Diagnosing Diabetes using Data Mining Techniques

P. Suresh Kumar and V. Umatejaswi*
Kakatiya Institute of Technology and Science, Warangal, TS, India
uma.vuda@gmail.com

Abstract- Diabetes is a disease which is affecting many people now-a-days. Most of research is happening in this area. In this paper, we proposed a model to solve the problems in existing system in applying data mining techniques namely clustering and classifications which are applied to diagnose the type of diabetes and its severity level for every patient from the data collected. This paper tries to diagnose diabetes based on the 650 patient's data with which we analyzed and identified severity of the diabetes. As part of procedure Simple k-means algorithm is used for clustering the entire dataset into 3 clusters i.e., cluster-0 - for gestational diabetes, cluster-1 for type-1 diabetes (juvenile diabetes), cluster-2 for type-2 diabetes. This clustered dataset was given as input to the classification model which further classifies each patient's risk levels of diabetes as mild, moderate and severe. Further, performance analysis of different algorithms has been done on this data to diagnose diabetes. The achieved results show the performance of each classification algorithm.

Index Terms- Classification, Clustering, Data Mining Techniques, Diagnosis of Diabetes, Expert Clinical System, Naive Bayes, Random Tree, C4.5, Simple Logistic.

1 INTRODUCTION

Diabetes is the condition that results from lack of insulin in a person's blood. There are other kinds of diabetes, like diabetes insipidus. However, when people say "diabetes", they usually mean Diabetes Mellitus (DM)[1]. People with DM are called "diabetics".

Symptoms of high blood sugar include frequent urination, increased thirst, and increased hunger. If left untreated, diabetes can cause many complications. Acute complications can include diabetic ketoacidosis, nonketotic hyperosmolar coma, or death.Serious long-term complications include heart disease, stroke, chronic kidney failure, foot ulcers, and damage to the eyes.When there is an increase in the sugar level in the blood, it is called pre-diabetes. The pre-diabetes is not so high than the normal value.

Diabetes is due to either the pancreas not producing enough insulin or the cells of the body not responding properly to the insulin produced. There are three main types of diabetes mellitus:

- Type 1 DM results from the pancreas's failure to produce enough insulin. This type was previously referred to as "insulin-dependent diabetes mellitus" (IDDM) or "juvenile diabetes". The cause is unknown. The type-1 diabetes is affected by the young people and below 20 years of age. In type 1 the pancreatic cells will get affected and fail to function. Because of nil secretion of insulin, the type-1 diabetic people suffer throughout their life and depend on insulin injection. The type1 diabetic patients should regularly follow exercises and healthy diet as suggested by dietitians.
- Type 2 DM begins with insulin resistance, a condition in which cells fail to respond to insulin properly.As the disease progresses a lack of insulin may also develop. This type was previously referred to as "non-insulin-dependent diabetes mellitus" (NIDDM) or "adult-onset diabetes". The most common cause is excessive body weight and not enough exercise.
- Gestational diabetesis the third main form and occurs when pregnant women without a previous history of diabetes develop high blood sugar levels. According to recent study of diabetes, it is found that around 18% of pregnant women have diabetes. Pregnancy during older age may have a risk of developing the gestational diabetes.

One of the main reasons for type-2 diabetes is Obesity. The type-2 diabetes can be controlled by doing proper exercise and taking appropriate diet. If the glucose level is not reduced by the above methods then medicines can be prescribed. National Diabetes Statistics Report 2014 says that 29.1 million people or 9.3% of the U.S. population have diabetes [2].

As of 2015, an estimated 415 million people had diabetes worldwide, with type 2 DM making up about 90% of the cases. This represents 8.3% of the adult population, with equal rates in both women and men. As of 2014, trends suggested the rate would continue to rise. Diabetes at least doubles a person's risk of early death. From 2012 to 2015, approximately 1.5 to 5.0 million deaths each year resulted from diabetes. The global economic cost of diabetes in 2014 was estimated to be US$612 billion. In the United States, diabetes cost $245 billion in 2012.

