INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 4, Issue 6)
Available online at: www.ijariit.com

Diabetes analysis using machine learning methods

Harwinder Kaur
harwinder.cu11@gmail.com
Chandigarh University, Ajitgarh, Punjab

Gurleen Kaur
gurleen.cse@cumail.in
Chandigarh University, Ajitgarh, Punjab

ABSTRACT

In this paper, various kinds of algorithms are explained that include Support Vector Machine. The aim is to improve efficiency in different parameters by describing the classification approach for detecting diabetes. In this, it will predict diabetes with SVM.SVM will classify the data into positive and negative data points. In this, we predict the diabetes of Type 1and Type 2.Type 1is a type of diabetes that has no cure. Type 2 diabetes is common diabetes. It develops from the child. Diabetes is the fastest growing problem with more health and economic results. The increasing rate is predicted to increase to 430 million. Different types of data mining techniques are used. With SVM it will predict better accuracy. When we will predict the result with SVM, it will give accuracy. With the prediction of different parameters, we can predict the target value. With diabetes, there can be eye blindness, stress and many more can happen. With the help of data mining, we can aware of diabetes. In this paper, mention all the data mining techniques, types of classifiers. In the end, In this paper describe the diabetes types and what we have done and accuracy of the data. Type 2 diabetes is not easy to predict all the effects.

Keywords— *Data mining, Support vector machine, Classification, Type 2 diabetes, Pattern discovery*

1. DATA MINING

Data mining is related to knowledge data discovery with a large volume of data sets. Main aspects of data mining are cluster analysis and associative rule. In data mining there are different rules, techniques and algorithms are used that plays a very important role. In current years 'Predictive Diabetic Data' problem has been eying consideration. By using concerned techniques, the large volume of information produced within the systemic area might be generated into data to guide organized variety making.[1] In India, there are four main critical noncommunicable diseases like cancer, diabetes, sickness, and breathing illness.

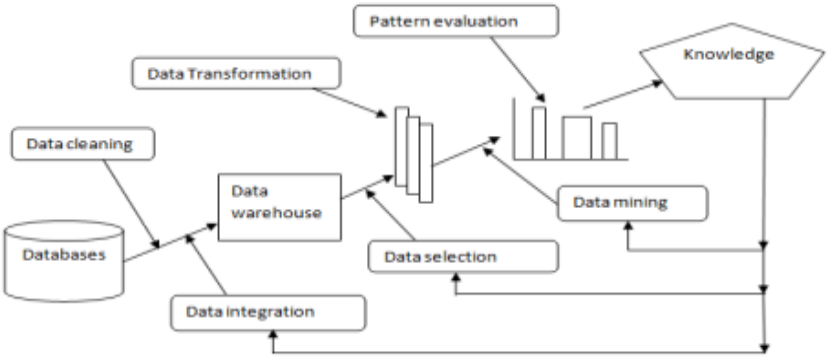

Fig. 1: Data mining process

1.1 Machine learning model

In the machine learning model, the learning process is divided into two steps:

- **Training**
- **Testing**

In training data, the machine know the patterns in the information, for the train the data make sure better accurateness and effectiveness of the algorithm cross-validation data is used. Test information is utilized to perceive how well the machine can anticipate new answers dependent on its preparation.

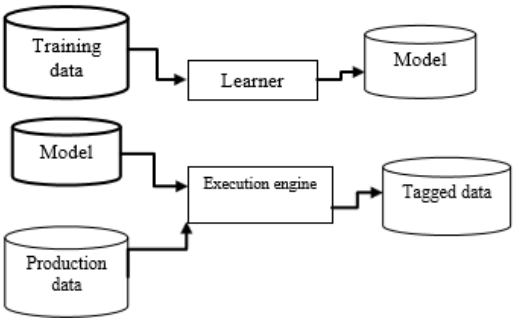

Fig. 2: Operational model of machine learning

2. INTRODUCTION

The main challenge of healthcare is diabetes. It increases day by day and affected the people in a very rapid manner. According to the World Health Organization, over the next 25 years, the growing number of people with diabetes is 130 million to 350 million. And the most unpleasant condition is that only one-half of the people attentive about diabetes. And from a medical point of view, diabetes leads to serve overdue problems. These problems contain micro and macrovascular changes and its results in renal problems, retinopathy, and heart disease. Like In the United State studies, diabetes is the fifth life-threatening disease but still no perfect treatment for this disease.

