data was pre-processed by corelating the data using particle swarm optimization. While using the Multi-layer perceptron (MLP) they obtained the accuracy of 77%. We have used the frequently used data set of the UCI machine learning which contains the Cleveland, Hungarian and Long Beach VA datasets. El-Baily et al. [9] conducted research on these by selecting 5 common variables. Two data mining techniques were applied namely Decision Tree (C4.5) and Fast Decision tree (FDT). The accuracy recorded were 69.5% for Long Beach VA using FDT and 78.54% using C4.5. The Table [3.] mentions the work we have referred for comparisons of accuracies.

## 3. Methodology

Our solution follows the KDD i.e. Knowledge Discovery in Databases process [6] which provides an orchestrated framework which eases the process of data mining. The fig [1] gives an overall idea how is the data cleansed and prepared for providing prediction and graphical representation.

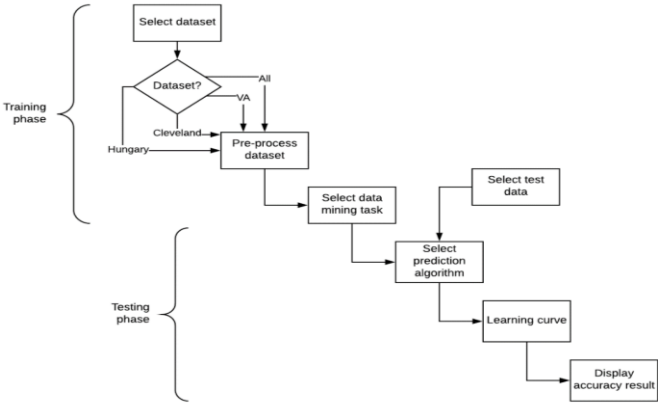


***Fig 1.*** Flow of data through Training and Testing Phases

### 3.1 Data Selection and Pre-processing

Section 1.3 already comments on the why this data set[1][4] was selected .As mentioned in the thesis [5] by Sunitha Sajja the missing values in the case of these datasets cannot be disregarded as 90% of the data set contain some or the other missing value. Disregarding the values will produce insignificant results. Table 2 contains the statistics of missing attributes in the datasets [1]. While pre-processing the missing values were replaced by null float values. Also, as we stated in our aim we are aiming to predict on the possibility i.e. whether a person is suffering from heart disease or not. Except the Hungarian Dataset the VA Long Beach and Cleveland data set has 4 outcome variables (Final Diagnosis) as mentioned in table (Table 2). We merged all the values i.e. 1,2,3,4 to value 1 which if predicted a person is susceptible to heart disease and value 0 to the already 0 predicted value. This categorization also helped us in reducing the processing time (Although insignificant) and improving the accuracy of trained model using Naïve Bayes Classification Algorithm and Logistic Regression. The Accuracy statistics and improvements will be discussed in further Section 3.

***Table 2.*** Percentage of missing values of attributes in datasets [5]

| Sr. No. | Attributes | Cleveland | Hungary | Long Beach VA |
|---|---|---|---|---|
| **1** | Age | | | |
| **2** | Sex | | | |
| **3** | Chest Pain | | | |
| **4** | Fasting Blood Sugar | | 3% | 4% |
| **5** | Restecg | | | |
| **6** | Exang | | | 27% |
| **7** | Slope | | 65% | 51% |
| **8** | CA | 1% | 99% | 99% |
| **9** | Thal | | 90% | 83% |
| **10** | Trestbps | | | 28% |
| **11** | Chol | | 8% | 4% |
| **12** | Thalach | | | 27% |
| **13** | Old peak ST | | | 28% |

### 3.2 Choosing Data Mining Task

This step involves the selection of a goal the KDD process. The goal can be Classification, Regression, clustering etc. We chose Naïve Bayes (Classification), and Logistic Regression (Regression).

### 3.3 Data Mining Algorithm

After selection of the Data Mining task i.e. Supervised Learning (Classification) we firstly choose Naïve Bayes Classification.

#### *2.3.1* Naïve Bayes

Naïve Bayes classifier is a supervised learning model. It is one of the data mining classifiers used in our system to forecast the diagnosis of the heart disease in a patient. The Naïve Bayes classifier is established on the Bayesian' theorem which determines the independent assumptions of the variables. According to the research statistics Naïve Bayes classifier is widely used in prediction systems as it is not very complex to build the model and easy to understand. After Naïve Bayes' law relates the bounded and minimal probabilities of 2 random events the Bayesian theorem is used to compute rearward probabilities for given observations. The probabilities are denoted by, S(a|b), from S(a), S(b), and S(b|a). The formula used for calculating the value of S(a|b) is given by

$$S(a|b) = S(b|a)\, S(a) / S(b)$$

Here,

$b = \{b_1, b_2, b_3 \ldots\ldots, b_n\}$ is a set of 'n' attributes and c is some hypothesis means.

