International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 07 | July 2022 www.irjet.net p-ISSN: 2395-0072

Forecasting Diabetes Mellitus at an Initial Stage using Machine Learning Methods

Seema Gaikwad[1], Prof. Sudhir Shikalpure[2]

[1]M.Tech Final Year Student, Department of Computer Science and Engineering, Government College of Engineering, Aurangabad, Maharashtra, India.
[2]Assistant Professor, Department of Computer Science and Engineering, Government College of Engineering, Aurangabad, Maharashtra, India.

Abstract - *Diabetes mellitus is an illness that arises whenever pancreas fails to generate enough insulin or if the body's glucose is inefficiently used. Hormone called insulin helps to stabilize blood sugar levels. Uncontrolled diabetes causes hyperglycaemia, or high blood sugar, which causes severe harm to various organ functions. Since the primary signs of the disease are difficult to recognise, a suitable technique of forecasting will aid individuals in self-diagnosis. In this study, symptoms are used for classification instead of performing an insulin test. Machine learning implementation becomes much easier with the availability of data on which models can be trained. Multilayer perceptron, K Nearest Neighbour, Gaussian Naïve Bayes, and Linear Discriminant Analysis are among machine learning models implemented. A diabetes diagnostic forecast model for high risk or early stage diabetes is developed selecting the one with the best accuracy. The accuracy of the Multilayer Perceptron on the testing data is 99 percent, making it the ideal model for forecasting diabetes.*

Key Words: Machine Learning, Diabetes Prediction, Multilayer Perceptron, Healthcare

1. INTRODUCTION

Diabetes is a severe public health issue that affects 463 million people worldwide. By 2045, it is expected to impact 700 million people globally. According to prevalence estimates, diabetes is spreading quickly in low- and middle-income nations than it is in high-income countries. Diabetes affects more than 77 million persons in India. According to researchers, this number will rise to 134 million people by the year 2045 [1].

When pancreas does not create any insulin, you are suffering from type 1 diabetes. When pancreas doesn't create enough insulin or your body can't utilise it properly, you are suffering from type 2 diabetes. Impaired glucose tolerance as well as impaired fasting glycaemia are intermediary states in the shift from normal blood glucose levels to diabetes (particularly type 2), although they are not always present. Gestational diabetes is a disorder that develops during pregnancy and increases the chance of developing type 2 diabetes in the future. Whenever blood glucose levels are above normal though not high enough to be diagnosed as diabetic, gestational diabetes is present. Cardiovascular disease, renal failure, visual loss, nerve damage, infections, foot difficulties, and cognitive impairments are all major effects of untreated diabetes [2].

Healthcare facilities are limited, and physicians can only identify considerable amount of patients in a given period of time. A quick and accurate diagnosis can aid patients in preventing diabetes and determining whether they have diabetes at a preliminary phase or they are risk of developing diabetes. Patients in our research were not required to undergo any medical tests, making our approach more intelligible and relevant. The aim of our research is to find a machine learning model which can predict diabetes risk or diabetes with good accuracy.

2. RELATED WORK

Author used six machine learning algorithms to develop diabetes prediction models: logistic regression, support vector machine, decision tree, random forest, boosting, and neural network. Logistic Regression, Support Vector Machine, and Decision Tree are the first three models, and they are basic and intuitive, with lower accuracy than the more sophisticated remaining three models. He compared their testing error results. Best model gave accuracy of 96.2% [3].

Authors created a diabetes prediction model using neural network. They employed characteristics like PG Concentration, Skin Thickness, Diastolic BP, etc. Many characteristics in that research need a skilled medical examination; it is not available to everyone [4].

Authors showed a Matlab implementation of Support Vector Machine and Naive Bayes algorithms in a dataset obtained from diabetic patients in Kosovo. Body Mass Index, Pre and Post meal glucose, Diastolic and Systolic blood pressure, Family history of diabetes, Regular diet, Physical activities are the attributes used in their research. Joint implementation of SVM and Naïve Bayes is performed to predict diabetes [5].

