International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

Diabetes Mellitus Prediction using Classification Techniques

Abdulhakim Salum Hassan, I. Malaserene, A. Anny Leema

Abstract: ***Diabetes is a metabolic* disease affecting people *in almost every country and it may lead to severe problems like stroke, kidney failure or premature death if it is not predicted at the early stage. To mitigate this many researchers are working to predict the diabetes at early stage using several methods. Different accessible conventional techniques are carried out to diagnose diabetes depend on physical and substance tests. Several data mining methods were designed to overcome these uncertainties. Classification techniques like Decision Tree, K-Nearest Neighbors, and Support Vector Machines are used to classify the patients with diabetes mellitus. The performance of these applied techniques are determined using the factors precision, accuracy, Sensitivity, and Specificity. The results obtained proved that SVM outperforms decision tree and KNN with highest accuracy of 90.23%. Performance analysis of these classification methods helps us to decide which appropriate technique to choose in future for analysing the given dataset.***

Key words: Data mining, KNN, SVM, Decision tree, Diabetes.

I. INTRODUCTION

Diabetes is a prolonged disease that creates complications in the body when the pancreas does not produce adequate insulin in the body. Currently, there more than 420 million peoples who live with diabetes. According to World Bank data shows Marshall Islands has the highest ratio with 3:10 people who have diabetes, while Benin has the lowest ratio of 1:10 people have diabetes in age between 20-79 in 2017 report. International diabetes federation shows there are more than 72 million adults who have diabetes in India in 2017. The data show that in 1980 there around 108 million people with diabetes now it's more than 400% with more than 420 million people. In the year 2012, 2.2 million deaths are caused by high blood glucose and 1.6 million deaths occurred in 2016 are caused by diabetes. Diabetes is divided into three main categories type 1, type 2 and gestational diabetes causes serious health concerns if it is not taken care properly. Type 1 diabetes was known insulin-dependent, childhood-onset or juvenile is characterized by body produce less insulin. People with type 1 diabetes require administration daily for insulin regulation of glucose in their blood. The person who has type 1 diabetes they cannot survive if they don't have access to insulin. Type 1 diabetes currently is not preventable because the causes is still unknown. Symptoms like weight loss, excessive urination fatigue, vision changes, and thirst can indicate the type 1 diabetes.

Revised Manuscript Received on March 30, 2020.
* Correspondence Author
Abdulhakim Salum Hassan, Department of Information Technology & Engineering Vellore Institute of Technology (VIT) Tamil Nadu, India.
I. Malaserene, Department of Information Technology & Engineering Vellore Institute of Technology (VIT) Tamil Nadu, India.
A. Anny Leema, Department of Information Technology & Engineering Vellore Institute of Technology (VIT) Tamil Nadu, India.

Type 2 diabetes also was known as non-insulin-dependent diabetes is due to improper food habits, over weight, not doing physical exercise. It is more common among diabetes patients. Symptoms of type 2 diabetes may be similar to those on type 1 diabetes. Type 2 diabetes was common to adults but now some cases occur to the child who was diagnosed with type 2 diabetes.

Gestational diabetes is occurred temporarily on a pregnant women and disappears after giving birth. They are in the risk of some complication during pregnancy and delivering time. This can be prevented by proper exercise and weight before they become pregnant.

II. LITERATURE REVIEW

Previous work on diabetes prediction segmentation generally uses the enhancement provided by the contrast agent in the prediction. There are tonnes of ways by which prediction can be performed giving us different types of results. Most of them have used the PIMA dataset to diagnose the diabetic result either it is positive or negative.
Data mining model is built to predict type 2 diabetes mellitus. It was adaptive and tested with more than one dataset. Pre-processing was done in WEKA applying various filters. The data is transformed for applying suitable data mining techniques. Numeric attribute was transformed to nominal attribute and the complication of the dataset was reduced. In Phase I, the improved K-means algorithm was used to form proper clustering group and this was used as input to the next level where logistic regression was used to classify the data. The performance of the model was verified with K–fold cross validation. [1]. Type 2 diabetes can be reduced by detecting it at earlier stage. SVM was used to diagnose the diabetes [2].
Diabetes is a most common disease in Saudi Arabia and data obtained from the world health organization web link were used to analyse the diabetes risk factors applying data mining techniques and it was consolidated into two age groups. Finally proper treatment plan was suggested as per their lifestyles. Control and proper planning of food intake, focussing to reduce the weight, proper physical exercise, avoid smoking were the various measures suggested for effective control of diabetes [3]. Hybrid system was used to discover the risk of diabetes at the early stage by applying machine learning techniques. Experimental analysis done on the on the dataset taken from UCI and obtained improved accuracy using PCA and NN [4]. Diabetic retinopathy causes damage to the retinal blood vessels and in order to diagnose this, dataset from UCI repository was used. Feedforward neural network is implemented by adjusting the weights of neural net which depends on the error rate obtained in previous epoch. Compared to the manual diagnosis the artificial neural network model was faster and accurate [5].