Diabetic retinopathy is a general problem of diabetes. Diabetes is more general disease, it effects on the eyes that causes blindness in the people. In developed countries as well as underdeveloped countries, the diabetes rate is increasing. It is observed that in developing countries 75% of people survive with diabetic retinopathy. In developing countries, there is no proper treatment of diabetes and their condition is so bad. In healthcare, they have estimated that 25% of people with diabetes are more suffered from blindness when compared to the individuals who have not diabetes. The most efficient healing of diabetes is in the early stages when it started. It can be prevented at the early stages with better treatment and screening machines.

Diabetes Mellitus (DM) is a continual, methodical and deadline disease. It happens when the pancreas does not undisclosed enough insulin or the body can't process it appropriately. The result can be irregular glucose intensity in the blood. Later on irregular glucose intensity cause harm to the blood vessel. This damage affects the body like kidney, eyes, nervous system, heart, and other organs. There are two types of diabetes. In Type1 diabetes, the human body fails to produce insulin [5]. Type 2 diabetes produces insulin but unable to use insulin properly.

For high performance and order methods, there is plenty of advances machines and sensors from different research fields producing information, including super goals advance microscope, mass spectrometry, magnetic resonance imagery and so forth. In spite of the fact that these advances create an abundance of information. The objective is to dig into the quickly collecting assemblage of natural information [1].

The power and adequacy of these methods are gotten from the capacity of comparable methods to extract patterns and make a model from information. The previously mentioned reality is especially noteworthy in the huge information period, particularly when the dataset can achieve terabytes or pet bytes of information. In such a field, a standout amongst the most vital research application is visualization and finding identified with human debilitating or potentially life quality decreasing diseases. One such infection is diabetes mellitus.

3. PREDICTIVE ANALYSIS ARCHITECTURE

This architecture includes many phases like data warehousing, data collection, processing analyzed reports and predictive analysis.

3.1 Data collection

The crude diabetes huge information or information index is feed as put into the framework. The formless large information can be acquired as of different Electronic health record (EHR)/Patient Health Record(PHR), Clinical frameworks and outer sources in various formats.

3.2 Data warehousing

At this stage, huge formless information warehoused into one unit and information from the different source is washed down, gathered and prepared for more handling.

3.3 Predictive analysis

The predictive analysis provides accurately healthcare providers and feedback to the patients' requirements. This gives the capacity to settle on monetary and clinical choices dependent on forecasts prepared by the framework. This framework uses the prescient examination calculation in Hadoop/Map Reduce condition to anticipate and arrange the kind of diabetes mellitus, entanglements related with it and the sort of cure to be given.

3.4 Hadoop

The open source Hadoop is appropriated information handling stage from apache. Hadoop can process the double role of information coordinator and examination instrument. Hadoop can process the very large volume of health information fundamentally by apportioning divided informational collections to various servers like bunches, every one of which tackles distinctive parts of the bigger issue and after that coordinate them for the last outcome. It contains two functions.

3.5 Pattern discovery

For diabetes management it is necessary to check the examples like plasma glucose focus, serum, insulin, diastolic circulatory starin, diabetes family, Body Mass Index (BMI), age, a number of time pregnant. It includes the following:

- **Association lead mining**: Relationship between diabetic kind and pages saw.
- **Clustering:** Grouping of similar examples of usage and so on.
- **Classification:** Classification of the well-being hazard an incentive by the level of patient wellbeing a condition. [4]

