Prognosis of Diabetes Mellitus using Machine Learning Techniques

Vidya J[1], Swastika T Jain[2], Shyamala Boosi[3], Bhanujyothi H C[4], Dr.Chetana Tukkoji[5]
[12345] Assistant Professor, CSE Dept., GITAM School of Technology, Bengaluru.

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: Diabetes mellitus is a condition caused due to increase in blood glucose level. More than 90% of people are diagnosed with Type 2 diabetes disease,T2D is a fast-growing, chronic disease caused by the imbalance in insulin function. Diabetes is a now the leading cause of heart disease, stroke, blindness, non-traumatic limb amputations and end-stage renal failure. Early detection may take a step towards keeping diabetes patients healthy and it also reduces the risk of such serious complications. Nowadays, the application of Machine learning in the medical field is gradually increasing. This can aid in improving the classification system used for disease diagnosis, that assist medical experts in detecting the fatal diseases at an early stage. This paper presents a performance comparison of the machine learning algorithms in diabetes detection. Techniques like SVM, Random forest, Gradient Boosting, Navie Bayes, Logistic regressionand KNN are used in this work.

Keywords: Diabetes detection, Machine learning, Random Forest, T2D.

I. Introduction

Diabetes mellitus (Diabetes) is a rising concern in India, with an estimated 8.7% of the diabetic patients aged 20 and 70 years [6]. It is a chronic disease, that occurs either when the pancreas fails to produce enough insulin (which is a blood sugar-regulating hormone) or when the produced insulin is not used efficiently by the body. Hyperglycaemia, or raised blood sugar, is often a common effect of diabetes that leads to severe damage to many of the organs, particularly the blood vessels and nerves, over time [1].

Diabetes is a global problem affecting many people. Around 9.3 percentage of the world adult population were diabetes patients in 2019 - by the year 2045 this number is expected to grow almost 11 percent[2]. It has no cure, one can take proper measures to manage diabetes and stay healthy[4].

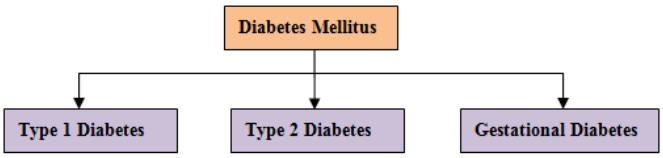

Figure 1. Types of Diabetes.

There are 3 types of diabetes as shown in figure 1. Gestational diabetes is a type of diabetes which can be seen in pregnant women when the body becomes less sensitive to insulin and it will be resolved after giving birth. These patients have greater chance of being affected by T2D later in life. Juvenile or type 1 diabetes is caused when the body fails to produce enough insulin, as the cells that makes insulin will be destroyed. Usually it is diagnosed in young adults and children; although it may appear at any age. These patients are insulin-dependent and must take artificial insulin on a daily basis to sustain [4]. T2D is caused when the cells of a body is not responding effectively, this is the most common type of diabetes which has strong link with obesity. People with lower BMI (Body Mass Index are likely to get affected with T2D [3]. It can be developed at any age. As time goes, it may lead to severe health complications like stroke, heart disease, eye problems, kidney problems, nerve damage, hearing problems, Alzheimer's disease [11], foot problems and also dental problems [4].

Normal blood sugar level sit in between 70 to 99 mg/dL, whereas a diabetic patient will have a fasting sugar level higher than 126 mg/dL. Some people have borderline diabetes or prediabetes whose blood sugar level will be in the range of 100 to 125 mg/dL (milligrams per deciliter) and they are at a high risk of developing T2D. Some of the risk factors are being overweight, family history of diabetes, having a sedentary lifestyle, age more than 45, history of PCOS and so on [5].

Medical Expert System is one of the active research areas where medical experts and the data analysts are collaborating continuously in order to make the prediction systems more accurate and useful in real life. Recent surveys by WHO indicates a rise in the count of diabetic patients and the demise that are attributed to blood glucose level each year[15].Early detection and treatment is very essential, as it is a vital cause for cardiovascular disease [10].

ML is a subfield of AI, which allows the system to learn based on the past examples, experience, history and data, it has been making great progress in many directions including the medical field to detect diseases. To diagnose diabetes more efficiently, an accurate detection technique and a good prediction model is required.