The author of the research suggested a model that divides type 2 diabetes treatment strategies into three categories:

insulin, food, and medicine. The model was built using data from the JABER_ABN_ABU_ALIZ clinic, which has 318 health records. The model was created with the WEKA tool along with J48 classifier, which has 70.8 percent accuracy [6].

The authors created a model to predict whether or not a person will acquire diabetes based on everyday healthy behaviours. The PIMA diabetic data set was utilized to make the prediction model, and the CART (Classification and Regression Trees) machine learning classifier was employed. The proposed model might have 75 percent accuracy rate [7].

Multilayer Perceptron is used to solve problems which are linearly inseparable with the help of backpropagation. Backpropagation is a technique for minimising the difference amongst the desired output and the actual output by adjusting the weighted values and threshold values continuously. It consists of input, output and hidden layer [8].

K Nearest Neighbor is non-parametric as it makes no assumptions about the distribution of data. It is a lazy learner. It calculates the distance between k neighbours [9].

Gaussian Naïve Bayes follows Gaussian distribution. For dimensionality reduction challenges, Linear Discriminant Analysis is a widely used approach [10, 11].

Accuracy states the efficacy of a classifier as a whole. Precision denotes the degree to which the data labels agree with the positive labels assigned by classifier. The efficacy of a classifier in identifying positive labels is measured by recall. The F1 score reflects the relationship between the positive labels assigned to data as well as those assigned by a classifier [12].

3. METHODOLOGY

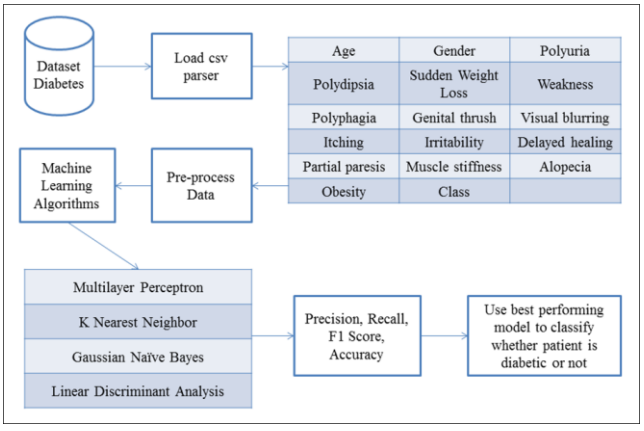

Figure 1: System Design

3.1 Data Description

We utilized data via the UCI machine learning repository. Data is gathered via Sylhet Diabetes Hospital patients. Dataset consists of 17 variables and 520 records. The dataset does not include any null values. Only number variable is age, while rest are categorical. Males make up 37% of the data set, while females make up 63%. 45% Males and 90% Females are insulin dependent in given dataset. Figure 2 indicates spread of diabetes in the dataset. 320 people are diabetic and 200 people are non-diabetic in our dataset i.e. 62% patients are diabetic and 38% patients are non-diabetic.

