Prediction of Heart Diseases in Diabetic Patients

[1]Bindushree D C, [2]Joshua S, [3]K A Ashik, [4]K R Rakshith Purshottama
[1,2,3,4]School of C&IT, REVA University, Bangalore

Article Info
Volume 83
Page Number: 4759-4764
Publication Issue:
May-June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *16 May 2020*

Abstract

In today's world we observe that a large group of people are affected with diseases like Diabetes, Cancer, Heart related issues, etc., due to which a substantial amount of raw data is being collected in the medical industry. This paper aim's to convert the obtained raw data into structured data using machine learning techniques in order to predict heart disease in diabetic patients. An Automated intelligent system to predict heart disease is developed using various data mining techniques such as Decision Tree, Logistic Regression, K nearest neighbor(Knn), Hybrid Algorithm and Random Forest. These data mining techniques on implementation provides a higher accuracy rate which is the primary requirement for better and faster prediction of heart disease in diabetic patients. The datasets used is obtained from PIDD (Pima Indians Diabetes Database) containing a number of instances with a set of attributes such as age, sex, blood sugar level, cholesterol etc. This system is developed to be User friendly by reducing the possibility of human errors and the time consumption.

Keywords: *Accuracy, Automated intelligent system, Knn, Logistic Regression, PIDD*

1. Introduction

The best thing about human mechanism is that it gives you signs that there is something wrong with your body but at times these signs are not enough to find out what type of problem it would be. In most of the cases it is observed that diabetic patients are mostly numb to most of these signs. Heart disease has been one of the most unpredictable disease which is affecting the current generation [11][12]. The reason for this is mainly the eating habits, stress, environmental condition, etc.

Studies say that diabetic patients are more prone to heart diseases and it becomes difficult to predict it in a very early stage. It becomes time consuming for doctors to predict heart disease especially when a patient has multiple disease of similar symptoms. We can make things better and get results faster using Machine Learning techniques which can make prediction of heart disease much faster and efficient [13].

The main aim of this paper is to make prediction of heart disease in Diabetic patient efficient and faster using machine learning techniques. There are multiple data mining techniques that are available to obtain the results. There are a few steps that are considered more important when compared to the other steps, such as data pre-processing, using classification techniques on the pre-processed data and data mining techniques which are used to achieve the required results which is a combination of evolutionary computation and data mining intelligence.

2. Literature Survey

Rao Muzamal Liaqat, et al. [1]. presented "Framework for Clustering Cardiac Patient's Records Using Unsupervised Learning Techniques", talks on the all the major steps used to separate the data got from AFIC (Armed Force Institute of Cardiology) which adds up to 1500 records including 36 attributes. The main Steps used by the researchers are achieving target data, getting pre-processed data from it leading to transformed data which in turn undergoes various data mining techniques to achieve required patterns/models to achieve the required data. The major drawback of this paper is that it does not provide any automated intelligence system.

Minyechil Alehegn, et al. [2]. presented "Analysis and prediction of diabetes Mellitus using machine learning Algorithm" that uses the raw data that we want for analysis and prediction which is safely collected, joined and prepared for investigation. The dataset required for this proposed system is obtained from public UCI repository PIDD (Pima Indian Diabetes Database) which is freely available online to analyze and predict diabetes Mellitus. Different parameters are taken such as Body Mass Index, Skin Thickness, Glucose, etc., The algorithms used here are the three most known prediction algorithms which were achieved from extent literature namely, Support vector machine (SVM), Naïve Net (NN), and Decision Stump (DS) classification algorithm. The predictions of these three algorithms are combined to one to increase the accuracy of the predictions using base learner. The disadvantage for this method is that it shows the accuracy or single algorithms rather than using ensemble method which might provide better predictions

