International Journal of Computer Trends and Technology Volume 67 Issue 4, 131-133, April 2019
ISSN: 2231-2803 / https://doi.org/10.14445/22312803/IJCTT-V67I4P126 © 2019 Seventh Sense Research Group®

Original Article

Data Mining Techniques for Prediction of Diabetes Disease

P. D. Bharsakle[1], G.R.Bamnote[2], P.K.Agrawal[3]

[1]*Prof. Ram Meghe Institute of Technology and Research, Badnera*
[2,3]*PRMITR, Badnera*

Abstract - *Nowadays, diabetes has become a common disease for all people, from young to old persons. The growth of the diabetic patients is increasing day by day due to various reasons such as viral or bacterial infection, toxic or chemical contents mixed with the food, autoimmune reaction, fatness, irregular diet, change in lifestyles, eating habits, environmental pollution, etc. therefore, diagnosing the diabetes is essential to save the human life from diabetes. A process of examining and identifying the hidden patterns from many data to conclude is data analytics. In health care, this analytical process uses machine learning algorithms to analyze medical data to build the machine learning models to carry out medical diagnoses. This paper exhibits a diabetes forecast framework to conclude diabetes. Additionally, this paper investigates the ways to deal with improving the exactness of diabetes expectations by utilizing medicinal information with different AI calculations and techniques.*

Keywords - *Support Vector Machine (SVM) K-Nearest Neighbor (K-NN) Prototype Nearest Neighbor Binomial Logistic Regression (BLR (PNN).*

I. INTRODUCTION

In this world, diverse sorts of Diabetes ailments, essentially named diabetes, a turmoil caused when the pancreas does not create Insulin or body cells never again react to Insulin. Our body cells are filled with Insulin which is one sort of hormone which goes about as a key that enables glucose from the blood to go into our cells. On the off chance that in the pancreas, the insulin-delivering beta cells are put down, at that point, the glucose in the blood isn't satisfactorily controlled. As a result, the blood glucose level increments unexpectedly, making an individual diabetic. On that point are principally four kinds of diabetes. Prediabetes is sorted by the glucose level higher than a typical yet not sufficiently high to be described as diabetes. Type1 diabetes is an immune system sickness that causes the insulin-delivering beta cells in the pancreas to be pulverized, restraining the body from having the capacity to yield enough Insulin to control the blood glucose levels successfully. Type 2 diabetes is a metabolic issue that outcomes from insulin obstruction; the body cells never again respond to insulin hormone, a circumstance in which cells neglect to use Insulin legitimately. Gestational diabetes alludes to higher than typical blood glucose levels during incubation in grown-up females who did not have diabetes before pregnancy. It is typically created between twenty-four and a twenty-multi week of gestation.[1]

Examining and identifying the hidden patterns from a large amount of data for concluding is called data analysis. This analytical process in health care uses machine learning algorithms to analyze the medical data for building machine learning models to carry out the medical diagnoses. A type of artificial intelligence (AI) called Machine learning enables a system to learn by itself and develop the knowledge models to make decisions by predicting the unknown data or label of the given data.

There are three rough categories of machine learning algorithms: supervised, unsupervised, and semi-supervised. Supervised learning algorithms are used when human expertise does not exist (navigating on Mars), but humans can not explain their expertise (speech recognition). The solution changes on time series basis (routing on a computer function), and the solution needs to be adapted to particular cases (user biometrics). The supervised learning algorithms are categorized into different types, such as probability-based, function-based,instance-based, etc. Unsurprised learning is used to describe or summarize the data. IT is a descriptive type of learning. Clustering and association rule mining are examples of unsupervised learning algorithms. The combination of supervised and unsupervised is called semi-supervised learning. The paper Reviews a diabetes prediction system to diagnose diabetics. o learn the diabetes data and to develop a diabetes prediction system for diagnosing diabetes, the supervised learning algorithm is used. The accuracy of this prediction system is improved using the pre-processing technique.

II. LITERATURE REVIEW

B.Tamilvanan, Dr .V. Murali Bhaskaran " An Experimental Study of Diabetes Disease Prediction System Using Classification Techniques" this paper, Explains the comparison of three different algorithms, and the results indicate the Naive Bayes algorithms can achieve a high accuracy rate along with minimum error rate when compared to other algorithms.

Aishwarya 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
In this paper, two algorithms, namely, J48 (decision tree algorithm) and Naïve Bayes, have been used to create the model for diagnosis. The data were divided into the training and test sets by the cross-validation and percentage split techniques. 10-fold cross-validation is used to prepare training and test data. After data pre-processing (CSV format), the J48 algorithm is employed on the dataset using WEKA (Java Toolkit for various data mining techniques), after which data are divided into "tested-positive" or "tested-negative" depending on the final result of the decision tree that is constructed. [2].

Basharat Naqvi, Arshad Ali, Muhammad Adnan Hashmi' and Muhammad Atif's " Prediction Techniques for Diagnosis of Diabetic Disease: A Comparative Study" In this paper uses the Knowledge Discovery Process (KDD) model, which consists of five steps of data selection, pre-processing, transformation, data mining, Interpretation. This paper introduces an expert system for the early prediction of a diabetic patient by using data mining classification techniques. This work is based on a dataset comprising 130-US Hospitals for the years 1999_2008 consisting of 50 attributes and more than 100,000 instances. The evaluation and comparison are performed using RapidMiner, a data science software platform that supports various machine learning steps, including results visualization. Random Forest, Decision Stump, Random Tree, and ID3 have been applied to mine the useful information from the data. The extracted information will assist the practitioner in writing the precise and wise prescription for diabetic patients.[3]

Classification is the process [2] of identifying a new observation category set based on a training set of data containing observations whose category is known. The cluster analysis [28] technique is used to group the objects according to their similarity. Studies have been made to compare the different techniques of classification which have been developed so far. In this study, the different classifiers have been compared, such as Naive Bayes, Decision Tree(c4.5), k-Means, Support vector machine(SVM), K-Nearest Neighbor Classifier(KNN), Prototype Neural Network (PNN), Binomial Logistic Regression(BLR), Multinomial Logistic Regression (MLR), partial least Square Regression-Discriminate Analysis (PLS-DA) Partial Least Square Partial Least Square-Linear-Linear Discriminate Analysis (PLS-LDA) and Apriori.

