Detection of diabetes mellitus using machine learning algorithms

Ramya. G. Franklin*1, Muthukumar B2

1Research Scholar, Department of Computer Science and Engineering, Sathyabama Institute of Science and Technology, Chennai, Tamil Nadu, India
2Department of Computer Science and Engineering, Chennai Institute of Technology, Chennai, Tamil Nadu, India

Article History:
Received on: 24 Sep 2019
Revised on: 06 Dec 2019
Accepted on: 18 Dec 2019

Keywords:
Machine learning, Diabetes Mellitus, Diagnosis, Prediction

ABSTRACT
The growth of technology has brought in sophistication in our day to day activities. This sophistication has brought in many health issues. One among the most important problems that has currently become a typical issue is Diabetes Mellitus. Diabetes Mellitus has affected over 246 million people worldwide with a majority of them being women. The WHO reports that by 2025 this number is expected to rise to over 380 million. Prevention is better than cure. The health care data is large and complex. It is better to predict the disease at an earlier stage which may save the life and also have a preventive measure in controlling the diseases. In this paper, we have taken up a heterogeneous data to analyze the various factors which are affecting this disease. The various machine learning algorithms used in this paper help us to decide the attributes which play a major role in diagnosis of Diabetes Mellitus.

*Corresponding Author
Name: Ramya. G. Franklin
Phone: 9790540045
Email: mikella.prabu@gmail.com

ISSN: 0975-7538
DOI: https://doi.org/10.26452/ijrps.v11i4.3662

Production and Hosted by
Pharmascopc.org © 2020 | All rights reserved.

INTRODUCTION

Chronic increase of glucose level in the blood is called Diabetes mellitus (DM), commonly referred to as diabetes. It is caused by the inability of the body to produce required amount of insulin for its own needs. It may be due to the marred secretion of insulin or impaired action or both. High blood sugar levels over a prolonged period leads to renal failure, loss of vision and several other tissue damages. The incidence of diabetes is increasing because female diabetics are able to have children. The incidence of diabetes is higher in persons above 40 years of age. Females, especially the married ones are at a higher risk in getting this disease. Obesity, dietary factors and heredity are the other contributory factors for diabetes. Alcoholic beverages increase appetite, encourage weight gain and when taken in excess damage the pancreas and thereby increase the risk of diabetes. In short, DM leads to several metabolic disorders in our body. DM can be classified into several types. Mainly there are two clinical types

1. The juvenile onset type or insulin dependent diabetes (IDDM),
2. The adult or maturity onset type or non-insulin dependent diabetes

The main causes of Type 2 DM are

Life style Physical activity Heredity Diet
90% of the diabetic people are affected by Type 2 DM . Only 10% of the people are affected by Type 1 DM. The other type of diabetes is gestational diabetes which happens only during pregnancy it may not be at higher risk. This type of diabetes may develop into Type 2 diabetes

Signs and Symptoms

© International Journal of Research in Pharmaceutical Sciences 6881
Figure 1: Data set

Figure 2: Distribution of attributes
Table 1 shows,

**MACHINE LEARNING ALGORITHMS**

The following are few of the Machine learning algorithms,

1. Decision Trees
2. Naive Bayes Classification
3. Support vector machines for classification problems
4. Random forest for classification and regression problems
5. Linear regression
6. Logistic Regression
7. Ensemble Methods
8. K-means for clustering problems
9. Apriori algorithm for association rule learning problems

Frequent urination Increased thirst. Increased appetite. Loss of weight. General weakness and fatigue. Pain in the legs. Irritability. Lack of concentration. Prone to infection. Delayed wound healing
Figure 5: Ranking of features using decision tree

| Feature                  | Rank | Score     |
|--------------------------|------|-----------|
| Glucose                  | 1    | 0.558787  |
| BMI                      | 2    | 0.260081  |
| Pregnancies              | 3    | 0.134401  |
| Age                      | 4    | 0.036539  |
| Insulin                  | 5    | 0.010191  |
| Diabetes Pedigree Function | 6 | 0.000000 |
| Skin Thickness           | 7    | 0.000000  |
| Blood Pressure           | 8    | 0.000000  |

Figure 6: ROC curve for KNN and Decision tree

Figure 7: Confusion Matrix decision tree
Table 1: Literature Survey

| Author & year            | Data mining Techniques                          | Accuracy  |
|--------------------------|------------------------------------------------|-----------|
| (Hayashi and Yukita, 2016) | Re-RXalgorithm, J48graft                        | 90%       |
| (Agarwal et al., 2016)   | fuzzy ontology-based Case-based reasoning        | 76.78%    |
| Tirunagari et al. (2015) | SOM                                             | 83.5%     |
| (B and Srinivasan, 2014) | C4.5, SVM, k-NN, PNN, BL R                      | 86%       |
| (Mohammadi et al., 2014) | Bayesian Network, Decision Tree                  | 90.4%     |
| (Rao and null Kumar, 2014)| Genetic algorithm                               | 80.5%     |
| (Vijayanv and Ravikumar, 2014) | KNN, K means, ANPIS algorithm               | 80%       |
| (Beloufa and Chikh, 2013) | Bee Colony                                      | 80%       |
| (Lakshmi, 2013)          | C4.5, SVM, k-NN, PNN, BLR, LDA, k-means & Apriori | 74.7%     |
| (Koklu and Unal, 2013)   | Multilayer Perceptron, J48 and Navie Bayes Classifier | 74%       |
| (Zarkogianni et al., 2013) | SOM                                             | 81%       |
| (Ganji and Abadeh, 2011) | Ant Colony, FCS-ANTMINER, LDA, MWSVM            | 78.5%     |
| (Çalişir and Doğantekin, 2011) | J-48, WEKA tool                             | 89.74     |

