ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

DIAGNOSING DIABETES USING DATA MINING TECHNIQUES

Srideivanai Nagarajan*, RM. Chandrasekaran
*Research Scholar, Annamalai University.Chidambaram, India.
Professor of Computer Science Engineering Department of Computer Science & Engineering Annamalai University, Chidambaram, India.

ABSTRACT

Diabetes Mellitus (DM) is an important health problem that affects many people including teenagers who get affected by it due to their sedentary lifestyle and food habits. Prevention or controlling diabetes is a big challenge as faced by diabetologists world over. Data mining is an important technique which can be used to find so far unpredicted pattern among huge heath care databases in a very effective and efficient manner. In fact, data mining is applied in many fields like education, retail, finance, healthcare, and many such fields. In this paper the rule based method named as PDC (Potential Diabetic Classifier) was used to diagnose diabetes from the healthcare database and it also diagnoses the type of diabetes from the diabetes dataset. Standard supervised machine learning algorithms like Naïve Bayes, SVM, Ada Boost, bagging, Decision Tree and Simple Cart were also used for diagnosing diabetes and also for predicting the type of diabetes like Gestational Diabetic Mellitus, Type-1 or Type-2 Diabetes. This paper compares the result of the standard classification methods with the PDC method and it was found that the PDC method gives better results when compared to other classification methods applied.

KEYWORDS: Data Mining, Diagnosing Diabetes, Supervised Machine Learning, Naïve Bayes, SVM, Ada Boost, Bagging, Decision Tree and Simple Cart, PDC.

INTRODUCTION

Diabetes is a metabolic disorder where there is an increase in blood sugar level for a prolonged period. Increased hunger, more thirst, weight loss and slow healing of wounds, blurry vision, fatigue and frequent urination are some of the symptoms of diabetes mellitus. The two reasons for diabetes mellitus are the pancreas is not able to produce enough insulin or the body cells may not respond to the insulin produced by the pancreas. DM is of three types as Type-1, Type-2 and Gestational Diabetic Mellitus (GDM). In the case of Type-1 DM, the pancreas fails to produce insulin. It is also called Insulin Dependent Diabetic Mellitus (IDDM) or Juvenile Diabetes (JD). Type-2 DM is due to the body cells not responding to the insulin produced by the pancreas. In due course of time, this Type-2 DM may also cause lack of insulin. This type of diabetes is also called as "Non-Insulin Dependent Diabetes Mellitus" (NIDDM). Some of the primary reasons for Type-2 diabetes are obesity and lack of exercise. When a pregnant woman has high blood sugar level only during her pregnancy time and not before, then she is said to suffer from Gestational Diabetic Mellitus (GDM) [1].

According to a survey conducted for the sixty sixth world health assembly, it was found that in the year 2014, 387 million people in the world had diabetes. Among these people, 316 million people had high risk due to this disease. It has been also predicted that by the year 2035 nearly 471 million people may be affected by DM. In 2013, nearly 5.3 million deaths were caused due to DM [2].

Data mining is an important tool for diabetes diagnosis and research. It is used to find hidden knowledge from diabetes data base and helps to improve the quality of treatment with respect to healthcare industry [3].

Classification is an important data mining technique which is used to assign classes or labels to groups. It works by creating a model to analyze the training dataset of a database. Then the constructed model is used for assigning class labels to each instance of the dataset [4]. The accuracy of the classification is estimated according to the percentage of test samples or test dataset that are correctly classified.

RELATED WORKS

The idea of using data mining techniques for diagnosing diseases has been on the increase nowadays. Researches in the area of data mining, machine learning, fuzzy logic and neural network have generated many automatic systems for diagnosing various diseases. Oracle data miner (ODM) tool with Support Vector Machine algorithm was used to predict the treatment of young and old diabetes patients. The author suggested postponing the treatment for young one in order to avoid the side effects and the treatment for old should me started at an earlier stage [5]. Diabetes has become one of the greatest problems of United States and the author could collect diabetes data from a data warehouse and applied classification and regression method (CART). The author found out that the disease is more with younger generation compared with older ones [6].

Decision tree was used to create a model for predicting Type-2 diabetes. The author tested the model with a highly populated sample dataset. The model was able to diagnose diabetes in some of the undiagnosed cases and also in other follow-up diabetic patients [7].