S(a|b) is the rearward probability of a class given prediction variable.

S(a) is the probability of class occurrence.

S(b|a) is the probability of a predictor given class of attribute.

S(b) is the probability of predictor.

One of the main advantages of using Naïve Bayes classifier is that it does not require a large amount of training data to assess the parameters. Naïve Bayes is operated in a way such that it is processed on the training set and then on the testing set. Naïve Bayes is a highly scalable and fast model. It considers attributes as independent of each other. Naïve Bayes accepts both binary and multiclass classification problems.

#### *2.3.2* Logistic Regression

Logistics Regression is a classification algorithm where the dependent variable is categorical. It is one regression algorithm which can be used for the purpose of predictive analysis. It is a statistical method for analyzing dataset. It is used to describe the relation between the dependent variable and one or more nominal

variable. It predicts the probability of a particular outcome. The outcome is measured in binary (1/0, True/False, Yes/No). The outcome is based on the use of several predictors. It produces a logistics curve whose value is limited between o and 1. The likelihood of event of an occasion is by fitting information to a logit function. It generates the variable of the formula to forecast the logit conversion of the type of interest-

$$\text{logit}(k) = z_o + z_1x_1 + z_2x_2 + z_3x_3 + \ldots + z_kx_k$$

where k is the probability of the characteristics of interest.

The logit transformation is-

$$\text{Odds} = \frac{k}{1-k} = \frac{\text{probability of presence of characteristic}}{\text{probability of absence of characteristic}}$$

And

$$\text{logit}(k) = \ln\left(\frac{k}{1-k}\right)$$

### *3.4 Finding Patterns Using Data Mining*

Graphical Visualization of data points by using data mining algorithms helps in analysing the effectiveness of a particular algorithm used. Learning curve is a plot of prediction error between training sets and validation sets over a range of training set. It is the graphical representation of the dataset. The curves give valuable information about the model training process. In some cases, learning curves can help to expand or reduce effort and expenses in gathering more data. The advantage of using learning curves is that we can learn about the data only through the fraction of its which is available. The learning curves are computationally intensive, as in fitting a single small model on a large dataset is possible in this case. Fig 2. Shows graphical representation of learning curve.

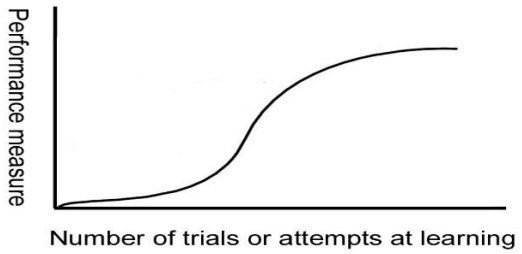


**Fig 2.** General Learning Curve

## 4. Performance Study of Algorithms

As we can infer from the number of attributes that heart disease diagnosis is a complex [7]. Data mining provides an orchestrated procedure which help us to analyse the data and provide the end user with a decision support system. Our goal can be directed possibly in two directions i.e. Descriptive and Predictive. A part of the descriptive analysis is already explained in section 3.4 where we explained how learning curves can be used for graphical representation of how a model fits the data set we are working on. The other method for providing descriptive analysis is by using statistical test. We have used Precision, Recall and accuracy as statistical test for descriptive analysis. Confusion matrix is used to give a view on the prediction results of various prediction algorithms. Confusion matrix is an n x n matrix which is used to calculate the actual and predicted classifications of the system based on the different classes in the systems. It can have binary classes or can be extended to the number of classes the system concludes the result. It is used to calculate various rates such as accuracy, precision, recall etc. The confusion matrix contains the following cases:

1. A case where the predicted result is true and the actual result is also true.
2. A case where the predicted result is false and the actual result is also false.
3. A case where the predicted result is true but the actual result is false.
4. A case where the predicted result is false but the actual result is true.

The confusion matrix is used to calculate the following results of our system:

1. The accuracy which is given as the ratio of total sum of instances where the actual result is same as predicted result and the total result. This gives the result of how often the classifier is correct.
2. The recall which is given as the ratio of the case where actual result and the predicted result are true and total summation of instances where actual result and the predicted result are true and the case where the predicted result is false but the actual result is true. This gives the probability of predicted true results to the total actual values.
3. The precision which is given as the ratio of the case where the actual result and the predicted result are true and total summation of instances where actual result and the predicted is true and the case where the predicted result is true but the actual result is false. This gives the probability of predicted true results to the total predicted values.

Predictive analysis is an important subdivision of advanced analytics which is used in making predictions about future events. The patterns found in historical and transactional data is used in identification of risks and predictions for the future. Predictive analytics model assesses relationships among various factors to calculate accuracy of prediction, recall etc. Predictive analysis' main functionality allows organisations and various sectors to become proactive, predict about the future outcome, anticipating outcomes and behaviours based upon the given dataset given.

## 5. Results & Discussions

### 5.1 Naïve Bayes

The **Table [3.]** Shows various accuracy found by authors over year

| Author | Year | Techniques | Dataset | Accuracy |
|---|---|---|---|---|
| Cheung et al. [10] | 2001 | Naïve Bayes | Cleveland Dataset | 81.48% |
| Yan et al. [11] | 2003 | Naïve Bayes | Cleveland Dataset | 78.56% |
| Andreeva P [12] | 2006 | Naïve Bayes | Cleveland Dataset | 95% |
| Sitar-Taut et al. [13] | 2009 | Naïve Bayes | Cleveland Dataset | 62.03% |
| Raj Kumar et al. [2] | 2010 | Naïve Bayes | Cleveland Dataset | 52.33% |
| Sunitha Sajja [1] | 2010 | Naïve Bayes | Cleveland, Hungarian, VA Long Beach Dataset | 63.97%, 65.74%, 38.42% |
| Srinivas et al. [14] | 2010 | Naïve Bayes | Cleveland Dataset | 84.14% |
| Shouman et al. [15] | 2012 | Naïve Bayes | Cleveland Dataset | 81.48% |
| František Babič et al.[16] | 2017 | Naïve Bayes | Cleveland, VA Dataset | 78.54%, 69.5% |

We tested the accuracy of the Naïve Bayes Model using various combination training and testing sets. Table 4. shows the best accuracy found for Naïve Bayes Model among all the various instances of training and testing sets.

**Table 4.** Naïve Bayes Descriptive Analysis

| Dataset | Accuracy | Precision | Recall |
|---|---|---|---|
| Cleveland | 85% | 86% | 86% |
| Hungarian | 80% | 82% | 81% |
| VA long beach | 77% | 76% | 78% |
| Combined | 85% | 86% | 86% |

For the Cleveland data set the best accuracy of the Naïve Bayes Model was found using the (90,10) data division to training and testing sets. Similarly, for Hungarian data set the best accuracy of model was found at division of (75,25) for training and testing set, (60,40) training and testing for VA Long Beach Data Set. Learning curves which are generated in our system were used to deter-

mine the correct ratio of training and testing for finding the best accuracy. Below tables [5,6,7,8] show individual descriptive statistics for each data sets individually and combined.

**Table 5.** Confusion matrix for Naïve Bayes using Cleveland dataset

| | Predicted | | Accuracy |
|---|---|---|---|
| | Predicted Yes | Predicted No | **85%** |
| Actual Yes | 10 | 2 | **Precision 86%** |
| Actual No | 2 | 14 | |
| | Recall **86%** | | |

**Table 6.**Confusion matrix for Naïve Bayes using Hungarian dataset

| | Predicted | | Accuracy |
|---|---|---|---|
| | Predicted Yes | Predicted No | **81%** |
| Actual Yes | **39** | **9** | Precision **82%** |
| Actual No | **5** | **21** | |
| | Recall **81%** | | |

**Table 7.**Confusion matrix for Naïve Bayes using Long Beach VA dataset

| | Predicted | | Accuracy |
|---|---|---|---|
| | Predicted Yes | Predicted No | **77%** |
| Actual Yes | **5** | **10** | Precision **76%** |
| Actual No | **8** | **57** | |
| | Recall **78%** | | |

**Table 8.** Confusion matrix for Naïve Bayes using combined dataset

| | Predicted | | Accuracy |
|---|---|---|---|
| | Predicted Yes | Predicted No | **85%** |
| Actual Yes | **69** | **11** | Precision **86%** |
| Actual No | **11** | **64** | |
| | Recall **86%** | | |

## 5.2 Logistic Regression

Similar to the descriptive analysis methodology used in the Naïve Classifiers different combinations of testing and training sets. The Table [11] depicts the maximum accuracy among all the combination of training and testing data.

**Table 9.** ***Logistic Regression Descriptive Analysis***

| Dataset | Accuracy | Precision | Recall |
|---|---|---|---|
| Cleveland | 89% | 90% | 89% |
| Hungarian | 82% | 83% | 82% |
| VA Long Beach | 77% | 76% | 78% |
| Combined | 84% | 87% | 87% |

**Table 10.** Confusion matrix for Logistic Regression using Cleveland dataset

| | Predicted | | Accuracy |
|---|---|---|---|
| | Predicted Yes | Predicted No | **89%** |
| Actual Yes | **11** | **1** | Precision **90%** |
| Actual No | **2** | **14** | |
| | Recall **89%** | | |

**Table 11.** Confusion matrix for Logistic Regression using Hungarian dataset

| | Predicted | | Accuracy |
|---|---|---|---|
| | Predicted Yes | Predicted No | **82%** |
| Actual Yes | **40** | **8** | Precision **83%** |
| Actual No | **5** | **21** | |
| | Recall **82%** | | |

**Table 12.** Confusion matrix for Logistic Regression using Long Beach VA dataset

| | Predicted | | Accuracy |
|---|---|---|---|
| | Predicted Yes | Predicted No | **77%** |
| Actual Yes | **5** | **10** | Precision **76%** |
| Actual No | **8** | **57** | |
| | Recall **78%** | | |

**Table 13.** Confusion matrix for Logistic Regression using combined dataset

| | Predicted | | Accuracy |
|---|---|---|---|
| | Predicted Yes | Predicted No | **87%** |
| Actual Yes | **69** | **11** | Precision **87%** |
| Actual No | **9** | **69** | |
| | Recall **87%** | | |

Below FIG [3, 4, 5 & 6] Provide comparative analysis of the metrics used for finding the details about both Naïve Bayes Model and Logistic Regression Model using graphical representations.

