International Journal of Scientific Research in Science, Engineering and Technology (www.ijsrset.com)
© 2020 IJSRSET | Volume 7 | Issue 1 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099

Analysis of Prediction of Diabetes by the help of Artificial Techniques

Gurwinder Singh[1], Mr. Siddharth Arora[2]

[1]Research Scholar, Department of Computer Science & Engineering, Guru Nanak Institute of Technology, Mullana- Ambala, Haryana, India

[2]Assistant Professor, Department of Computer Science & Engineering, Guru Nanak Institute of Technology, Mullana- Ambala, Haryana, India

ABSTRACT

Diabetes mellitus (DM) is a metabolic disease characterized by high blood sugar. The main clinical types are type 1 diabetes and type 2 diabetes. Now, the proportion of young people suffering from type 1 diabetes has increased significantly. Type 1 diabetes is chronic when it occurs in childhood and adolescence, and has a long incubation period. The early symptoms of the onset are not obvious, which may lead to failure to detect in time and delay treatment. Long-term high blood sugar can cause chronic damage and dysfunction of various tissues, especially eyes, kidneys, heart, blood vessels and nerves. Therefore, the early prediction of diabetes is particularly important. In this paper, we use supervised machine-learning algorithms like Support Vector Machine (SVM), Naive Bayes classifier and LightGBM to train on the actual data of 520 diabetic patients and potential diabetic patients aged 16 to 90. Through comparative analysis of classification and recognition accuracy, the performance of support vector machine is the best.

Keywords : Support Vector Machine, LightGBM, Naive Bayes classifier

I. INTRODUCTION

Predictive analytics use statistical or machine learning method to make a prediction about future or unknown outcomes [1]. It uses text mining for unstructured data, answers the question “what is next step?” It uses historical and present data to predict future regarding activity, behaviour and trends. To do this it makes use of statistical analysis techniques, analytical queries and automated machine learning algorithms. Predictive analytics need experts to build predictive models. Tese models are used for prediction. Tere are many applications of predictive analytics, out of which one is health care. A most common disease now a day’s is diabetes. People are sufering with it and the patient number increases day by day. Te World Health Organization (WHO) predicts that by 2030 there will be approximately 350 million people worldwide afected by diabetes [2, 3]. Mostly whatever food we eat is converted into glucose or sugar. Now, this glucose or sugar is used for energy. Glucose is transported to body cells through insulin. If the body does not produce sufcient insulin or does not make proper use of insulin then it leads to diabetes. Tere are four types of diabetes which are TYPE 1, TYPE 2, GESTATIONAL, PRE DIABETES. TYPE 1 diabetes is also known as insulin dependent diabetes [4] where the pancreas does not produce the hormone insulin. TYPE 2 diabetes is also known as noninsulin dependent diabetes [4] where adequate

insulin is produced but the body cannot make use of insulin. Gestation diabetes is a type of diabetes which occurs during pregnancy [5]. Pre diabetes refers to a situation where blood glucose levels are higher than normal but not so high to diagnosis as diabetes [6]. Diabetes is a disease in which blindness, nerve damage, blood vessel damage, kidney disease and heart disease can be developed [7]. By the use of predictive analytics in the feld of diabetes, diabetes diagnosis, diabetes prediction, diabetes self-management and diabetes prevention can be achieved as per the literature survey. Future the paper is organized into three sections. “Related work” gives the background of predictive analytics. “Clinical prediction model” describes diferent predictive models used in health care particularly for diabetes, follow.