This paper presents a performance comparison work of different machine learning algorithms for forecasting diabetes.
The residuum of this paper is arranged as follows: Section II describes the related work previously done. Section III describes the dataset attributes. Section IV explains the methodologyand the different algorithms used for diabetes detection. Section V describes the evaluation metrics and the results, and section VI concludes the work.

II. Related work

Pahulpreet Singh Kohli et al, have used various classification algorithms on three datasets: Breast Cancer, Heart and Diabetes for early forecast of diseases. Selection of features for each dataset was done by backward modeling using p-value measure. Firstly data is explored in Python environment, next missing values are identified and they are replaced by mean value in case of a categorical variable or a continuous variable. In feature selection step, based on the p-value the attributes are eliminated. The attributes with p-value more than 0.05 were removed and the model will be refitted with the rest of the variables. This was repeated until all the variables came to a significant level. In order to measure the proportion of difference described by the independent variables which contributes to the prediction of target variable, R square value will be observed after every iteration. For selected features algorithms like Decision trees, Random forest, SVM, Logistic regression and Adaptive boosting were applied and prediction accuracy was compared through Train/Test split method. These steps can be automated in future and for data preprocessing pipeline structure can be used to improve the results.[**7**]
Adel Al-Zebari et al, have used different machine learning algorithms for forecasting diabetes at an early stage. Discriminant Analysis, K-Nearest Neighbors, Support Vector Machine, Logistic Regression and Ensemble learners supervised machine learning algorithms have been used for classification. MCLT (Matlab Classification Learner Tool) has been used in their work, for data classification, dimension reduction, selection of feature, feature analysis and evaluation of performance. The dataset is considered in 10-fold cross validation manner. The Logistic Regression method gave the best accuracy score. [**8**]
G. A .Pethunachiyarhas used Support Vector Machine with various kernel functions to forecast the diabetes at an early stage. The dataset has been taken from UCI machine learning repository. The detection process involves 5 steps: Initially, data is selected and errors such as missing values, wrong information and inconsistency in data are rectified. Next 70% of the data is considered as training data and remaining as testing data. Using SVM a model is built for training data. To make predictions on the resulting value generated, test data are applied to the built model. Linear, Polynomial and Radial kernel functions are used. Support Vector Machine with Linear kernel produced highest accuracy value. [**9]**
M.Shanthi et al, have proposed a model for diagnosing T2D through ELM (Extreme Learning Machine) method. The mathematical model ELM has a single hidden layer feed forward network, hidden nodes can be generated randomly. Initially, parameters are generated for the hidden nodes randomly. Next output matrix is calculated and then the optimal weight of the network will be given as output. The output is obtained from the features, input weight and the activation functions. The available activation functions are sine, triangular basis, sigmoid and hard-limit. This EKM model assist medical experts to forecast the type 2 diabetes.[**10]**
Md. Kowsher et al. have proposed a prediction model for type 2 diabetes. They have emphasized on machine learning algorithms like KNN, logistic regression, Decision tree, random forest, Navie Bayes, ANN and Linear Discriminant Analysis for diabetes prediction. The workflow is divided into 4 parts: data collection, preprocessing of collected data, training the data and making predictions. The 80% of the dataset is chosen as training dataset and remaining as testing dataset. The data is preprocessed on order to convert it into a recognizable format. Missing values are replaced by mean. Features are selected, the features that have no impact on removed. Next feature scaling is done. Dimensionality reduction is done to minimize random variables that avoids overfitting. The training dataset is applied to the algorithm to assess the model performance and to find out medications. **[11]**
Mr. Gaurav Shetty et al presents a comparative analysis of various algorithms like XGBoost classifier, Decision tree-based ensemble classifier, Random forest, and AdaBoost classifier for forecasting type-2 diabetes. Initially the dataset is pre-processed, missing values are replaced with the median. Next the dataset is divided into training and testing data in order to avoid underfitting and overfitting problems. The above mentioned four classifiers are used to train the model. These classifiers has different features and the models were trained with different combinations of these features. The preprocessing technique increased the accuracy. **[12]**
Jaimin Shah et al, have used classification algorithms to predict diabetes in distributed environment. Classification algorithms like Support Vector Machine (SVM), Artificial Neural Networks (ANN) and Navie Bayes are used in this work. Apache Spark is used for classification that could handle big data applications very efficiently. This could help organizations to develop the data solutions fast and efficiently. The parameters considered for classification are BMI, insulin level, BP and so on. The results are compared based on the accuracy, recall, precision etc. **[13]**