PROPOSED SYSTEM

Attributes Details

Pregnancies
No. of times pregnant

Glucose - Plasma Glucose Concentration
A 2 hour in an oral glucose tolerance test (mg/dl) A 2-hour value between 140 and 200 mg/dL is called impaired glucose tolerance. This is called "prediabetes." It means you are at increased risk of developing diabetes over time. A glucose level of 200 mg/dL or higher is used to diagnose diabetes.

Blood Pressure
Diastolic Blood Pressure (mmHg) If Diastolic B.P > 90 means High B.P (High Probability of Diabetes) Diastolic B.P < 60 means low B.P (Less Probability of Diabetes)

Data Preparation
Data cleaning and transformation of data
The zero values or the null values in the following (Figure 1) cannot be zero. So the values have been replaced by the mean of that
Removing outliers

The below graph (Figure 2) shows the distribution of data set of different attributes. By carefully studying the graph, we can figure out that insulin and skin thickness have outliers.

The outliers are removed using IQR (Inter Quartile Range) method.

The algorithms used are decision tree algorithm and KNN algorithm

The decision tree model (Figure 3) was applied on the training dataset. The depth of the tree is 4 and the total numbers of nodes are 21. With the help of the decision tree we are able to select the features that play an important role in the detection of diabetes. The following (Figures 4 and 5) shows the important features and their ranking

RESULTS AND DISCUSSION

Comparison between Decision tree and KNN

A Receiver Operating Characteristic curve, i.e., ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve plotted for True Positive Rate and False Positive rate give us a good result (Figure 6). The curve area is 0.72, whereas, the curve area for KNN is 0.67. The dashed line in the diagonal represents the ROC Curve of a random predictor. It is a baseline to check if the model is useful or not. The confusion matrix is tabulated for both the methods (Figures 7 and 8).

CONCLUSIONS

The decision tree model has achieved 76% accuracy. After considering various options to improve the accuracy, we were able to achieve the desired accuracy by removing outliers, categorizing data and keeping the tree depth to 4. During this process, only few attributes out of the eight attributes play an important role. Glucose, BMI, Pregnancies, Age and Insulin were important. And also the factors like skin thickness, Diabetes Pedigree function and blood pressure had negligible effect. Hence we conclude decision tree model best among the two when compared with KNN.

Conflict of Interest

The authors declare that they have no conflict of interest for this study.

Funding Support

The authors declare that they have no funding support for this study.

REFERENCES

Agarwal, V., Podchiyska, T., Banda, J. M., Goel, V., Leung, T. I., Minty, E. P., Shah, N. H. 2016. Learning statistical models of phenotypes using noisy labeled training data. Journal of the American Medical Informatics Association, 23(6):1166–1173.

B, R., Srinivasan 2014. Predicting Diabetes by consequencing the various Data mining Classification.

Beloufa, F., Chikh, M. A. 2013. Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm. Computer Methods and Programs in Biomedicine, 112(1):92–103.

Çalişır, D., Doğantekin, E. 2011. An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier. Expert Systems with Applications, 38(7):8311–8315.

Ganji, M. F., Abadeh, M. S. 2011. A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis. Expert Systems with Applications, 38(12):14650–14659.

Hayashi, Y., Yukita, S. 2016. Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. Informatics in Medicine Unlocked, 2:92–104.

Kaur, G., Chhabra, A. 2014. Improved J48 Classification Algorithm for the Prediction of Diabetes. International Journal of Computer Applications, 98(22):13–17.

Koklu, M., Unal, Y. 2013. Analysis of a Population of Diabetic Patients Databases with.

Lakshmi, K. R. 2013. Utilization of Data Mining Techniques for Prediction and Diagnosis of Tuberculosis Disease Survivability. International Journal of Modern Education and Computer Science, 5(8):8–17.

Mohammadi, M., Hosseini, M., Tabatabaee, H. 2014. Using Bayesian Network for the Prediction and Diagnosis of Diabetes. BEPLS Bull. Env. Pharmacol. Life Sci. vol, 2(5):892–902.

Rao, V. S., null Kumar 2014. Applying Data mining Technique to predict the diabetes of our future.

Tirunagari, S., Poh, N., Hu, G., Windridge, D. 2015. Identifying Similar Patients Using Self-Organising. Vijayav, V., Ravikumar, A. 2014. Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus. International Journal of Computer Applications.

Zarkogianni, K., Litsa, E., Vazeou, A., Nikita, K. S. 2013. Personalized glucose-insulin metabolism
model based on self-organizing maps for patients with Type 1 Diabetes Mellitus. 13th IEEE International Conference on BioInformatics and BioEngineering, pages 1–4.