Diagnosis of diabetes is considered a challenging problem for quantitative research. Some parameters like A1c [9], fructosamine, white blood cell count, fibrinogen and hematological indices [10] were shown to be ineffective due to some limitations. Different research studies used these parameters for the diagnosis of diabetes [11–13]. A few treatments have thought to raise A1C including chronic ingestion of liquor, salicylates and narcotics. Ingestion of vitamin C may elevate A1c when estimated by electrophoresis but levels may appear to diminish when estimated by chromatography [14]. Most studies have suggested that a higher white blood cell count is due to chronic inflammation during hypertension [15]. A family history of diabetes has not been associated with BMI and insulin [16]. However, an increased BMI is not always associated with abdominal obesity [17]. A single parameter is not very effective to accurately diagnose diabetes and may be misleading in the decision making process. There is a need to combine different parameters to effectively predict diabetes at an early stage

Literature Review :

Shetty et al. [18] used KNN and the Naïve Bayes technique has been used for the prediction of diabetes. Their technique was implemented as an expert software program, where users provide input in terms of patient records and the finding that either the patient is diabetic or not. Singh et al. [19] applied different algorithms on datasets of different types. They used the KNN, random forest and Naïve Bayesian algorithms. The K-fold cross-validation technique was used for evaluation. Ahmed [20] utilized patient information and plan of treatment dimensions for the classification of diabetes. Three algorithms were applied which were Naïve Bayes, logistic, and J48 algorithms. Antony et al. [21] utilized medical data for diabetes prediction. Naïve Bayes, function-based multilayer perceptron (MLP), and decision tree-based random forests (RF) algorithms were applied after pre-processing of the data. A correlation based feature selection method was employed to remove extra features. A learning model then predicted whether the patient was diabetic or not. Using a pre-processing technique, results were improved when employing Naïve Bayes as compared with other machine learning algorithms. Amina et al. [22] compared different data mining algorithms by using the PID dataset for early prediction of diabetes. Sellappan Palaniappan et al. [23] proposed a heart disease prediction system by using the Naïve Bayes, ANN and decision tree algorithms. Delen et al. [24] used logistic regression, ANN, and decision trees to predict breast cancer using a large dataset. Shadab Adam Pattekari and Asma Parveen [25] developed a web based application for prediction of myocardial infarction using Naive Bayes. Anuja Kumari and R. Chitra [26] used the SVM model to diagnose diabetes using a high-dimensional medical dataset.

As per literature, there are many forms of describing predictive analytics. It is inductive [8]. It doesn’t expect anything about data but it allows the data lead the way. It uses statics, machine learning, neural computing, robotics, computational mathematics and artifcial intelligence to explore all data and fnd meaningful relationships and patterns. Predictive analytics is a set of business intelligence (BI) technologies that uncover relationships and patterns within large volumes of data that can be used to

predict behaviour and events. To be more clear see the Fig. 1. Machine learning used by predictive analytics is a technique to train algorithm which can predict an output based on some input value. Tis leads to correlations and not to conclusions [9]. From previous work, there are two major types of predictive analytics such as supervised learning and unsupervised learning [8, 10]. Supervised learning is a process of creating predictive models using a set of historical data and produce predictive results. Examples are classifcation, regression and time-series analysis where as in Unsupervised learning does not use the previously known result to train its models. It uses descriptive statics. It identifes clusters or groups [8]. Further classifcation of predictive models are of nine types business rules, classifcation and decision trees [11–13] naive Bayes, linear regression [10–12], logistic regression [11, 12, 14], neural networks (NNs) [11–13], machine learning, support vector machines (SVMs), natural language processing (NLP) [11]. In paper [13] the author has described seven types of regression models, each holding importance of its own He listed seven types of regression models as linear regression, logistic regression, polynomial regression, stepwise regression, ridge regression, Lasso regression, elastic net regression. More versions of predictive models are described in two ways Smooth Forecast Model which describe smooth variable outcome for example proft and Scoring Model which describe binary outcome for example whether the blood report indicates disease or normal condition [9]. Another list of predictive models is linear models, decision trees, neural networks, clusters models, support vector machines, expert systems.

Methodology :

The dataset used in this study, is originally taken from the National Institute of Diabetes and Digestive and Kidney Diseases (publicly available at: UCI ML Repository [29]). The main Objective of using this dataset was to predict through diagnosis whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Many limitations were faced during the selection of the occurrences from the bigger dataset. The type of dataset and problem is a classic supervised binary classification. The Pima Indian Diabetes (PID) dataset having: 9 = 8 + 1 (Class Attribute) attributes, 768 records describing female patients (of which there were 500 negative instances (65.1%) and 268 positive instances (34.9%)).