Apriori association algorithms were used for classifying Type-2 diabetes [8]. The author generated association rules for the class value "yes" as well as for the class value "no". Data preprocessing was also applied to improve the quality of the data.

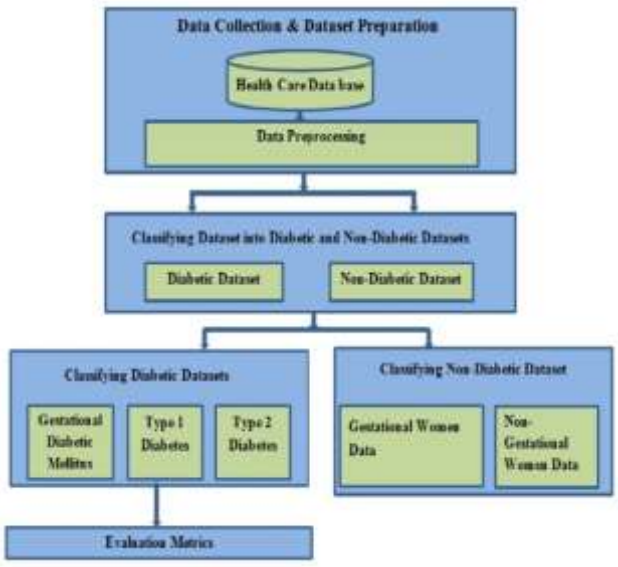
K-means algorithm was used to remove the noisiness in the data and after preprocessing, Support Vector machine classification algorithms was applied with the Pima Indian and diabetes dataset. The research proved that K-means algorithm was better when compared to Support Vector Machine for diagnosing diabetes among the pregnant women of Pima [9.

METHODOLOGY

System Design

Figure 1 shows the outline of the proposed system.

Figure1 – System Design



The model proposed in this paper has following steps.

1. Data Collection was done from many patients from many hospital sources.
2. Preprocessing was done in order to remove the data with zero and null values.
3. Classification model was applied for the data set to diagnose diabetes among the patients. The data set was divided into two sets as diabetes and non-diabetes data sets.
4. Classification model was applied to the diabetes data set to predict the type of diabetes as Gestational Diabetic Mellitus, Type 1 diabetes or Type 2 diabetes.
5. Classification model was again applied to the non-diabetes data set to group that into two clusters as Gestational women dataset and Non-gestational women dataset.

6. Evaluation of the performance of the system was done by using the accuracy measures like precision and recall.

ALGORITHMS APPLIED

Classification Methods

The model developed for this paper uses supervised machine learning algorithms like Naïve Bayes, Decision Tree, SVM and K-Nearest Neighbor to diagnose diabetes among the patients in the healthcare database.

Naïve Bayes

Naïvebayes method uses probability theory for classification. This method assigns probabilities for all class assignments without assigning a single classification. Naive Bayes is a simple and most extensively used probabilistic supervised learning algorithm. This algorithm learns from the training data and from the conditional probability that is assigned for a class label. Naïvebayes algorithm works with principal that all attributes are independent with respect to the class label C. Bayesian rule is applied to compute the probability of C and the class with the highest posterior probability is calculated [10].

Support Vector Machines

Support Vector Machines (SVMs) are supervised machine learning classifiers which use hyper-planes to distinguish between positive and negative instances. It is a powerful classifier which uses statistical learning theory for several classification tasks. This classifier calculates the hyper plane maximizing the minimum distance between the plane and the training points. The SVM classifiers minimize the classification errors and increase the geometric margin. This is a significant feature of SVM classifier. It is called as maximum margin classifier because of this feature [11].

Ada Boost

AdaBoost, is the short form for Adaptive Boosting. With the aim of improving the performance, this algorithm can be used in combining with other algorithms. The weighted sum of all algorithms' output will be represented as Ada Boost's output. It is an adaptive algorithm. This algorithm is sensitive to noisy and erroneous data. Even though the performance of each algorithm is weak, the final model is stronger and gives better output. [12].

