Prediction of Type2 Diabetes Mellitus Based on Data Mining

D. Jeevanandhini , E. Gokul Raj ,V. Dinesh Kumar, N. Sasipriyaa ,
Assistant Professor,
Department of Computer Science,
Kongu Engineering College.

***Abstract*:- Diabetes Mellitus is a one of the common and growing Chronic Disease due to high blood glucose level. Nearly, half of all diabetes have household heredity factors, which is one of the most important features of data mining .Failure of the pancreas to produce enough insulin and the body inefficient use of insulin are both pathologic causes of diabetes mellitus .Diabetes is growing in several countries and all of them are working to prevent this disease at early stage by predicting the symptoms of diabetes using several method. The main aim of this paper is to compare the performance of the algorithm using the data mining techniques.**

Performance analysis are considered by two stages. In stage one using K-means clustering algorithm remove incorrect data by data cleaning .In stage two the remaining data will be used as the input to the classification algorithm .Classification algorithm used here are SVM , KNN, J48, Random forest. The Pima Indian Diabetes dataset are taken from UCI Repository are consider for the analysis. Finally we obtained that SVM has higher accuracy comparing to other three algorithm . In further, to make the model adapt for various Diabetes mellitus dataset.

1. INTRODUCTION

Diabetes is one of the common and growing disease in several countries and its causes many health problems . Diabetes mellitus is classified as four types : type 1, type 2, gestational diabetes and other specific types[2]. All forms of diabetes increase the risk of long-term complications. People with diabetes have an increased risk of developing a number of serious health problems. Consistently high blood glucose levels can lead to serious diseases affecting the heart and blood vessels, eyes, kidneys, nerves and teeth. In addition, people with diabetes also have a higher risk of developing infections. In almost all developed countries, diabetes is a leading cause of cardiovascular disease**,** blindness**,** kidney failure[3]**.** Now it is very important to develop predictive model using the risk factors for the development of diabetes. The International Diabetes Federation presents the latest data on DM in the Diabetes Atlas .It shows that in 2015, the number of Diabetics worldwide was close to 415 million. Data mining also known as Knowledge Discovery in Database (KDD), is defined as the computational process of discovering patterns in large datasets involving methods at the intersection of artificial intelligence, machine learning , statistics, and database systems[4].

Data mining contains a series of steps disposed automatically or semi automatically in order to extract and discover interesting, unknown, hidden features from large quantities of data[6]. Data mining problems are often solved using different approaches from both computer sciences, such as multi-dimensional databases, machine learning, soft computing and data visualization and statistics, including hypothesis testing, clustering, classification, and regression techniques[7]. As we all know, the number of diabetics is large, and it is continuously increasing. Additionally, most people know little about their health quality. In particular , we have focused on T2DM. Section 2 details the literature reviews, Section 3 describes the tools, methods and dataset. Section 4 details the result of the experiment. Section 5 concludes the paper.

2. LITERATURE REVIEW

In recent years, using the data mining technique has been used with increasing frequency to predict the possibility of disease. Many algorithms and toolkits have been created and studied by researchers. These have highlighted the tremendous potential of this research field. In this section, a few important works that are closely related to the proposed issue are presented. Based on several studies, we found that a commonly used dataset was the Pima Indians Diabetes Dataset from the University of California, Irvine (UCI) Machine Learning Database[9]. Patil [10] proposed a hybrid prediction model (HPM), which used a K-means clustering algorithm aimed at validating a chosen class label of given data and used the C4.5 algorithm aimed at building the final classifier model, with 92.38% classification accuracy. Ahmad [11] compared the prediction accuracy of multilayer perception (MLP) in neural networks against the ID3 and J48 algorithms. The results showed that a pruned J48 tree performed

with higher accuracy, which was 89.3% compared to 81.9%. Marcano-Cedeño [12] proposed artificial metaplasticity on multilayer perceptron (AMMLP) as a prediction model for diabetes, for which the best result obtained was 89.93%. All the studies presented above used the same Pima Indians Diabetes Dataset as the experimental material. The Waikato Environment for Knowledge Analysis (WEKA) toolkit was the primary tool which most researchers chose.