The recent estimates by the International Diabetes Federation (IDF), with type2 there are about 366 million people in 2011 that got affected and by 2030 it may be increased to 552 million. Almost 80% of the diabetic people belong to middle- and low-income countries. The high blood sugar patient can have heart disease, kidney failure, strokes, and diabetic retinopathy [3]. The number of persons affected by type2 will be increased by 2025. In India the occurrences of diabetes mellitus are reduced by 2.7% in rural area when compared to urban area[4]. The prehypertension is affiliated with overweight, obesity and diabetes mellitus. The Indian Diabetic Risk Score (IDRS) found that a person who has normal blood pressure but with high Indian diabetic risk score is said be hypertensive or diabetic [5].

Among all diabetes patients, 90% of cases are type-2 diabetes, and the other 10% as type-1 and gestational diabetes[6].

Data mining techniques such as clustering and classification can be used to study the health conditions of diabetic patients. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in different fields in such a way that includes machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics[7].

Classification is a supervised machine learning technique that assigns target classes to different objects or groups[8]. It is a two-step process: The first step is model construction, which is used to evaluate the training dataset of a database. The second step is model usage, where the constructed model is used for classification. According to the percentage of test samples or test dataset that are classified, the accuracy of the classification is estimated.

In this paper, the data mining techniques such as clustering and classification are applied to diagnose the type of diabetes and its severity level for every patient.

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 followed by experimental results in section 4 and finally section 5 concludes the work.

2 Related Works

According to the World Health Organization (WHO) there are about 350 million people suffering from diabetes mellitus (DM) and diabetes will become the seventh leading cause of death worldwide by 2030. The deaths due to diabetes are expected to rise by 50% during the next 10 years. The number of diabetic persons is increasing in every country, 4 out of 5 people with diabetes live in low and middle income countries and half of diabetics don't know they suffer from this disease. This global epidemic could be largely attributed to the rapid increase in the rates of overweight, obesity and physical inactivity.

Gao et al., [9] presented an approach known as CoLe which tracks the diabetes in its early stage. CoLe which is a multi-agent system framework that runs multiple miner agents as well as a combination agent. The main aim of CoLe is to achieve a better knowledge in which it presentsthe data in different methods.

Rajesh et al., [10] used various classification algorithms like ID3, C4.5, LDA, Naïve Bayes, K-NN for diagnosing diabetes for the given dataset. The author concluded that C4.5 is the best algorithm with less error rate of 0.0938 and more accuracy value of 91%.

Afrand et al., [11] presented Artificial Intelligence to design an advanced clinical system for diagnosing diabetes. The author presented Extended Classifier System (XCS) in which we got high accuracy when compared with the other data mining techniques.

Adidela et al., [12] presented the type of diabetes by using Fuzzy ID3 method. The author uses the system for predicting the disease from data set as it initially clusters the data and applies the classification algorithms on clustered data. The author presented a combination of classification method where they developed EM algorithm for clustering and fuzzy ID3 algorithm to attain decision tree for each cluster.

Patil et al., [13] applied Apriori algorithms to classify type-2 diabetes. The author presented four association rules for the class value "yes", and for the class value "no", the author presented ten association rules. For increasing the dataset quality, the preprocessing methods are applied.

Aljarullah et al.,[14] proposed J48 algorithm to diagnose type-2 diabetes which is used for constructing a decision tree. The accuracy of the model is 78.68%.

Jaya Rama Krishnaiah et al., [15] presented a new framework known as duo-mining tool which is used for diagnosing diabetes. The author also applied many classification algorithms like KNN, SVM, decision Tree for type-2 diabetes. Among all algorithms SVM algorithm has the highest accuracy value of 96.39%.

Mandal et al., [16] used hierarchical clustering algorithm to discover the different models for controlling diabetes mellitus.