Various data mining techniques are used to diagnose the diabetes at the early stage of life were discussed [6].
Prevention of diabetes is a challenging task in society and prediction of this is increasing in healthcare. J48 was used to classify whether the patient has diabetes mellitus risk factors or not. Adaboost ensemble method was also applied and found comparatively it was better than J48 [7,21].
Diabetes leads to other risks like harm in blood vessel, loss of eye sight, heart failure, damages in nerve, kidney disease etc. PIMA data set was used for their analysis and J48 was used and obtained the highest accuracy of 99.87% but they insisted that the proposed algorithm has to be tested with larger data set because PIMA data set has only 768 data. [8]
Dataset collected from private medical diabetes containing 540 patients details was used for the analysis and different methods like J48 algorithm, classification and regression tree (CART), Support vector machine (SVM) and K-Nearest Neighbor (KNN) algorithm were applied. Among those methods J48 produce the highest accuracy about 67.16%. They state that to improve the overall accuracy they need more data set [9].
Diabetes can be caused by obesity, being overweight and being inactive. The diabetes can cause heart disease, stroke, kidney failure, blindness and premature death. The author used PIMA data set for analysing and implemented ANN, Decision tree, Naive bays, SVM, C4.5, ID3 and CART. Among all methods CART provide the highest accuracy of 83.2% [10]. This paper discusses the various types of diabetes and its symptom. Regular exercise and proper food control can control type-1 and type-II diabetes and reduces the risk factor if it is diagnosed at the earlier stage. Various data mining techniques and statistical methods are used to predict the diabetes and neural Fuzzy Networks are implemented so that its performance can be compared with other data mining techniques [11].
Data Mining plays a vital role in diagnosing the types of diabetes. Ensemble machine learning approach was built using K-Nearest Neighbor (KNN), Naïve Bayes (NBs), Random Forest (RF), J48 and reduced bagging. [12]. Survey was done on diabetes, its types and prognosis were also discussed [13]. Multilayer perceptron neural network is good for pattern classification and prediction [14, 19]. K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) algorithms tested on the PIMA data set and research found that KNN produce the highest accuracy of 85.8%.compared to SVM [15]. Different prediction methods of diabetes was presented and future direction was given for the severity of diabetes. It was suggested to undergo the test HbA1c and check cholesterol level every year. Doctor recommends the diabetes patient to undergo pathology tests such as an electrocardiogram (ECG) because Heart disease and blood vessel disease are the common problems to the diabetic people[16].
T. Monika and Rakhi Wagji [17] they state that diseases like blindness, blood preassure, heart disease, kidney disease and nerve damage are disease where the person with diabetes are in danger to get. In their work they use Pima dataset and they achieve accuracy of 65% by using Naive Bayes method. They recommend their proposed algorithm to be tested on a larger dataset.
Diabetes can be caused by overweight, obesity, lack of physical activity, poor diet and leads to various disease like kidney failure, loss of eye sight, cardiac arrest etc. Dataset was obtained from US hospital for the analysis purpose Naive Bayes and J48 methods were used and compared the accuracy of both the methods. J48 accuracy is 79.68% which is better than Naïve Bayes [18]. Pre-processing was done to the clinical dataset to remove the inconsistencies and the most relevant features were selected to predict the disease in more accurate manner. Pima Indian Diabetes Dataset from the UCI learning repository of 768 instances, eight attributes are considered for analysis. Hybrid classification model was built using ensemble technique [20]. Study performed on various data mining techniques used for predicting the diabetes at earlier stage and also determined the accuracy of the applied data mining techniques. J48 was improvised and the modified J48 provided the high accuracy in diagnosing the disease [22].