Bagging

Bootstrap sampling technique is used in Bagging algorithm. This algorithm is also called as Bootstrap Aggregating algorithm. It works by taking weak algorithms and promote them into their optimal level. It starts by working with training data. A sample data called bootstrap sample will be created from the training dataset. This bagging algorithm trains the classifier with this bootstrap sample data. This is repeated for many times to find the weighted majority value of all learned classifiers. The result is an ensemble classifier [13].

Decision Tree

Decision Tree is one of the most popularly used supervised machine learning algorithms. It classifies the given data set by constructing a tree. It works by searching a feature space and this feature space to the tree structure. The main aim of this approach is to select a minimal set of features that effectively partitions the feature space into classes of observations and to form them into trees. Every feature selected in the process of searching is represented as a node in the tree. The nodes represent a choice point among a number of different possible values of a feature. This process of constructing the decision tree continues until all the training dataset values are taken into consideration. The tree pruning is done to eliminate some nodes and make the tree fit for the current dataset. The test dataset instances are tested by suing the decision tree [14].

Simple Cart

A combination of classification and regression algorithms is used by Simple Cart classifier. Simple Cart uses cross validation technique or a set of test samples for constructing the tree. The best tree will be selected during tree pruning process. This algorithm selects the maximum intolerant feature at each stage of the process. Because of this feature it is called as greedy algorithm. It creates a binary tree at each stage of classification. [15].

OneR

OneR is a rule based classification technique in which the classification rules are based on the value of a single predictor. This method gives ranks to each attributes according to the error value which are generated from the training

set. The numerical valued attributes are treated as continuous valued attributes and it divides the attributes range into several disjoint intervals. The missing values are handled as genuine values [16] [17].

PDC (Potential Diabetic Classifier) Model

PDC model stands for Potential Diabetic Classifier. This classification method applies induction rule generation method to classify the data set. The input is given in the form of rows and columns. The model generates rules for the given data set. It can be used for classifying subjects into appropriate class labels and for prediction. [18].

RESULTS AND DISCUSSIONS

Dataset Used

The Dataset collected from known sources was a clinical dataset containing records of 2650 patients of all age group. Table 1 shows all the attributes used for the research

Table 1 – The attributes used in the experimentation

Attribute	Description	Type
gender	Male or Female	Numeric
Insulin dependent	100% Insulin dependent	Numeric
plasma	Plasma glucose concentration - oral GTT	Numeric
HbA1c	glycated haemoglobin	Numeric
systolic	blood pressure (Systolic)	Numeric
diastolic	blood pressure (Diastolic)	Numeric
bmi	Body Mass Index	Numeric
bg	Blood group	Nominal
age	Patient's age	Numeric
pedigree	Family history details	Numeric
Prevpreghis	Previous Pregnancy history about GDM (only for the female patients who are pregnant)	Numeric
preg1	Only for the female patients 1. Pregnant 0. Others	Numeric
preg	Number of pregnancies (only for the female patients who are pregnant)	Numeric
lifestyle	Patient's Life style 1. Sedentary 2. Normal	Numeric

Data Preprocessing

As the collected data contain some inconsistencies data preprocessing was done to remove the inconsistencies. During this study, instances which had zero values for the attributes – Pregnant, Plasma, lifestyle and blood group were removed. In this study for data preprocessing, supervised attribute filtering technique was used. Discretize filter was used for deriving good intervals of data. After pre-processing, only 2599 valid instances remained out of 2650.

The study was done in two stages. In the first stage, classification algorithms were applied to the data set and the entire dataset was divided into two segments as diabetes dataset and non-diabetes dataset. After analyzing the dataset, it was found that 1679 patients were having diabetes and 920 patients were non-diabetic. In stage 2, the 1679 diabetic dataset was given as input to the supervised machine learning algorithms to find the type of diabetes. The result showing the type of diabetes is shown in Table 2.

Table 2- Diabetes Types

S. No	Type of Diabetes	Number of patients	(%)
1	Gestational Diabetic Mellitus (GDM)	712	42.4 %
2	Type-1 Diabetes	51	3.03 %
3	Type-2 Diabetes	916	54.55 %

From the Table 2, it was found that 712 (42.4%) of the patients had Gestational Diabetic Mellitus, 51 (3.03%) of them had Type-1 diabetes and 916 (54.55%) of them had Type-2 diabetes.

