

Role of Data Mining In Prediction Diabetes

Neelmani Kumar¹, Kavita², Sahil Verma³, Anup Lal Yadav⁴

¹ Research Scholar, ,Lovely Professional University
neelmanikumax1@gmail.com

^{2,3}Department of Computer Science and Engineering,
Lovely Professional University, Phagwara, India
kavita.21914@lpu.co.in, sahil.21915@lpu.co.in

⁴Department of Computer Science and Engineering,
SRM Institute of Engineering and Technology, Haryana, India
anupsaran@gmail.com

*Corresponding author: sahil.21915@lpu.co.in

Abstract- Health Organizations around the globe are suggesting that Diabetes is a matter of global concern. In a study by the World Health Organization (W.H.O), there were 153 million people suffering from diabetes in 1980 which raised up to 422 million in 2016. It can be cured properly if it gets detected in the early stage. This paper is focused on a prediction of Diabetes in its early stage using different parameters. We have used a predictive data mining model for the research. These findings will help physicians to be able to take a proper decision of medication for their patients.

1. Introduction

Diabetes is a condition in which blood glucose (or blood sugar) level becomes too high. It is divided into two types in Type 1 Diabetes body do not produce the required level of Insulin which helps glucose to move into cells to make energy. Whereas in the case of Type-2 Diabetes, insulin is not getting produce or use as a result glucose level in the body starts increasing. A Higher level of glucose (or sugar) in the body leads to prediabetes. Prediabetic people have high chances of getting diabetes. High levels of glucose in the body can have a negative impact on kidneys, leads to loss of vision and can even cause heart stroke. Gestational Diabetes which occurs during pregnancy in women affects more than 1 million women in India.

Diabetes is one of the fastest-growing diseases worldwide. India holds 49 percent of diabetic people globally. Whereas 72.9 million people in India don't know they have diabetes. According to a W.H.O(World Health Organization), in India, this figure will rise up to 101.2 million diabetic people by 2030.

According to a study the amount of insulin which is used for the treatment of Diabetes will increase by 20 percent globally in the next 12 years.

Our objective is to design and implement a model that can predict the level of glucose in diabetes patients with the help of different parameters given to them. For this, we are going to use a regression-based data mining technique on the dataset. By predicting the level of glucose inside the body one can plan their daily habits and food as per the requirement. Our paper will help in the design of a new prediction framework based on data mining.

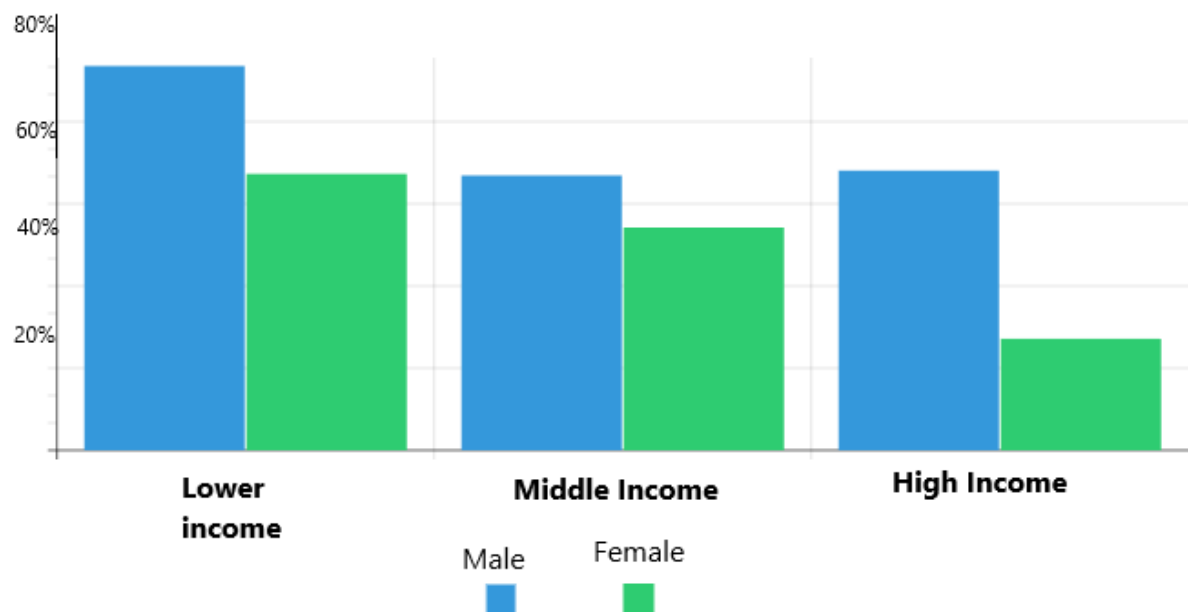


Figure 1: Income chart

2. Literature Review

The application of Data mining in medical fields is recently been gaining popularity. There is a large stretch of Data Mining in analyzing and enhancing the Medical field.