In order to obtain more useful and meaningful data, we realized that the preprocessing methods and parameters should be chosen rationally. Vijayan V. [13] reviewed the benefits of different preprocessing techniques for predicting DM. The preprocessing methods were principal component analysis (PCA) and discretization. It concluded that the preprocessing methods improved the accuracy of the naive Bayes classifier and decision tree (DT), while the support vector machine (SVM) accuracy decreased. Wei [14] analyzed risk factors of T2DM based on the FP-growth and Apriori algorithms. Guo [15] proposed the receiver operating characteristic (ROC) area, the sensitivity, and the specificity predictive values to validate and verify the experimental results. On the basis of an effective prediction algorithm, we need an appropriate way to make the model convenient for everyone [16].

We found that Sowjanya [17] had developed an android application-based solution to overcome the deficiency of awareness about DM in his paper. The application used the DT classifier to predict diabetes levels for users. The system also provided information and suggestions about diabetes.

3. PROPOSED WORK

In this section dataset description , clustering and classification algorithm are discussed. The following algorithm like K-means for clustering and classification algorithm as SVM, KNN, J48 and RandomForest are taken for this analysis .

Matlab are used for the model for predictive analysis of the diabetes dataset.

3.1 Dataset Description

The Pima Indian Diabetes Dataset consists of information on 768 patients (268 tested positive instances and 500 tested negative instances)[18]. Tested positive and tested negative indicates whether the patient is diabetic or not, respectively. Each instance is comprised of 8 attributes, which are all numeric. Attribute details are listed below

- Number of times pregnant (preg)
- Plasma glucose concentration at 2 hours in an oral glucose tolerance test (plas)
- Diastolic blood pressure (pres)
- Triceps skin fold thickness (skin)
- 2-hour serum insulin (insu)
- Body mass index (bmi)
- Diabetes pedigree function (pedi)
- Age (age)
- Class variable (class)

3.1.1.Data Preprocessing

The quality of the data, to a large extent, affects the result of prediction. This means that data preprocessing plays an important role in the model [19]. We determined that the number of pregnancies has little connection with DM [10]. The value 0 indicates non-pregnant and 1 indicates pregnant.

The complexity of the dataset was reduced by this process .There are some missing and incorrect values in the dataset due to errors or deregulation. Most of the inaccurate experimental results were caused by these meaningless values. For example, in the original dataset, the values of diastolic blood pressure and body mass index could not be 0, which indicates that the real value was missing. To reduce the influence of meaningless values, we used the means from the training data to replace all missing values. The unsupervised normalize filter for attribute was used to normalize all the data.

3.2 *Model Description*

The model consists of Double level algorithm .In the first level, we used the K-Means clustering algorithm to remove incorrectly clustered data and optimized data was used as input to classification algorithms[4].

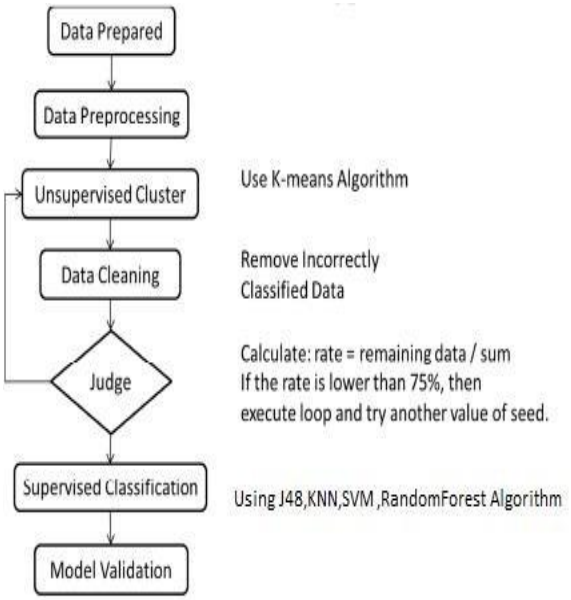

Figure 2. Algorithm Model

3.2.1 K-Means Clustering Algorithm

K-means cluster is one of the most popular cluster algorithm. This is the distance based cluster algorithm, the smaller distance between objects shows the greater similarity. Working principle of K-means algorithm are as follows[6]:

1) Select K from number of initial cluster center , initially take the value of K as 2.

2) Calculate distance between each object and cluster centroid. Cluster every object to the nearest cluster according to the distance .

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \le \|x_p - m_j^{(t)}\|^2 \forall j, 1 \le j \le k\}$$

3) Recalculate every cluster center to verify whether they are changed.

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

4) Repeat the step 2 and step 3 until the new cluster center is same as the original one.