3.6 Types of classifiers

- **K-Nearest Neighbor:** K-means clustering algorithm is defined as the basic partitioning based method in which different clustering tasks has been utilized in order to perform a function within the low dimensional data sets. K is referred to as the parameters and by partitioning n objects the k clusters are generated. Within one cluster similar types of objects are grouped and in the separate clusters, dissimilar objects are placed. With the help of this algorithm, it is feasible to identify the cluster centres. At each data point, it is necessary to reduce the sum of the squared distances to the nearest centre of the cluster which is required.
- **Bayesian Classifier:** The most generalized approach for the supervised parametric classifiers theory is quadratic discrimination. These classifiers obtained the decision boundaries when it is required to deal with d-dimensions at the end this process becomes complicated. All the discriminate function has been done off-line for the computational generation. Due to dimensionality, this approach is more affected as there is quadratic discriminate in a large number of parameters which is necessary to manage. Its performance is affected drastically by the small training samples.
- **Multi-layer Perceptron (MLP):** In the"artificial neural network", the multi-layer perceptron classifier is the basic step. In order tom simple the process, a single hidden layer has been utilized initially after which for better classification performance they move towards two hidden layers. For each data set, the hidden units are chosen in a different manner. After various numbers of attempts, a number of hidden neurons is found out. In order to identify the number of hidden neurons a rule of thumb has been utilized as a total net weight is around n/10, wherever n is the total number of training points. The back-propagation algorithm is utilized in order to train the neural network [6].
- **SVM Classification:** SVM stands for "support vector machine", a classification algorithm that is based on optimization theory. As it maximizes the margin it is also known as a binary classifier. All the data points of an individual class are separated by the best hyper plane; this can be identified through the classification provided by SVM. In the SVM the largest the best hyper plane is described by the largest margin between the two classes. There are no interior data points when there is maximum width between the slabs parallel to the hyper plane which is also known as margin. The maximum margin in the hyper plane is separated by the SVM algorithm.

4. LITERATURE REVIEW

Han Wu, et.al (2018) proposed a novel model based on data mining techniques for predicting type 2 diabetes mellitus (T2DM). Nowadays, the influence of diabetes mellitus is increased and it affects more families. Due to this disease millions of people in the worldwide us suffering. The main objective of this paper is to improve the accuracy of the prediction model and to more than one dataset model is made adaptive in nature. The proposed model comprised of two parts based on a series of pre-processing procedures [11]. These two parts are improved K-means algorithm and the logistic regression algorithm. In order to compare the results with other methods, the Pima Indians Diabetes Dataset and the Waikato Environment was utilized for Knowledge Analysis toolkit. As per performed experiments, it is concluded that the proposed model show better accuracy as compared to other methods and also provide sufficient dataset quality. In order to evaluate the performance of the model, it is applied to another diabetes dataset, in which good performance is shown by both the methods.

Prova Biswas, et.al (2018) proposed a plasma methodology in which glucose-insulin homoeostasis model and plasma glucose measurement with a Bayesian non-linear filter are fused with each other. The proposed method was utilized to estimate the number of masses present in the stomach, intestine, plasma and tissue; insulin masses in the portal vein, liver, plasma, and interstitial fluid. There is Uncertainty present in the measurement of glucose [15]. By adding the process noise to the homeostasis model there is incorporation in the art model over the individual variations. This estimation is carried out for healthy people as well as type 2 diabetes mellitus patients. The truth followed by the estimator accurately in the process of simulation for both the cases. In this paper author compared the performance of two non-linear filters such as a Kalman filter (KF) and cubature quadrature KF in terms of root mean square error. This proposed methodology provides various useful applications such as observe a patient's insulin-glucose profile and for any hyperglycaemic patients it calculates the drug dose and third, for automated insulin delivery system it develops a closed loop controller.

Yu-Xuan Wang, et.al, (2017) analyzed various applications that provide significance of the data mining and machine learning in different fields. different data mining and machine learning techniques have been utilized to analyze the huge amount of data. It creates more commercial values in high-end enterprise systems. It becomes easy with the advancement in technology to use data mining and machine learning on personal computers or embedded systems that are typically low-end systems [13]. Research on the management designs of different components of the system is proposed as most of the work is done on the characteristics of the system that varies from time to time. The performance of the system with static or statically adaptive is optimized with the help of the proposed method in order to design system. The author in this paper proposed a method to design an operating system that uses the support of data mining and machine learning. With the help of this, it becomes possible to discover a new, automatized way by which optimization of the complex algorithms become simple and easy to use. For the validation of the proposed method cache design was utilized that automatically control the replacement of cached contents to make decisions. All the collected data from the system was analyzed when a reply is obtained from a data miner. As per performed experiments, it is concluded that the proposed method provides effective results.