III. IMPLEMENTATION

A. Data mining

It is the process of extracting the hidden patterns in large data sets by following different steps like data cleaning and integration, data transformation, applying suitable data mining techniques, extract pattern and visualization.
Combining the data from various separate sources, selections of the data to discover the knowledge is an essential task. After selection of the attributes, we need to clean the data before transformation to the suitable format. Then various data mining techniques are applied to extract pattern and finally pattern is evaluated to remove the redundant patterns. The last one is knowledge presentation where the knowledge is visualized and presented.
Data cleaning is the process of making sure the data is clear as depicted in Fig:1 and make sure it is ready for the further process, filling missing data is one of the processes on a cleaning of data. Techniques like Binning methods, Regression and clustering are some of the methods used for removing noisy data.
Data transformation is the process of transforming data to the suitable form for mining process, in a transformation of data there are different methods for data transformation like normalization which work by scaling data in between -1.0 to 1.0 or 0.0 to 1.0, discretisation method which replaces raw data values of numeric by internal or conceptual.

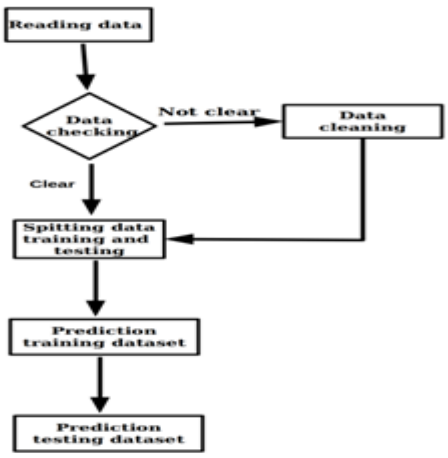


Fig: 1 Pre-processing the Data

B. Dataset

PIDD (Pima Indian Diabetic Dataset) which collected from females of age from 21 and above are used in implementation of different methods. The various features of Pima dataset is given in Table:1.

Table: 1 Features of PIMA dataset

Pregnancies	Number of times pregnant
Glucose	Plasma glucose
Blood pressure	Diastolic blood pressure (mm Hg)
Skin thickness	Triceps skin fold thickness (mm)
Insulin	2 hours serum insulin (mu U/ml)
BMI	Body Mass Index
Diabetic pedigree function	Score likelihood of diabetes based on family history
Age	Age(Years)
Outcome/class	0 – Non-diabetic, 1 - Diabetic

C. Methodology

Decision tree is the one of the classification methods which is commonly used for classifying data. In a decision tree we need to have the class which will be used for the final output. According to other attributes the decision will be made according to the gain of each attribute, the higher gain will be the root and repeat the same process until reaches to the leaf node through internal node. Information gain is how much information the feature give about the class. After calculating the information gain we need to find the gain of each attribute by

Gain (T,X) = Entropy(T) – Entropy(T,X)

The attribute with higher gain will be the root node.

Then it will split the dataset for training and testing, after splitting it will go directly to the prediction part for training and testing.

K-Nearest Neighbour is considered one of the lazy learning methods. KNN depends on neighbour to predict the unknown tuple. KNN identify k neighbours for unknown tuple; it uses distance measure like Euclidean distance to calculate the distance between data points.

$$dist(X_1, X_2) = \sqrt{\sum(x_1 i - x_2 i)^2}$$

First, reading data then data checking if there is any missing value's or not, if null value found then it need to be filled. Next is data transformation by scaling which improve the accuracy. After transforming the data then method will predict the dataset and give out the results.

Support Vector Machine (SVM) is a supervised learning method where it categorize new example with an optimal hyperplane. Because of significant accuracy support vector machine is highly preferred with less computational power.

