International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 4

Various Data Mining Techniques for Diabetes Prognosis: A Review

Misba Reyaz, Gagan Dhawan
Modern Institute of Engineering & Technology,
Kurukshetra University, Kurukshetra, Haryana, India

ABSTRACT

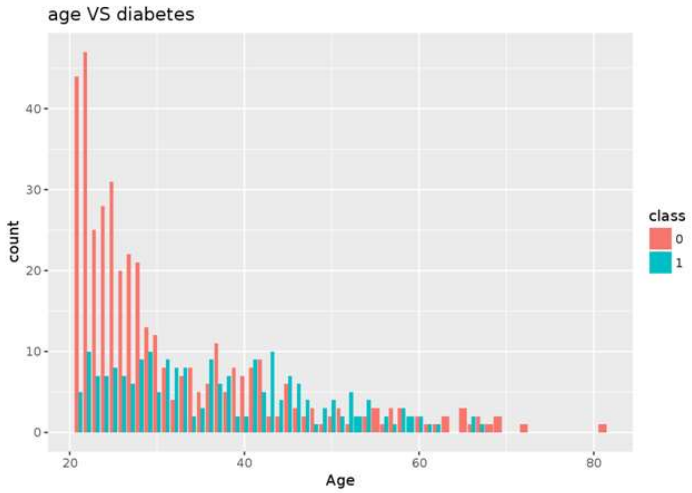
Most of the food we eat is converted to glucose, or sugar which is used for energy. When you have diabetes, your body either doesn't make enough insulin or cannot use its own insulin as well as it should. This causes sugar to build up in your blood leading to complications like heart disease, stroke, neuropathy, poor circulation leading to loss of limbs, blindness, kidney failure, nerve damage, and death.

Data mining adopts a series of pattern recognition technologies and statistical and mathematical techniques to discover the possible rules or relationships that govern the data in the databases. Data mining plays an important role in data prediction. There are different types of diseases predicted in data mining namely Hepatitis, Lung Cancer, Liver disorder, Breast cancer, Thyroid disease, Diabetes etc… This paper analyzes the Diabetes predictions.

Keywords: *Data Mining Techniques, Diabetes, classification, Type1 diabetes, Type2 diabetes*

I. INTRODUCTION

Data mining is described as the process of discovering correlations, patterns and trends to search through a large amount of data stored in repositories, databases, and data warehouses. The development of finding useful patterns or importance in raw data has been called KDD (knowledge discovery in databases). Data mining plays an important role in data prediction. The global report on diabetes published by WHO, in year 2017, mentions that more than 422 million people are suffering from diabetes. So this is such kind of scenario one cannot ignore. Most of the food we eat is converted to glucose, or sugar which is used for energy. The pancreas secretes insulin which carries glucose into the cells of our bodies, which in turn produces energy for the perfect functioning of the body. When you have diabetes, your body either doesn't make enough insulin or cannot use its own insulin as well as it should. This causes sugar to build up in your blood leading to complications like heart disease, stroke, neuropathy, poor circulation leading to loss of limbs, blindness, kidney failure, nerve damage, and death.



Types of Diabetes

Type 1 - Diabetes also called as Insulin Dependent Diabetes Mellitus (IDDM), or Juvenile Onset Diabetes Mellitus is commonly seen in children and young adults however, older patients do present with this form of diabetes on occasion. In type 1 diabetes, the pancreas undergoes an autoimmune attack by the body itself therefore; pancreas does not produce the hormone insulin. The body does not properly metabolize food resulting in high blood sugar (glucose) and the patient must rely on insulin shots.

Type I disorder appears in people younger than 35, usually from the ages 10 to 16.

Type II - Diabetes is also called as Non-Insulin Dependent Diabetes Mellitus (NIDDM), or Adult Onset Diabetes Mellitus. Patients produce adequate insulin but the body cannot make use of it as there is a lack of sensitivity to insulin by the cells of the body. Type II disorder occurs mostly after the 40.India has the dubious distinction of being the diabetic capital of the world. Home to around 33 million people with diabetes, 19% of the world's diabetic population is from India. Nearly 12.5% of Indian's urban populations have diabetes. The number is expected to escalate to an alarming 80 million by the year 2030. Amongst the chronic diabetic complications, diabetic foot is the most devastating result. Over 50,000 leg amputations take place every year due to diabetes in India. Diabetes patients can often experience loss of sensation in their feet. Even the smallest injury can cause infection that can be various serious. 15% of patients with diabetes will develop foot ulcers due to nerve damage and reduced blood flow. Diabetes slowly steals the person's vision. It is the cause for common blindness and cataracts.

