Diabetes Prediction using different Machine Learning Classifiers

Tanya Gupta
Electronics & Communication
Engineering(ECE)
Vellore Institute of
Technology,Vellore.
Vellore, India.
tanya.gupta2020@vitstudent.ac.in

Mithul Raaj A T
Electronics & Communication
Engineering(ECE)
Vellore Institute of
Technology,Vellore.
Vellore, India.
mithulraaj.at2020@vitstudent.ac.in

Rajesh Kumar M
Senior Member, IEEE School of
Electronics and Engineering
Vellore Institute of Technology
Vellore, India
mrajeshkumar@vit.ac.in

***Abstract*— Diabetes is a chronic disease caused by either insufficient insulin production by the pancreas or inefficient use of insulin by the body. 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. Different machine learning classifier methods such as Logistic Regression, Naive Bayes, K-nearest neighbors, Decision tree, Support Vector Machines and Random Forest, corresponding accuracy score gained using each algorithm is compared.**

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

I. Introduction

Diabetes is a chronic (long-term) medical condition that affects how the body converts food into energy. Body breaks down food into sugar (glucose) and releases it into your bloodstream. When blood sugar rises, it signals the pancreas to release insulin. Insulin acts as the key to taking blood sugar into the body's cells to use it as energy. In diabetes, the body cannot make or use enough insulin. When there isn't enough insulin, or when cells don't respond to insulin, excess blood sugar stays in the bloodstream. Over time, this can lead to serious health problems such as heart disease, vision loss, and kidney disease. In 2014, 8.5% of adults over the age of 18 had diabetes. In 2019, diabetes was the direct cause of 1.5 million deaths, and 48% of all diabetes deaths were under 70s. Diabetes is responsible for another 460,000 deaths from kidney disease, and elevated blood sugar accounts for approximately 20% of cardiovascular deaths.

A. *Diabetes type one*

Type 1 diabetes, formerly known as insulin-dependent diabetes or juvenile diabetes, is a chronic disease. In this state, the pancreas produces little or no insulin. Insulin is a hormone that controls the blood sugar level. Various factors can cause his type 1 diabetes, including heredity and some viruses. Type 1 diabetes usually begins in childhood or adolescence, but it can also occur in adults. Type 1 diabetes is incurable, even after much research has been done. Treatment aims to control the amount of sugar in the blood with insulin, diet and lifestyle to prevent complications

B. *Diabetes type two*

Type 2 diabetes is a disorder of the way the body regulates and uses sugar (glucose) for fuel. This long-term (chronic) condition causes excessive circulation of sugar in the bloodstream. Finally, high blood sugar levels can lead to disorders of the circulatory, nervous and immune systems. There are two main related issues at work in type 2 diabetes. The pancreas does not produce enough insulin (the hormone that regulates the movement of sugar into the cells), and the cells respond poorly to insulin and absorb less sugar.

C. *Health impact*

With time, diabetes can damage the heart, kidneys, nerves, blood vessels, and eyes. Adults with diabetes are two to three times more likely to have a heart attack or stroke. Foot neuropathy (nerve damage), combined with reduced blood flow, increases the risk of foot ulcers, infections, and ultimately limb amputation. Diabetic retinopathy is an important cause of blindness and results from long-term, cumulative damage to the small blood vessels of the retina. About 1 million people are blind due to diabetes. Diabetes is one of the main causes of kidney failure.

II. Related work

This section presents recent related work on diabetes prediction using machine learning algorithms and remote diabetes prediction using machine learning algorithms and remote diabetes healthcare monitoring systems.

Random Forest method implementation for early-stage prediction has been explained in detail in [1]. The various methods to categorize raw data and information about extracting required data from them using bagging, oversampling, repeated random sub sampling are given. An accuracy of 81.17 was achieved after pre-processing data and using ML algorithm.

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 techniques.

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 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.