Ms. Nilam chandgude, et al. [3]. presented "A survey on diagnosis of diabetes using various classification algorithm" illustrates a model using various classification algorithms to diagnose diabetes from data sets collected from various online sources such as PIMA Indian diabetes dataset from UCI repository, Pub Med, WebMD and Medline. This models work flow executes four main steps to achieve the desired result. The model exhibits process namely, Dataset collection based on various attributes which undergoes post-processing which in turn is moved to the training system which is the main brains of the model. It implements various steps such as reading of all input data set and using particular algorithm such as Back Propagation, Feed Forward Neural Network, Naïve Bayes, SVM etc. Then the datasets are passed to the testing system phase to check whether the system is diagnosing the disease properly based on the symptoms or not. The disadvantage of this paper is that a higher accuracy with a less amount of time is not achieved.

R. Sivanesan, et al. [4]. Presented "Review on Diabetes Mellitus diagnosis using classification" illustrates the classification done using general techniques used in medical data mining, these medical data mining techniques include the performances of algorithms such as J48 decision tree which predicts the target value using 768 instances containing 9 attributes. The major drawback of this paper is that a model to give best accuracy in Naïve Bayes classifiers using Gini Index based fuzzy classification for diagnosing diabetes mellitus is not found.

R.S.Suryakirani , et al. [5]. Presented "Comparative study and analysis of classification algorithm in data mining using diabetic datasets" The performance of the algorithm used using various measures shows that the J48 algorithm has better accuracy level than the rest of the algorithms used. The disadvantage of this paper is that the focus on using other classification algorithms of data mining was not considered which would have led to better accuracy.

Sadri Sa'Di et al. [6]. Presented "Comparison of Data Mining Algorithms in the diagnosis of type II diabetes" this paper involves in diagnosing type II diabetes using data mining algorithms. The disadvantages of this paper is diagnosing diabetes with methods such as Neuro-fuzzy networks and comparison of the algorithms used in this paper was not executed.

3. Methodology

The development of the model involves 6 stages to achieve the automated intelligent system which is shown in below Fig.1

A. Data collection
B. Data preprocessing
C. Analysis
a. Principal component analysis
b. Co-relation analysis
D. Modeling/Training
E. Testing
F. Validation