Samrat Kumar Dey et al, have developed an application to predict diabetes. For disease prediction, KNN, ANN, SVM and Navie Bayes algorithms are used. The dataset has been divided into 2 subparts one for training and another part for testing. In order to increase accuracy, Min Max Scaler (MMS) normalization method is used. To implement machine learning model tensorflow is used. PHP language is used for backend development and Javascript is used for frontend design.To train the ANN model, dataset values are collected from SQL database. For disease prediction user has to enter some information like serum insulin, BP, BMI and so on. The application will predict whether the test result is positive or negative. **[14]**
Maham Jahangir et al, have presented a diabetes prediction framework which is an application of automatic-multilayer perceptron (AutoMLP) that is combined with an enhanced class outlier detector. This is auto-tunable and can optimize the parameters automatically during the training process. This system consists of 2 phases: pre-processing involves outlier detection based on class factor and outlier-free dataset is used for training AutoMLP. In the second stage AutoMLP will classify the diabetic patients. The attributes used for diabetes prediction are plasma glucose level, BP and number of times pregnant. The proposed system gave 88.7% accurate results. **[15]**

III. Dataset description

The dataset contains few medical predictor variables like BMI, insulin, number of pregnancies and so on and a target variable outcome. The table 1gives the description of the attributes. Based on these attributes, diabetes will be predicted.

Table 1. Description of the attributes.

Sl.No	Attribute	Explanation
1.	Pregnancies	No. of times the patient was pregnant
2.	Glucose	Concentration of glucose
3.	Blood Pressure	Diastolic BP (mm Hg)
4.	Skin Thickness	Triceps skin fold thickness in mm
5.	Insulin	serum insulin (mu U/ml)
6.	BMI	Body mass index (weight in kg/(height in m)^2)
7.	Diabetes pedigree function	scores likelihood of diabetes on family history basis
8.	Age	Age of the patient (years)
9.	Outcome	Class variable (0 or 1) 268 of 768 are 1, the others are 0

IV. Methodology

Step 1: Initially Diabetes patient's data is collected.
Step 2: In this step data is pre-processed for eliminating wrong data, redundant data, filling missing data and so on.
Step 3: Dataset is divided into Training and testing dataset.
Step 4: Different algorithms are used to diagnose the disease.
Step 5: Result is compared to identify the training model that gives more accuracy.
The below figure 2, is the proposed model for detection of diabetes.

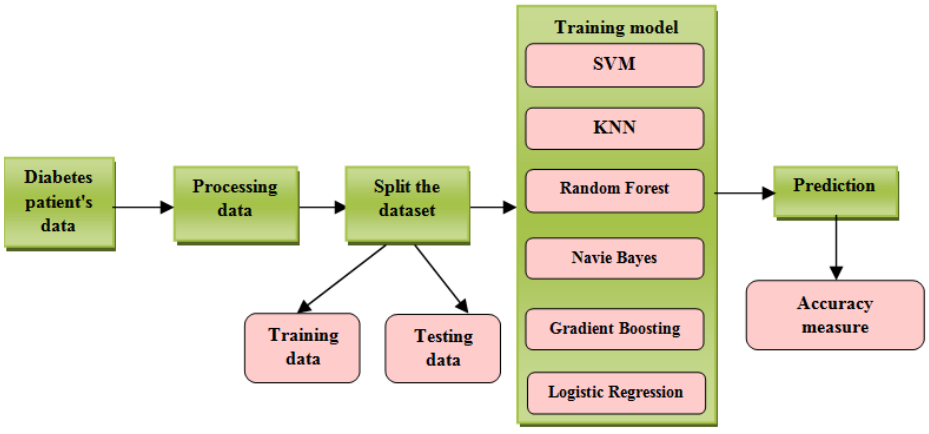

Figure 2. Diabetes detection process.

Training models
The process of training involves machine learning algorithm, here 6 different machine learning algorithms have been used along with the training data to learn from. During the training process the algorithm will find the

pattern in the training data which maps the input data to the target and it produces the model that could capture these patterns. The below figure 3, shows the machine learning techniques used for diabetes prediction.

