Diabetes Mellitus Prediction Using Machine Learning Techniques

G. G. Rajput1, A. Ashwini 2

1, 2 Department of Computer Science, Karnataka State Akkamahadevi Women’s University,

Vijayapura, Karnataka-586108

1[ggrajput@yahoo.co.in](mailto:ggrajput@yahoo.co.in), 2ashwinialashetty@gmail.com

**Abstract**

Diabetes Mellitus (DM) is a disease characterized by the body’s inability to metabolize glucose. Diabetes is found in every population in the world and in all regions and that the number of people with diabetes is steadily rising [1]. Machine learning tools enable the early detection and diagnosis of DM through an automated process which is more effective compared to manual diagnosis. The goal of this takes a look at changed into to construct a powerful predictive version with excessive sensitivity and selectivity to higher become aware of Canadian sufferers prone to having Diabetes Mellitus primarily based totally on affected person demographic statistics and the laboratory consequences for the duration of their visits to clinical facilities. Decision tree, K-NN, logistic regression and random forest techniques and ensemble of these are considered in the study applied on standard datasets namely Pima Indians Diabetes Database and Diabetes 130-US hospitals Data Set.

**Keywords:** Machine Learning, logistic regression, KNN, Diabetes Mellitus.

**I Introduction**

Epidemiology and international burden of diabetes is located in each populace withinside the global and in all regions, inclusive of rural components of low- and middle-earnings countries. The quantity of humans with diabetes is gradually rising, with WHO estimating there have been 422 million adults with diabetes global in 2014[2]. Medications and medicines couldalso be wont tocontrol it. Detecting the disease at the early stage helps reduce the risk of patients having more complicated health issues. The diabetes is diagnosed with the 2-hr post load plasma glucose being at least 200mg/dl [3] and the necessity of dying diabetes timely calls in various studies about diabetes recognition [4]. There are two common types of diabetes, Type1 diabetes and Type 2 diabetes. Type 1 Diabetes, known as insulin dependent diabetes, is commonly referred to as “Juvenile” diabetes because it is more common in children and adolescents but sometime it can be seen in people as young as 40 years. Type2 diabetes usually occurs in adults and mainly in overweight individuals. This problem appears when the pancreas fails to meet the high insulin requirement when body cells do not respond properly to insulin. There are few national assessments of diabetes care in the hospitalized patient which they could serve as a baseline for change. Analysis of a large clinical data base was undertaken to examine the patterns of diabetes care in patients with diabetes and to inform future directions might lead to improvements in patient safety [5]. Machine learning is the subfield of Artificial Intelligence that solves the real-world problems by providing learning skills to computer [6]. The first attempt of machine learning is the year of 1952 when Arthur Samuel developed the first game playing program for checker champion, later in 1957 Frank Rosenblatt created an electronic device which has the ability to learn how to solve complex problems by imitating the process in human brain [7]. Development of machine learning contributed to the greater use of computers in medical field [8]. Machine Learning (ML) and Artificial Intelligence (AI) tools helps in disease detection and diagnosis at an initial stage by an automated process and is more probable and efficient than the manual DM recognition method [9]. Several approaches have been proposed using ML and AI methods for DM control. A brief review of methods is presented in the following section.

II Related Work

Berina Alic et.al [10] presented overview of machine learning techniques in classification of diabetes and cardiovascular diseases (CVD) using Artificial Neural Networks (ANNs) and Bayesian Networks (BNs). they have performed comparative analysis on selected papers that are published within the amount from 2008 to 2017. In their study, they reported that Na'ive Bayesian network showed the absolute best accuracy values for classification of diabetes as 99.51%. They also reported that the calculation of mean accuracy of observed networks has shown better results using ANN.

In the study presented by Sidong Wei and Xuejiao [11], the authors explored deep neural networks (DNN) and Support Vector Machines(SVM) to identify diabetes. These techniques were examined by the accuracy of cross validation on Pima Indian dataset. They reported 77.86% accuracy using 10-fold cross validation using DNN.

Kemal Polat et. al [12] proposed a cascade learning system based on Generalized Discriminant Analysis and Least Square Support Vector Machine. Within the primary stage, Generalized Discriminant Analysis was used to discriminant feature variables between healthy and patient (diabetes) data as pre-processing process. LS-SVM was used to identify diabetes. LS-SVM resulted in 78.21% classification accuracy using 10-fold cross validation, whereas proposed system called GDA–LS-SVM yielded 82.05% classification accuracy using 10-fold cross validation.