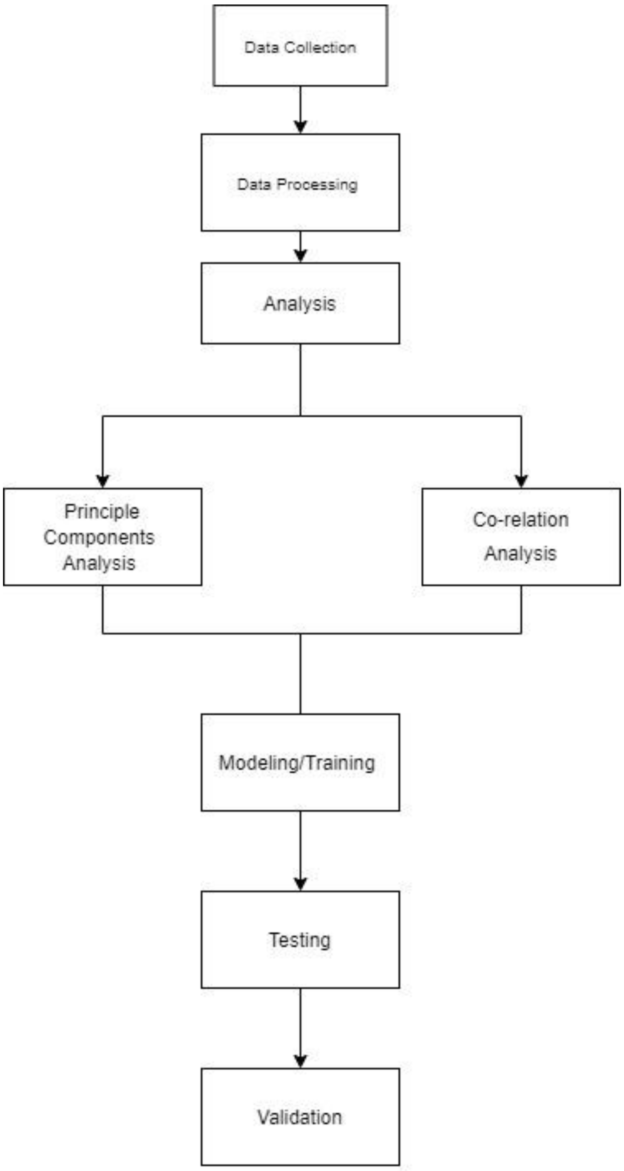


Figure 1: Proposed Methodology flowchart

A. Data collection

This is the initial stage for the development of our automated intelligent system which involves collection of data. One of the main ways to collect data is through medical reports but these data are mostly available in the raw data form. The data collected for this process has been collected from PIDD (Pima Indian Diabetic Dataset). These collected data are used in the next stage to develop it into a structured form.

B. Data Preprocessing

This stage involves converting the raw data into structured data. The main motivation of this stage is to make these data compatible to the machine learning algorithm. Preprocessing involves removal of junk data, demoralization of the data in order to give structure to the data.

C. Analysis

Analysis is an important stage in the methodological process as it involves the cleansing of data. The analysis of this methodology can be observed under two aspects.

a. Principle component analysis
Principle component analysis is mostly used as a tool in exploratory data analysis which finds relatedness between the datasets. At this stage it involves removal of unwanted data and obtain useful data in order to effectively structure the data.

b. Co-relation analysis
This stage of analysis is used to measure how strongly pairs of variables are related. For example, age and blood sugar; older the age most possibly affected by blood sugar.

D. Modeling/Training

At this stage of modeling/training it involves usage of various machine learning algorithms such as support vector, decision tree, random forest, etc. In order to train the automated intelligent system. Deep learning, reinforcement learning are both a part of machine learning which in turn is a part of wider set of artificial intelligence tool. It includes 80% of the data collected to train the data and the rest 20% for testing the accuracy of the result obtained.

E. Testing

Testing is a very important stage in which accuracy of the system is being measured. The accuracy is measured using confusion matrix, receiver operating characteristic. Receiver operating characteristic is useful to compare the overall accuracy.

F. Validation

Validation is mainly used when the accuracy is not up to the mark. In this case optimization is involved to increase the performance characteristics such as execution speed, code size, etc.

4. Experimental Results

The structured data collected is used in various algorithms like Knn, Decision Tree, Hybrid Algorithm, Random Forest and Logistic Regression. The result best obtained is given out as a predicted output through a user interface. The expected results can be viewed in two aspects,

- Back-end

Feature analysis is done between the two variables to check how strong the variables are related to each other. which are shown in the below given Fig. 2, Fig. 3, Fig. 4.

