2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN) | 979-8-3503-4798-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/VITECON58111.2023.10157531

Diabetes Prediction using different Machine Learning Classifiers

Tanya Gupta, Mithul Raaj A T, Rani C, Rajesh Kumar M
Vellore Institute of Technology, Vellore, India
tanya.gupta2020@vitstudent.ac.in, mithulraaj.at2020@vitstudent.ac.in, crani@vit.ac.in,mrajeshkumar@vit.ac.in

***Abstract*— Diabetes is a persistent medical condition caused due either when pancreas doesn't secrete as much insulin as the body needs or the body is unable to use insulin efficiently. Diabetes greatly increases the risk of many heart diseases. Early diabetes diagnosis can result in more effective therapy. This paper proposes a method to predict diabetes using different machine learning algorithm methods. The accuracy obtained after using different machine learning models, that is Logistic Regression, KNN, Naive Bayes, SVM, Decision Tree and Random Forest, has been compared.**

Keywords*— Classifier, Outliers, Decision tree, Interquartile range.

I. INTRODUCTION

Diabetes is a life-long medical condition that affects human body adversely. Inefficient insulin affects body phenomenon of breaking down food into energy. Firstly, body breaks down food into glucose which is released in bloodstream. Once sugar level rises, it triggers pancreas to secrete insulin. Insulin acts as a carrier to take glucose to body cells as per the energy requirement. A diabetic body cannot can't secrete or use sufficient insulin. Deficiency or insulin increases blood sugar level in the bloodstream. Over a period of time, increased sugar level can cause serious health issues such as kidney disease, reducing eyesight, and heart disease.

Diabetes is responsible for around 460,000 deaths due to kidney disease, and blood pressure. In 2019, over 1.5 million people died due to diabetes, out of which 48% were under 70 years of age.

A. Diabetes type one

Diabetes type one (insulin-dependent diabetes), is a persistent medical condition that arises when the pancreas generates inadequate or no insulin. Insulin is a hormone which is responsible for controlling blood sugar level. Several factors, including heredity and certain viruses, can contribute to the development of type 1 diabetes. Although it commonly begins during childhood or adolescence, adults can also be affected by this condition. Unfortunately, type 1 diabetes remains incurable despite extensive research efforts. Treatment focuses on managing blood sugar levels using insulin, dietary modifications, and lifestyle changes to prevent complications.

B. Diabetes type two

Diabetes type two is a medical condition which affects the synthesis and use of glucose in the body. It is the most common diabetes, which usually develops in adults. Due to this condition, an excess of sugar circulates in the body, which over time, causes diseases related to nervous, circulatory and immune system. Causes of type two diabetes can be broadly divided into two categories. Firstly, inability of pancreas to produce sufficient insulin, the hormone which maintains blood sugar level in the body. Secondly, the cells do not respond well to insulin and are not able to absorb sugar efficiently.

C. Health Consequences

Over time, diabetes can have detrimental effects on various parts of the body, that can be eyes, heart, blood vessels, nerves, and kidneys. People suffering from diabetes have a greater risk of experiencing heart attacks or strokes, with a likelihood that is two to three times higher than those without diabetes. Foot neuropathy, a condition of nerve damage and a series of consequences such as depreciated blood flow, increases the risk of foot ulcers, which ultimately causes limb amputation.

Diabetic retinopathy, a condition which occurs due to prolonged damage to blood vessels in the retina, is a significant cause of blindness, affecting approximately one million people worldwide. Furthermore, diabetes is one of the leading causes of kidney failure.

II. RELATED WORK

In this section, we will review about recent studies related to diabetes prediction using different machine learning methods.

This paper [1], proposes a new super ensemble learning model for early diagnosis of diabetes. This algorithm uses results of more than one machine learning algorithm to make the most accurate prediction. The proposed model claimed to attain an accuracy of 99.6%.

