ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

A Survey on Prediction of Diabetes Using Data Mining Technique

K.Priyadarshini[1], Dr.I.Lakshmi[2]

PG.Scholar, Department of Computer Science, Stella Maris College, Teynampet, Chennai, Tamil Nadu, India [1]

Assistant Professor, Department of Computer Science, Stella Maris College, Teynampet, Chennai, Tamil Nadu, India[2]

ABSTRACT: This paper helps in predicting diabetes by applying data mining techniques. Data mining is a process of obtaining the information from a dataset and transforms it into unambiguous structure. Medical data mining has been a great capability for finding hidden patterns in the data sets of the medical domain. Diabetes is one of the major global health issues.There are many Data mining techniques for the prediction of diseases like heart diseases,cancer,kidneystones,etc.Prediction of Diabetes is an emerging and fastest growing technology. Using various Data mining techniques we can predict Diabetes from the data set of a patient. This research paper concentrates on the overall survey related to various Data mining techniques for predicting diabetes.

KEYWORDS: Diabetes, Data mining,Prediction,Classification,Clustering,Navie Bayes,Decision Tree, KDD, Association,Regression.

1. INTRODUCTION

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 [1].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. Data mining techniques are used for variety of applications like education domain, banking etc.

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 can be controlled with the help of insulin 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 childrens and young peoples [7]

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

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 [3]

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 [6]

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

1.2. Classification

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[3].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. ame elements called as clusters. 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.

1.3. 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.

1.4. 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 [6].

2.5. Knowledge discovery database

Knowledge discovery in databases (KDD) is the process of discovering knowledge and information from the collection of data. This process is also called as a data mining technique. It includes the data preparation, data selection, and data cleansing and establishing knowledge on the data sets and then interpreting the right solutions from the observed result. Data mining is an important step of KDD. It is used for the mining of the data from the database.KDD consists of an iterative sequence of Data integration, Data mining pattern recognition and knowledge presentation

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

2.6. Regression

Regression is a widely used data mining function. Regression is similar to classification. Regression is used to predict relentless values that can predicts a number, Age, weight, temperature,disease.All these can be predicted using regression techniques. Regression tasks can solve many medical domain problems. Linear regression and logistic regression are the most popular regression methods [8].

2.7. Association

Association is one of the eminent data mining function that invents the probability of the items in a collection. The relationships between coexisting and the items are demonstrated as association rules. Association is also called market based analysis. In terms of association, each values or more generally each attributes or value pair is considered as an item. The association task has two goals, one is to find frequent item sets and other is to find association rules. Association modeling has used as an important practice in other domains as well. Association is also related to other application sectors such as bioinformatics, medical diagnosis, Web mining and scientific data analysis. Association rules are created by examining data for repeated patterns and using the measure for support and confidence to recognize the most extensive relationships. Support represents the frequently of the items visible in the database. Confidence represents the number of times the statements have been found is to be true [8]

III. LITERATURE SURVEY

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 Research paper, "Disease Prediction in Data Mining Technique"– A Survey. The disease prediction plays an important role in data mining. 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 different algorithm of data mining are used in the field of medical prediction are discussed in this paper[1].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,"ASurvey 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[3].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 methods 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[6].

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

IV. PROPOSED METHODOLOGY

Mining patient's data to predict diabetes disease or to make decision by using patient's personal and medical information, it requires steps to be followed. Steps for proposed model are given below.

4.1. Data collection

The first step consists of data collection from healthcare organizations such as hospitals, laboratories and medical centers. This data consists about patient's personal information such as patient name, age, address, contact number, height, weight and medical information such as blood pressure level, size of the body, sugar level and their laboratory results.

4.2. Data preprocessing

Data preprocessing is a widely used data mining technique. It is a process that involves transforming raw data into an understandable way. The Data which is collected is mostly incomplete, inconsistent,ambigious and may contain many errors. Data preprocessing is a proven method for rectifying such issues. Data goes through a series of steps during preprocessing. There are several data preprocessing techniques.

4.2.1. Data Cleaning

First and foremost the raw data should be cleansed in such a way that the data should not contain any missing values, errors and noises. And removing the inconsistent values or data.

4.2.2. Data Integration

Data with the different representations and formats are integrated together within the merged data.

4.2.3. Data Transformation

Data is scale down into a normalized form

4.2.4. Data Reduction

This technique is used to reduce the unwanted representation of the data into the usable data.

4.2.5. Data Discretization

This technique involves the reduction of the number of values of an attribute by dividing the range of attribute intervals.

The Second step proposed of transforming and preprocessing of data. Data collected from healthcare organizations are obtained into one single form understandable by data mining tool. Data comes from different resources each with its own different form. This different form of data needs transformation and preprocessing. This transformation and preprocessing consists of the following steps.

Data cleaning
Matching
Combining
Removing noisy and relevant data
Standardization

4.3. Data storage

The third step consists of data storage process. In data storage, converted data is stored into a single database with the same format. So, this data can be used to apply the data mining techniques.

4.4. Data analysis

After storage of the data, the other step is data analysis. Data analysis is the most crucial phase in this proposed model. It consists of the following procedure:

- First it includes the applying data mining techniques or classification algorithms on patient's data being loaded dataset into data mining tool.
- Next it computes classification results of algorithm.

4.5. Classification results

The last step of proposed model composed of classification results of algorithm and it includes computing the best classifier for the dataset obtained according to the study. After classifying the result, comparison of this result will be appear and according to different parameters of algorithm like accuracy, efficiency, best algorithm will be chosen which will be helpful to doctors, physicians to predict diabetes and help them to make decision for patients data [7].

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

V. CONCLUSION

This survey paper concentrates about various data mining techniques and methods which are used for the early prediction of a various diabetes from the medical data set of the patient. Since diabetes is a chronic disease. An early prediction of the disease will save the patient life. 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] S.Vijiyarani S.Sudha,“ Disease Prediction in Data Mining Technique”– A Survey,International Journal of Computer Applications & Information Technology Vol. II, Issue I, January 2013 (ISSN: 2278-7720)
[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] 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
[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] 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.
[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] Web resource: http://conexion.it-nova.co/wpcontent/uploads/2014/12/DataMiningSQLServer2008.pdf