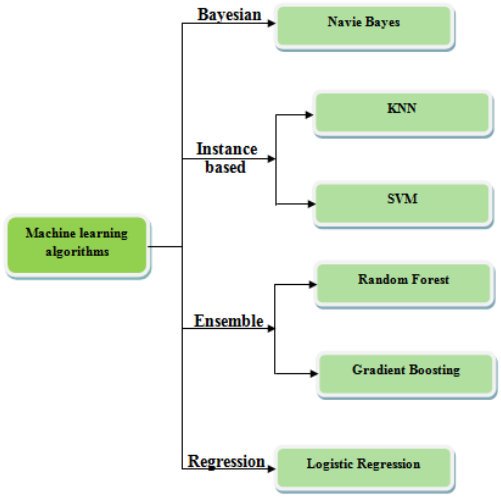

Figure 3. Training models used.

1. Logistic Regression

Logistic Regression is one of the supervised learning method used in classification problems, it is based on the concept of probability. It will assign the observations to a discrete set of classes and transforms the output using the sigmoid function and it returns a probability value. It can be used in healthcare applications, online transactions fraud and so on.

2. Support Vector Machine (SVM)

SVM is one of the supervised machine learning algorithms, usually used for classification purpose and can also be used with regression challenges.

3. K-Nearest Neighbor

It is one of the simplest supervised machine learning algorithm that could be used to solve regression and classification problems. It assumes that the similar things exists in close proximity.

4. Random forest

Random forest algorithm is one of the supervised learning model which uses labeled data to learn how to classify unlabeled data. By using this algorithm both regression as well as classification problems can be solved. It can be used in banking sector, stock market, e-commerce, medicines and so on.

5. Navie Bayes

It is one of the popular classification algorithm that is most widely used to get the base accuracy of the dataset. It makes an assumption that all the variables present in the dataset are Navie (not correlated to each other). It can be used in real-time prediction, multi-class prediction, spam filtering, sentimental analysis, text classification, recommendation system and so on.

6. Gradient Boosting

It is one of the machine learning techniques used for classification and regression problems, it produces prediction model in the form of ensemble of weak prediction models.

V. Evaluation metrics and Results

1. Accuracy: The accuracy score computes the accuracy, the fraction or count of correct predictions.

$$accuracy = \frac{TP + TN}{(FP + FN + TP + TN)}$$

2. ROC AUC curve: A Receiver Operating Characteristic (ROC) curve is a graph that shows the performance of a classification model at all the thresholds. It plots 2 parameters: True positive rate (TPR) and False positive rate (FPR).

$$TPR = \frac{TP}{(TP + FN)} and FPR = \frac{FP}{(FP + TN)}$$

3. Cross validation: This technique is used to train the model with the subset of dataset and then evaluating using complementary subset of the dataset.

4. Confusion matrix: It is an NxN (where N represents the number of target classes) matrix used to evaluate the performance of a classification model.

	Actual values

		Positive	Negative
Predicted values	Positive	TP	FP
	Negative	FN	TN

Figure 4. Confusion matrix.

where, TP indicates that the actual and the predicted values are matching and both are positive.
TN indicates that the actual and the predicted values are matching and both are negative.
FP indicates that the model predicted positive value while the actual is negative.
FN indicates that the model predicted negative value while it was positive actually.

The attributes like BMI, insulin, pregnancies, skin thickness, glucose level, BP, diabetes pedigree function and patient's age. The values of all these attributes are numbers. The dataset includes 2000 subjects. We opted to have a 10-fold cross-validation for evaluating the result. The accuracy score and ROC results are given in the below table 2. All the techniques used produces accuracy score around 70% and the results shows that Random forest gives highest accuracy score of 97%. The accuracy score and ROC AUC curves are shown in figure 5 and figure 6 respectively.

Table 2. Accuracy scores and ROC values obtained by various ML techniques.