In [2], Diabetes is predicted using a list of classifiers like KNN, RF, DT, NB, AB, XGBoost using the pipelining method. Parameters like neuron initializers, batch size, number of epochs and learning rates were used for determining the hidden layers. Tuning of results to achieve better results were done using grid search methodologies.

XG Boost method's strength is shown in this paper [3] and it is because of the highly organized form of the dataset used here. But getting a dataset like that is not possible every time.

Apart from different ML methods, [4] discusses about ANN which is actually comes under the category of Deep Learning method. Even then, it doesn't give that process a success as it falls short to 4th place in accuracy when compared with other methods like XB, SVM, AT, AB, KNN. Also, the advantages of using Amazon Web Services as a cloud application and its ability to become an IoT based

978-8-3503-4798-2/23/$31.00 ©2023 IEEE

SaaS method to enable commercial implementation easily was mentioned.

Feed Forward neural networks using single hidden layer using linear activation function and non-linear function to differentiate between two to N observations was explained in [5]. It has several benefits like smaller training error and a faster learning process than traditional multi-layered networks.

In [6], they have explained the use of K fold cross validation techniques and the result change observed before and after using it. Logistic regression turns out to be a winner with 77.22% accuracy showing that it responds well to the technique rather than Gradient Boosting method which had highest accuracy at the beginning.

Case study review of different data mining techniques was done in [7] and the steps to pre-process data like in terms of correlation were discussed in detail. The challenges faced during dataset selection with relevant attributes may tend to have direct consequence on final values.

This study [8] developed an efficient method for classifying diabetes using the SMO algorithm for position estimation. The proposed method outperformed existing methods when working with real-time data and was found to be robust and reliable due to its ability to handle weak input cases and being free from outliers. Further research is needed to validate its effectiveness in larger datasets. This study contributes to the development of accurate methods for diagnosing diabetes, important for effective healthcare management and prevention.

In [9], they have proposed an approach based on Euclidian distance parameter using NDDM approximation to predict the risk of type 2 diabetes in patients. Data is pre-processed and cleaned with better approximation and accuracy has model has increased drastically as compared to other regression models.

Result of different machine learning classifiers has been compared in this paper, [10]. Six machine learning classifiers namely LR, NB, DT, KNN and RF is used and corresponding results have been compared and the best model has been used for early prediction of diabetes in a patient.

For this study, dataset has been collected from a hospital in Nagpur. The motive of the paper [11] is to diagnose diabetes in early stage to save money and time of the patient. In this paper four machine learning algorithms namely Random forest, support vector machine, Naïve bayes and Logistic regression have been used for early detection of diabetes.

The paper [12], uses StackingCVClassifier to ensemble four models LR, KNN, AdaBoost, and Multilayer perceptron. The model is applied on two different datasets, which gives an accuracy of 91 percent and 93 percent respectively.

In [13], machine learning methods have been used for early detection of diabetes. The algorithms used in the paper are SVM, LR, KNN, NB, DT, and RF. Among these algorithms, support vector machine outperforms with a accuracy score of 81.21.

The paper [14], uses dual characteristic variable selection method which is based on regression and LightGBM, which uses the features present in the dataset. Three algorithms namely XGBoost, ResNet and GA^{2}Ms have been used which gave an accuracy of 85.3, 88.8 and 87.5 respectively.

Comparative analysis of ML and DL algorithms is done in this paper [15], dataset with 17 attributes is used which is collected from UCI repository. The accuracy obtained from different ML and DL algorithms have been compared, XGBoost gives the best accuracy among all algorithms that is, 100.0%.

III. Proposed Methodology

Diabetes prediction has been done with the methodology depicted in the flowchart in Figure 1.

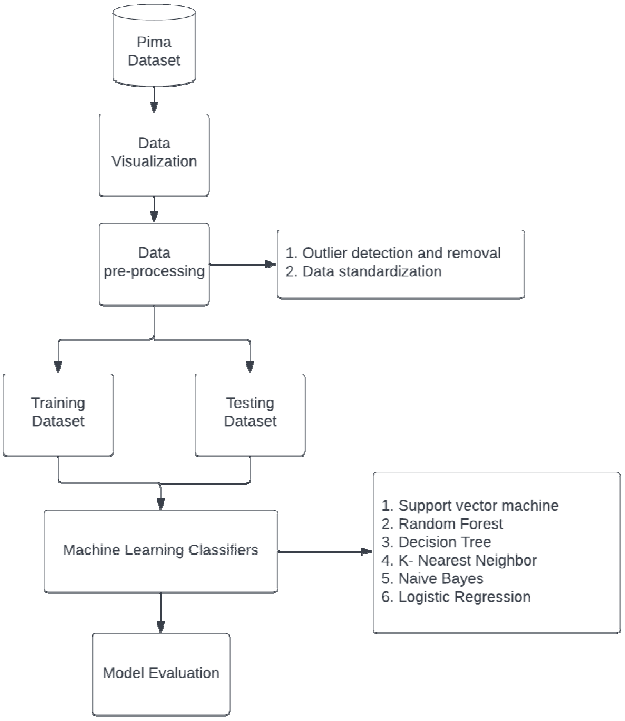

Figure 1. Proposed model

A. Dataset Description

The purpose of this data set is to diagnose whether a patient has diabetes using the specific diagnostic instruments in the data set. Selecting these instances from a large database had some limitations. All patients are Pima women of Indian descent over the age of 21.The dataset comprises information about 768 female patients who have diabetes, with 268 patients identified as diabetic and 500 patients identified as non-diabetic. Table 1 provides a brief description of each feature of the dataset.

Table 1. Dataset Description

Data Feature	Data Description
Pregnancies	Number of pregnancies
Glucose	Blood glucose level
BloodPressure (mm Hg)	Diastolic pressure
SkinThickness	skinfold thickness
Insulin	2-hour insulin level
BMI	Body mass Index
DiabetesPedigreeFunction	Pedigree Function
Age	Age of the patient
Outcome	Label

B. Data Visualization

1. Histogram- Histogram shows the graphical distribution of data in each feature of the dataset is shown in Figure 2.

2. Pie chart- Pie chart visualizes the labels of the dataset and shows the number of diabetic and non-diabetic patients is shown in Figure 3.

3. Pairplot- The seaborn Pairplot draws pairwise relationships between different features in the dataset based on the results, as shown in Figure 4.

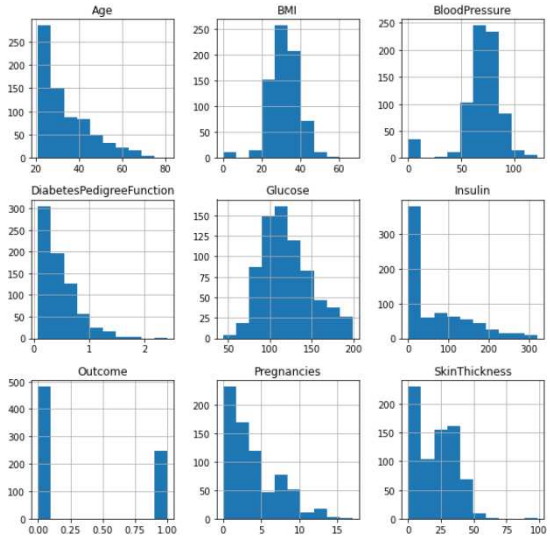

Figure 2. Histogram

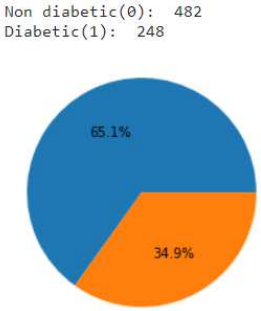

Figure 3. Pie chart

Figure 4. Pair plot

C. Data Pre-processing

Datasets are pre-processed and standardized to check for missing values, noisy data, and other inconsistencies before running the algorithm. Missing values and other contaminants are common in health data and reduce data validity. Data pre-processing is thus necessary to remove redundant data before applying a machine learning algorithm. Such redundant data is called outliers.

IQR(interquartile range) is used as a method of removing outliers from a dataset. First, the dataset is grouped according to the results and the corresponding mean values are calculated.

$$IQR\ (a) = \begin{cases} b, & if\ Q1 - 1.5 \times IQR \leq a \leq Q3 + 1.5 \times IQR \\ reject, & otherwise \end{cases} \quad (1)$$

Equation (1) can be applied to detect outliers, where 'y' represents an n-dimensional feature vector and 'Rn' represents the number of dimensions in the vector.
Since we observed significant differences in glucose and insulin levels in diabetic and non-diabetic subjects, the outliers for these two features were removed, reducing the dataset from 768 to 730 entries.

The original dataset is split into two parts using the train/test split method, with a ratio of 8:2 for the training set and testing set, as illustrated in Figure 5.

Figure 5. Train/test split

As shown in Figures 6 and 7, the dataset is evenly distributed into train and test data sets to preserve the ratio of positive and negative score entries in the respective datasets. Both graphs have the same nature, but differ in the density of the data.

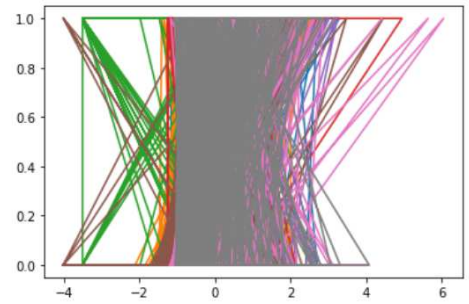

Figure 6. Graph for Training dataset

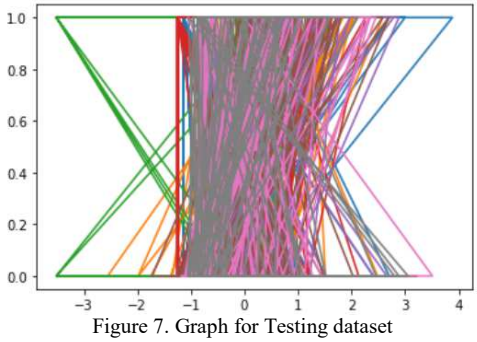

Figure 7. Graph for Testing dataset

D. Machine Learning Classifiers

1. Logistic Regression

Logistic regression is a machine learning classifier method that is widely used due to its simplicity and high effectiveness for binary classification tasks. This technique is employed to analyze data sets that are dependent on dependent variables present in the dataset. The primary aim of the classifier is to determine the best model that fits the relationship between dependent and independent fields in the dataset.

***2**. K- nearest neighbour*

The k-nearest neighbour (KNN) is not a parametrized machine learning classifier which relies on proximity to classify and predict the clustering of individual datapoints. By retaining all available data, KNN can classify new data points based on their similarities to existing data. This allows for easy categorization of new data when it is introduced using KNN algorithms. The KNN model stores only the training dataset and, upon receiving new data, classifies it into categories that closely resemble the new data.

3. Naive Byes

The Naive Bayes algorithm is a supervised learning technique that leverages Bayes' theorem to solve classification problems. It is commonly utilized in text classification applications that involve high-dimensional training data sets. The Naive Bayes Classifier is an efficient and straightforward classification algorithm that enables the creation of rapid machine learning models capable of generating quick predictions. As a probabilistic classifier, it utilizes probabilities to make predictions based on subject probabilities.

4.Support vector machine

The objective of the SVM algorithm is to construct ideal decision boundaries or lines that can partition the n-dimensional space into distinct categories, allowing for the accurate classification of future data points. These decision boundaries are referred to as hyperplanes. SVM identifies extreme or boundary cases, known as support vectors, to generate these hyperplanes. Therefore, the classifier is known as support vector machine.

