International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-6, Issue-9)

Research Article | **September 2017**

Data Mining Techniques for Diagnosis of Diabetes: A Review

Tushar Deshmukh
Fergusson College, Pune,
Maharashtra, India

Dr. H. S. Fadewar
S.R.T.M. University, Nanded,
Maharashtra, India

Abstract—

This Diabetes is such a common dieses found all over the globe, in which blood glucose or in normal terminology the sugar level in blood is increased. It is the condition of the body in which the insulin which is required for the metabolism of the food is not created or body cannot use the insulin produced properly. Doctors say that diabetes can be controlled if it is detected in its early stages.
Data mining is the process in which the data can be used for the prediction based on historic data. The intention here is to analysis how various researchers have used the data mining for better prediction of diabetes so that it could be controlled and possible even cured.

Keywords— Data Mining Techniques, Diabetes, classification, Type1 diabetes, Type2 diabetes

I. INTRODUCTION

The global report on diabetes published by WHO, in year 2016, mentions that more than 422 million people are suffering from diabetes in 2014[1]. So this is such kind of scenario one cannot ignore. Diabetes is nothing but disorder in metabolism. Generally whatever we eat is converted into glucose. And body secretes insulin to use this glucose. In diabetes either body stops producing insulin, or loses the ability to use insulin. So the glucose stays in blood which is dangerous. Technically there are 2 types of diabetes: [2]

Type1 diabetes:

In this type of diabetes, body stops producing insulin. This type of diabetes is found in early age of patient. Once suffered, patient has to take insulin injections for the rest of life. Amongst all the patients, 10% patients are suffering from Type1 Diabetes.

Type2 diabetes:

In this type of diabetes, the insulin is produced but in low quantity or sometimes body stops interacting with insulin. Most of the patient cannot control and the dieses go on increasing. Most of the diabetes patients are affected by this type of diabetes. Generally only adult human beings are the victims of this type of diabetes. Overweighted or excessive fat people have higher risks of Type2 diabetes.

Gestational diabetes

This type of diabetes occurs only in female patient to that only during pregnancy, so sometimes called as pregnancy induced diabetes. This may lead to complications in pregnancy. Here the patient's body cannot produce required amount of insulin.If un-noticed, the child may bear with some complications or in size greater than expected. Such type of diabetes is temporary one most of the times, after pregnancy the patient may recover to normal glucose levels. But in some cases the gestational diabetes automatically turned to be Type 2 diabetes after the pregnancy. The symptoms of diabetes are easily observed in few cases, but most of the symptoms go un noticed.

So whatever may be the type of diabetes it is very much important to detect the diabetes in early stage because suffering for long time period it may cause severe problems. Undetected or uncontrolled diabetes may damage the eyes, kidneys or nerves. It may even lead to heart attack. That's the reason to predict the diabetes is at most important. This paper is an attempt to find out how to predict the diabetes using different data mining techniques.

II. RELATED WORK

Tahani Daghistani, Riyad Alshammari[3] proposed how different classification methods are useful for the prediction of diabetes. For their study they have chosen three classification methods. a) SOM(Kohonen map) which analyzes the heterogeneous data sets by means of unsupervised learning. It processes by dimensionality

reduction technique. b) C4.5, which is a decision tree algorithm. C4.5 uses entropy information to decide the splitting criteria and classify the data set correctly. c) RandomForest, which is another decision tree algorithm which is considered to be the fast algorithm. RandomForest doesn't face the problem of overfiting. After comparing these methods the author finds that Radom forest method is best suited for prediction as it gives highest Recall and precision. The paper also puts a light on risk factors from the multiple attribute like age, BMI, blood pressure etc.

Harleen, Dr. PankajBhambri[4] worked on the concept of data mining for Diabetes Mellitus. The researcher uses naïve Bayes method for classification of data. In this paper the researcher also mentions the details about the KDD (Knowledge Discovery in Database), which is iterative process for transferring the raw data into useful information. The paper talks about how to handle null values and noise in the data. It gives the vivid description of what classification means. It uses WEKA software for classification. The study makes a comparison of two well known algorithms, Naive Bayes algorithm and J48algorithm. The accuracy the algorithm has approached is 81%.

In their study,K. Rajesh and V. Sangeetha[5] uses the Pima Indians Diabetes Database of National Institute of Diabetes and Digestive and Kidney Diseases from UCI Machine Learning Repository, which is very rich dataset and 768 tuples and around 8 different continuous attribute. When different classification algorithms were applied, C4.5 has returned with maximum accuracy, nearly about 91%. But when feature relevance technique was used the accuracy was dropped to 88%. So the researchers recommended C4.5 as the best classifier, whereas the RND tree has suffered from over fitting problem. The paper also listed out all the classification rules generated by C4.5 algorithm.