Krati Saxena, et.al [13] used K-nearest neighbor algorithm for the diagnosis of diabetes mellitus and calculated accuracy and error rates for k=3,5. The results reported shows that because the worth of K increases, accuracy rate and error rate will also increase.

N. Sneha , et.al[14] proposed the analysis of the features in the diabetic dataset and selects the optimal features based on the correlation values. The decision tree algorithm and random forest yielded the absolute best specificity of 98.20% and 98.00% respectively. It holds the analysis of diabetic data. SVM and NB techniques gave the accuracy of 77.73% and 73.48%, respectively from the prevailing method and thus the proposed method improves the accuracy of the classification.

K.c Tan et.al [15] proposed Genetic algorithms (GAs) and support vector machines (SVMs) integrated i.e., GA-SVM hybrid approach for attribute selection. GA-SVM hybrid is subsequently validated using datasets obtained from the UCI machine learning repository. Further, a correlation measure between attributes as a fitness measure is used to reinforce the classification accuracy.

III Proposed methodology

The dataset was obtained from UCI machine repository [16]. The Health Facts database (Cerner Corporation, Kansas City, MO) is a national data warehouse that collects comprehensive clinical records across hospitals throughout the United States and also from the India.

Health facts data represents 10 years of clinical care at 130 hospitals and integrated delivery networks throughout the United States. The data set was created in two steps. Firstly, encounters of interest were extracted from the database with 55 attributes. Second, preliminary analysis and pre-processing of the data was performed resulting in retaining only features and encounters that could be used in further analysis, that is, contain sufficient information. The dataset consists of several medical predictor variables and one target variable that is outcome. Predictor variables include the number of pregnancies the patient had, their insulin level, BMI, age and so on.

Another data set namely Pima data set[ 16 ] consisting of eight attributes is also used in the study.

The proposed method consists of following steps. First step is to find the best data pre-processor for each classifier we have chosen, while the second step is to optimize the parameters of each classifier, third step is to compare these techniques of diabetes identification in terms of their accuracy and after that we will consider the features. The dataset consists of several medical predictor variables and one target variable, that is outcome.

Data Pre-processors

Pre-processing refers to the transformation applied to the infobefore feeding it to the algorithm. There are two essential data pre-processors called Principal Component Analysis [17] and Linear Discriminant analysis [18]. within thedata set of Clinical Database Patient Records in pre-processing steps another problem is missing data, lack of missing information of some features. as we willsee during thisdataset, we've50 attributes. Namely, as patient number, race, gender, age, admission type, time in hospital, medicineof admitting physician, number of lab test performed, HbA1c test result, diagnosis, etc. during thisdata set we'vesome missing data, namely Weight column has 96% of missing values followed by medical\_specialty(49%) and payer code(39%). quite80% of knowledgewere missing in these columns, hence we cannot work on 20% data, for this reason here we cannot apply mean values or the otherimputation technique.

hence, we are dropping these columns for the greatresult. In pima Indian data we used the Feature scaling method to limit the range of variables inorder thatit is oftencompared with the footing. It fetches the peripheryand it fits the range of the limit. during thisdata set Glucose, vital sign**,** Skin thickness, insulin and BMI attributes have missing values, for solving these missing data, we used the mean of all existing data to fill the missing values, and this method is named as imputation. We propose a classification model with boosted accuracy to predict diabetic patient. Here during this paper we used popular classifiers like KNN, Logistic Regression, Random Forest and Ensemble. The main focus is on increasing the accuracy by using resample techniques.

Classification: Classification may be apredictive modelling problem where a category label is predicted for a given example of input file. it’sa supervised learning mechanism which predicts a category for an input variable. Linear classification approach includes Logistic regression and support Vector Machines. Nonlinear classification models include K nearest neighbors, kernel support vector machines, Naïve Baye’s, Decision tree classification and random forests [ 22 ]. During thispaper, we use Random Forest, K nearest neighbor, Logistic regression and ensemble of those to review diabetics prediction. a quick description these classifiers are presented below.

Random Forest: Random Forest model are oftenthought of as bagging when deciding where to separate and the way to formdecisions, bagging decision trees have the complete disposal of features to settle onfrom. Random forest model implements A levelof differentiation because each tree will split supported different features. This level differentiation provides a greater ensemble to aggregate over and it produces a more accurate prediction. Applied sk-learn models of coaching, testing and cross validation methods. By taking the bulkvote are going to be held during thisclassification tree.

