International Journal of Innovative Research in Technology & Science

ISSN: 2321-1156 Volume VII Issue I, January 2019

A REVIEW ON DIABETES DATASET DETECTION ANALYSIS USING DATA MINING ALGORITHMS

Dipak R Nemade, Computer Science & Engineering Department, SRK University, Bhopal,India;Nemade.dipak@gmail.com

Abstract

Data Mining is used for numerous functions in several applications like industries, medical etc. this can be used for extracting the helpful data from the large quantity of information set. Health observance is additionally used the information mining idea for predict the diagnosing of the diseases. In health observance diabetes is that the common health problem these days, that affects peoples. There are numerous data processing techniques and rule is employed for locating the polygenic disorder. Neural Network, Artificial neural fuzzy interference system, K-Nearest-Neighbor (KNN), Genetic rule, SVM and call Tree etc. These techniques and therefore the algorithms offer the higher result to the individuals and therefore the doctors concerning the diagnosing of the diabetes. From these results the individuals will predict he's affected with the diabetes or non-diabetes additional, performance analysis of various algorithms has been done on this information to diagnose diabetes. The achieved results show the performance of every classification rule.

Keywords: - Data Mining, Neural Network, SOM, Clustering Algorithm-Nearest-Neighbor (KNN), Machine Learning, Partitioning Clustering, Information Extraction, Hierarchical Clustering, classification.

I. INTRODUCTION

Nowadays data processing techniques and tools are wide utilized in almost each field like promoting, E -business, Retails, health care Systems etc. health care System is one in all the new emerging analysis areas wherever we will apply data processing techniques and tools. Our health care systems are made in intonation however poor in information thus there's large want of getting techniques and tools to extract the intonation from the large information set in order that diagnosis is done. diagnosis is thought to be a vital yet sophisticated task that must be executed accurately and expeditiously [1]. Diabetes is that the condition that results from lack of insulin in an exceedingly person's blood. There are different kinds of diabetes, like diabetes insipid USA. However, once individuals say "diabetes", they typically mean diabetes (DM)[2]. individuals with DM are referred to as "diabetics". Symptoms of high glucose embrace frequent elimination, increased thirst, and increased hunger. If left untreated, diabetes will cause several complications. Acute complications will embrace diabetic acidosis, nonketotic hyperosmolar coma, or death. Serious semi-permanent complications embrace heart condition, stroke, chronic renal failure, foot ulcers, and injury to the eyes. Once there's a rise within the sugar level within the blood, it's known as pre-diabetes. The pre-diabetes isn't thus high than the conventional worth. diabetes is because of either the exocrine gland not manufacturing enough insulin or the cells of the body not responding properly to the insulin created.

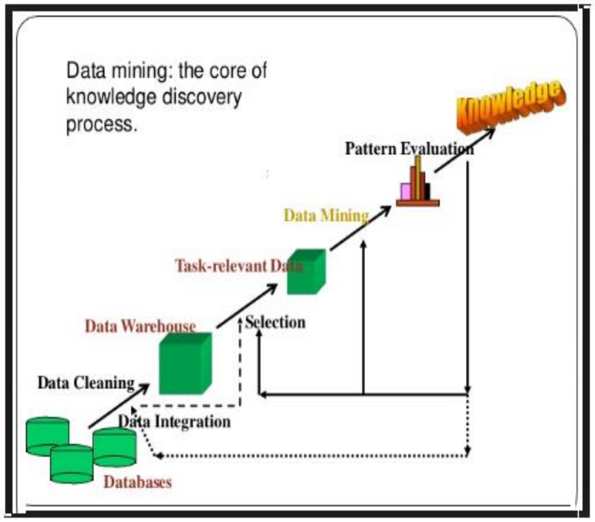

Fig 1 Knowledge Discovery Process

II. TYPES OF DIABETES MELLITUS

There are three main types of diabetes mellitus

(I) DM results from the pancreas's failure to supply enough internal secretion. this kind was previously named as "insulin-dependent diabetes mellitus" (IDDM) or "juvenile diabetes". The cause is unknown. The type-1 diabetes is suffering from the children and below twenty years older. In kind one the pancreatic cells can get affected and fail to perform. as a result of null secretion of internal secretion, the type-1 diabetic individuals suffer throughout their life and depend upon internal secretion injection. The type1 diabetic patients ought to frequently follow exercises and healthy diet as advised by dietitians.

(II) DM begins with internal secretion resistance, a condition within which cells fail to reply to internal secretion properly. because the disease progresses a scarcity of internal secretion may develop. this kind was antecedently cited as "non-insulin-dependent diabetes mellitus" (NIDDM) or "adult-onset diabetes". the foremost common cause is excessive weight and not enough exercise.