Sl.No.	Algorithm	Accuracy	ROC
1.	Navie Bayes	0.737	0.672
2.	KNN	0.845	0.805
3.	SVM	0.782	0.699
4.	Random Forest	0.972	0.963
5.	Gradient Boosting	0.882	0.845
6.	Logistic Regression	0.800	0.714

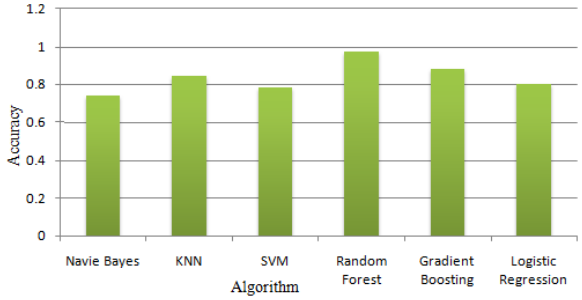

Figure 5. Accuracy score graph.

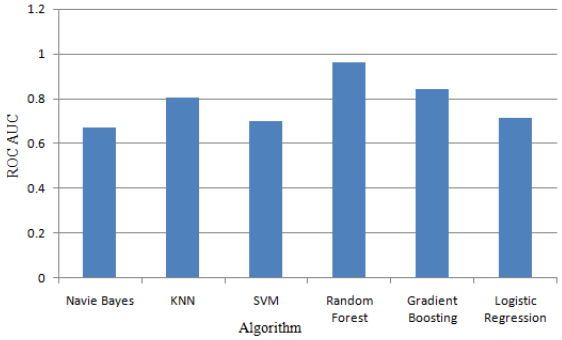

Figure 6. ROC AUC curve.

VI. Conclusion:

As diabetes may lead to other maladies like heart diseases, blindness, stroke, and so on. Early diagnosis of diabetes is very essential as it may help the patients to stay healthy. So, in this paper different machine learning algorithms like KNN, SVM, Gradient Boosting, Random Forest, Logistic regression, and Naive Bayes have been used for diabetes prediction, and among these techniques, Random forest gives more accurate results for diabetes detection. In future, deep neural networks can be applied to increase the accuracy of classification.

References

1. https://www.who.int/news-room/fact-sheets/detail/diabetes Accessed on: 23 -Sep-2020
2. https://www.statista.com/statistics/271464/percentage-of-diabetics-worldwide/ Accessed on: 23 -Sep-2020
3. https://www.medicalnewstoday.com/articles/323627#types accessed on 23/9/2020
4. https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes Accessed on: 23-sep-2020
5. https://www.medicalnewstoday.com/articles/323627#how-insulin-problems-develop Accessed on: 23-sep-2020
6. http://origin.searo.who.int/india/topics/diabetes_mellitus/en/ Accessed on 30/9/2020
7. Pahulpreet Singh Kohli, Shriya Arora, "Application of Machine Learning in Disease Prediction", 2018 4th International Conference on Computing Communication and Automation (ICCCA).
8. Adel Al-Zebari, AbdulkadirSengur, "Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection", 2019 IEEE.
9. G. A .Pethunachiyar, "Classification Of Diabetes Patients Using Kernel Based Support Vector Machines", 2020 International Conference on Computer Communication and Informatics (ICCCI -2020), Jan. 22-24, 2020, Coimbatore, INDIA.
10. M.Shanthi, RamalathaMarimuthu, S.N.Shivapriya, R.Navaneethakrishnan, "Diagnosis of Diabetes using an Extreme Learning Machine Algorithm based Model".
11. Md. Kowsher, MahbubaYesminTuraba, Tanvir Sajed, M MMahabubur Rahman, "Prognosis and Treatment Prediction of Type-2 Diabetes Using Deep Neural Network and Machine Learning Classifiers", 2019 22nd International Conference on Computer and Information Technology (ICCIT), 18-20 December 2019.
12. Mr. Gaurav Shetty, Dr. Vijay Katkar, "Type-II Diabetes detection using Decision-tree based Ensemble of Classifiers", 2019 5th International Conference on Computing Communication Control and Automation (ICCUBEA).
13. Jaimin Shah, Raj Patel, "Classification techniques for Disease detection using Big-data", 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT).
14. Samrat Kumar Dey, Ashraf Hossain, Md. Mahbubur Rahman, "Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm", 2018 21st International Conference of Computer and Information Technology (ICCIT), 21-23 December, 2018.
15. Maham Jahangir, Hammad Afzal, Mehreen Ahmed, KhawarKhurshid, Raheel Nawaz, "An Expert System for Diabetes Prediction using Auto Tuned Multi-Layer Perceptron", Intelligent Systems Conference 2017 7-8 September 2017| London, UK.