Kavitha et al., [17]initiated CART Method for predicting the type of Diabetes. The algorithm shows the differences between high risk and low risk patients. The accuracy of this algorithm is 96.37%

Ferreira et al., [18] used different classification algorithms like SimpleCart, J48, Simple Logistics, SMO, NaiveBayes and BayesNet for diagnosing neonatal jaundice in type1 diabetes. Among all algorithms, it was found that Simple Logistics as the best algorithm.

Ananthapadmanaban et al., [19] developed the SVM and Naïve Bayes classification algorithms for speculating diabetic retinopathy and found out that the Naive Bayes algorithm has got the accuracy rate of 84%.

SantiWulanPurnami et al., [22] proposed a diagnosis model for diagnosing breast cancer by considering feature selection methods and classification techniques.

PardhaRepalli[23] tries to predict the diabetes of a patient by applying different data mining techniques and analyzed the data based on mining strategies to give predictions to the patients.

Joseph L. Breault [24] proposed a diagnosis database for data mining techniques in which the data is clustered and classified by using different algorithms like c4.5, Naive Bayes

G. Parthiban et al, [25] applied Naïve Bayes method to diagnose heart related problems which are occurring in diabetic patients.

P. Padmaja[26] proposed a model where in different clustering techniques was used to characterize diabetes data and analyzed it to get different evaluations.

National Center for Chronic Disease Prevention and Health Promotion presented gestational diabetes [20] which shows the Centers for Disease Control and Prevention.

Type-2 diabetes complications [21] are controlled by doing proper exerciseand taking appropriate diet. If the glucose level is not reduced by the above methods then medicines can be prescribed

3 Methodology

In the proposed model, Simple K-means clustering algorithm is used for predicting the type of diabetes. The classification algorithms like Random Tree, Naive Bayes, C4.5 and simple

Logistics are used to predict the risk levelsof diabetes patients who are aware of their diabetes type.

3.1. Discussion

The model proposed in this paper has three stages.

- Stage-1: Data pre-processing.
- Stage-2: Applying Simple K-Means algorithm to the dataset for clustering the data into three clusters as cluster-0 (gestational diabetes), cluster-1 (type-1 diabetes), and cluster-2 (type-2 diabetes).
- Stage-3: Applying Classification algorithms to classify the patient's risk level of diabetes

3.2. Dataset Used

The data which is used in this project has records of 650 total diabetic patients of all age group and Table 1 shows all the attributes used for this work.

Table 1: The attributes used in Data Set for this work

Attribute	Description	Type
Gender	Considered as Male=1 Female=0	Numeric
Insulin dependent	Considered as min=50and max=500	Numeric
Plasma	Considered as min=2 and max=11	Numeric
HbA1c	Considered as min=3 and max=19	Numeric
Systolic	blood pressure (Systolic)Considered as min=30 and Max=370	Numeric
Diastolic	blood pressure (Diastolic) Considered as min=60 and max=350	Numeric
Mass	BMI Considered as min=1 and max=200	Numeric
Bg	Blood groupConsidered as 0= 'O',1= 'A',2 = 'B',3 = 'AB'	Nominal
Age	Considered as min=1 and max=125	Numeric
Pedigree	Considered as 0= no family history and 1= family history	Numeric
Pregnancy	Considered as 1= yes 0= no	Numeric
Living area	ConsideredLiving area as 0=Urban and 1= Rural	Nominal
Job type	Considered as 0= stressed job and 1= unstressed job	Numeric
Food habit	Considered as 0=Healthy and 1=Moderate and 2= Junk Food	Nominal

3.3. Data Preprocessing

The Dataset used in this work is clinical dataset which may have some inconsistencies. To remove those inconsistencies data preprocessing is done. In data preprocessing, supervised attribute filtering technique was used. Discretize filter was used for obtaining good intervals of data.

We got only 620 values as valid instances out of 650 total values after data preprocessing. It has eliminated all the null and invalid data from the dataset which we have used as input in this research.