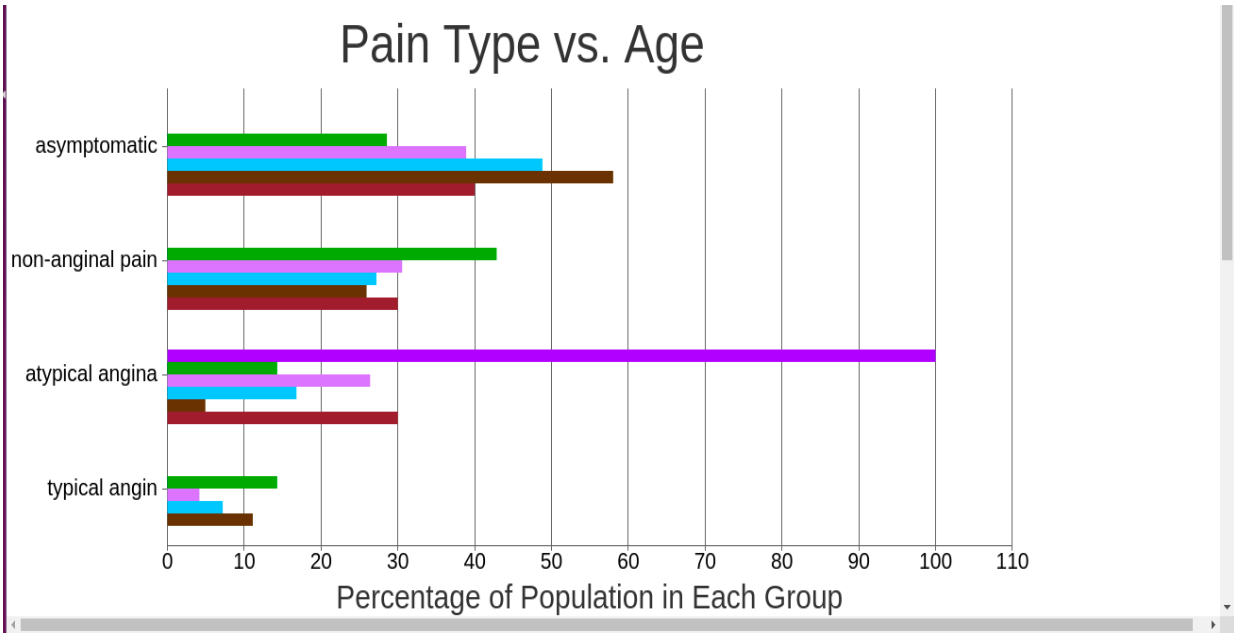


Figure 2: A bar graph representation for the variables Pain Type vs Age

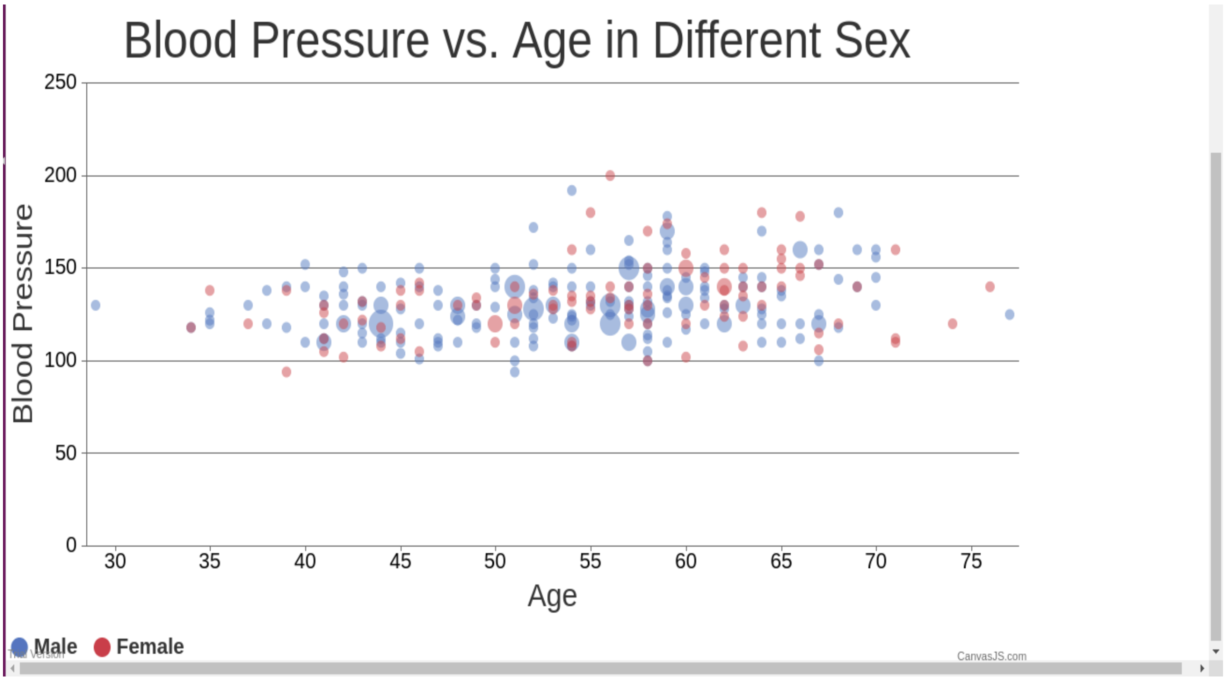


Figure 3: A scatter plot representation for the variables Blood Pressure vs Age in Different Sex