The different types of prognostic biomarkers have been clearly tabulated in [8]. The attributes with their ranks for each selection features including Chi Square, mRMR, RFE-RF are mentioned and is then applied on different algorithms. Some of the attributes which play important role are Polydipsia, Polyuria, Sudden Weight Loss, Partial Paresis and Gender of the person. It also measures sensitivity, specificity, MCC value, AUC value for each method giving a more mathematical outlook on diabetes prediction.

III. Proposed Methodology

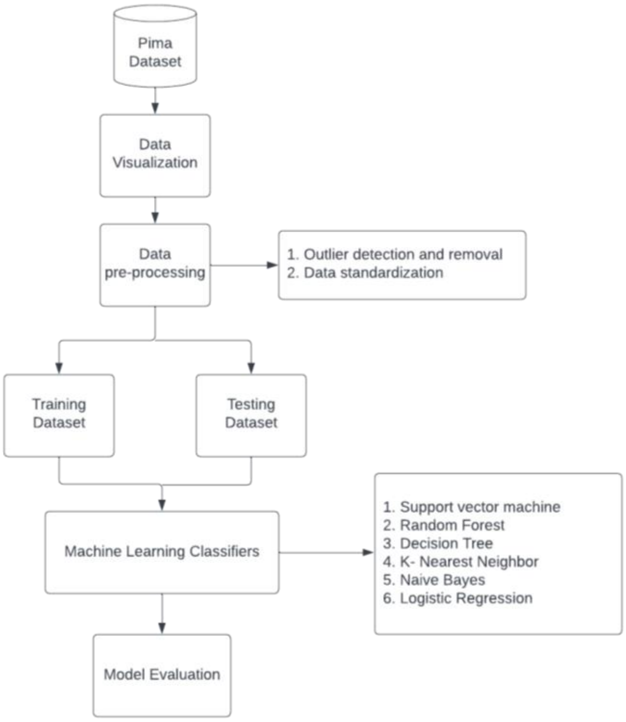


Figure 1. Proposed model

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 database contains datasets of 768 female diabetic patients, which includes 268 correctly identified diabetic patients and 500 non-diabetic patients. Each feature of the dataset is explained briefly in Table 1.

Data Feature	Data Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration
BloodPressure	Diastole blood pressure
SkinThikness	Triceps skinfold thickness
Insulin	2-hour serum insulin
BMI	Body mass Index
DiabetesPedigreeFunction	Diabetes Pedigree Function
Age	Age (years)
Outcome	Class variables (0 or 1)

Table 1. Dataset Description

B. *Data Visualization*

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

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

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

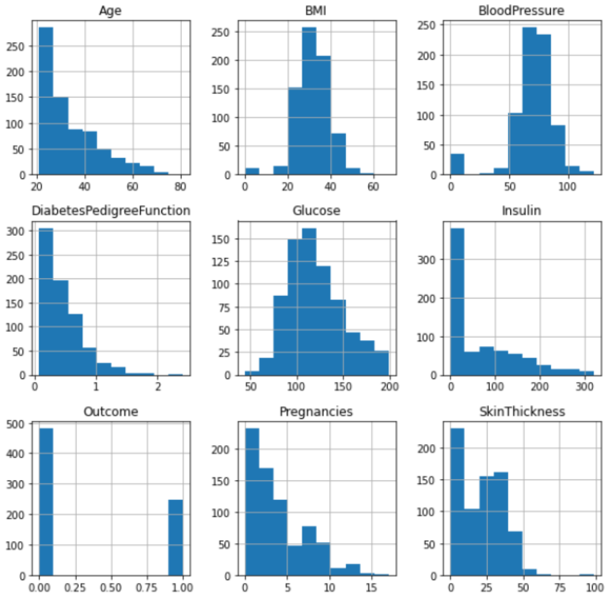
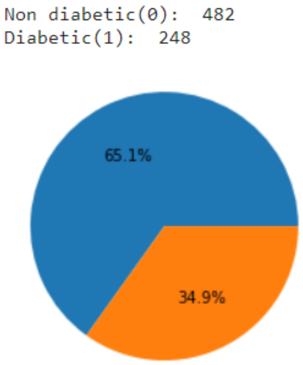