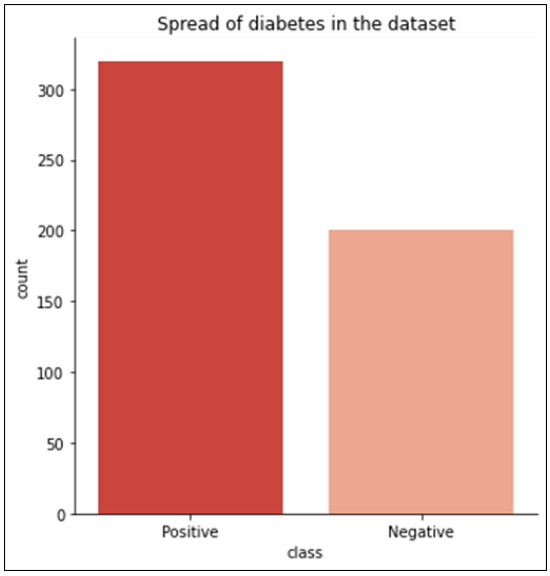

Figure 2: Spread of diabetes in the dataset

3.2 Preprocessing

Data pre-processing refers to the actions that must be taken to encode or alter data so it will be easily interpreted by a machine. First step is to get dataset. Then import necessary libraries. Then import dataset. Then we need to handle missing values present in dataset. Our dataset doesn't consist of any missing value. To prepare the data for analysis, the dataset was translated from categorical labels to numeric labels. A heatmap is visual representation of data that uses colours to display the matrix's value. Brighter colours are utilised to indicate more common values, whereas lighter colours are favoured to indicate less common values. According to Figure 3, a heatmap is plotted to find relationship between different attributes. Polyuria has 0.67 as correlation coefficient and Polydipsia has 0.65 as correlation coefficient with respect to diabetes, which indicates it has strongest link to diabetes. Diabetes and age revealed a minimal connection with 0.11 as correlation coefficient. To see if there is a link between age and diabetes, the age field in dataset is converted to categorical data type. As per Figure 4 diabetes is shared evenly across all age groups, there is no proper association between age and diabetes. Next step is to distribute our data into training and

testing sets respectively. The last stage in data pre-processing is feature scaling. It is a strategy for standardising the independent variables in dataset in a particular range.