K-Nearest Neighbors: KNN algorithm are going to beused for regression also as classification problems. It’salso called as lazy learner algorithm because it doesn’t learn from the training dataset immediately, instead it stores the info set and at the time of classification it performs an action on the data set [18]. The classifier selects the amount k of the neighbors. Calculate the Euclidean distance (or other distance measures [ 22 ]) of k number of neighbors. Takes the k nearest neighbors as per the calculated Euclidean distance, among these k neighbors, counts the amount of the infopoints in each category, and assigns the new data points there tocategory that the amountof the votes is maximum.

Logistic Regression: logistic regression is predictive analysis algorithm and supportedthe concept of probability. The hypothesis of logistic regression tends to limit the valuefunction between 0 and 1. 0≤hⱺ(x)≤1 this is oftenthe logistic regression hypothesis expectation. The independent variables can each be a binary variable ( two classes, coded by an indicator variables). The corresponding probability of  the worth labelled “1” can vary between 0 and 1, hence it’sthe labelling; the function that converts log-odds to probability is that thelogistic function of the name. the unit of log measurement for the log-odd scale is understoodas logit, from logistic unit; hence analogous models with a specialsigmoid function rather thanthe logistic function also canbe used like **the** probity model, the defining characteristics of the logistic model is that increasing one among the independent variables multi scales the chances of the given outcome at a continuing rate with each  variable having its own parameter [19].

Ensemble learning helps improve classification results by combining several classification models. It combines techniques into one optimal predictive model so as to decrease variance, bias or improve predictions. During thispaper, for prediction of diabetics we combine KNN, logistic regression and random forest techniques to get a replacementmodel. There are three differing typesof ensemble techniques namely, Max Voting, Averaging and weighted averaging. During this paper we used Max Voting method, the Max Voting method used for classification problems. The prediction of everymodel is going to beconsidered as a ‘Vote’. Prediction of the generalmajority of the models are wont to be final prediction.

IV. Results and Discussions

This study uses the data obtained from the Center for Machine Learning and Intelligent Systems at University of California, Irvine [23] It contains clinical records from over 100,000 individual encounters corresponding to over 60,000 distinct patients and contains several attributes which correspond to the times of admission and discharge of diabetic patients. These records contain information about various laboratory tests and procedures, diagnosis, and medications that were administered in the duration of the hospital stay. Following are the variables selected for modelling: The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medications, diabetic medications, number of outpatients, inpatient, and emergency visits in the year before the hospitalization, readmit etc.

PIMA dataset used in this study is from the National Institute of Diabetes and Digestive and Kidney Diseases. The records belong to female patients at least 21 years old of Pima Indian heritage [24]. The prediction is done based on the parameters; Glucose, IBM, insulin, thickness, of the skin, blood pressure, and age.  In this experiment five machine learning algorithms were used. These algorithms are Decision Tree, Random Forest, Logistic Regression, KNN and Ensemble model. All these algorithms were applied on pima Indian dataset and HbA1c readmission rates. Data was divided into two portions, training data and testing data, both these portions consisting 80% and 20% data respectively. All these five algorithms were applied on same data set. Predicting accuracy is the main evaluation parameters that we used in this experiment. Below table.1 shows the Accuracy of the given data sets namely, Pima Indian Diabetes and HbA1c readmission rates. The accuracy of the prediction for each of these models including ensemble model is shown in the table 1 and table 2. for PIMA and HbA1c readmission rates datasets, respectively. From the table, it is evident that random forest and ensemble yielded similar results with better accuracy compared to KNN and decision tree techniques.

Table.1. Prediction results of classifiers for Pima Indian Diabetes dataset consisting of 768 records with 9 attributes

|  |  |
| --- | --- |
| **Models** | **Accuracy** |
| KNN | 85% |
| Logistic Regression | 81% |
| Random Forest | 87% |
| Ensemble | 88% |

Table 2. Prediction results of classifiers for Clinical Database Patient Records with 100000 records and 50 attributes

|  |  |
| --- | --- |
| **Models** | **Accuracy** |
| Logistic regression | 88.6% |
| Random Forest | 88.7% |
| Decision Tree | 83.9% |
| Ensemble model | 88.7% |

The experimental results obtained in terms of accuracy with the proposed approach are compared with results of other methods in literature. The model used, dataset , and accuracy is shown in table 3.