- **Data preprocessing**

In real-world data there can be missing values and/or noisy and inconsistent data. If data quality is low then no quality results may be found. It is necessary to preprocess the data to achieve quality results. Cleaning, integration, transformation, reduction, and discretization of data are applied to preprocess the data. It is important to make the data more appropriate for data mining and analysis with respect to time, cost, and quality [30].

- **Data cleaning** Data cleaning consists of filling the missing values and removing noisy data. Noisy data contains outliers which are removed to resolve inconsistencies [31]. In our dataset, glucose, blood Pressure, skin thickness, insulin, and BMI have some zero (0) values. Thus, all the zero values were replaced with the median value of that attribute.
- **Data reduction** Data reduction obtains a reduced representation of the dataset that is much smaller in volume yet produces the same (or almost the same) result. Dimensionally reduction has been used to reduce the number of attributes in a dataset [32]. The principal component analysis method was used to extract significant attributes from a complete dataset. Glucose, BMI, diastolic blood pressure and age were significant attributes in the dataset.
- **Data transformation** Data transformation consists of smoothing, normalization, and aggregation of data [33]. For the smoothing of data, the binning method has used. The attribute of age has been useful to classify in five categories, Blood glucose concentration in patients who do not have diabetes is different from patients with diabetes. Glucose values have been divided into 5 categories [34]. A strong association has been found between healthy and

diabetic patients regarding their blood pressure levels [35]. Blood pressure has been divided into five different categories. The relationship between BMI and diabetes prevalence is consistent. The prevalence of diabetes and obesity is increasing concurrently worldwide. Furthermore, previous studies have shown that BMI is the most important risk factor for type 2 diabetes [36]. BMI values have been categorized into five classes. For the completion of the preprocessing task, selection of significant attributes and transformation of significant attributes into bins are done after data cleaning.

Association rule mining

Data mining techniques are also used to extract useful information to generate rules. Association rule mining is an important branch to determine the patterns and frequent items used in the dataset. It contains two parts:

1) Determine the frequent item set,

2) Generate rules.

An association rule mining approach was developed by Agrawal and Srikan in 1994 which was based on the performance analysis of a Walmart supermarket, buying products with the Apriori algorithm. Association rule mining plays an important role in medical as well as in commercial data analysis to detect and characterize interesting and important patterns. There are several methods to generate rules from data using association rule mining algorithms such as the Apriori algorithm, Tertius and predictive Apriori algorithms. Mostly, association rule-based algorithms are linked with Apriori, which make it a state-ofthe-art algorithm. Apriori works as an iterative method to identify the frequent item set in a given dataset, and to generate important rules from it. To determine the association between two item sets X and Y, there is a need to set the minimum support of that fraction of transactions which contains both X and Y called minsupp. The other important task is to set the minimum confidence that measures how often items in Y appear in transactions that contain X, known as minconf, to determine frequent item sets [37]. There were only 268 patients with diabetes in dataset, so only those instances were used to generate rules among them. To develop rules from a given dataset, set minimum support as 0.25 and minimum confidence as 0.9 to generate the following three different rules.The association of blood glucose, blood pressure, age, and BMI with diabetes also depended on socio economic, geographic, and clinical factors [38].

Modeling

Three models were used for early prediction of diabetes, following.

- **Artificial neural network (ANN)** The Artificial neural network (ANN) is a research area of artificial intelligence and an important technique which is used in data mining. The ANN has three layers: input, hidden, and output layer. The hidden layer consists of units that transform the input layer to the output layer. The output of one neuron works as the input for another layer. ANN detects complex patterns and learns on the basis of these patterns. The human brain contains billions of neurons. These cells are connected to other cells by axons and a single neuron is called as perceptron. Input is accepted by dendrites which is taken as stimuli. Similarly, the ANN is composed of multiple nodes that are connected with each other. The connection between units is represented by a weight. The objective of ANN is to convert input into significant output. Input is the combination of a set of input values that are associated with the weight vector, where the weight can be negative or positive. There is a function that sums the weight and maps the result to the output, such as y = w11 22 x wx + . The influence of a unit depends on the weighting; where the input signal of neurons meets is called the synapse. ANN works for both supervised and unsupervised learning techniques. Supervised learning was used in our study because the output is given to the model. In supervised learning, both input and output are known. After processing, the actual output with compared with required outputs. Errors are then back propagated to the system for adjustment. During

training, the data is processed many times, so that the network can adjust the weights and refine them [39].