A. Data Mining model types

We have two types of existing Models in data mining: Descriptive model and a predictive model. A descriptive model is responsible for the finding in the patterns in the available and relationships between the different aspects(attributes) of the dataset. The predictive model is responsible for the prediction of those

occurrences which is likely to outcome based on the data retrieved from the database. This type of model has higher worthiness as a comparison to the other as it helps to prevent the occurrence. In this paper, we have adopted predictive classification techniques.

B. Data Mining in Health Care

Apriori algorithm is a type of classical algorithm in data mining which is used to evaluate the association between attributes. Ilayaraja and Meyyappan, used it in 2013 to find frequent diseases. In their dataset 1246 patients, records were used for the test of their system. They used data mining software called WEKA for finding the rate of occurrence of diseases hitting people in a year and the number of people affected by each disease.

Jianchao Han; Rodriguez, J.C., (2008) conducted Analysis of data on the PIMA dataset and build a decision tree predictive model using RapidMiner, they removed the columns that have a large number of missing values in the dataset.

According to (Srinivas,2010) The gain of information in the medical field can be achieved by taking the help of data mining techniques and for taking valuable insight from that. The relationship between cause and effect among symptoms, disorders, and effects can be brought and interpreted to make numerous predictions with the help of appropriate software.

In 2018 Ammar Aldallal et al. [13] found that Artificial Neural Network techniques can be used for prediction of Diabetes after preprocessing and by making clusters of available dataset. A network of multi-layer Perceptron Neural Network along with Backpropagation is adopted for training.

In 2018, Xuejiao Zhao et al. [2] used techniques like Deep Neural Network(DNN) and Support Vector Machine for processing of data which is the Pima Indian Data set. They made small changes in the parameters which enhanced the accuracy which came out to be 77.86% which was higher than any previous findings.

In the year of 2018, Popova and Pustozarov [3] put forward a system to track gestational diabetes. Which proved to be beneficial for physicians. This system analyses those data which are based on blood glucose, insulin injection, physical activity, diet and figures gathered from different centers. The insights gained are used in two kinds of proposals: At the point of self-government Personal consultation on food habits, made by the models based on glucose level in blood

and physician's fundamental proposals. Taking consideration of the insights gained from the patient's record and various journals.

In the year 2017, Verma and Mishra [4] used data classification by using the dataset of diabetes and cancer from the UCI repository. They used a suite of machine learning algorithm WEKA Explorer and Weka Experimenter interface. The duo took the leverage of REP tree classification, SMO, Naïve-Bayes, J48, and Machine Learning Predictive algorithms to perform the task of classification. The mentioned algorithms were employed to measure their performance on the datasets of diabetes by training, validating and testing process. The analysis of all the algorithms was for Naïve Bayes algorithm the accuracy of 72.70% for breast cancer and SMO yield an accuracy of 76.80% on the diabetes dataset.

In the year of 2017, Bhatia et al. [5] to remove the inconsistency in data he used k*-means clustering algorithm. Support Vector Machine (SVM) technique along with a genetic algorithm was employed for optimum attribute classification. The accuracy of the proposed model with the help of the Pima Indians Diabetes dataset got incremented by the exactness of 1.35%. The presented method scored more desirable result as compared to the present models.

In the year 2018, Thirumalai et al. [6] demonstrated diabetes a technique that could predict diabetes in the early stage. This is useful for people at various stages of life. This used a control chart as well as cost optimization. As we know diabetes is a condition that is caused by the increment of glucose beyond a defined level in blood. Diabetes can be determined by inspecting glycated hemoglobin test and the protein measured in the blood.

In the year 2017, Ara and Ara [7] proposed that the huge amount of data which is generated by medical services such as pharmaceuticals, research journals. A huge amount of structured and unstructured information is constantly generating out from medical devices, like motion monitors, implantable defibrillators, machines that measure sugar levels. The upcoming advanced gadgets and modern applications have made the generation of data at a faster pace. They have analyzed Zion China which is a specialized arrangement e-follow-up that was based on knowledge taken from on the site and various storages. In their quest, they have required quick, ingenious and sharp methods which bolster data gained from these machines to the storage place. The findings provide a comprehension of the method to deal with such huge data magnitude effectively.