After the K-means process the result of the data belonged to which cluster, cluster 1 or cluster 2 value and the class label value are compared then mismatch rows are removed the incorrect data. From, the removal procedure the remaining 589 correctly classified data will serve as input to classification.

NUMBER	LABEL	COUNT
1	CLUSTER 0	458
2	CLUSTER 1	310

Table 1 Result of the K-means cluster

3.2.2 Classification algorithm

The following algorithms are considered for comparison analysis for prediction of diabetes.

a. Decision Tree J48.
b. KNN Classifier.
c. Random Forest.
d. Support Vector Machine

a. *Decision tree j48*

The following j48 is the decision tree algorithm[7][8].

Pseudo code of J48

1) Check for base cases
2) For each attribute *a*
a) It checks for normalized information gain on *a*.
3) Select the attribute which has highest information gain
4) It creates a decision node with that attribute.
5) This process is repeated with sub list of the nodes and added to its child node.

b. *K- Nearest Neighbor*

K Nearest Neighbors (KNN) is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its KNN measured by a distance function. The Euclidean distance between two points x and y is given by the equation[7][8] .

$$\text{Euclidean} = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

The value of *k* (the positive integer) is determined by inspecting the data set. Cross-validation is another way to retrospectively determine a good *k* value by using an independent data set to validate the *k*. Here we have taken the values (*k* =1, 3 and 5) and it produces good result at *k* =5. This implies that the *k* value gets larger the result will be more accurate. In most cases the optimal *k* value will be between 3 and 10.

c. *Support Vector Machine*

Support Vector Machine (SVM) can also be used as a regression method, maintaining all the main features that characterize the algorithm . The main function of this algorithm is to predict the class membership for categorical target by tasks by constructing hyper planes in a multidimensional space that separates cases of different class labels.

SVM supports maximum prediction accuracy that avoids over fit and it also supports text data and sparse transactional data. The SVM provides empirically good performance in the field of bioinformatics, text and image recognition. The SVM is primarily a classier method that performs classification. The SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables[7][8].

d. *Random Forest*

Random forest algorithm is the statistical and machine learning algorithm which uses multiple learning algorithms to obtain better predictive performance than others. This algorithm has two parts[7][8].

a. Tree bagging
b. From tree bagging to random forest

Each tree is grown as follows:

1. If the number of cases in the training set is N, sample N cases at random –but with replacement,
From the original data .This sample will be the training set for growing the tree.

2. If there are M input variable ,a random number of attributes area selected and the best split used to split the node. The value of M is held constant during the forest growing.

3. Each tree is grown to the largest extent possible. There is no pruning

4. RESULT

CLASSIFICATION TECHNIQUE	ACCURACY
KNN	67.4479
J48	71.0937
SVM	77.9947
Random Forest	66.0156

Table 2 Accuracy Table

CLASSIFICATION TECHNIQUE	ERROR RATE
KNN	32.5520
J48	28.9062
SVM	22.0052
Random Forest	33.9843

Table 3 Error rate Table

5. CONCLUSION

In this section performance analysis is made for type2 diabetes mellitus dataset to improve the accuracy by using clustering and classification algorithm .we compared the four prediction model using 8 important attributes .From this studies concludes that Support Vector Machine (SVM) classifier achieves higher accuracy of 77.82 % than other three classifiers. This study can be used to select best classifier for predicting diabetes. For future work, it is necessary to bring in hospital's real and latest patients' data for continuous training and optimization and also the quantity of the dataset should be large enough for training and predicting.

REFERENCES:

[1] International Diabetes Federation (IDF) DIABETES ATLAS (Seventh Edition), 2015.
[2] https://www.sciencedirect.com/science/article/pii/S2352914817301405.
[3] *The International Diabetes Federation (IDF)* [Internet].http://www.idf.org/complications-diabetes.
[4] http://en.wikipedia.org/wiki/Data_mining#cite_note- acm-1.
[5] Wu H, Yang S, Huang Z, He J, Wang X, Type 2 diabetes mellitus prediction model based on data mining,*Informatics in Medicine Unlocked*(2018),doi:10.1016/j.imu.2017.12.006.
[6] https://www.sciencedirect.com/science/article/pii/S2352914817301405.
[7] https://www.sciencedirect.com/science/article/pii/S1877050915004500.
[8] https://ac.els-cdn.com/S1877050915004500/1-s2.0-S1877050915004500-main.pdf?_tid=f721250c-d935- 497c-b84c- 986b803ab30&acdnat=1520326508_3495c7cac8e512ab149acea41f03627f.