- **Random forest (RF)** The random forest method is a flexible, fast, and simple machine learning algorithm which is a combination of tree predictors. Random forest produces satisfactory results most of the time. It is difficult to improve on its performance, and it can also handle different types of data including numerical, binary, and nominal. Random forest builds multiple decision trees and aggregates them to achieve more suitable and accurate results. It has been used for both classification and regression. Classification is a major task of machine learning. It has the same hyper parameters as the decision tree or bagging classifier. The fact behind random forest is the overlapping of random trees, and it can be analyzed easily. Suppose if seven random trees have provided the information related to some variable, among them four trees agree and the remaining three disagree. On the basis of majority voting, the machine learning model is constructed based on probabilities. In random forest, a random subset of attributes gives more accurate results on large datasets, and more random trees can be generated by fixing a random threshold for all attributes, instead of finding the most accurate threshold. This algorithm also solves the overfitting issue [40].
- **K- means clustering** Clustering is the process of grouping similar objects together on the basis of their characteristics. It is an unsupervised learning technique, in which we determine the natural grouping of instances given for unlabeled data. The clusters are similar to each other. However, the objects of one cluster are different from the objects of other clusters. In clustering, intra clustering similarity between objects is high and inter cluster similarity of objects is low. There are many type of clustering, such as partitioning and Hierarchal clustering but in this study, the kMeans clustering method was used. K-Means clustering is relatively simple to implement and understandable, and works on numerical data, in which K is represented as centers of clusters. Taking the distance of each datapoint from the center it assigns each instance to a cluster, and moves cluster centers by taking the means of all the data points s in a cluster and repeating until the cluster center stops moving

Results and Discussion :

Different classification algorithms were applied on our dataset, and results for all techniques were slightly different as the working criteria of each algorithm is different. The results were evaluated on the basis of accuracy and the AUROC curve. The accuracy of models was predicted with the help of a confusion matrix. First, the random forest algorithm was applied. Experiments were done to tune the model with respect to the number of decision tress and the maximum depth of the decision trees. In the first iteration, the number of decision trees was 8 and the depth of the trees were 4. Again while tuning the model and increasing the number of trees, the results were effective as compared to prior results. Increasing the number of decision trees could be used to obtain improved results, but when the number of trees reached 50, performance diminished. We obtained a best accuracy of 74.7% and an AUROC curve value of 0.806 when the number of decision trees was 32 and the depth of the decision trees was 4. and the confusion matrix is also shown in Fig. 4(A). After the random forest algorithm, the ANN was applied to obtain better results. The model was tuned on the basis of number of hidden neurons, number of learning iterations as well as value of initial learning weights. In first iteration, when number of hidden neurons were 50, number of learning iterations were 100 as well as the value of initial learning weights were 0.1, the model has provided satisfactory results. When the values of the tuned parameters were increased, the results worsened. In the 3rd iteration, the values of tuned parameters were decreased; then better results were obtained as compared to the 1st iteration. In the 4th iteration, results were obtained which were most effective when the number of hidden neurons was 5, the number of learning