3.4. Accuracy Measures

RandomTree, Naive Bayes, C4.5 and Simple Logistics algorithms were used for this work.The tests are performed by means of internal cross validation 10-folds. Accuracy of each algorithm shows how the datasets are being classified. Recall and precision are the accuracy measures used for this work.

Precision = TP/ (TP + FP).
Recall = TP/ (TP + FN).
Accuracy = (TP + TN)/ (TP + TN + FP + FN).

TP - Positive tuples.
TN - Negative tuples.
FP - Incorrectly classified positive tuples.
FN - Incorrectly classified negative tuples.

The corresponding classifiers precision and recall values are listed in Table 2.

Table 2: Results of precision and recall for different classifiers

Classifier	Precision			Recall		
	Mild	Moderate	Severe	Mild	Moderate	Severe
Naïve Bayes	1	0.84	0.89	1	0.85	0.88
RandomTree	0.947	0.945	0.986	0.953	0.947	0.979
C4.5	1	1	1	1	1	1
Simple Logistics	1	0.996	1	1	0.995	1

4 Results and Discussion

The Proposed system was executed on WEKA tool in three stages, they are 1) First the entire dataset was preprocessed by applying Simple K-means algorithm 2) After preprocessing the dataset was clustered into 3 types as type-1, type-2 and gestational diabetes to find the type of diabetes for each patient 3)The clustered dataset was classified into three classes as mild, moderate and severe. Classification is done in order to predict the risk levels of diabetes for each patient.

4.1. Performance of the Simple K-Means Algorithm

The Simple k-means algorithm clusters the whole dataset into 3 clusters as

- Cluster-0 for gestational diabetes
- Cluster-1 for type-1 diabetes
- Cluster-2 for type-2 diabetes.

The time taken to build the model was 0.15 seconds. Among the 620 instances of the data after preprocessing, 146 were in cluster-0, 115 were in cluster-1and 359 in cluster-2.

4.2. Classifiers Performance

In the classification model, the clustered dataset is given as input in which each patient's risk levels of diabetes is classified as mild, moderate and severe. Next it uses all classification algorithms discussed in section 3. Table 3 shows the results of the classification algorithms.

Table 3. The results of the Risk levels in each type

Diabetes Type	Number of Patients	Risk Level	Number of Patients
Type-1 Diabetes	115	Mild	25
		Moderate	54
		Severe	36
Type-2 Diabetes	359	Mild	56
		Moderate	168
		Severe	135
Type-0 Diabetes (Gestational)	146	Mild	61
		Moderate	57
		Severe	28

The Error rate and accuracy value of each algorithm are shown in Table 4.

Table 4. Classifiers error rate and accuracy values

Classifier	Error Rate	Accuracy Value
Naive Bayes	0.091	90.9
Random Tree	0.036	096.3
C4.5	0	100
Simple Logistics	0.1	0.99

From the Table 4, the diabetes dataset consists of 650 tuples with 14 attributes were analyzed to find the accuracy and error rate by using various classification algorithms. From the above analysis, it was found that C4.5 algorithm is the best onewhen compared to other classifiers for diagnosing diabetes because C4.5 algorithm has more accuracy valueand less error rate.

4.3. Random Tree

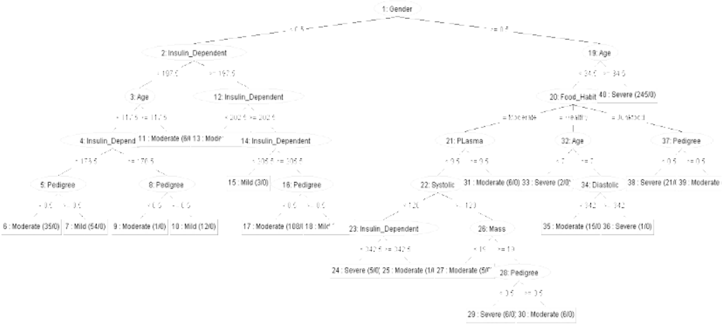

Figure 1: Random Tree.