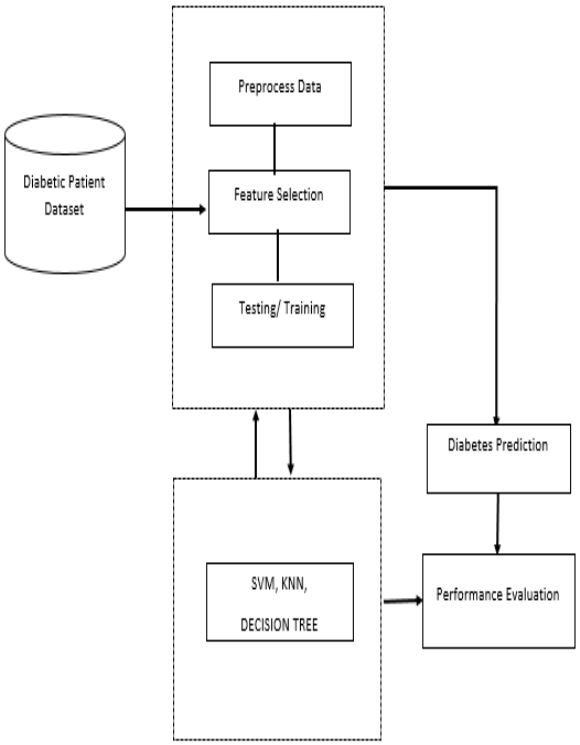


Fig: 2 Architecture Diagram of the Proposed System

Hyperplane can be written as the set of points $\vec{w}$ satisfying

$$\vec{w} * \vec{x} - b = 0$$

And the hard margin

$$\vec{w} * \vec{x} - b = 1$$

For anything on or above the boundary

$$\vec{w} * \vec{x} - b = -1$$

For anything on or below the boundary.

Soft margin is used for training purpose where it allow support vector machine to make certain mistakes for keeping margin as wide as possible while other points can still be classified correctly. The proposed system is an implementation of three different methods, Decision tree, K-Nearest Neighbor and Support Vector Machine(SVM) for predicrtion of the diabetes. The architecture diagram of the proposed system is given in Fig: 2

D. Evaluation

Confusion matrix is used to visualize the performance of the algorithms which cross tabulates the observed and predicted classes with associated statistics, evaluation metrics like sensitivity, specificity, precision and accuracy are used to evaluate the performance of the method. Factors like True Positive (TP), True Negative (TN), False Positive (FP) and False Negative are used.

Actual class	Positive	Negative
Positive	True Positive	False Positive
Negative	False Negative	True Negative

IV. RESULT

The Performance of the classification algorithms are tested based on various measures and it was found SVM has better accuracy compared to Decision tree and KNN. Data transformation improves the performance of the KNN while it does not have impact on Decision tree but in SVM tuning has improved the performance. Scaled data improves the performance time for KNN. Decision tree for training dataset produces 71.17% of accuracy. Confusion matrix table for training dataset

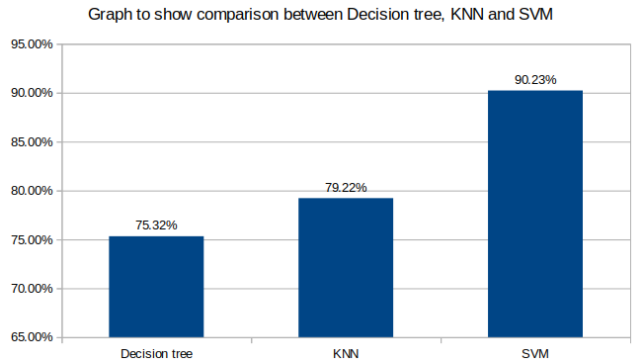
	0	1
0	339	61
1	116	98

and for testing it produce 75.32% of accuracy.
Confusion matrix table for testing dataset

	0	1
0	89	11
1	27	27

In KNN it the accuracy is different for transformed datasets and dataset without data transformation.
For the non-normalized dataset it produces accuracy of 72.72% when k=11 and it is the highest.
But after data transformation the accuracy is improved and the highest accuracy is when k=14 it produce 75.97% of accuracy compare to 72.72% without data transformation.
On SVM the accuracy improved after tuning where the accuracy before tuning is 82.42% and after tuning is 90.23%
Data mining and machine learning algorithms in health sector extracts different hidden patterns from the medical data. They can be utilized for the examination of significant clinical parameters, expectation of different illnesses, guaging undertakings in medication, extraction of therapeutic information, treatment arranging backing and patient administration. There are lot of the methods available which can be deployed, in this project three choosen methods are deployed but the outcome is the SVM method outperform other methods by provide the highest accuracy better than other methods.

Method	Accuracy
SVM	90.23%
KNN	75.97%
Decision Tree	75.32%



V. CONCLUSION AND FUTURE WORK

Classification is one of the important task which helps the researcher to predict the categorical valued functions. In order to diagnose the diabetes, various classification techniques like SVM, KNN and Decision tree algorithms are implemented to the chosen dataset. The performance metrics namely accuracy, specificity, sensitivity, precision are computed and the result shows that the SVM algorithm outperforms the other two. In future it is planned to improve the accuracy prediction by testing our classification techniques with huge data set and its performance can be improved .