iterations was 10, and the value of initial learning weights was 0.4. The AUROC curve of ANN is shown in Fig. 2(B), which has a value of 0.816 and an accuracy of 75.7%, calculated from confusion matrix The complete results of ANN for all iterations is described in The K-means clustering method was used after the RF and ANN implementation. To apply K-means clustering in our dataset, we normalized the dataset attributes by using the Min-Max normalization technique. Significant attributes were normalized, having the range of 0–1. K-Means clustering was applied by initially setting the value of K = 2, (as in our dataset only two types of patients exist), one for patients with diabetes and the second for patients without diabetes. When the number of clusters was increased, then accuracy decreased. The KMeans clustering predicted 273 to have a value of 1 (positive) and 495 as 0 (Negative). To evaluate the accuracy of K-means clustering, the results were compared to the target class, which shows 203 instances were classified incorrectly, as noted in the confusion matrix of Fig. 4(C). Both clusters were shown in Fig. 3, in which circles in the image show the incorrect instances. Incorrectly classified instances were 26.43% which show that the accuracy of K-means clustering method was 73.6%. Accuracy of the proposed models has been compared. The random forest method provided an accuracy of 74.7%, ANN gave 75.7% and Kmeans clustering method has given 73.6% accuracy. ANN is a nonlinear model that is straightforward and used for comparing statistical methods.

Table 1 Performanace evaluation of ANN

Number of Hidden nodes	Number of learning Iteration	Initial learning Weight	Accuracy	AUROC
50	100	0.1	74.2	0.799
100	1000	0.6	72.8	0.758
20	50	0.4	75.2	0.803
5	10	0.4	75.7	0.816

It is a nonparametric model, while the majority of statistical techniques are parametric and require a higher foundation of statistics. The main benefit of utilizing ANN over other statistical techniques is its capacity to capture the non-linear relationship among the concerned variables [42]. The primary weakness of the random forest method is that numerous trees can make the algorithm slow and inadequate for prediction in real time. This algorithm is quick to train, yet very moderate to make predictions once it is trained. A gradually more precise prediction requires more trees, which results in a slower model. Hence, these are the main reasons leading to ineffective results in our study [43].

Conclusion

Machine learning and data mining techniques are valuable in disease diagnosis. The capability to predict diabetes early, assumes a vital role for the patient's appropriate treatment procedure. In this paper, a few existing classification methods for medical diagnosis of diabetes patients have been discussed on the basis of accuracy. An classification problem has been detected in the expressions of accuracy. Three machine learning techniques were applied on the Pima Indians diabetes dataset, as well as trained and validated against a test dataset. The results of our model implementations have shown that ANN outperforms the other models. Using association rule mining, the results have shown that there is a strong association of BMI and glucose with diabetes. The limitation of this study is that a structured dataset has been selected but in the future, unstructured data will also be considered, and these methods will be applied to other medical domains for prediction, such as for different types of cancer, psoriasis, and Parkinson's disease. Other attributes including physical inactivity, family history of diabetes, and smoking habit, are also planned to be considered in the future for the diagnosis of diabetes.

II. REFERENCES

[1]. Brown DE, et al. Predictive analytics. Washington: IEEE Computer Society; 2015.

[2]. http://www.predictiveanalyticsworld.com/patimes/intro-to-machine-learning-algorithms-for-it-professionals-0620152/5580/. Accessed 2 July 2017.

[3]. http://www.who.int/diabetes/publications/en/screening_mnc03.pdf. Accessed 29 Mar 2017.

[4]. Sanakal R, Jayakumari T. Prognosis of diabetes using data mining approach-fuzzy C means clustering and support vector machine. Int J Comput Trends Technol. 2014;11(2):94–8.

[5]. Lakshmi KR, Kumar SP. Utilization of data mining techniques for prediction of diabetes disease survivability. Int J Sci Eng Res. 2013;4(6):933–40.

[6]. Repalli P. Prediction on diabetes using data mining approach. Stillwater: Oklahoma State University; 2011.

[7]. Motka R, et al. Diabetes mellitus forecast using diferent data mining techniques. In: Computer and communication technology (ICCCT), IEEE, 4th international conference. New York: IEEE; 2013.

[8]. Eckerson WW. Predictive analytics. Tdwi Research. 2006.

[9]. http://data-magnum.com/types-and-uses-of-predictive-analytics-what-they-are-and-where-you-can-put-them-towork/. Accessed 15 Apr 2017.