II. Literature Survey

P.Yasodha et al [1] Data classification in diabetic patients data set is developed by collecting data from hospital repository consists of 249 instances with 7 different attributes. The instances in the Dataset are pertaining to the two categories of blood tests, urine tests. WEKA tool is used to classify the data and the data is evaluated using 10-fold cross validation and the results are compared. The main purpose of the system is to guide diabetic patients during the disease. Diabetic patients could benefit from the diabetes expert system by entering their daily glucoses rate and insulin dosages; producing a graph from insulin history; consulting their insulin dosage for next day. The diabetes expert system is not only for diabetic patient, but also for the people who suspect if they are diabetic. It's also tried to determine an estimation method to predict glucose rate in blood which indicates diabetes risk.

K. Rajesh et al [2] This project aims for mining the relationship in Diabetes data for efficient classification. They have applied many classification algorithms on Diabetes dataset and the performances of those algorithms have been analyzed. A classification rate of 91% was obtained for C4.5 algorithm.

Sukhjinder Singh et al [3] The various techniques are discussed for predicting the diagnosis of diabetes. Using the data mining technique the health care management predicts the disease and diagnosis of the diabetes and then the health care management can alert the human being regarding diabetes based upon this prediction. The Principal Component Analysis (PCA) is also the technique used for the analysis. The PCA is the feature extraction technique has more act upon on the accuracy of classification techniques. But when the PCA combined with the Neural Networks for classification achieved the best classification accuracy and the PCA performs better for non-diabetic samples than the diabetic samples when combined with Neural Networks. Classification speed of ANFIS is not better than the Neural Networks.

Veena Vijayan V. et al [4] The main data mining algorithms discussed in this paper are EM algorithm, KNN algorithm, K-means algorithm, amalgam KNN algorithm and ANFIS algorithm. EM algorithm is the expectation-maximization algorithm used for sampling, to determine and maximize the expectation in successive iteration cycles. KNN algorithm is used for classifying the objects and used to predict the labels based on some closest training examples in the feature space. K means algorithm follows partitioning methods based on some input parameters on the datasets of n objects. Amalgam combines both the features of KNN and K means with some additional processing. ANFIS is the Adaptive Neuro Fuzzy Inference System which combines the features of adaptive neural network and Fuzzy Inference System. From the observation EM possess the least classification accuracy and amalgam KNN and ANFIS provide the better classification accuracy results. Amalgam KNN comprises both the feature of KNN and K means. ANFIS in cooperates both the features of adaptive neural both ANFIS and amalgam KNN is used. Co active ANFIS was extended with some capabilities of its predecessor ANFIS to provide better classification and prediction accuracy. Classification shows better accuracy when the k value is increased to a large value.

P.Yasodha et al [5] The objective of this study is to evaluate and investigate FIVE selected classification algorithms based on WEKA. The best algorithm in WEKA is J48 classifier with an accuracy of 70.59%

that takes 0.29 seconds for training. They are used in various healthcare units all over the world. The analysis had been carried out using a standard blood group data set and using the J48 decision tree algorithm implemented in WEKA. The research work is used to classify the diabetic patient's based on the gender, age, height & weight, blood group, blood sugar(F), blood sugar(PP), urine sugar(F), urine sugar(PP). The J48 derived model along with the extended definition for identifying regular patients provided a good classification accuracy based model.

Aiswarya Iyer et al [6] This paper aims at finding solutions to diagnose the disease by analyzing the patterns found in the data through classification analysis by employing Decision Tree and Naïve Bayes algorithms. This paper shows how Decision Trees and Naïve Bayes are used to model actual diagnosis of diabetes for local and systematic treatment, along with presenting related work in the field. Experimental results show the effectiveness of the proposed model.

Sabreena Jan et al [7] This paper focuses on the analysis of diabetes data by various data mining techniques which involve Naïve Bayes,J48,Multi layer Perceptron and K-star. The main goal of the research was to identify the most common data mining algorithms, implemented in modern Medical Decision Support Systems, and evaluate their performance on diabetic dataset. Four algorithms were chosen: Multilayer Perceptron, NaïveBayes, j48 and k-star. For evaluation, UCI diabetes database was used. Several performance metrics were utilized: percent of correct classifications, True/False Positive rates, AUC, Precision, Recall, F-measure and a set of errors. The results showed that for the present diabetes dataset, the percentage of correctly classified cases were highest in k-star followed by j48, multilayer perceptron and naïve bayes respectively. The root mean squared error was found minimum in j48.precision and recall was highest in multilayer perceptron. The underlying reason for such a research was the fact that no work was found which would analyze these four algorithms under identical conditions.

P. Suresh Kumar et al [8] This paper tries to diagnose diabetes based on the 650 patient's data with which we analyzed and identified severity of the diabetes. As part of procedure Simple k-means algorithm is used for clustering the entire dataset into 3 clusters i.e., cluster-0 - for gestational diabetes, cluster-1 for type-1 diabetes (juvenile diabetes), cluster-2 for type-2 diabetes. This clustered dataset was given as input to the classification model which further classifies each patient's risk levels of diabetes as mild, moderate and severe. Further, performance analysis of different algorithms has been done on this data to diagnose diabetes. The achieved results show the performance of each classification algorithm.

Saman Hina et al [9] In this research different classifying algorithms such as Naïve bayes, MLP, J.48, ZeroR, Random Forest, and Regression were applied to depict the result. The conducted research aims to extract knowledge from the given set of data and to generate comprehensive and intelligent results. In terms of performance, it was found that multi layer perception function is most effective hence it shows fewer errors however it takes too much processing time because it requires calculation of weights of each node. ZeroR is useful to determine baseline performance for others classification method. Naïve Bayes is also very efficient as it gives a predominant result after each validation but its performance is not quit impressive. J4.8 gives a graphical image of the precedence of the attribute as it calculates the priority of each attribute with other and yet it also predicts accurate results with least error hence it requires time.

S.Selvakumar et al [10] In this paper classification techniques such as Binary Logistic Regression, Multilayer Perceptron and K-Nearest Neighbor are classified for diabetes data and classification accuracy were compared for classifying data. From the analysis, it was examined that the formation of classifications will be different for classification methods. From the histogram, it was seen that the Binary Logistic Regression accuracy is 0.69, Multilayer Perceptron accuracy is 0.71 and KNN gives the accuracy of 0.80.k-Nearest Neighbor is higher than the accuracy of Binary Logistic Regression and Multilayer Perceptron.

III. Data mining techniques for diabetes prognosis

A. K-Star

K Star (K*), developed in 1995 by John G. Cleary and Leonard E. Trigg, provides a reliable approach to handle symbolic attributes, smoothness problems ,dealing with mixed values, real valued attributes and missing values. Space required for the storage is very large as compared to other algorithms. The K* algorithm can be defined as a method of cluster

analysis which mainly aims at the partition of „n" observation into „k" clusters in which each observation belongs to the cluster with the nearest mean. We can describe K* algorithm as an instance based learner which uses entropy as a distance measure. Instance-based (IB) learners, also called Memory-based ones, store the training instances in a lookup table and interpolate from these. IB learners are able to learn quickly from a very small dataset. Also, is important to note that they can use continue valued features and predict numeric valued classes because they retain each instance as a separate concept. New data instances, *x*, are assigned to the class that occurs most frequently amongst the k-nearest data points, yj, where j = 1, 2…k. Entropic distance is then used to retrieve the most similar instances from the data set. By means of entropic distance as a metric has a number of benefits including handling of real valued attributes and missing values. The K* function can be calculated as:

$$K^*(y_i, x) = -ln\ P^*(y_i, x)$$

Where *P** is the probability of all transformational paths from instance *x* to *y*. It can be useful to understand this as the probability that *x* will arrive at *y* via a random walk in IC feature space.

B. Multilayer perceptron neural network (MLPNN)

Multilayer perceptron neural networks (MLPNNs) are the most commonly used feed forward neural networks due to their fast operation, ease of implementation, and smaller training set requirements. It works on how different attributes results process and interact with one another and alter their results in such a way that the final outcome is the filtered version of each node (neuron). Multi-Layer perception bestows great advantages as it is used for pattern classification, recognition, prediction and approximation. In this study, we used a MLPNN model with single hidden layer of 5 hidden neurons.

In Figure **1**, a network of different layers namely input layer, hidden layer and output layer consisting of input nodes (green) or "neurons", output nodes (yellow) and some hidden nodes (red) some of them are visible.

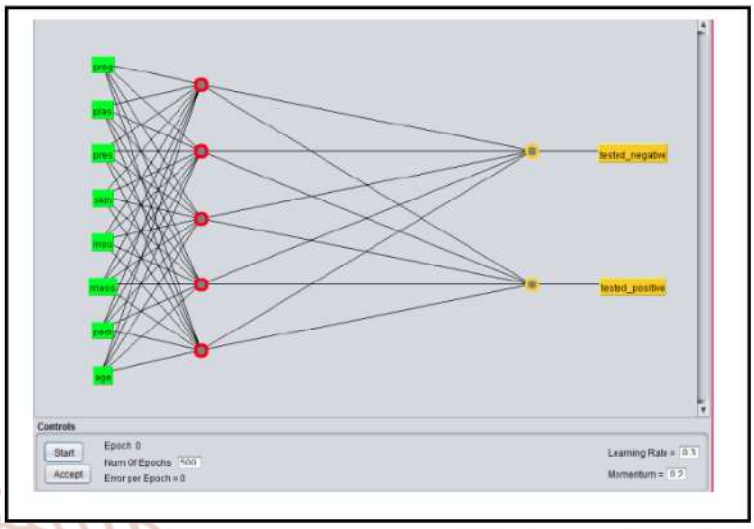


Figure 3: Neural network (MLPNN).