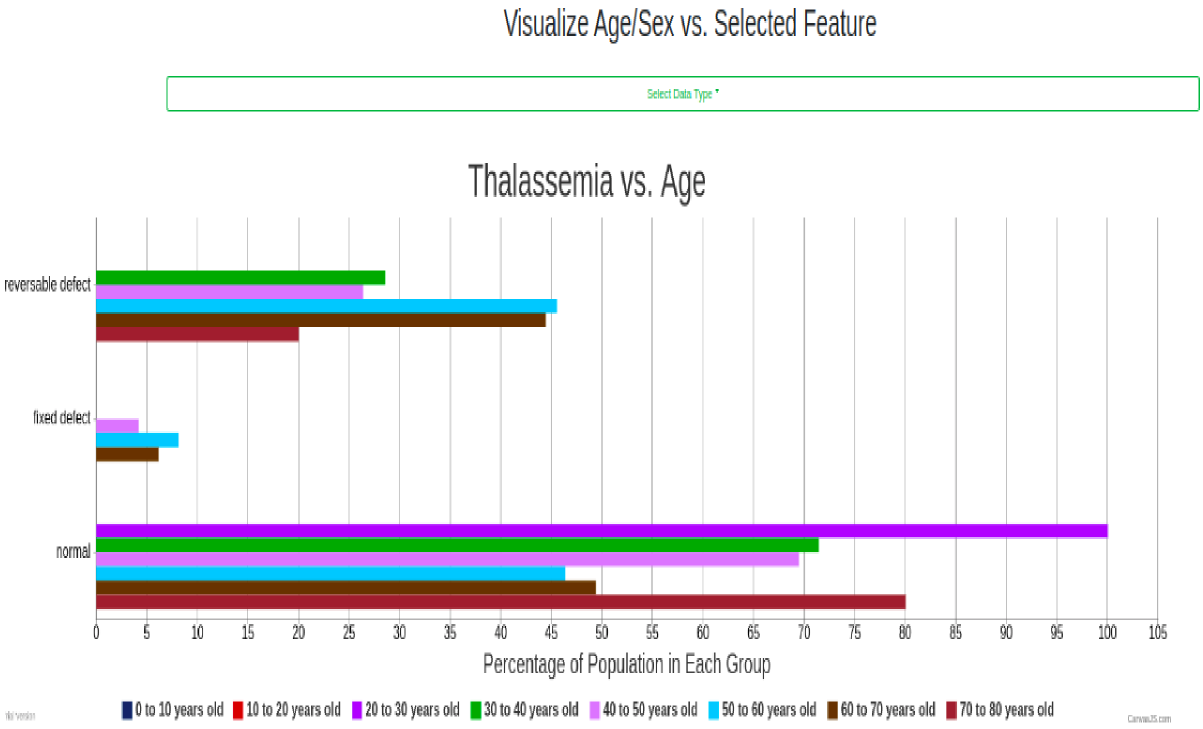


Figure 4: A bar graph representation for the variables Thalassemia vs Age

Prediction of relation between the variables are given in the form of Potential Factor Ranking (Method drop) which are obtained by Co-relation analysis.