[10]. https://link.springer.com/chapter/10.1057%2F9781137379283_6#page-1. Accessed 5 July 2017.

[11]. Kalechofsky H. A simple framework for building predictive models. 2016.

[12]. Tevet D, et al. Introduction to predictive modeling using GLMs a practitioner's viewpoint.

[13]. https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/. Accessed 20 Apr 2017.

[14]. Gemson Andrew Ebenezer J. Big data analytics in healthcare: a survey. ARPN J Eng Appl Sci. 2015;10(8).

[15]. http://www.dummies.com/programming/big-data/data-science/data-science-for-dummies-cheat-sheet/.Accessed 30 Mar 2017.

[16]. Predictive modeling, Julie Chambers, the 56th annual Canadian reinsurance conference.

[17]. Abbott Analytics. Strategies for building predictive models. 2014.

[18]. Predictive analytics: poised to drive population health White Paper, Optum.

[19]. Duncan I. Introduction to predictive modeling. 2015.

[20]. https://www.linkedin.com/pulse/4-types-predictive-analytics-models-mark-rabkin. Accessed 3 July 2017.

[21]. http://234w.tc.tracom.net/healthcare/Pages/Diabetes-Readmission-Predictive-Analytics.aspx. Accessed 25 Mar 2017.

[22]. Lee YH, et al. How to establish clinical prediction models. Seoul: Korean Endocrine Society; 2016.

[23]. Lee J, et al. Development of a predictive model for type 2 diabetes mellitus using genetic and clinical data. Osong Public Health Res Perspect. 2011;2(2):75–82.

[24]. Plis K, et al. A machine learning approach to predicting blood glucose levels for diabetes management. Association for the Advancement of Artifcial Intelligence. 2014.

[25]. Yang Y, et.al. Forecasting potential diabetes complications. In: Proceedings of the twenty-eighth AAAI Conference on artifcial intelligence, Copyright c. Association for the Advancement of Artifcial. 2014.

[26]. Patil BM, et al. Hybrid prediction model for type-2 diabetic patients. Expert Syst Appl. 2010;37(12):8102–8.

[27]. Sarojini Ilango, B. et al. A hybrid prediction model with F-score feature selection for type ii diabetes databases. In: A2CWiC. 2010.

[28]. Temurtas H., et al. A comparative study on diabetes disease diagnosis using neural networks. Expert Syst Appl. 2009;36(4):8610–5.
[29]. Divya et al. Predictive model for diabetic patients using hybrid twin support vector machine. In: Proc. of int. conf. on advances in communication, network, and computing, CNC. Amsterdam: Elsevier; 2014.
[30]. Ahmed TM. Developing a predicted model for diabetes type 2 treatment plans by using data mining. J Theor Appl Inf Technol. 2016;90(2):181–7.
[31]. Devi MN, et al. Developing a modifed logistic regression model for diabetes mellitus and identifying the important factors of type II DM. Indian J Sci Technol. 2016; 9(4).Thirugnanam M, et al. Hybrid tool for diagnosis of diabetes. IIOAB J. 2016;7(5).
[32]. Osman AH, et al. Diabetes disease diagnosis method based on feature extraction using K-SVM. Int J Adv Comput Sci Appl. 2017;8(1).
[33]. Anand A. Prediction of diabetes based on personal lifestyle indicators. In: 2015 1st international conference on next generation computing technologies (Ngct-2015) Dehradun, India, 4–5 September 2015.
[34]. Jakhmola S. A computational approach of data smoothening and prediction of diabetes dataset. New York City: ACM; 2015.
[35]. AlJarullah AA. Decision tree discovery for the diagnosis of type II diabetes. In: International conference on innovations in information technology. New York: IE

Cite this article as :

Gurwinder Singh, Mr. Siddharth Arora, "Analysis of Prediction of Diabetes by the help of Artificial Techniques", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 7 Issue 1, pp. 348-355, January-February 2020.
Journal URL : https://ijsrset.com/IJSRSET229127