(III) Gestational diabetes is that the third main type and happens once pregnant girls while not a previous history of diabetes develops high blood glucose levels. in keeping with recent study of diabetes, it's found that around eighteen of pregnant girls have diabetes. pregnancy throughout older age could have a risk of developing the physiological state diabetes. type-2 polygenic disorder is often controlled by doing correct exercise and taking acceptable diet. If the glucose level isn't reduced by the on top of ways then medicines are often prescribed. National diabetes Statistics Report 2014 says that twenty-nine.1 million individuals or 9.3% of the U.S. population have diabetes [3].

As of 2015, a calculable 415 million individuals had diabetes worldwide, with kind two DM creating up regarding ninetieth of the cases. This represents eight.3% of the adult population, with equal rates in each ladies and men. As of 2014, trends instructed the speed would still rise. Diabetes a minimum of doubles an individual's risk of early death. From 2012 to 2015, more or less 1.5 to 5.0 million deaths annually resulted from diabetes. the world economic value of diabetes in 2014 was calculable to be US$612 billion. within the us, diabetes value $245 billion in 2012.

The recent estimates by the International diabetes Federation (IDF), with type2 there are regarding 366 million individuals in 2011 that got affected and by 2030 it should be accumulated to 552 million. virtually eightieth of the diabetic individuals belong to middle- and low-income countries. The high blood glucose patient will have cardiovascular disease, nephritis, strokes, and diabetic retinopathy [4]. the quantity of persons littered with type2 are accumulated by 2025. In Bharat the occurrences of DM are reduced by two.7% in geographic region when put next to urban area[5]. The prehypertension is attached with overweight, blubber and DM. The Indian Diabetic Risk Score (IDRS) found that someone WHO has traditional pressure however with high Indian diabetic risk score is alleged be hypertensive or diabetic [6]. Among all diabetes patients, ninetieth of cases are type-2 diabetes, and also the different 100 percent as type-1 and physiological state diabetes[7].

III. TYPES OF DATA MINING TECHNIQUES

Data mining techniques like cluster and classification will be accustomed study the health conditions of diabetic patients. during this paper, the information mining techniques like cluster and classification are applied to diagnose the kind of diabetes and its severity level for each patient.

(I) unsupervised machine learning technique :clustering could be an unsupervised machine learning technique that Cluster analysis or cluster is that the task of clustering a group of objects in such the simplest way that objects within the same group (called a cluster) are additional the same as different one another than to those in other teams (clusters). It's a main task of alpha data processing, and a typical technique for applied math information analysis, utilized in totally different fields in such the simplest way that has machine learning, pattern recognition, image analysis, data retrieval, bioinformatics, information compression, and lighting tricks [8].

(ii) supervised machine learning technique: Classification could be a supervised machine learning technique that assigns target categories to totally different objects or teams [9]. It's a two-step process: the primary step is model construction, that is employed to evaluate the training dataset of an information. The second step is model usage, wherever they made model is employed for classification. in step with the proportion of check samples or check dataset that are classified, the accuracy of the classification is calculable.

IV. RELATED WORK

Further, this paper comprises the following sections. The study of related work is presented in section 2. The proposed methodology is shown in section 3 and expect outcome in section 4 and finally section 5 concludes the work.

P Repalli et al. [10], in their research work predicted how likely the people with different age groups are affected by diabetes based on their life style activities. They also found out factors responsible for the individual to be diabetic. Statistics given by the Centers for Disease Control states that 26.9% of the population affected by diabetes are people whose age is greater than 65, 11.8% of all men aged 20 years or older are affected by diabetes and 10.8% of all women aged 20 years or older are affected by diabetes. The dataset used for analysis and modeling has 50784 records with 37 variables. They computed a new variable age_new as nominal variable, dividing in to three group's young age, middle age and old age and the target variable diabetes_diag_binary is a binary variable. They found 34% of the population whose age was below 20 years was not affected by diabetes. 33.9% of the population whose. age was above 20 and below 45 years was not affected by diabetes. 26.8% of the population whose age was above 45 years was not diabetic.

P. Thangaraju et al [11] Data mining is the practice of examining large pre- existing databases in order to generate new information. There are different kinds of data mining techniques are available. Classification, Clustering, Association Rule and Neural Network are some of the most significant techniques in data mining. In Health care industries, Data mining plays a significant role. Most frequently the data mining is used in health care industries for the process of forecasting diseases. Diabetes is a chronic condition. This means that is lasts for a long time, often for someone's whole life . This paper studies the comparison of diabetes forecasting approaches using clustering techniques. Here we are using three different kinds of clustering techniques named as Hierarchical clustering; Density based clustering, and Simple K-Means clustering. Mat lab is used as a tool.