The nodes in the network are all sigmoid. Each connected network has some value in it which will be pass on to other nodes and each nodes perform a weighted sum of its input and pass it on until it generate some results. Hidden layer depends upon the complexity of the data.

C. K-means algorithm

Unsupervised algorithms are those algorithms that operate on unlabelled samples. That means the output is unknown even if the input is known. K means algorithm is one among the unsupervised learning algorithm. They take input parameter, number of clusters and n object data set partition into k clusters. Algorithm select k objects randomly. Based on the closeness of each object with corresponding cluster, each object is assigned to one cluster. Next step is to find the points that are closest to each other. To assign the object to the closest center, Euclidean distance is preferred. Once the objects are distributed to k clusters, the new k cluster centers are found by taking the mean of objects of k clusters respectively. The process is repeated till there is no change in k cluster centers.

Algorithm:
The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster

Input:
K: the number of clusters
D: a data set containing *n* objects.

Output: A set of K clusters.

Method:

(1) Arbitrarily choose k objects from Das the initial cluster centers;
(2) Repeat
(3) (re)assign each object to the cluster to which the objects is the most similar, based on the mean value of the objects in the cluster;
(4) Update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) Until no change;

D. Zero R

ZeroR is the simplest classification method. It is that type of classification method which would lean on the target and ignore other attributes invasion. The baseline for both classification and regression problems is called the Zero Rule algorithm. For a regression predictive modeling problem where a numeric value is predicted, the Zero Rule algorithm predicts the mean of the training dataset. Also called ZeroR or 0-R.ZeroR classifier simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods. It is really important to have a performance baseline on your machine learning problem. It will give you a point of reference to which you can compare all other models that you construct. For a classification predictive modeling problem where a categorical value is predicted, the Zero Rule algorithm predicts the class value that has the most observations in the training dataset.

V. CONCLUSION

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 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. In this paper, the various techniques are discussed to predict the diagnosis of diabetes. Using the data mining technique the health care management predicts the disease, diagnosis diabetes and then the health care management can alert the human being regarding diabetes based upon this prediction.

REFERENCES

1) P.Yasodha, M.Kannan . Analysis of a Population of Diabetic Patients Databases in Weka Tool. International Journal of Scientific & Engineering Research Volume 2, Issue 5, May-2011.
2) K. Rajesh, V. Sangeetha. Application of Data Mining Methods and Techniques for Diabetes Diagnosis. International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 3, September 2012.
3) Sukhjinder Singh, Kamaljit Kaur. A Review on Diagnosis of Diabetes in Data Mining. International Journal of Science and Research (IJSR). 2013.
4) Veena Vijayan V, Aswathy Ravikumar. Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus. International Journal of Computer Applications (0975 – 8887) Volume 95– No.17, June 2014.
5) P.Yasodha, N.R. Ananthanarayanan.Comparative Study of Diabetic Patient Data's Using Classification Algorithm in WEKA Tool . International Journal of Computer Applications Technology and Research Volume 3– Issue 9, 554 - 558, 2014 .
6) Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly. DIAGNOSIS OF DIABETES USING CLASSIFICATION MINING TECHNIQUES. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.
7) Sabreena Jan, Vinod Sharma, A Study of various data mining techniques for diabetic prognosis. International Journal of Modern Computer Science (IJMCS) ,Volume 4, Issue 3, June, 2016.
8) P. Suresh Kumar and V. Umatejaswi* . Diagnosing Diabetes using Data Mining Techniques. International Journal of Scientific and Research Publications, Volume 7, Issue 6, June 2017 .

9) Saman Hina*, Anita Shaikh and Sohail Abul Sattar. Analyzing Diabetes Datasets using Data Mining. Journal of Basic & Applied Sciences, 2017, 13 .

10) S.Selvakumar, K.Senthamarai Kannan and S.GothaiNachiyar. Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques. International Journal of Statistics and Systems. Volume 12, Number 2 (2017) , pp. 183-188 © Research India Publications http://www.ripublication.com