Tanja Dujić Dijana Sejdinovic, Lejla Gurbeta, Almir Badnjević, Maja Malenica, Adlija Čaušević, Tamer Bego, Lejla Divović Mehmedović[6] uses Artificial Neural Network for classification. The prior intention was to focus on Type2 diabetes and prediabetes. The researchers have applied 2 layered, feed forward neural networks. The numbers of neurons required in hidden layer were 15. The neural network was formed with 2 important parameters glucose level at fasting and HbA1c test. For the experiment the training data and test data were separate. The accuracy they got was almost 94.1 % for prediabetes and around 93.3 % for Type 2 diabetes.

In another study, Krati Saxena, Dr. Zubair Khan, Shefali Singh[7] used a dataset from Stanford University which consists of 100 records and 11 different attribute. The researcher has used two different sets for test data with 50 records each. The algorithm they have used was K-nearest neighbor. The accuracy was tested for two different values of K, k=3 and K=5. The experiment results show that as value of k increases the accuracy also increases. Matlab was used for the simulation of this classification.

Poonguzhali.E, Poonguzhali.E, Sandia Kannan, Sivagami.P[8] has made a research on diabetes diagnosis using Neural Network. Their study was limited to Type 2 diabetes only. The dataset was generated live through survey method, which has 13 different columns. So in all the input layer of the network was feed with 13000 inputs. The network which was designed has a synaptic layer with 13 nodes, 2 hidden layers and one response layer.

In a research done by Baratam Yasaswi, Bodapati Prajna[**9]** the dataset was taken from UCI repository and has 9 different attributes. From the various algorithm available the researches has chosen C4.5 algorithm for their study. C4.5 algorithm is decision tree technique, in which the tree is recursively built to improve the calculation of accuracy. They found that the algorithm could predict the results with high level of accuracy. The accuracy of the algorithm was tested on entirely different testing data set which was not used for the training of the algorithm.

The last paper reviewed here was from Dr. M. Renuka Devi, J. Maria Shyla[10]. The researchers analysed the various classification techniques in data mining using Weka and MATLAB. The dataset they have used was PIMA Indian data set, consisting of 768 rows. The analysis made by these researchers' shows that the modified J48 got the highest accuracy as the prediction algorithm for detection of diabetes. J48 algorithm is a decision tree algorithm, actually it is the implementation of ID3 algorithm.

III. CONCLUSIONS

There has been lot of research made in this domain i.e. prediction of diabetes using data mining techniques. There are lot many different techniques are available like K- nearest neighbor, naïve Bayes method, artificial neural network, association rule mining, decision tree etc. Each researcher has made an attempt to work with different data sets, different types of diabetes as well as different conditions of patients. The intension was very clear; if the diabetes is diagnosed at early stage the patient can be saved from lot many severe problems.

REFERENCES

[1] World health organization, "Global report on daibetes," France., 2016.
[2] (2004-2017) www.medicalnewstoday.com. [Online]. https://www.medicalnewstoday.com/info/diabetes
[3] Riyad Alshammari Tahani Daghistani, "Diagnosis of Diabetes by Applying Data Mining Classification Techniques," vol. Vol. 7, no. 7, 2016.

[4] Dr. Pankaj Bhambri Harleen, "A prediction Technique in Data Mining for Diabetes Meelitus," vol. 4, no. 1, 2016.
[5] K. Rajesh and V. Sangeetha, "Application of Data Mining Methods and Techniques for Diabetes Diagnosis," vol. 2, no. 3, 2012.
[6] Lejla Gurbeta, Almir Badnjević, Maja Malenica, Adlija Čaušević, Tamer Bego, Lejla Divović Mehmedović Tanja Dujić Dijana Sejdinovic, "CLASSIFICATION OF PREDIABETES AND TYPE 2 DIABETES USING ARTIFICIAL NEURAL NETWORK," , Badnjevic A., 2017.
[7] Dr. Zubair Khan,Shefali Singh Krati Saxena, "Diagnosis of Diabetes Mellitus using K Nearest Neighbour Algorithm," vol. 2, no. 4, 2014.
[8] Poonguzhali.E, Sandia Kannan, Sivagami.P Poonguzhali.E, "Diagnosis of Diabetes Mellitus Type 2 using Neural Network," vol. 4, no. 2, 2014.
[9] Bodapati Prajna Baratam Yasaswi, "The Early Augmentation for Diabetes Diagnosis Using Data Mining Approaches ," International Journal of Computer Sci ence And Technology, vol. 7, no. 3, september 2016.
[10] J. Maria Shyla Dr. M. Renuka Devi, "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus," vol. 11, no. 1, 2016.