3. Issues Findings

In our study the following problem has been identified:

1. For better class wise division and grouping of data, we need better clustering and classification techniques.
2. Multiple algorithms and classification techniques could have been applied for better results.
3. For finding flawless information we have a requirement of better procedures such as advanced algorithms so that absolute and missing sets can be distinguished.
4. There is a need for better procedures to handle missing values and biased set.

4. Description

The presented Table depicting the analysis between different Approaches that were made and about the methods and their limitations that were raised.

Table 1: Earlier Works Comparisons

Sl. No.	Sources	Approaches	Techniques used	Limitations
1.	[8]	Use of Digitization for clinical protocols for diabetes management.	Proposed a method to make diabetes management protocols, this helps experts for treatment.	Different cases and parameters are considered.
2.	[9]	KNN mean classification in Diabetes Type 1	They used a KNN algorithm of classification algorithm to predict postprandial glucose profile.	They gave suggestions working on corrections of individualization
3.	[10]	Attribute Extraction	It was used to analyze how people with different stages of life are being affected by diabetes based on their daily activities and to discover the factors responsible for people to	Feature selection with attribute categorization.

			get diabetic.	
4.	[11]	Exploration Mining in Diabetic patients	They have analyzed finished data on diabetic patients, the issues surfaced and on the application.	There is a requirement of applying a level-wise classification.
5.	[12]	Data mining approach	data mining techniques used in finding useful information to diminish the test and enhancement of accuracy rate. Diabetes is mostly perceived by young adults. The objective is to decrease the masses with diabetes.	It still needs to do more research in the real world

5. Conclusion

We have presented different approaches for reviewing methodologies for prediction of diabetes disease. This paper analyzed the different works that have been done in the field of Diabetes Prediction using Data Mining. It highlights the in-depth study and missing areas of previous works. In the future, there could be a web app in which people can enter their parameters and they can determine diabetes instantly. In the future, these works can also be elaborated on different diseases like Heart diseases, Depression.

References

- [1]. Mohammadi M, Hosseini M, Tabatabaee H. Using Bayesian Network for the prediction and Diagnosis of Diabetes. Bull. Env. Pharmacol. Life Sci. 2015; 4:109-14.
- [2]. Wei S, Zhao X, Miao C. A comprehensive exploration to the machine learning techniques for diabetes identification. In Internet of Things (WF-IoT), 2018 IEEE 4th World Forum on 2018 Feb 5 (pp. 291-295).IEEE

- [3]. Pustozarov E, Popova P. Mobile-based decision support system for gestational diabetes mellitus. In 2018 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBREIT) 2018 May 7 (pp. 45-48). IEEE.
- [4]. Verma D, Mishra N. Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques. In 2017 International Conference on Intelligent Sustainable Systems (ICISS) 2017 Dec 7 (pp. 533-538). IEEE.
- [5]. Bhatia K, Syal R. Predictive analysis using hybrid clustering in diabetes diagnosis. Recent developments in control, automation & power engineering 2017 (pp. 447-52). IEEE.
- [6]. Thirumalai C, SaiSharan GV, Krishna KV, Senapathi KJ. Prediction of diabetes disease using control chart and cost optimization-based decision. In Trends in Electronics and Informatics (ICEI), 2017 International Conference on 2017 May 11 (pp. 996-999). IEEE.
- [7]. Ara A, Ara A. Case study: Integrating IoT, streaming analytics and machine learning to improve intelligent diabetes management system. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) 2017 Aug 1 (pp. 3179-3182). IEEE.
- [8]. Thirumal PC, Nagarajan N. Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study. ARPN Journal of Engineering and Applied Science. 2015 Jan; 10(1).
- [9]. Aiello EM, Toffanin C, Messori M, Cobelli C, Magni L. Postprandial Glucose Regulation via KNN meal classification in Type 1 Diabetes. IEEE Control Systems Letters. 2018 Jun 5.
- [10] Thangaraju P, Nancy Bharathi G. Data Mining Approaches for Diabetes using Feature selection. International Journal of Computer Science and Information Technologies. 2014; 5(4): 5939-43.
- [11] Hsu W, Lee ML, Liu B, Ling TW. Exploration mining in diabetic patients databases: findings and conclusions. In Proceedings of the sixth ACM SIGKDD international conference of Knowledge discovery and data mining 2000 (pp. 430-436). ACM.

- [12] Girdhar Gopal Ladha, Ravi Kumar Singh Pippal. A computation analysis to predict diabetes based on data mining: A review, 2018 3rd International Conference on Communication and Electronics Systems (ICCES), 2018
- [13] Ammar Aldallal, Amina Abdul Aziz Al-Moosa."Using Data Mining Techniques to Predict Diabetes and Heart Diseases", 2018 4th International Conference on Frontiers of Signal Processing (ICFSP), 2018