5.Decision tree

The decision tree is a classifier with a tree-like structure, in which root nodes depicts characteristics of the data set, branches of internal nodes indicate decision rule, and each of the leaf node reflects an outcome. Decision tree is composed of two types of nodes: decision nodes and leaf nodes. Decision node contains multiple branches, which are responsible for making decision on the basis of attributes of the dataset. Leaf nodes, on the other hand, represent the final decision or outcome resulting from those decisions and do not have any additional branches.

6. Random Forest

The random forest is a classification method that utilizes multiple decision trees, each of which is created using a different subset of the dataset, to improve prediction accuracy. Rather than depending solely on decision trees, random forests consider the predictions from each tree and use a majority vote approach to arrive at the final output. The accuracy of the model is enhanced with an increase in the number of trees, while overfitting issues can be avoided. Random forests are a prime example of bagging techniques because each model employs a distinct portion of the data to make predictions.

IV. MODEL EVALUATION AND RESULT

All machine learning models were analysed on the given dataset and the corresponding training accuracies and accuracy scores are shown in Table 2. Fig. 8 shows accuracy scores of different machine learning classifiers.

Table 2. Result of machine learning classifiers

Model Name	Train Accuracy	Accuracy Score
Logistic Regression	78.082	78.767
KNN	82.876	76.027
Naive Bayes	75.513	78.082
SVM	82.876	80.821
Decision Tree	100.0	73.287
Random Forest	100.0	81.506

Classification reports are utilized to assess the effectiveness of classification algorithms' predictions. To generate a classification report, indicators for classification such as true positive, false positive, true negative, and false negative are utilized. These indicators are then used to calculate the parameters of classification report, such as precision, recall, and F-1 score are computed using equations (2), (3), and (4), respectively.

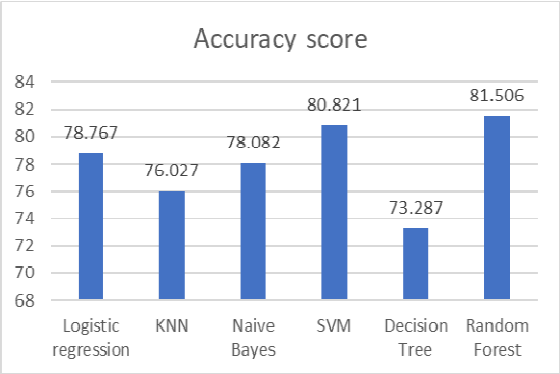

Figure 8. Accuracy of different machine learning classifiers

$$precision = \frac{TruePositive}{TruePositive+FalseNegative} \quad (2)$$

$$recall = \frac{TruePositive}{TruePositive+FalseNegative} \quad (3)$$

$$f1-score = \frac{precision \times recall}{precision + recall} \quad (4)$$

Table 3 displays the precision and other performance measures computed for various classifiers employed in forecasting the accuracy of diabetic forecasts.

Table 3. Accuracy, precision and recall of different machine learning classifiers

Model Name	Accuracy	Precision	Recall	F1 - score
Logistic Regression	0.78	0.78	0.79	0.78
KNN	0.76	0.75	0.76	0.75
Naive Bayes	0.78	0.78	0.78	0.78
SVM	0.80	0.80	0.81	0.80
Decision Tree	0.73	0.73	0.73	0.73
Random Forest	0.81	0.81	0.82	0.81

V. CONCLUSION AND DISCUSSION

In this study, machine learning classifiers were utilized to evaluate the precision of various models. The approach involves the implementation of several classifiers, including Logistic Regression, K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Machines (SVM), Decision Trees, and Random Forests, to compare and analyze their accuracy. The result shows thar the Random Forest achieved highest accuracy of 81.506. The research results could help health professionals in making early diabetes prediction.