[9] http://archive.ics.uci.edu/ml/datasets/Pima+Indians+ Diabetes.
[10] B.M. Patil, Hybrid prediction model for Type-2 diabetic patients. Expert Systems with Applications 37 (2010) 8102–8108.
[11] Aliza Ahmad and Aida MustaphaH, Comparison between Neural Networks against Decision Tree in Improving Prediction.Accuracy for Diabetes Mellitus. ICDIPC 2011, Part I, CCIS 188, pp. 537–545, 2011.
[12] Alexis Marcano-Cedeño , Joaquín Torres, and Diego Andina, A Prediction Model to Diabetes Using Artificial Metaplasticity. IWINAC 2011, Part II, LNCS 6687, pp. 418–425, 2011.
[13] VeenaVijayan V. and Anjali C., Decision Support Systems for Predicting Diabetes Mellitus –A Review. Proceedings of 2015 Global Conference on Communication Technologies (GCCT 2015).
[14] Zhe Wei , Guangjian Ye and Nengcai Wang. Analysis for risk factors of type 2 diabetes mellitus based on FP-growth algorithm. China Medical Equipment, 2016. 13(5):45-48.
[15] Yirui Guo. Application of artificial neural network to predict individual risk of type 2 diabetes mellitus. Journal of Zhengzhou University, 2014.49(3):180-183.
[16] Shuaishuai Li, Enke Zhang, Min Li and Wei Pan, Research on the Effectiveness of Application of Diabetes Management APP, China Medical Devices, 2015. Vol 30. No.08.
[17] Ms. K Sowjanya, MobDBTest: A machine learning based system for predicting diabetes risk using mobile devices. 2015 IEEE International Advance Computing Conference (IACC).
[18]]http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes.
[19] Karim M. Orabi1 , Yasser M. Kamal, and Thanaa M. Rabah. Early Predictive System for Diabetes Mellitus Disease. ICDM 2016, LNAI 9728, pp. 420– 427, 2016.
[20] Guojun, G., Chaoqu, M. and Jianhong, W.. Data clustering theory algorithm and application (1st Ed.). ASA-SIAM.M (2007).
[21] *Diagnostic Criteria and Classification of Hyper glycaemia First Detected in Pregnancy* – WHO Publications,2013
[22] *The International Diabetes Federation (IDF)*[Internet].http://www.idf.org/complications- diabetes.
[23] Jia Z, Zhou Y, Liu X, Wang Y, Zhao X, Wang Y, Liang W, Wu S. Comparison of Different Anthropometric Measures as Predictors of Diabetes Incidence in a Chinese Population. *Diabetes Research and Clinical Practice*, 2011; **92**:267-271.
[24] *Encyclopedia of Data Warehousing and Mining, Edited by John Wang*- Idea Group Publishing, PCK Edition 2005.
[25] Lily T, Hossein M, Omid H, Jalal P. Real-Data Comparison of Data Mining Methods in Prediction of Diabetes in Iran. *Healthcare Information Research*, 2013; **19**:177-185.
[26] Mehta SR, Kashyap AS, Das S. Diabetes Mellitus in India: The Modern Scourge, *Medical Journal Armed Forces India*, 2009; **65**:50-54.
[27] Karthikeyani V**,**Parvin Begum I**,**Tajudin K**,**Shahina Begam I Comparative of Data Mining Classification Algorithm (CDMCA) in Diabetes Disease Prediction**,** *International Journal of Computer Applications*, 2012; **60**:2631.
[28] Olaiya F. Comparative Study of Different Data Mining Techniques Performance in Knowledge Discovery from Medical Database, *International Journal of Advanced Research in Computer Science and Software Engineering*, 2013; **3**:11-15.
[29] Nirmala Devi M, Appavu alias Balamurugan S, Swathi UV. An Amalgam KNN to Predict Diabetes Mellitus, *IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology (ICECCN 2013).*
[30] http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes.
[31] Riccardo, B., & Blaz, Z. (2008). Predictive data mining in clinical medicine: Current issues and guidelines, International Journal of Medical Informatics, 77, 81–97.
[32] Mechelle Gittens, Reco King, Curtis Gittens and Adrian Als, Post-diagnosis Management of Diabetes through a Mobile Health Consultation Application, 2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom).