International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869 (O) 2454-4698 (P) Volume-8, Issue-12, December 2018

Review on Prediction of Diabetes Mellitus using Data Mining Technique

Karnika Dwivedi, Dr. Hari Om Sharan

***Abstract*— Data mining plays a vital role in prediction of diseases in health care industry. Diabetes is one of the major health issues in the world. According to World Health Organization 2014 report, around 422 million people worldwide are suffering from diabetes. Diabetes is considered as one of deadliest and chronic disease which causes an increase in blood sugar. Many complications occur if diabetes remains untreated and unidentified. Data mining is a process of obtaining the information from a dataset and transforms it into unambiguous structure. Medical Data mining techniques are used to find hidden patterns in the data sets of medical domain for medical diagnosis and treatment. There are various data mining techniques for prediction of diseases like heart diseases, cancer, and kidney etc. Prediction of diabetes is a fastest growing technology. This paper helps in predicting polygenic disorder by applying data processing techniques. Using various data mining techniques we can predict Diabetes from the data set of a patient. This paper concentrates on the overall survey related to data mining techniques for predicting diabetes.**

***Index Terms*— Diabetes Mellitus, Data mining, Prediction, Classification, Clustering, Decision Tree**

I. INTRODUCTION

Diabetes Mellitus may be a chronic sickness that there's no illustrious cure except in terribly specific things management concentrates on keeping blood glucose levels as about to traditional as possible without causing hypoglycemia. This can be controlled with diet, exercise and use of applicable medications. Diabetes Mellitus happens throughout the planet and it's a lot of in developed countries. The increase in rates in developing countries follows the trend of urbanization and life style changes, including a "western-style" diet. This is because of less awareness.

Data Mining is used to invent knowledge out of data and exhibiting it in a condition that is easily understandable to humans. It is a process to inspect large amounts of data collected. Information technology plays a vital role for implementing the Data mining techniques in various sectors like banking, education, etc. [2] .In the field of medical domain data mining can be effectively used for the prediction of diseases by using various data mining techniques. There are two predominant goals of data mining tend to be prediction and description. Prediction involves some variables or fields in the data set to predict unknown or future values of other variables of interest. Description focuses on finding the patterns detailing the data that can be interpreted by humans. Basic conception of growth and characteristics affecting diabetes from external sources is very much essential before constructing predictive models. The idea is to predict the diabetes and to find the factors responsible for diabetes using data mining methods [3]. Data mining techniques can be used for early prediction of the disease with greater quality in order to save the human life and it will also reduce the treatment cost. According to International Diabetes Federation stated that 382 million people are affected with diabetes worldwide. By 2035, this will be doubled as 592 million [5].This paper analyzes the Diabetes prediction using various DM techniques. Some of the most important and popular data mining techniques are classification, clustering, prediction, Naive Bayes, Decision Tree are analyzed to predict the diabetes disease.

The purpose of information mining is to extract helpful information from massive databases or data warehouses. Data mining applications square measure used for commercial and scientific sides. Data mining is method of choosing, exploring and modeling large amounts of data in order to discover unknown patterns or relationships which provide a clear and useful result to the data analyst. KDD method might consists many steps: like information choice, data cleaning, data transformation, pattern searching i.e. Data mining, finding presentation, finding interpretation and finding evaluation [6].

1.1. Diabetes: Diabetes is a long lasting disease and its affects people worldwide. It happens when the body is not capable of producing enough insulin. Insulin which is secreted by pancreas, one of the most important hormones in the body, which is needed to maintain the level of glucose. Diabetes may be controlled with the assistance of hormone injections, a healthy diet and regular exercise. Diabetes leads to much other disease such as blindness, blood pressure, heart disease, and kidney disease etc. There are four types of diabetes

Type 1 Diabetes: It occurs when the pancreas is not capable of producing insulin. Insulin is a hormone produced by the pancreas. Type 1 diabetes can occur at any age. It will occur most commonly among children and young people [7]

Type 2 Diabetes: It occurs when the amount of insulin is not sufficient for the body needs. Due to family heredity, old age, obesity increases the risk of getting type 2 diabetes. Mostly occurs at the age of 40 [9]

Gestational Diabetes: It is the third main form, majorly occurs with the pregnant women due to excess blood sugar level in the body [4]

Pregestational Diabetes: Pregestational diabetes occurs when the insulin-dependent diabetes before becoming pregnant [8]

Karnika Dwivedi, M. Tech Scholar, Department of Computer Science, Rama University Kanpur, Uttar Pradesh, India

Dr. Hari Om Sharan, Assistant Professor, Department of Computer Science, Rama University Kanpur, Uttar Pradesh, India

Diabetes Mellitus is a chronic disease for which there is no known cure except in very specific situations management concentrates on keeping blood sugar levels as
close to normal as possible without causing hypoglycemia.
This can be controlled with diet, exercise and use of appropriate medications.
Diabetes Mellitus occurs throughout the world and it is more in developed countries. The increase in rates in developing countries follows the trend of urbanization and life style changes, including a "western-style" diet. This is because of less awareness.
The purpose of data mining is to extract useful information from large databases or data warehouses. Data mining applications are used for commercial and scientific sides [1].
Data mining is process of selecting, exploring and modeling large amounts of data in order to discover unknown patterns or relationships which provide a clear and useful result to the data analyst [2].
KDD process may consists several steps: like data selection, data cleaning, data transformation, pattern searching i.e. data mining, finding presentation, finding interpretation and finding evaluation [3].

www.advancejournals.org

Open Access Scientific Publisher

The Diabetes Prediction plays an important role in data mining technique. There are various Data mining technique applied for the prediction of diabetes. These are the Data mining techniques which is given below have been applied for predicting diabetes.

II. EXISTING DATA MINING TECHNIQUES

Decision Tree: Decision tree is possibly the most popular data mining technique. Decision tree is one of most important classifier which is easy and simple to implement. Decision tree uses a decision tree as a predictive model used in data mining. In this research, decision tree is used to predict disease from patients data along with classification technique. Decision trees are quick to construct and easy to interpret. Prediction based on decision trees is well ordered. It handles the huge amount of extent data. It is more suitable for searching the knowledge discovery. Finally the results attained from Decision Tree are easier to clarify and read [8].

Naive Bayes: Naive Bayes is based on Bayes theorem with speculation between predictors. The Naive Bayes enables to quickly create models that provide predictive abilities and also provide a new method of traversing and understanding the data. This technique can be applied to predictive analysis when building a predictive model with Naive Bayes.For this all the input attributes must be comparatively independent. A Naive Bayesian model is easy to construct and it has no complexive iterative parameters. Naive Bayes can be a powerful predictor. This technique is very useful for very large data sets. Naive Bayes performs persistently before and after lowering of attributes with the same representation of construction time. The envision provided for Naive Bayes are easy to understand [3].

K-nearest neighbor's algorithm (k-NN): is the one of the important method for classifying objects based on closest training data in the feature space. It is simplest among all machines learning rule however, the accuracy of K-NN rule is degraded by presence of yelling options.

Classification via Clustering: Classification is one of the procedure used for the prediction of diabetes. Classification is the most eminent data mining tasks. Large amount of business and medical data sets usually involves classification. Classification is a data mining function that can allocates the items in a collection to target categories. Classification refers to assigning cases into division based on a predictable attribute. Each case contains a set of attributes one of which is the class attribute i.e. predictable attribute. The job requires finding a model that describes the class attribute as a function of input attributes. In data mining tools classification measures with discovering the problem by distinguishing the features of diseases between patients and diagnose or predict which algorithm shows best performance [9].The main purpose of classification is to exactly predict the target class for each case in the data. This technique may be used as a preprocessing step before storing the data into the classifying model. To find a group of data we need to cluster in order to gather everything based on their characteristics. And aggregating them according to their similarities. Clustering is also called as segmentation. It is used to locate unprocessed groupings of cases based on a set of attributes. Cases within the same group having more or less matching attribute values. Clustering is an individual data mining task. Which means no single attribute is used to model the training process. All the input attributes are treated equally. The attribute values need to be formalized before clustering to eliminate the higher value attributes influencing the lower value attributes.

Neural Network: Artificial neural networks are well suited to tackle problems that people are not good at solving, like prediction and pattern recognition. Neural networks have been applied within the medical domain for clinical diagnosis, image analysis and interpretation, signal analysis and interpretation and drug development [6].

III. LITERATURE SURVEY

Gyorgy J. Simon, Pedro J.Caraballo, et al., [1] projected the strategy of spacing association rule mining to spot sets of risk factors and also the corresponding patient subpopulations that square measure considerably enlarged risk of progressing to diabetes. And to find sets of risk issue, here uses bottom up summarization algorithmic program that produces most fitted outline that describes subpopulations at high risk of polygenic disease. The Subpopulation identified by this summary covered most high risk of patients, had low overlap and were at very high risk. This methodology is employed for once the patient having high risk. Dr. Zuber khan, shaifali sing, et al., [10] worked on the concept of Diabetes Mellitus using K-nearest Neighbor algorithm which is most Important technique of Artificial Intelligence. The accuracy rate is showing that how many outputs of the data of the test dataset are same as the output of the data of different features of the trained dataset. The error rate observation that what percentage outputs of the information of the take a look at dataset aren't same because the output of the information of various options of the training dataset. The result they showed that because the worth of k will increase, accuracy rate and error rate will increase. K-Nearest Neighbor