Chang-Shing Lee et al.[12]. Proposed A Fuzzy Expert System for Diabetes Decision Support Application in which is a five-layer fuzzy ontology that includes different fuzzy layers to describe the knowledge with uncertainty. The layers such as a fuzzy knowledge layer, fuzzy group relation layer, fuzzy group domain layer, fuzzy personal relation layer, and fuzzy personal domain layer. The author developed semantic decision-making process in diabetic disease diagnosis. Even though the techniques are effective it has certain limitations such as the application was tested with a single dataset, the adaptation of the technique should be evaluated. The fuzzification approach is only applied in the fuzzy expert system is still more important rather than the ontology model. The approach suffers from the accuracy in disease diagnosis.

Santi Wulan Purnami et al. [13], in their research work used support vector machine for feature selection and classification of breast cancer and also emphasizes how 1-norm SVM can be used in feature selection and smooth SVM (SSVM) for classification. Two problems addressed here are, the first is to identify the importance of the parameters on the breast cancer. The second research problem is to diagnose breast cancer based on nine attributes of Wisconsin breast cancer dataset. To identify the importance of the parameters, the 1-norm SVM of the original data was done. The stronger parameters are as follows: parameter 1 (Clump thickness), parameter 3(Uniformity Of Cell shape), parameter 6 (Bare Nuclei), parameter 7 (Bland Chromatin), and parameter 9(Mitoses), while parameter 2 (Uniformity Of Belsize), parameter 4 (Marginal Adhesion), parameter 5(Single Epithelial Cell Size) and parameter 8 (Normal Nucleoli) are weaker. The obtained training and testing classification accuracy using 10-fold cross validation were 97.52% and 97.01% respectively. When one of the weak parameters was removed both training and testing shows a little decrease in accuracy

.

K. Polat et al. [14]. proposed a two-stage diagnostic system with the accuracy of 89.47%. In the first stage input features were reduced by using Principal component analysis (PCA). In the second stage adaptive neuro-fuzzy inference system was used for diagnosis.

Joseph L. Breault [15], In his research work used the publicly available Pima Indian diabetic database (PIDD) at the UCIrvine Machine Learning Lab. They tested data mining algorithms to predict their accuracy in predicting diabetic status from the 8 variables given. Out of 392 complete cases, guessing all are non-diabetic gives an accuracy of 65.1%. Rough sets as a data mining predictive tool applied rough sets to PIDD using ROSETTA software. The test sets were classified according to defaults of the naïve Bayes classifier, and the 10 accuracies ranged from 69.6% to 85.5% with a mean of 73.8% and a 95% CI .The accuracy of predicting diabetic status on the PIDD was 82.6% on the initial random sample, which exceeds the previously used machine learning algorithms that ranged from 66- 81%. Using a group of 10 random samples the mean accuracy was 73.2%.

G. Parthiban et al. [16] *The main objective of their research paper is to predict the chances of diabetic patient getting heart disease. In this study, we are applying Naïve Bayes data mining classifier technique which produces an optimal prediction model using minimum training set. They proposed a system which predicts attributes such as age, sex, blood pressure and blood sugar and the chances of a diabetic patient getting a heart disease. They used Naïve Bayes Classifier. It is a term dealing with simple probabilistic classifier based on applying Bayes Theorem with strong independence assumptions. The data set used in their work was clinical data set collected from one of the leading diabetic research institute in Chennai and contain records of about 500 patients. The clinical data set specification provides concise, unambiguous definition for items related to diabetes. The WEKA tool was used for Data mining. They used 10-fold cross validation. They found most of the diabetic patients with high cholesterol values are in the age group of 45 – 55, have a body weight in the range of 60 – 71, have BP value of 148 or 230, have a Fasting value in the range of 102 – 135, have a PP value in the range of 88 – 107, and have a A1C value in the range of 7.7 – 9.6.*

M. Durairaj et al [17] Neural Networks are one of the soft computing techniques that can be used to make predictions on medical data. Neural Networks are known as the Universal predictors. Diabetes mellitus or simply diabetes is a disease caused due to the increase level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. The Artificial Neural Networks (ANNs) based system can effectively applied for high blood pressure risk prediction. This improved model separates the dataset into either one of the two groups. The earlier detection using soft computing techniques help the physicians to reduce the probability of getting severe of the disease. The data set chosen for classification and experimental simulation is based on Pima Indian Diabetic Set from (UCI) Repository of Machine Learning databases**.** In this paper, a detailed survey is conducted on the application of different soft computing techniques for the prediction of diabetes. This survey is aimed to identify and propose an effective technique for earlier prediction of the disease.