Performance Measures

In this research Naïve Bayes, Support Vector machine, Decision Tree and K-Nearest neighbor algorithms were used. The tests were performed by means of internal cross validation 10-folds. Each algorithm's accuracy indicates how far the datasets are being classified. Recall and precision are the accuracy measures used for this study.

Precision = TP / (TP + FP), Recall = TP / (TP + FN)

Accuracy = (TP + TN) / (TP + TN + FP + FN)

TP – Correctly classified positive tuples

TN – Correctly classified negative tuples

FP - Incorrectly classified positive tuples

FN - Incorrectly classified negative tuples

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

Table 3 – Accuracy Measures Precision and Recall

Classifiers	Precision			Recall		
	GDM	Type-1 Diabetes	Type-2 Diabetes	GDM	Type-1 Diabetes	Type-2 Diabetes
Naïve Bayes	0.988	1	0.998	1	0.972	0.991
SVM	0.99	1	1	1	1	0.992
Ada Boost	0.98	1	0.99	1	0.98	1
Bagging	1	0.96	0.99	1	0.98	1
J48 Decision Tree	1	0.96	1	1	0.93	1
Simple Cart	0.99	0.94	1	1	9.94	0.99
OneR	0.986	0	0.947	1	0	0.989
PDC	**1**	**0.98**	**1**	**1**	**0.99**	**1**

From table 3 it was found that PDC method gives a maximum precision and recall value for all three types of diabetes with precision and recall values 1 GDM, Type-2 diabetes. J48 Decision Tree, SVM stands second in highest precision

and racall values. For this Diabetes dataset naïve Bayes and Simple Cart methods gives less precision and recall values for all types of diabetes.

The Error rate and accuracy value of each classification methods are shown in table 4.

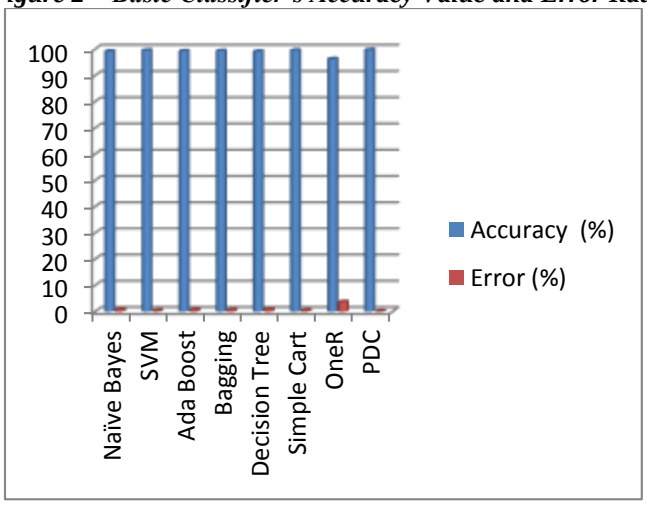
Classifier Performances

Table 4 – Classifiers Accuracy Values and Error Rates

Classifier	**Accuracy Value (%)**	**Error Rate**
Naïve Bayes	99.2	0.8
SVM	99.6	0.4
Ada Boost	99.3	0.7
Bagging	99.4	0.6
Decision Tree	99.2	0.8
Simple Cart	99.6	0.4
OneR	96.3	3.7
PDC	**99.8**	**0.2**

From the table 4, it is clear that PDC (Potential Diabetic Classifier) is having the highest accuracy value (99.8%) and the least error rate (0.2%). Simple Cart and SVM algorithm's outputs are same, which gives accuracy value of 99.6%. Bagging gives the next better accuracy value of 99.4%, which is then followed by Ada Boost with an accuracy value of 99.3%. One R method give the least accuracy value of 96.3% when compared to other machine learning classification methods.

Figure 2 – Basic Classifier's Accuracy Value and Error Rate



CONCLUSION

Diabetes is one of the frequently occurring conditions nowadays in developing countries like India. As diabetes leads to other health related problems, it is necessary to control it. Type-1 and Type-2 diabetes may lead to heart problems, kidney diseases and eye related ailments. Gestational Diabetes Mellitus (GDM) may disappear after delivery of child,

but GDM women are seven times prone for Type-2 diabetes than the non GDM women. GDM mother's child may have the risk for obesity and Type-1 diabetes. These problems can be controlled by controlling or preventing diabetes. This study found that data mining techniques can be used most suitably for diagnosing diabetes and to diagnose the type of diabetes among the patients.