A. Decision Tree

Milan Zormanaet al [5] addressed the problem of mining rules from the diabetes database using a combination of decision trees and association rules. About 1251 different cases from the original database, with the help of association rule approaches, different trees are built and converted into a different set of rules. These rules were further reduced and filtered. The main objective of this rule was to analyze the number of rules generated and how many rules will be balanced after performing filtering and reduction. It also analyzes how many rules will be generated employing the association rule approach on the same database. The conclusion is that the sets of rules built by decision trees were much smaller than the results of association rules.

B. Support Vector Machine (SVM)

Support Vector Machines (SVM) are a moderately new-fangled type of learning algorithm originally introduced. Naturally, SVM aims at the point for the hyperplane that most excellent separates the data classes. SVMs have confirmed the capability to separate entities into correct classes accurately and identify instances whose established classification is not supported by the comparatively insensitive defined distribution of training examples of each class. SVM can be extended to perform numerical calculations. The first goal is to produce a linear function [6] that can make a fairly accurate target function in two such extensions. An extra extension is to learn to rank elements rather than produce a classification for individual elements. Ranking can be reduced to comparing pairs of instances and producing a+1 estimate if the pair is in the correct ranking order and -1 otherwise.

C. K-Nearest Neighbor (K-NN)

It is the nearest neighbor algorithm [6]. The k-nearest neighbor's algorithm is a technique for classifying objects based on the next training data in the feature space. It is the simplest among all machine learning algorithms. This algorithm is initialized by selecting k points in kd as the initial k cluster representatives or "centroids." Techniques for selecting the primary seeds include sampling at random from the dataset setting them as the solution of clustering a small

subset of the data, or perturbing the global mean of the data k-times. Then the algorithm iterates between two steps to the junction:
Step1: In Data Assignment, each data point is assigned to its adjoining centroid, with ties broken arbitrarily. This results in a partitioning of the data.

Step2: Relocation of "means" each group representative is relocated to the center of all data points. If the data points come with a possibility measure, then the relocation is to the expectations of the data partitions. In "kernelize," k-means though margins between clusters are still linear in the embedded high-dimensional [3]space, they can become non-linear when projected back to the original space, thus allowing kernel k-means to deal with more complex clusters. Dillon [5] has shown a close connection between kernel k-means and spectral clustering. The k-medoid algorithm is similar to k-means except that the centroids have to belong to the data set being clustered.

D. Prototype Nearest Neighbor (PNN)

Prototype NN [36] classification is easy to understand and easy-to-implement classification techniques. Despite its simplicity, it can perform well in many situations. The new prototype p is simply the average vector of weighted p1 and p2. The new prototype class is the same as the one of p1 and p2. Continue the merging process until the number of incorrect classifications of patients in T starts to increase.

E. Binomial Logistic Regression (BLR)

Predictive analysis in health care is primarily used to determine which patients are at risk of developing certain conditions, like diabetes, asthma, heart disease, and other lifetime illnesses. Additionally, sophisticated clinical decision support systems incorporate predictive analytics to support medical decision-making at the point of care. Logistic regression is a generalization of linear regression. It is used primarily for predicting binary or multi-class dependent variables.

III. CONCLUSION

The main goal of medical data mining algorithms is to get the best algorithms that describe given data from multiple aspects. These algorithms are very necessary to intend an automatic classification tool. The PLS-DA was the best one among the eleven. In the future, the data mining techniques have to be done more intrinsically to correlate diabetes with other diseases enabling the doctors and patients to a much more accurate and early detection of diabetes

REFERENCES

[1] B.Tamilvanan1, Dr.V. Murali Bhaskaran2, "An Experimental Study of Diabetes Disease Prediction System Using Classification Techniques”

[2] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly,” DIAGNOSIS OF DIABETES USING

[3] CLASSIFICATION MINING TECHNIQUES”, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015

[4] Basharat Naqvi, Arshad Ali, Muhammad Adnan Hashmi and Muhammad Atif ” Prediction Techniques for Diagnosis of Diabetic Disease: A Comparative Study,”

[5] Kaveeshwar, S.A., and Cornwall, J., 2014, "The current state of diabetes mellitus in India." AMJ, 7(1), pp. 45-48.

[6] Dean, L., McEntyre, J., 2004, “The Genetic Landscape of Diabetes [Internet]. Bethesda (MD): National Center for Biotechnology Information (US);. Chapter 1, Introduction to Diabetes. 2004 Jul 7.

[7] Mohammed, A.K., Sateesh, K. P., Dash G. N., 2013, “A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases” International Journal of Advanced Research in Computer Science and Software Engineering, 3(8), pp. 149-153.

[8] Chunhui, Z., Chengxia, Y., 2015, "Rapid Model Identification for Online Subcutaneous Glucose Concentration Prediction for New Subjects with Type I Diabetes," IEEE Transactions on Biomedical Engineering, 62 (5), pp. 1333 – 1344

[9] Vaishali, A., Harsh, K., Anil, KA, 2016, "Performance Analysis of the Competitive Learning Algorithms on Gaussian Data in Automatic Cluster Selection," 2016 Second International Conference on Computational Intelligence & Communication Technology.