VI. REFERENCES

[1] Doğru A, Buyrukoğlu S, Arı M. A hybrid super ensemble learning model for the early-stage prediction of diabetes risk. Medical & Biological Engineering & Computing. 2023;61(3):785-797. doi:10.1007/s11517-022-02749-z

[2] N. K. Trivedi, V. Gautam, H. Sharma, A. Anand and S. Agarwal, "Diabetes Prediction using Different Machine Learning Techniques," 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2022, pp. 2173-2177, doi: 10.1109/ICACITE53722.2022.9823640.

[3] K. C. Infancy, P. M. Bruntha, S. Pandiaraj, J. Joshiba Reby, A. Joselin and S. Selvadass, "Prediction of Diabetes Using ML Classifiers," 2022 6th International Conference on Devices, Circuits and Systems (ICDCS), 2022, pp. 484-488, doi: 10.1109/ICDCS54290.2022.9780830.

[4] M. Omkar and K. Nimala, "Machine Learning based Diabetes Prediction using with AWS cloud," 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), 2022, pp. 1-7, doi: 10.1109/ICSES55317.2022.9914160.

[5] N. Elsayed, Z. ElSayed and M. Ozer, "Early Stage Diabetes Prediction via Extreme Learning Machine," SoutheastCon 2022, 2022,pp.374-379, doi: 10.1109/SoutheastCon48659.2022.9764032.

[6] S. S et al., "A Comparative Analysis of Diabetes Prediction Models using Machine Learning Algorithms," 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), 2022, pp. 261-265, doi: 10.1109/ICACCS54159.2022.9785280.

[7] K. S. Anil and R. Jain, "Data Mining Techniques in Diabetes Prediction and Diagnosis: A Review," 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI), 2022, pp. 1696-1701, doi: 10.1109/ICOEI53556.2022.9776754.

[8] A. K. Shrivastava, K. V, K. S and S. M, "Early Diabetes Prediction using Random Forest," 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), 2022, pp. 1154-1159, doi: 10.1109/ICESC54411.2022.9885683.

[9] Omana J(1), Moorthi M(2). Predictive Analysis and Prognostic Approach of Diabetes Prediction with Machine Learning Techniques. Wireless Personal Communications. 2022;127(1):465-478. doi:10.1007/s11277-021-08274-w

[10] T. R. Mahesh, V. Vivek, V. V. Kumar, R. Natarajan, S. Sathya and S. Kanimozhi, "A Comparative Performance Analysis of Machine Learning Approaches for the Early Prediction of Diabetes Disease," 2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), 2022, pp. 1-6, doi: 10.1109/ACCAI53970.2022.9752543.

[11] Y. Dubey, P. Wankhede, T. Borkar, A. Borkar and K. Mitra, "Diabetes Prediction and Classification using Machine Learning Algorithms," 2021 IEEE International Conference on Biomedical Engineering, Computer and Information Technology for Health (BECITHCON), 2021, pp. 60-63, doi: 10.1109/BECITHCON54710.2021.9893653.

[12] F. R. Liza et al., "An Ensemble Approach of Supervised Learning Algorithms and Artificial Neural Network for Early Prediction of Diabetes," 2021 3rd International Conference on Sustainable Technologies for Industry 4.0 (STI), 2021, pp. 1-6, doi: 10.1109/STI53101.2021.9732413

[13] J. Kumar, R. K. Tiwari and V. Pandey, "Diabetes prediction using machine learning tools," 2021 4th International Conference on Recent Trends in Computer Science 263 -267, doi: 10.1109/ICRTCST54752.2022.9781963.

[14] B. Chen, M. Yan, H. Zhong and B. He, "Prediction Model of Diabetes Based on Machine Learning," 2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT), 2021, pp. 128-136, doi: 10.1109/ACAIT53529.2021.9731180.

[15] M. A. R. Refat, M. A. Amin, C. Kaushal, M. N. Yeasmin and M. K. Islam, "A Comparative Analysis of Early Stage Diabetes Prediction using Machine Learning and Deep Learning Approach," 2021 6th International Conference on Signal Processing, Computing and Control