algorithm is one of the most important techniques of AI which is used widely for diagnostic purposes. Through KNN more Accurate results can be obtain. This methodology is incredibly effective for the coaching knowledge set that is incredibly massive. The objective of the research paper, "Predicting Diabetes by consequence the various Data Mining Classification Techniques" describes the various Data Mining Classification Techniques. There are many classification techniques used in this paper for predicting diabetes [2]. The analysis paper, "Disease Prediction in Data Mining Technique"– A Survey. The malady prediction plays a vital in data processing. This paper analyzes about various diseases like Heart disease prediction, Breast cancer prediction, Diabetes by using many techniques like Classification, Clustering, Decision Tree, Naive Bayes methods in order to predict the diabetes disease. This paper also tells about predictive and descriptive type about the data. Prediction involves some fields in the data set to predict the values of other variables. On the other hand Description focuses on finding patterns of the data that can be interpreted by humans. .The Research paper, "Analysis of various Data Mining Techniques to Predict Diabetes Mellitus", concentrates about overall population affected by diabetes worldwide. This paper also predicts about the overall population affected by diabetes will also double the rate of diabetes of the population by the upcoming years. This Paper aims about the early prediction of the diabetes will save the life of the human. The paper analyzes about the three types of diabetes and their causes. It also uses the prediction, classification technique .This provides the higher accuracy for the disease prediction [5]. The research paper, "Review on Prediction of Diabetes using Data Mining Technique", elaborates about detailed review of existing data mining methods used for prediction of diabetes. It also gives about the types of diabetes disease Type1, type2, and type3.The aim of the diabetes is to predict the diabetes with the help of Data mining methods such as the K-Nearest Neighbor Algorithm, Bayesian Classifier, Naive Bayesian Classifier, Bayesian Network, all the methods are used for prediction of diabetes. This paper also mentions about the effects of diabetes on patients [7]. The research paper, "A survey on Naive Bayes Algorithm for Diabetes Data Set Problems", explores about various Data mining algorithm approaches of data mining that have been utilized for diabetic disease prediction. In this paper Classification and Naive Bayes is one of the most used algorithm for the prediction of disease[9].The research paper," Prediction of Diabetes Mellitus Techniques", describes about the Decision Tree, Naive Bayes, K-nearest neighbor's algorithm (k-NN),Classification and Clustering. By using this effective algorithm method diabetes prediction can be done [4]. The research paper, "A Comparative Study of Classification Algorithms for Disease Prediction in Health Care. This paper describes about Diseases Prediction; Classification algorithm; Data Mining, Decision tree. The main aim of this paper is to find out best classifier from different classification algorithm that can be used to predict disease on applying data set of the patients [8].

IV. CONCLUSION

Different approaches for the prediction of diabetes and its varieties square measure focused during this study. This paper concentrates about various data mining techniques and methods which are used for the early prediction of various diabetes. Data mining is a techniques used to extract useful information from existing large volume of data which enables you to gain more knowledge. Therefore applying Data mining methods and techniques will helps to predict the diabetes and also reduces the treatment cost. In this way data mining techniques are applied in medical data domain in order to predict diabetes and to find out efficient ways to treat them as well

REFERENCES

[1] GyorgyJ.Simon,Pedro J.Caraballo,Terry M. Therneau,Steven S. Cha, M. Regina Castro and Peter W.Li "Extending Association Rule Summarization Techniques to Assess Risk Of Diabetes Mellitus," IEEE Transanctions on Knowledge and Data Engineering ,vol 27,No.1,January 2015

[2] P. Radha , Dr. B. Srinivasan, "Predicting Diabetes by cosequencing the various Data Mining Classification Techniques",IJISET - International Journal of Innovative Science, Engineering & Technology Vol. 1 Issue 6, August 2014

[3] K.Priyadarshini1 , Dr.I.Lakshmi2 "A Survey on Prediction of Diabetes Using Data Mining Technique" International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization) Vol. 6, Special Issue 11, September 2017.

[4] Haldurai Lingaraj, Rajmohan Devadass, Vidya Gopi, Kaliraj Palanisamy," PREDICTION OF DIABETES MELLITUS USING DATA MINING TECHNIQUES":A REVIEW, Journal of Bioinformatics &Cheminformatics,Feburary 19,2015

[5] Dr.M.Renuka Devi, J.Maria Shyla,"Analysis of various Data Mining Techniques to Predict Diabetes Mellitus", International Journal of Applied Engineering Research ISSN 0973-4562 Vol 11, Number 1(2016)

[6]WEBSOURCE:https://www.researchgate.net/publication/273023827_PREDICTION_OF_DIABETES_MELLITUS_USING_DATA_MINING_TECHNIQUES_A_REVIEW

[7] Vrushali Balpande, Rakhi Wajgi," Review on Prediction of Diabetes using Data Mining Technique", International Journal of Research and Scientific Innovation (IJRSI) |Volume IV, Issue IA, January 2017 | ISSN 2321–2705.

[8] Isha Vashi, Prof. Shailendra Mishra,"A Comparative Study of Classification Algorithms for Disease Prediction in Health Care", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 9, September 2016.

[9] Nilesh Jagdish Vispute, Dinesh Kumar Sahu, Anil Rajput,"ASurvey on naive Bayes Algorithm for Diabetes Data Set Problems", International journal for research in Applied Science & Engineering Technology (IJRASET),Volume 3 issue XII,December 2015

[10] Dr.Zuber khan, shaifali singh and Krati Sexena," Diagnosis of Diabetes Mellitus using K- Nearest Neighbor Algorithm," in proceeding of International Journal of Computer Science Trends and Technology, vol.2 , July-Aug 2014