REFERENCES

1. Han Wu, Shengqi Yang , Zhangqin Huang, Jian He, Xiaoyi Wang, “Type 2 diabetes mellitus prediction model based on data mining”, Informatics in Medicine Unlocked, 2017
2. Nahla Barakat , Andrew P. Bradley,Mohamed Nabil H. Baraka. "Intelligible Support Vector Machines For Diagnosis of Diabetes Mellitus IEEE transaction (2010)", International Conference on Computational Intelligence and Data Science(ICCIDS 2018), 2018
3. Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, " Application of data mining: Diabetes health care in young and old patients ", Journal of King Saud University – Computer and Information Sciences, 2012.
4. Mehrbakhsh ,Othman, Mohammad, and Leila "Accuracy improvement for diabetes disease classification: A case on a public medical dataset", Fuzzy Information and Engineering. Elsevier,vol(9),345-357,2017.
5. Adefemi A. ADEKUNLE , Adnan KHASHMAN, Ebenezer O. OLANIYI, Oyebade K. OYEDOTUN, "Diabetic Retinopathy Diagnosis Using Neural Network Arbitration", Bulletin of the Transilvania University of Braşov • Vol 10(59), No. 1 - 2017
6. N. Jayanthi, B.Vijay Babu, N. Sambasiva Rao, "Data Mining Techniques for CPD of Diabetes", IJECRT- International Journal of Engineering Computational Research and Technology, 2016
7. Sajida Perveen, Muhammad Shahbaz, Aziz Guergachi, Karim Keshavjee, "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes", Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia
8. Gaganjot Kaur, Amit Chhabra, "Improved J48 Classification Algorithm for the Prediction of Diabetes", International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014
9. K. Saravananathan, T. Velmurugan, "Analyzing Diabetic Data using Classification Algorithms in Data Mining", Indian Journal of Science and Technology, Vol 9(43), DOI: 10.17485/ijst/2016/v9i43/93874, November 2016
10. Ms. Nilam chandgude(Author), Prof. Suvarna pawar, "A survey on diagnosis of diabetes using various classification algorithm", International Journal on Recent and Innovation Trends in Computing and Communication Volume: 3 Issue: 12 ISSN: 2321-8169 6706 - 6710
11. DR. M. Mayilvaganan, R. Deepa, P. Nandakumar, "A study on data mining and statistical methods used in diabetes mellitus diagnosis", International Journal of Advanced Research (2016), Volume 4, Issue 7, 447-452
12. Bhavana N , Meghana S Chadaga, "A review of ensemble machine learning approach in prediction of diabetes diseases", International Journal on Future Revolution in Computer Science & Communication Engineering Volume: 4 Issue: 3 ISSN: 2454-4248 463 – 466, 2018
13. Misba Reyaz, Gagan Dhawan, "Various Data Mining Techniques for Diabetes Prognosis: A Review", International Journal of Trend in Scientific Research and Development (IJTSRD) International Open Access Journal ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 4, 2018
14. Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, "Diagnosis of diabetes using classification mining techniques", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015

15. Saima Bano, Muhammad Naeem Ahmed Khan, "A Framework to Improve Diabetes Prediction using k-NN and SVM", International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 11, November 2016
16. 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
17. T.monika Singh, Rajashekar shastry, "Prediction of Diabetes Using Probability Approach", International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 02 | Feb -2017 www.irjet.net e-ISSN: 2395 -0056 p-ISSN: 2395-0072
18. J.Anitha, Dr.A.Pethalakshmi, "Comparison of Classification Algorithms in Diabetic Dataset", International Journal of Information Technology (IJIT) – Volume 3 Issue 3, May - Jun 2017
19. K.Saravananathan , T.Velmurugan, "Impact of Classification Algorithms in Diabetes Data: A Survey", The 3 rd International Conference on Small & Medium Business 2016 January 19 - 21, 2016, Nikko Saigon Hotel, Hochiminh, Vietnam
20. N.Deepika, Dr.S.Poonkuzhali, "Design of Hybrid Classifier for Prediction of Diabetes through Feature Relevance Analysis", IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 10, October 2015. www.ijiset.com ISSN 2348 – 7968
21. S. Peter, "An Analytical Study on Early Diagnosis and Classification of Diabetes Mellitus", Bonfring International Journal of Data Mining, Vol. 4, No. 2, June 2014
22. Mr. R. Sengamuthu , Mrs. R. Abirami , Mr. D. Karthik, "Various data mining techniques analysis to predict diabetes mellitus", International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 05 | May-2018 p-ISSN: 2395-0072