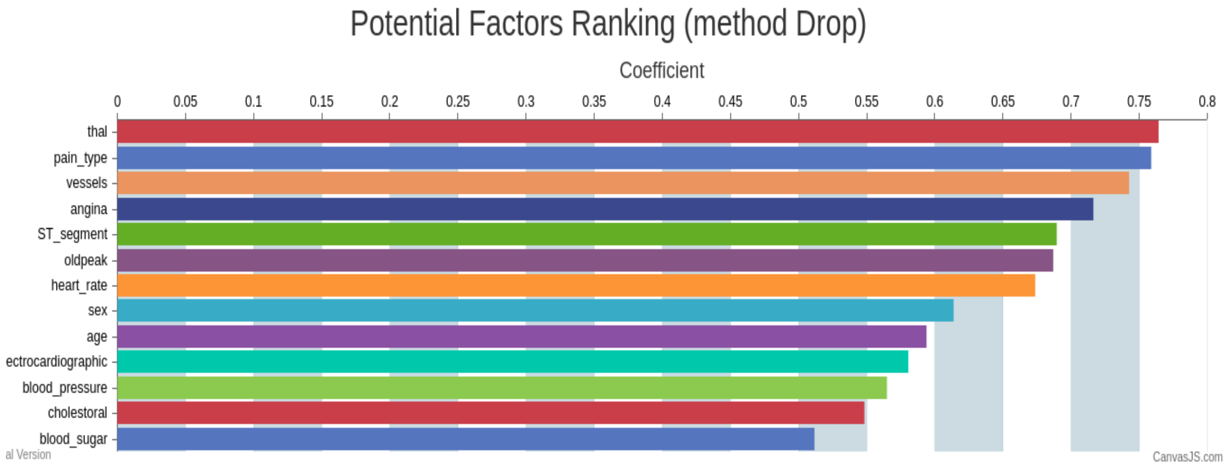
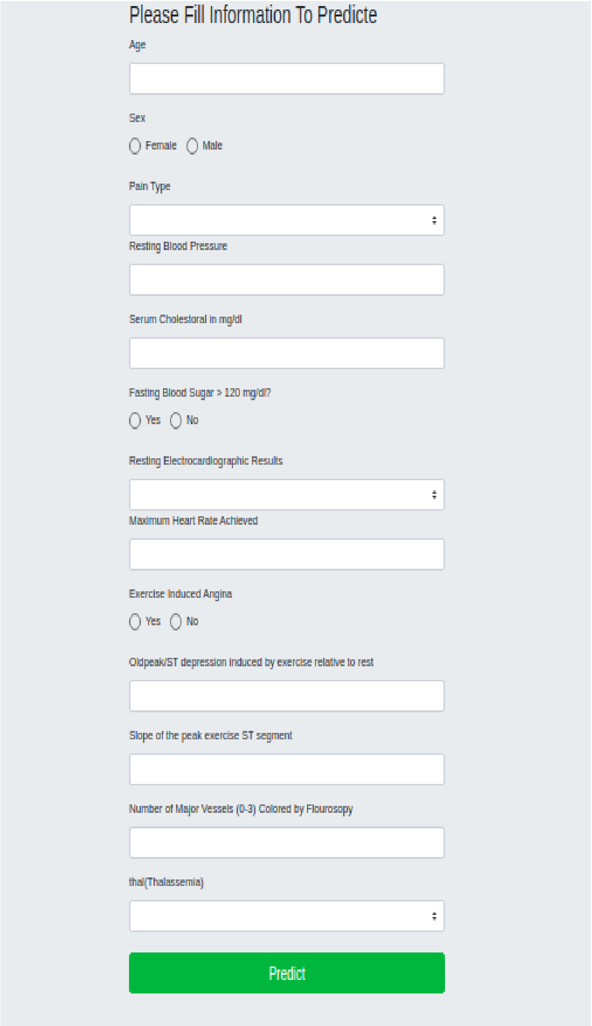


Figure 5: Rank of features regarded correlation to target

- Front-end

The Front-end is a User Interface which is a web page developed using HTML, Flask server and Node Manager.



Please Fill Information To Predicte

Age

Sex

○ Female ○ Male

Pain Type

Resting Blood Pressure

Serum Cholestoral in mg/dl

Fasting Blood Sugar > 120 mg/dl?

○ Yes ○ No

Resting Electrocardiographic Results

Maximum Heart Rate Achieved

Exercise Induced Angina

○ Yes ○ No

Oldpeak/ST depression induced by exercise relative to rest

Slope of the peak exercise ST segment

Number of Major Vessels (0-3) Colored by Flourosopy

thal(Thalassemia)

Figure 6: User interface to input information for prediction

The previous paper referred has the study and analysis of four classification algorithms and experimental result shows that j48 classifier gives better accuracy 73.28%. Whereas this paper involves five classification algorithms and the experimental results shows that Logistic regression gives the highest accuracy of 86% which is comparatively higher in accuracy than the previous paper referred.

5. Conclusions

Diabetes is an epidemic which makes it a noticeable factor to come up with cures that rapidly erases its effects as a fact that at this point of time there are almost 1 out of 11 people in a country who are affected by it. The said epidemic disease has no cure as of date and just contains a history of controlling and managing it from a personal perspective. Henceforth, this developed automated intelligent system provides a solution to the above obstacle by improving the accuracy rates of the data mining algorithms used in this proposed system namely The Random Forest algorithm, Decision Tree, Knn Algorithm, Hybrid Algorithm, Logistic Regression that involves the raw datasets being converted to useful datasets which in turn helps in saving human errors and primarily time. In the near future, this proposed system can be improvised using better methods to treat patients directly mapped to the automated system to yield even more faster results and resolve its symptoms at its earlier stages.

References

[1] Rao Muzamal Liaqat,"Framework for Clustering Cardiac Patient's Records Using Unsupervised Learning Techniques" 2016 6th international conference on current and future trends of information and communication technologies in health care(ICTH 2016)

[2] Minyechil Alehegn,"Analysis and prediction of diabetes Mellitus using machine learning

Algorithm" symbiosis international university, Pune, Maharashtra, 412115,India

[3] Ms. Nilam chandgude"A survey on diagnosis of diabetes using various classification algorithm" International Journel on Recent and Innovation Trends in computing and communication volume: 3 issue:12

[4] R. Sivanesan,"Review on Diabetes Mellitus diagnosis using classification" 2017 International Journal of Advanced Research in Computer Science and Management Studies

[5] R.S.Suryakirani "Comparative study and analysis of classification algorithm in data mining using diabetic datasets" 2018 IJSRST, voulume:4 issue:2

[6] Sadri Sa'Di ,"Comparison of Data Mining Algorithms in the diagnosis of type II diabetes" 2015 International Journal on Computational Science and Applications (IJSCA) volume:5, no.5, October 2015

[7] K.Saravananathan, "Impact of Classification Algorithms in Diabetes Data:A Survey" 2016 The Third International Conference on Small & Medium Business 2016 January 19-21

[8] Aiswarya Iyer, "Diagnosis of Diabetes Using Classification Mining Techniques" 2015 International journal of Data Mining & Knowledge Management Process (IJDKP) vol.5, no.1, January 2015

[9] Mr.R.Sengamuthu, "Various Data Mining Techniques Analysis to Predict Diabetese Mellitus" 2018 International Research Journal of Engineering and Technology (IRJET)

[10] Pooja Sharma, "Survey on Classification Algorithms Using Big Data Set" International Journal of Mondern Trends in Engineering and Research

[11] Bindushree D C, Dr. Udayarani V, "Prediction Of Cardiovascular Risk Analysis And Performance Evaluation Using Various Data Mining Techniques: A Review", International Journal of Engineering Research ISSN:2319-6890) (online), 2347-5013(print)

[12] Bindushree D C, Dr. Udayarani V, "A Review On Using Various DM Techniques For Evaluation Of Performance And Analysis Of Heart Disease Prediction", 978-1-5386-0569-1$31.00_c 2017 IEEE

[13] Bindushree D C, Dr. Udayarani V, "An Analysis of Heart Disease for Diabetic Patients Using Recursive Feature Elimination with Random Forest", Journal of Computer Science 2020, 16 (1): 105-116.