From figure 1, it was found that among the 14 attributes, insulin dependent, plasma, age, pedigree and food habit played an essential role in diagnosing diabetes. The tree reveals that for a diabetic patient with insulin dependent value less than 197.5mmol, the diabetic level will be mild and if insulin dependent value is greater than 197.5mmol, it may lead to moderate level of diabetes. When the age of the patient is less than 35 with pedigree =1, then the patient may have moderate level of type-2 diabetes. If the patient's age is greater than 35 with pedigree=0 and HbA1c value greater than 5%, it may lead to severe level of diabetes. The tree also shows the food habit of the patient and with age value greater than 35, it will lead to severe level of diabetes.

4.4. C4.5 Tree

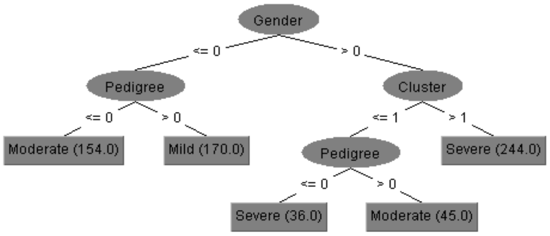

Figure 2: C4.5 Tree

From Figure 2, it was found that the attributes which are used in this tree are gender, pedigree, junk food. The tree reveals that if gender =0 and pedigree =0 and if age <35 then the severity will be mild. If the age >35 with gender=0 and pedigree =0 then the severity will be moderate. It shows when gender=1 and pedigree =0 with age<35 then the severity will be mild, and with age>35 shows the severity as moderate. And it shows the severity level as severe when the gender=1 and pedigree =1 with age<35.

5 Conclusion

Diabetes is most commonly occurring disease. Preventing, controlling and creating awareness about diabetes is important as it leads to other health problems. Type-1 and type-2 diabetes may lead to heart problems, kidneydiseases and eye related problems.It is important to prevent or control gestational diabetes because Gestational Diabetes Mellitus (GDM) may go away after pregnancy, but women who have GDM seven times more are likely to develop type-2 diabetes than women who do not have GDM in pregnancy. The children of the GDM mother have the risk of obesity and type-2 diabetes. All of these difficulties can be handled by controlling the blood sugar levels. From this study, it was found out that data mining techniques can be used for predicting the type and risk levels of diabetes.Through this study it is found that the datamining techniques are important and it leads to valid approaches for predicting the risk of gestational diabetes. So it is our recommendation to use new techniques like data mining for decision making in medical fields, which improves the diagnosis of diseases like gestational diabetes. This research helps the doctors and health organizations in using the datamining techniques in the medical field which helps in predicting the type of diabetics and risks levels associated with it. Thus the proposed model helps in improving the diagnosis of the diseases which indeed helps in early cure of disease in the patients.

6 References

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

2 National Diabetes Statistics Report. 2014. Available from: http://www.cdc.gov/diabetes/pubs/statsreport14/national-diabetes-report-web.pdf

3 Type-2 diabetes in India: Challenges and possible. solutions. Available from: http://www.apiindia.org/medicine_update_2013/chap40.pdf

4 JaliMV, HiremathMB. 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 Han, Kamber M. Data mining concepts and techniques. 2nd ed. Burlington, Massachusetts: Morgan Kaufmann; 2006. p. 285–8.

9 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

10 Rajesh K, Sangeetha V. Application of data mining methods and techniques for diabetes diagnosis. International Journal of Engineering and Innovative Technology (IJEIT). 2012; 2(3):224–9.

11 AfrandP,Yazdani NM, Moetamedzadeh H, NaderiF,Panahi MS. Design and implementation of an expert clinical system for diabetes diagnosis. Global Journal of Science, Engineering and Technology; 2012. p. 23–31. ISSN:2322-2441.