Figure 2. Histogram



```
Non diabetic(0):  482
Diabetic(1):  248
```

Figure 3. Pie chart

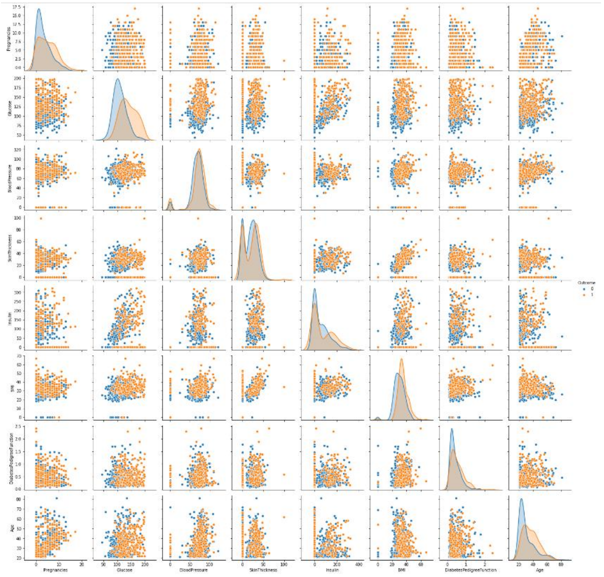


Figure 4. Pair plot

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.

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

Equation (2) is useful for finding outliers. where y is the n-dimensional spatial instance of the feature vector and R_n is the dimensional number.

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 training/test split method splits the data set into two parts in the ratio 8:2 (training set and testing set) as shown in Fig. 5.

Figure 5. Train/test split

As shown in Figures 6 and 7, the training and test data sets are evenly distributed to preserve the ratio of positive and negative score entries in the training and test data sets. Both graphs have the same nature, but differ in the density of the data.

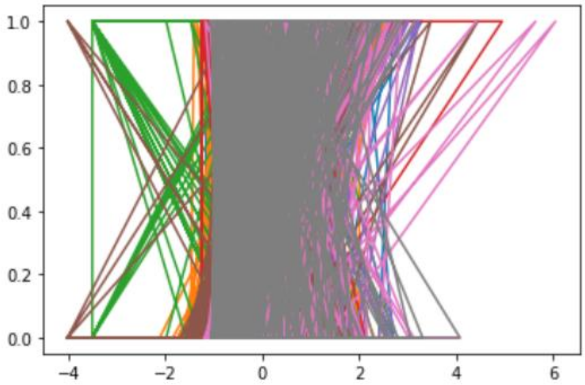


Figure 6. Graph for Training dataset

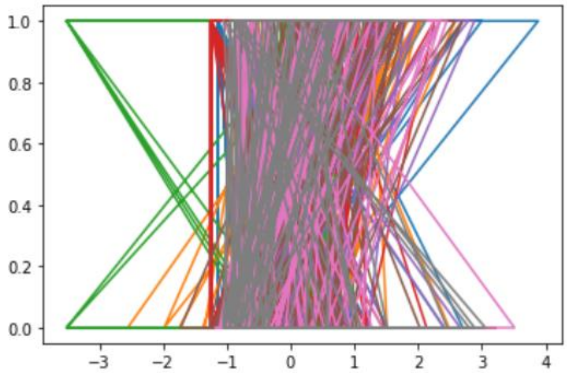


Figure 7. Graph for Testing dataset

D. *Machine Learning Classifiers*

1. Logistic Regression

Logistic regression is a classification technique borrowed from statistical machine learning. Logistic regression is a simple but highly effective classification algorithm and is therefore widely used for many binary classification tasks. Logistic regression is a statistical technique for analyzing data sets that have one or more independent variables that determine the outcome. The goal of using logistic regression is to find the model that best describes the relationship between the dependent and independent variables.