Table 3. Comparison with other methods in literature

|  |  |  |  |
| --- | --- | --- | --- |
| Authors | Prediction model | Dataset Used | Accuracy (%) |
| A.iyer (2015) | Decision tree | Pima Indian diabetes data | 74.8% |
| N. Gupta(2013) | Decision Tree | Pima Indian diabetes data | 81.33% |
| Lee(2014) | Decision Tree | National Health and Nutrition Survey | 67% |
| Naveen Kishor G, V. Rajesh (2020) | KNN | National institute of diabetes and digestive disease. | 71.3% |
| Naveen Kishor G, V.Rajesh(2020) | Random Forest | National institute of diabetes and digestive disease. | 74.4% |
| Naveen Kishor G, V. Rajesh (2020) | Logistic Regression | National institute of diabetes and digestive disease. | 72.39 |
| Proposed | Ensemble method | Pima Indian Diabetes dataset  Clinical Database Patient Records | 88.00%  88.7% |

**VI. Conclusion**

In this paper, we proposed method to discriminate between diabetes effected patients and non-diabetes patient using machine learning algorithms. Experiments evaluated on world data extracted by the pima Indian population and HbA1c measurement on hospital Readmission rates. By using all four machine learning algorithms we had measures different parameters within the data set and we had come through better accuracy rate with Random Forest and Ensemble model with nearly 88.7% This work can be extended by adding any other algorithms which can give better accuracy then Random Forest and Ensemble model. we applied 10 cross validations for both the dataset and compared individual algorithms.

**References**

1. Beata Strack, Jonathan P.Deshazo, Chris Gennings, cios and John N.clore, “impact of HbA1c measurement on hospital readmission rates: analysis of 70000 clinical database patient records”, vol 2014, pp: 1-11.
2. Sidong wei, Xuejiao Zhao, Chunyan Mia, “ a comprehensive exploration to the machine learning techniques for diabetes identification” pp 291-295.
3. World Health organization, “report of a study group: diabetes mellitus” world health organization technical report series, Geneva, 727, 1985
4. A.Ghaheri, S. Shoar, M. Naderan and S.S.Hoseini, the applications of genetic algorithms in medicine. Oman medical journal 2015, 30(6), 406..
5. Pang-ning tan, Michael Steinbach, vipinkumar, “introduction to data mining” text book.
6. Kemal al.polat salih gunes and Ahmet Arslan,“ a cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine” expert systems with applications, vol. 34.1, January, 2008, pp. 482-487.
7. Beata strack, Jonathan p.deshao, chris gennings, juan l.olmo, “impact of HbA1c measurement on hospital readmission rates: analysis of 70000 clinical database patient records. Vol 2014, pp 1-11.
8. N.sneha and tarun gangil, “analysis of diabetes mellitus for early prediction using optimal features selection” pp 1-19, 2019.
9. Wold s, eshensen k. and geladi p., “principal component analysis” chemometrics and intelligen laboratory systems, vol. 2. Pp 37-52, 1987.
10. Berina Alic, lejla Gurbeta and almir Badjevic, “machine learning techniques for classification of diabetes and cardiovascular diseases”, vol 2017, IEEE.
11. Sidong wei, Xuejiao Zhao, Chunyan Mia, “ a comprehensive exploration to the machine learning techniques for diabetes identification” pp 291-295.
12. Kemal al.polat salih gunes and Ahmet Arslan,“ a cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine” expert systems with applications, vol. 34.1, January, 2008, pp. 482-487.
13. Krati saxena, zubai khat, et. “Diagnosis of diabetes mellitus using K-nearest neighbor algorithm” vol.2 issue 4, July 2014, pp 36-42.
14. N.sneha and tarun gangil, “analysis of diabetes mellitus for early prediction using optimal features selection” pp 1-19, 2019.
15. K.C.Tan, E.J.Teoh, Q. Yua, K.C. Goh, “A hybrid evolutionary algorithm for attribute selection in data mining”, 2008 Published by Elsevier Ltd.
16. https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008
17. N.sandhya, k.R. charanjeet A review on machine learning techniques, international journal on recent and innovation trends in computing and communication, 2016, ISSN: 2321-8169, 395-399.
18. S.saru and s. subashree, “ analysis and prediction of diabetes using machine learning”, vol.5, April 2019. Pp 167-175.
19. Krati saxena, zubai khat, et. “Diagnosis of diabetes mellitus using K-nearest neighbor algorithm” vol.2 issue 4, July 2014, pp 36-42.
20. Kumari, M, Yohra, R., and Arora, A(2014). Prediction of Diabetes using Bayesian Network.
21. R.M Bergenstal, J.L.Fahrbach, et. “Preadmission gycemic control and changes to diabetes mellitus treatment regimn after hospitalization,” vol.18, no.3, pp.371-375, 2012.
22. Data Mining: Concepts and Techniques Book by Jiawei han.
23. [https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008 and Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014].
24. [Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *In Proceedings of the Symposium on Computer Applications and Medical Care* (pp. 261--265). IEEE Computer Society Press.]