Ioannis Kavakiotis, et.al (2017) presented a study related to Diabetes mellitus (DM) in order to detect which several techniques were developed [11]. Today, one of the greatest health challenges being faced is DM and for the prediction diagnosis and biomarker identification, lots of research has been done. There has been a huge increment in the data being generated on daily bases with the growth in biotechnology over the years. Thus, to diagnose and treat DM several machine learning and data mining approaches were proposed by researchers. The datasets were generated here by collecting data from the clinics and other biological fields. Improvements were made in techniques to provide better diagnostic results.

P. Suresh Kumar, et.al (2017) proposed a model that overcome all the problems such as clustering and classifications from the existing system by applying data mining technique. This method is used to diagnose the type of diabetes and from the collected data a security level for every patient. There are various effects of this disease due to which most of the research is done in this area [9]. All the collected data of the 650 patient's was used in this paper for the investigation purpose and its effects are identified. In order to cluster the entire dataset Simple k-means algorithm was used. It is divided into three datasets such as cluster-0 - for gestational diabetes, cluster-1 for type-1 diabetes, cluster-2 for type-2 diabetes. In the classification model, this clustered dataset was used as an input that is used for the classification process such as the patient's risk levels of diabetes as mild, moderate and severe. In order to diagnose diabetes, performance analysis of different algorithms was done. On the basis of the obtained result, the performance of each classification algorithm is measured.

Bayu Adhi Tama, et.al (2016) presented in this paper a chronic disease that causes major causalities in the worldwide that is Diabetes. As per International Diabetes Federation (IDF) around the world estimated 285 million people are suffering from diabetes [12]. This range and data will increase in nearby future as there is no appropriate method till date that minimizes the effects and prevents it completely. Type 2 diabetes (TTD) is the most common type of diabetes. The major issue was the detection of TTD as it was not easy to predict all the effects. Therefore, data mining was used as it provides the optimal results and helps in knowledge discovery from data. In the data mining process, support vector machine (SVM) was utilized that acquire all the information extract all the data of patients from previous records. The early detection of TTD provides the support to take an effective decision.

Jahin Majumdar, et.al, (2016) presented the most popular research areas in computer science that is data mining and machine learning is utilized in order to provide essential data or information. A huge amount of data is present in today's world hence it is difficult to collect only useful data as the size of data is getting increased. Therefore, it is necessary to invent a method that extracts useful information from data that will be helpful in industry and markets. Currently, the primary data mining algorithms have been utilized such as k-means, Apriori, PageRank and many more but these methods can be enhanced using machine learning as it knows the complex patterns more easily [15]. The SFS and SBS approaches are the optimal approaches and preferred as its uses with forwarding selection. SVM techniques are used by the proposed heuristic model as it provides the accuracy and heavy in the computational functions. The accuracy level of SVM is measured with the help of dataset. In order to improve the data classification and pattern recognition in Data Mining mainly feature selection various existing approaches were focused and experimented. As per performed experiments, it is concluded that comparison between the existing techniques was done in order to find out the best method. The theoretical limitations of existing algorithms were overcome by the proposed method.

Alexis Marcano-Cede˜no, et.al (2016) presented that diabetes is commonly found in all age group and it is a fatal disease. There are many issues are created by this disease in the body such as heart disease, kidney disease, blindness, nerve damage, and blood vessel damage. The major problem in the classification is that diabetes data was used via proper interpretation for the diagnosis of diabetic disease. In order to overcome the issue of diabetes, various techniques were applied that is based on artificial intelligence [14]. The main objective of this paper was to detect diabetes by applying the artificial met plasticity on multilayer perception (AMMLP) as a data mining (DM) technique. In order to check the performance of the proposed model AMMLP, the Pima Indians diabetes was used. On the basis of obtained results from the proposed model was compared with a decision tree (DT), the Bayesian classifier (BC) and other algorithms. The classification accuracy, analysis of sensitivity and specificity, confusion matrix and 10-fold cross-validation method was utilized to measure the robustness of the algorithm. It is concluded that the proposed method provides better performance as compared to DT and BC.