**Comparative Analysis Of Algorithms Used Using Graphs**

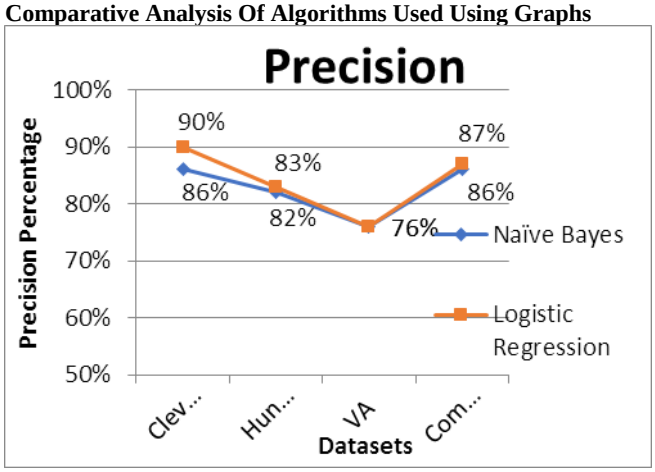


**Fig 3. Precision Analysis**

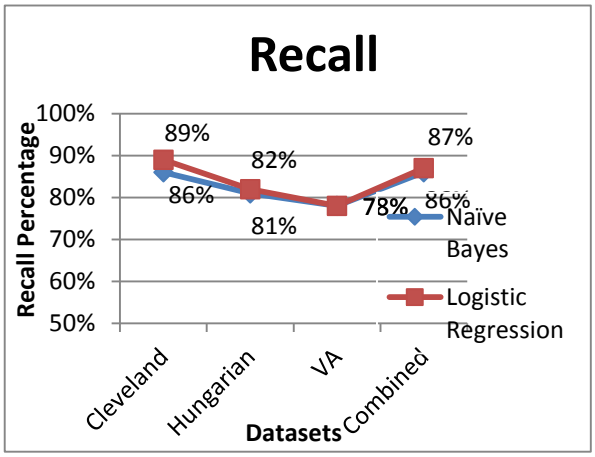


**Fig 4. Recall**

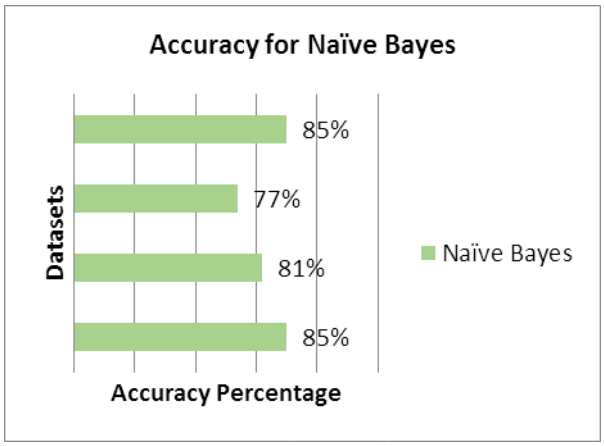


**Fig 5.** Accuracy of Datasets for Naïve Bayes

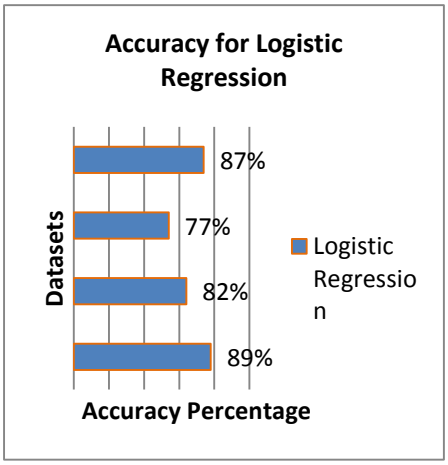


**Fig 6.** Accuracy of Datasets for Logistic Regression

## 6. Conclusion

This paper portrays methodology which can be used for the application of various data mining method and statistical method in order to create a software solution which provides accurate prediction model. We applied Naïve Bayes Classifier and Logistic Regression models for prediction and used learning curves in order to find the correct ratio of training to testing dataset. In comparison to existing studies available [Table 3.] for UCI repository datasets we have found plausible, comparable and even better results.

## References

[1] Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

[2] Rajkumar, Asha, and G. Sophia Reena. "Diagnosis of heart disease using datamining algorithm." Global journal of computer science and technology 10.10 (2010): 38-43.

[3] Kelley, Deeanna. "Heart disease: Causes, prevention, and current research." JCCC Honors Journal 5.2 (2014):

[4] Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.,University Hospital, Zurich, Switzerland: William Steinbrunn, M.D. , University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D. , V.A. Medical Center, Long Beach and Cleveland Clinic Foundation:Robert Detrano, M.D., Ph.D.

[5] Sajja, Sunitha. "Data mining of medical datasets with missing attributes from different sources." PhD diss., Youngstown State University, 2010.

[6] Rajput, Anupama, M. Ramachandran, V. D. Gotmare, and P. P. Raichurkar. "Recent Bioactive Materials for Development of Eco-friendly Dippers: An Overview." Journal of Pharmaceutical Sciences and Research 9, no. 10 (2017): 1844-1848.

[7] Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview",in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34

[8] Babič, František, et al. "Predictive and descriptive analysis for heart disease diagnosis." Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on. IEEE, 2017.

[9] L. Verma, S. Srivastaa, and P.C. Negi, "A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data", Journal of Medical Systems, vol. 40, no. 178, 2016, doi: 10.1007/s10916-016-0536-z.

[10] R. El-Bialy, M. A. Salamay, O. H. Karam, and M. E. Khalifa, "Feature Analysis of Coronary Artery Heart Disease Data Sets", Procedia Computer Science, ICCMIT 2015, vol. 65, pp. 459–468, doi: 10.1016/j.procs.2015.09.132

[11] Cheung, Marian C., et al. "Antioxidant supplements block the response of HDL to simvastatin-niacin therapy in patients with coronary artery disease and low HDL." Arteriosclerosis, thrombosis, and vascular biology 21.8 (2001): 1320-1326.

[12] Yan, Wei-wu, and Hui-he Shao. "Application of support vector machines and least squares support vector machines to heart disease diagnoses." Control and Decision 18.3 (2003): 358-360.

[13] Andreeva, Plamena. "Data modelling and specific rule generation via data mining techniques." International Conference on Computer Systems and Technologies-CompSysTech. 2006.

[14] Sitar-Taut, D. A., et al. "Using machine learning algorithms in cardiovascular disease risk evaluation." Journal of Applied Computer Science & Mathematics 3.5 (2009): 29-32

[15] Deepali Mor, M Ramachandran, Pramod Raichurkar, "Optimization of Solid Wastes Disposal Strategy by Fuzzy Topsis Method", Nature Environment and Pollution Technology 16(1):247-250, 2017.

[16] Shouman, Mai, Tim Turner, and Rob Stocker. "Using data mining techniques in heart disease diagnosis and treatment." Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on. IEEE, 2012.

[17] Babič, František, et al. "On patient's characteristics extraction for metabolic syndrome diagnosis: Predictive modelling based on machine learning." International Conference on Information Technology in Bio-and Medical Informatics. Springer, Cham, 2014.

[18] Siddharth. J et al. Int. Journal of Engineering Research and Application ISSN : 2248-9622, Vol. 7, Issue 4, ( Part -2) April 2017, pp.14-19