REFERENCES

[1] RSSDI Text Book of Diabetes, under the sub heading: "Clinical Types of diabetes" (Classification) pp 15 – 18.
[2] International Diabetes Federation Atlas, Sixth Edition pp. – 07.
[3] Miroslav Marinov, M.S, Abu Saleh Mohammad Mosa, M.S Illhoi Yoo, Ph.D and Suzanne Austin Boren, Ph.D., MHA1, 2011,"Data-Mining Technologies for Diabetes: A Systematic Review", Journal of Diabetes Science and Technology, Vol 5(6).
[4] Data Mining Concepts and Techniques – Jiawei Han and Micheline Kamber, Second edition, ELSEVIER Publisher, pg- 285 - 288.
[5] Abdullah Aljumah A, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, 2013, "Application of data mining: Diabetes healthcare in young and old patients", Journal of King Saud University – Computer and Information Sciences Vol 25, pp – 127 – 136.
[6] Joseph L. Breaulta,B, Colin R. Goodall.C.D, Peter J. Fose,B, 2002, " Data mining a diabetic data warehouse", "Artificial Intelligence in Medicine" vol -26, pp- 37–54.
[7] Azra Ramezankhani A, Omid Pournik B,C, Jamal Shahrabi D, Davood Khalili A, E, Fereidoun Azizi F, Farzad Hadaegh A, 2014, " Applying decision tree for identification of a low risk population for type 2 diabetes- Tehran Lipid and Glucose Study", Diabetes research and C.linical Practice, vol 5, pp- 391 – 398.
[8] Patil.B.M, Joshi.R.C, Durga Toshniwal, 2010, "Association rule for classification of type-2 diabetic patients", IEEE - Second International Conference on Machine Learning and Computing, DOI 10.1109/ICMLC, Pg-67.
[9] Santhanam, T,Padmavathi.M.S, 2015, "Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis", , Procedia Computer Science, vol – 47, pp-76 – 83.
[10] Qiong Wang, George Garrity, M, James Tiedje, M and James Cole R, "Naïve Bayesian Classifier for Rapid Assignment of RNA Sequences into the New Bacterial Taxonomy", Applied and Environmental Microbiology, vol 73 pp- 5261 b- 5267.
[11] John Platt, C, 1998, " Fast Training of Support Vector Machines Using Sequential Minimal Optimization", Advances in Kernel Methods - Support Vector Learning.
[12] Yubo Wang, "Real Time Facial Expression Recognition with Adaboost", " Proceedings of the 17th International Conference on patter Recognision",Vol – 3, pp-926-929,2004
[13] Xia Rui, Zong Chengqing, Li Shoushan," Ensemble of feature sets and classification algorithms for opinion classification". Inf Sci 181:1138–1152, 2011.
[14] Ted Pedersen, 2001, "A Decision Tree of Bigrams is an Accurate Predictor of Word Sense", In Proceedings of the Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics.
[15] Aruna, S, Rajagopalan, S.P, Nandakishore, L,V, "An Empirical Comparison of Supervised Learning Algorithms in Disease Detection", "International Journal of Information Technology Convergence and Services (IJITCS)", Vol-1(4), 2011 pp- 81 – 92, 2011.
[16] R.Sujatha, D. Ezhilmaran,"Evaluation of classifiers to Enhance Model Selection", International Journal of Computer Science & Engineering Technology", Vol – 4, No – 01, pp- 16 – 21.
[17] Robert C. Holte, Very Simple Classification Rules Perform Well on Most Commonly Used Datasets, Machine Learning 11, pp- 63 – 91.
[18] Srideivanai Nagarajan, Chandrasekaran, RM, 2015, "Supervised Machine Learning Techniques for predicting the Risk Levels of Gestational Diabetes mellitus", International Journal of Applied Engineering Research, Vol – 10, No – 18 , pp- 38729 – 38732.