5. CONCLUSION

A chronic disease that causes major causalities in the worldwide that is Diabetes. As per International Diabetes Federation (IDF) around the world estimated 285 million people are suffering from diabetes. This range and data will increase in nearby future as there is no appropriate method till date that minimizes the effects and prevents it completely. Type 2 diabetes (TTD) is the most common type of diabetes. The major issue was the detection of TTD as it was not easy to predict all the effects. Therefore, data mining was used as it provides the optimal results and helps in knowledge discovery from data. In this work, it is concluded that diabetes prediction is applied using the approach of classifications. To implement prediction analysis, whole data is divided into training and test sets. The SVM classifier is applied for the prediction of diabetes. The performance of the classifier is analyzed in terms of certain parameters.

6. REFERENCES

[1] Cho, N., Shaw, J., Karuranga, S., Huang, Y., da Rocha Fernandes, J., Ohlrogge, A. and Malanda, B. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Research and Clinical Practice, 138, pp.271-281

[2] Kashyap, N., Singh, D. K., & Singh, G. K. (2017, October). Mobile phone-based diabetic retinopathy detection system using ANN-DWT. In Electrical, Computer and Electronics (UPCON), 2017 4th IEEE Uttar Pradesh Section International Conference on (pp. 463-467). IEEE.

[3] Bamgbose, S. O., Li, X., & Qian, L. (2017, October). Closed loop control of blood glucose level with neural network predictor for diabetic patients. In e-Health Networking, Applications and Services (Healthcom), 2017 IEEE 19th International Conference on (pp. 1-6). IEEE.
[4] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I. and Chouvarda, I. (2018). Machine Learning and Data Mining Methods in Diabetes Research.
[5] Barry, E., Roberts, S., Oke, J., Vijayaraghavan, S., Normansell, R., & Greenhalgh, T. (2017). Efficacy and effectiveness of screen and treat policies in the prevention of type 2 diabetes: systematic review and meta-analysis of screening tests and interventions. BMJ, 356, i6538.
[6] Eswari, T., Sampath, P., & Lavanya, S. (2015). Predictive methodology for diabetic data analysis in big data. Procedia Computer Science, 50, 203-208
[7] Yu, J.B. Byun, H.G., SO, M.S., & Huh,J.S.(2005).Analysis of diabetic patient's braeth with conducting polymer sensor array. Sensors and Actuators B: Chemical, 108(1-2), 305-308.
[8] Prasad, S. T., Sangavi, S., Deepa, A., Sairabanu, F., & Ragasudha, R. (2017, February). Diabetic data analysis in big data with the predictive method. In Algorithms, Methodology, Models, and Applications in Emerging Technologies (ICAMMAET), 2017 International Conference on (pp. 1-4). IEEE
[9] 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.
[10] M. Sharma, G. Singh, R. Singh, "Stark Assessment of Lifestyle Based Human Disorders Using Data Mining Based Learning Techniques", Elsevier, vol. 5, pp. 202-222, 2017.
[11] Han Wu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang, "Type 2 diabetes mellitus prediction model based on data mining", ScienceDirect, Vol. 11, issue 3, pp. 12-23, 2018.
[12]] Bayu Adhi Tama,1 Afriyan Firdaus,2 Rodiyatul FS, "Detection of Type 2 Diabetes Mellitus with Data Mining Approach Using Support Vector Machine", Vol. 11, issue 3, pp. 12-23, 2008.
[13] Yan Luo, Charles Ling, Ph.D., Jody Schuurman, Robert Petrella, MD, "GlucoGuide: An Intelligent Type-2 Diabetes Solution Using Data Mining and Mobile Computing", 2014 IEEE International Conference on Data Mining, Vol. 9, issue 8, pp. 12-23, 2014.
[14] 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, 2015.
[15] Alexis Marcano-Cede˜no, Diego Andina, "Data mining for the diagnosis of type 2 diabetes", IEEE, Vol. 11, issue 3, pp. 9-19, 2016.
[16] Jahin Majumdar, Anwesha Mal, Shruti Gupta, "Heuristic Model to Improve Feature Selection Based on Machine Learning in Data Mining", 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), vol. 3, pp. 73-77, 2016.
[17] Prova Biswas1,2, Ashoke Sutradhar3, Pallab Datta, "Estimation of parameters for plasma glucose regulation in type-2 diabetics in presence of meal", IET Syst. Biol., 2018, Vol. 12 Iss. 1, pp. 18-25, 2018.