2. **K- nearest neighbour**

The k-nearest neighbour algorithm, is a non-parametric supervised learning classifier that uses proximity to make classifications or predictions about clustering of single data points. The K-NN algorithm stores all available data and classifies new data points based on similarity. This means that when new data appears, it can be easily placed into the appropriate category using K-NN algorithms. The KNN algorithms stores only the training phase dataset, and upon receiving new data, classifies the data into categories that are very similar to the new data.

3. Naive Byes

The naive bayes algorithm is a supervised learning algorithm based on Bayes' theorem and is used to solve classification problems. It is primarily used in text classification with high-dimensional training data sets. Naive Bayes Classifier is one of the simplest and most effective classification algorithms that help you build fast machine learning models that can make fast predictions. It's a probabilistic classifier, meaning it makes predictions based on subject probabilities.

4. Support vector machine

The goal of the SVM algorithm is to create optimal lines or decision boundaries that can divide the n-dimensional space into classes so that new data points can be easily placed in the correct category in the future. This optimal decision boundary is called a hyperplane. SVM selects extrema/vectors to help create hyperplanes. These extreme cases are called support vectors, and the algorithm is called a support vector machine.

5.Decision tree

It is a tree-structured classifier, with internal nodes representing characteristics of the data set, branches representing decision rules, and each leaf node representing a result. The decision tree has his two nodes, a decision node and a leaf node. Decision nodes are used to make decisions and have multiple branches, while leaf nodes are the result of those decisions and contain no further branches. A decision or test is made based on the characteristics of a particular data set.

6. Random Forest

A random forest is a classifier that takes a set of decision trees over different subsets of a given dataset and takes an average to improve the prediction accuracy of that dataset. Instead of relying on decision trees, random forests take predictions from each tree and predict the final output based on the majority vote of the predictions. Higher number of trees in the forest gives better accuracy and avoids overfitting problems. Random forests are also a classic example of bagging approaches. This is because each model uses a different subset 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.

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

Table 2. Result of different machine learning classifiers

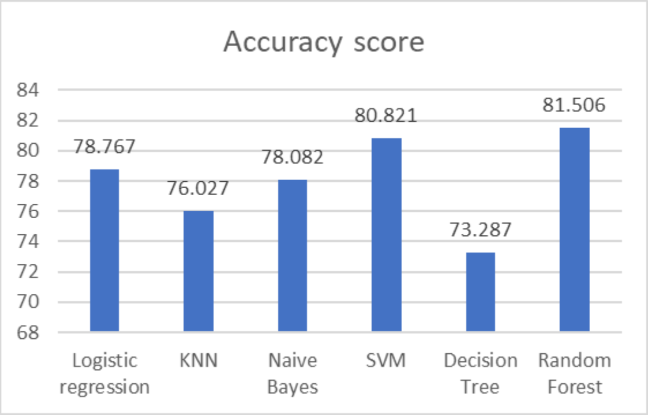


Figure 8. Accuracy of different machine learning classifiers

Classification reports are used to measure the quality of the classification algorithm's predictions. True positives, false positives, true negatives, and false negatives are used to predict indicators for classification reports. Classification report shows precision, recall and F-1 score where precision, recall and F-1 score are calculated using (2), (3) and (4) respectively.

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

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

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

Table 3 shows the calculated accuracy and other metrics for the different classifiers used to predict the quality of the diabetic predictions.

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

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

V. CONCLUSION AND DISCUSSION

This paper used machine learning classifiers to compare and analyze accuracy of different machine learning classifier models. The proposed method uses Logistic Regression, KNN, Naive Bayes, SVM, Decision Tree and Random Forest classifiers. 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.

REFERENCES

[1] 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.

[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] U. Das, A. Yakin Srizon, M. Ansarul Islam, D. Sikder Tonmoy and M. Al Mehedi Hasan, "Prognostic Biomarkers Identification for Diabetes Prediction by Utilizing Machine Learning Classifiers," 2020 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), 2020, pp. 1-6, doi: 10.1109/STI50764.2020.9350498.

[9] 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.

[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 and Technology (ICRTCST), 2022, pp. 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