12 Adidela DR, Lavanya DG, Jaya SG, Allam AR. Application of fuzzy ID3 to predict diabetes. Int J AdvComput Math Sci. 2012; 3(4):541–5.

13 PatilBM, Joshi RC, Toshniwal D. Association rule for classification of type-2 diabetic patients. 2nd International Conference of IEEE on Machine Learning and Computing; 2010. p. 67. DOI 10.1109/ICMLC.

14 AljarullahAA. Decision tree discovery for the diagnosis of type II diabetes. International Conference on Innovative in Information Technology; 2011. p. 303–7.

15 Jaya Rama Krishnaiah VV, Chandra Shekar DV, Satya Prasad R, Rao KRH. An empirical study about type-2 diabetes suing duo mining approach. International Journal of Computational Engineering Research. 2012; 2(6):33–42.

16 Mandal S, Dubey V. Implementation and evaluation of diabetes management system using clustering technique. Special Issue of International Journal of Computer Science and Informatics. 2(2):33–6.

17 Kavitha K, Sarojamma RM. Monitoring of diabetes with data mining via CART Method. International Journal of Emerging Technology and Advanced Engineering. 2012; 2(11):157–62.

18 Ferreira D, Oliveira A, Freitas A. Applying data mining techniques to improve diagnoses in neonatal jaundice. BMC Med InformatDecis Making. 2012; 12:143. DOI: 10.1186/1472-6947-12-143.

19 Ananthapadmanaban KR, Parthiban G. Prediction of chances - diabetic retinopathy using data mining classification techniques. Indian Journal of Science and Technology. 2014 Oct; 7(10):1498–503.

20 National Center for Chronic Disease Prevention and Health Promotion. Gestational Diabetes. Centers for Disease Control and Prevention. U.S. Department of Health and Human Services; 2011. Available from: http://www.cdc.gov/

21 Type-2 diabetes complications. Available from:http://www.mayoclinic.org/diseases-conditions/type-2 diabetes/basics/complications/con-20031902

22 SantiWulanPurnami, S.P. Rahayu and AbdullahEmbong, "Feature selection and classification of breast cancer diagnosis based on support vector machine", IEEE 2008.

23 PardhaRepalli, "Prediction on Diabetes Using Data Mining Approach". Dept. of Computer Sciences, Purdue University, 2050 N University St, West Lafayette, IN 47907-2066.

24 Joseph L. Breault., "Data Mining Diabetic Databases: Are Rough Sets a Useful Addition? JamiaHamdard University, New Delhi, Proceedings of the 4th National Conference, INDIA Com-2010 Computing for Nation Development, February 25-26, 2010 BharatiVidyapeeth's Institute of Computer Applications and Management, New Delhi.

25 G. Parthiban, A. Rajesh, S.K.Srivatsa, "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method ", International Journal of Computer Applications (0975 – 8887) Volume 24– No.3, June 2011.

26 P. Padmaja, "Characteristic evaluation of diabetes data using clustering techniques", IJCSNS International Journal of Computer Science and Network Security, VOL.8 No.11, November 2008.

Authors

Dr. Peddoju Suresh Kumar has an overall professional experience of nearly 13 years and working currently with Department of Computer Science and Engineering, Kakatiya Institute of Technology & Science, Warangal, India. He obtained his Master's degree from Osmania University, Hyderabad and Doctoral degree from JNTU, Hyderabad. His areas of research include Big Data, Internet of Things (IoT), Cloud Computing and Mobile Computing. He is a member in various International/National Technical societies such as IEEE, IACSIT, IAENG, CSI and ISTE. He published over 15 publications in standard publishers like ACM/IEEE/Springer, including peer-reviewed journals and conferences. He is contributing as an editorial board member and reviewer for many international journals and conferences.

V. Umatejaswiis currently doing her Master's in Kakatiya Institute of Technology & Science, Warangal, India. She obtained her Bachelor of Degree from JNTU, Hyderabad. She is a member in various International/National Technical societies such as IEEE and CSI.