Sumathy et al. [18]. Diabetes mellitus, in simple terms called as diabetes, is a metabolic disease, where a person is affected with high blood glucose level. Diabetes is a metabolic disorder caused due to the failure of body to produce insulin or to properly utilize insulin. This condition arises when the body does not produce enough insulin, or because the cells do not respond to the insulin that is produced. Blood glucose test is the crucial method for diagnosing diabetes. Also, there have been numerous computerized methods proposed for diagnosis of diabetes. All these methods have some input values which would be the result of different tests that should be carried out in hospitals. This paper proposes a methodology that aims to ease the patients undergoing various medical tests, which most of them consider as a tedious task and time consuming. The parameters identified for diagnosing diabetes have been designed in such a way that, the user can predict if he is affected with diabetes himself. Back Propagation algorithm is used for diagnosis.

IV. CONCLUSION

This paper presents a scientific review of literature involved with data processing ways and cluster techniques for diabetes} disease diagnosing .diabetes could be a chronic illness and a significant public health challenge. Day by day diabetic patients are increasing altogether over the globe. From this review, it's discovered that numerous procedure intelligent techniques are used for diabetes classification. Among those rules based mostly classification algorithms are extremely utilized by the researchers for diagnosing. Since these call rules are simply explainable and perceivable. For the long run analysis work, it's urged that for call rule generation roughest theory may also be wont to handle noisy, missing and uncertainty information. This survey paper concentrates regarding numerous data processing techniques and ways that are used for the first prediction of assorted diabetes from the medical information set of the patient. Since diabetes} could be a chronic disease. An early prediction of the illness can save the patient life. So applying data processing ways and techniques can helps to predict the diabetes and additionally scale back the treatment value. During this means data processing techniques are applied in medical information domain so as to predict diabetes and to search out economical ways that to treat them in addition.

REFERENCES

[1]. Soni, Jyoti, Ujma Ansari, Dipesh Sharma, and Sunita Soni. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." International Journal of Computer Applications, pp: 43-48, 2011.

[2]. Type-1 diabetes. Available from: http://www.diabetes.org/diabetes-basics/type

[3]. National Diabetes Statistics Report. 2014. Available from: http://www.cdc.gov /diabetes/pubs /statsreport14 /national-diabetes-report-web.pdf Type-2 diabetes in India: Challenges and possible. solutions. Available from: http://www.apiindia.org /medicine_update_2013 /chap40.pdf.

[4]. JaliMV, Hiremath MB. Diabetes. Indian Journal of Science and Technology. 2010 Oct; 3(10).

[5]. LanordM, Stanley J, Elantamilan D, Kumaravel TS. Prevalence of Prehypertension and its Correlation with Indian Diabetic Risk Score in Rural Population. Indian Journal of Science and Technology. 2014 Oct; 7(10):1498–503.

[6]. Diseases and conditions with subheading women's health. Available from: http://www.thehealthsite.com.

[7]. Han Kamber M. Data mining concepts and techniques. 2nd ed. Amsterdam, Netherlands: Elsevier Publisher; 2006. p. 383–5.

[8]. Gao, Denzinger J, James RC. CoLe: A cooperative data mining approach and its application to early diabetes detection. Proceedings of the 5th International Conference on Data Mining (ICDM'05); 2005

[9]. Repalli, Pardha. "Prediction on diabetes using data mining approach." Oklahoma State University (2011).

[10]. P. Thangaraju, B. Deepa, T. Karthikeyan, "Comparison of Data mining Techniques for Forecasting Diabetes Mellitus", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue No. 8, August 2014.

[11]. Lee, Chang-Shing, and Mei-Hui Wang. "A fuzzy expert system for diabetes decision support application." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 41.1 (2011): 139-153.

[12]. Santi Wulan Purnami, S.P. Rahayu and Abdullah Embong, "Feature selection and classification of breast cancer diagnosis based on support vector machine", IEEE 2008.

[13]. Polat, K., Gunes, S. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digital Signal Processing,17(4),(2007)702–710.

[14]. Breault, Joseph L., Colin R. Goodall, and Peter J. Fos. "Data mining a diabetic data warehouse." Artificial intelligence in medicine 26.1-2 (2002): 37-54.

[15]. G. Parthiban, A. Rajesh, S. K. Srivatsa, "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method", International Journal of Computer Applications, Volume 24– No.3, June 2011.

[16]. M. Durairaj, G. Kalaiselvi, " Prediction Of Diabetes Using Soft Computing Techniques- A Survey", International Journal of Scientific & Technology Research, Vol. 4, Issue No.3, March 2015.

[17]. Sumathy, Mythili, Dr. Praveen Kumar, Jishnujit T M, K Ranjith Kumar, "Diagnosis of Diabetes Mellitus based on Risk Factors", International Journal of Computer Applications, Vol.10, Issue No.4, November.2010 .