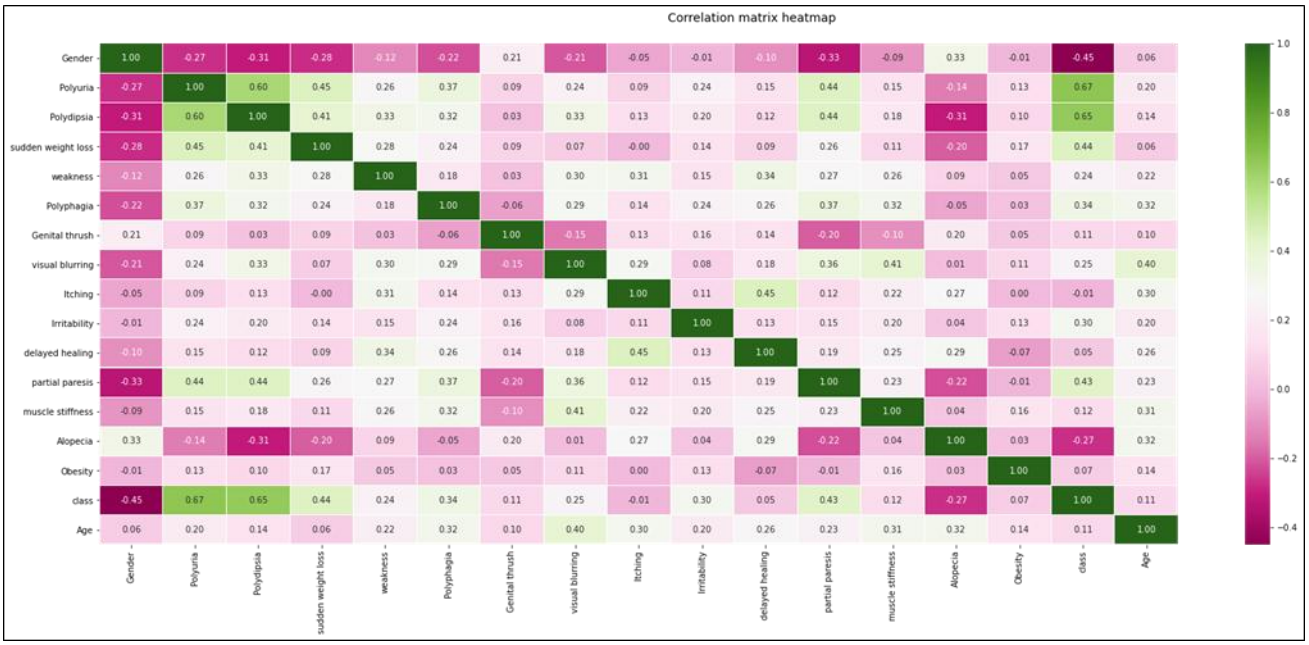

Figure 3: Heatmap to Find Correlation among Attributes

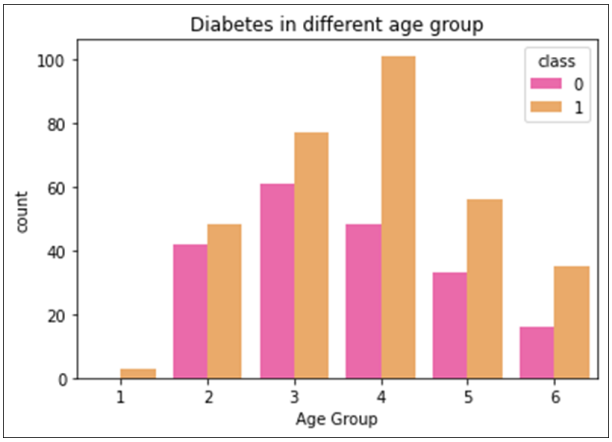

Figure 4: Diabetes in Different Age Group

3.3 Machine Learning Models

The diabetes class is predicted using all 16 characteristics. 80% data is used for training the model and 20% data is used for testing the model. Once the data is trained, test data predictions are generated, and the algorithm's performance is measured.

3.3.1 Multilayer Perceptron

A fully linked multilayer perceptron comprises of input, output, and hidden layers. It's often used in healthcare to identify complicated disease states. The sizes of the hidden layers are changed to three layers with eleven nodes. From Figure 5, as per confusion matrix, 103 patients are correctly classified and 1 is misclassified. From Figure 6, classification report indicates accuracy and F1 score of 99%.

```
[[40  0]
 [ 1 63]]
```

Figure 5: Multilayer Perceptron Confusion Matrix

	precision	recall	f1-score	support
0	0.98	1.00	0.99	40
1	1.00	0.98	0.99	64
accuracy			0.99	104
macro avg	0.99	0.99	0.99	104
weighted avg	0.99	0.99	0.99	104

Figure 6: Multilayer Perceptron Classification Report

3.3.2 K Nearest Neighbor

It sorts data points into groups depending on which data points are closest to it. We have set neighbors size to 5. The distance metric is set to minkowski. From Figure 7, as per confusion matrix, 100 patients are correctly classified and 4 are misclassified. From Figure 8, classification report indicates accuracy and F1 score of 96%.

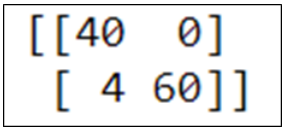

Figure 7: K Nearest Neighbor Confusion Matrix

	precision	recall	f1-score	support
0	0.91	1.00	0.95	40
1	1.00	0.94	0.97	64
accuracy			0.96	104
macro avg	0.95	0.97	0.96	104
weighted avg	0.97	0.96	0.96	104

Figure 8: K Nearest Neighbor Classification Report

3.3.3 Gaussian Naïve Bayes

It is an example of Generative Model. A Gaussian distribution is followed by every class. The characteristics are assumed to be independent so covariance matrices are diagonal in nature. From Figure 9, as per confusion matrix, 93 patients are correctly classified and 11 are misclassified. From Figure 10, classification report indicates accuracy and F1 score of 89%.

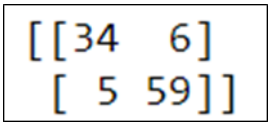

Figure 9: Gaussian Naïve Bayes Confusion Matrix

	precision	recall	f1-score	support
0	0.87	0.85	0.86	40
1	0.91	0.92	0.91	64
accuracy			0.89	104
macro avg	0.89	0.89	0.89	104
weighted avg	0.89	0.89	0.89	104

Figure 10: Gaussian Naïve Bayes Classification Report

3.3.4 Linear Discriminant Analysis

It is used for dimensionality reduction to solve classification issues. A common covariance matrix is assumed by Linear Discriminant Analysis. From Figure 11, as per confusion matrix, 90 patients are correctly classified and 14 are misclassified. From Figure 12, classification report indicates accuracy and F1 score of 87%.

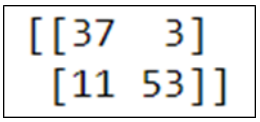

Figure 11: Linear Discriminant Analysis Confusion Matrix

	precision	recall	f1-score	support
0	0.77	0.93	0.84	40
1	0.95	0.83	0.88	64
accuracy			0.87	104
macro avg	0.86	0.88	0.86	104
weighted avg	0.88	0.87	0.87	104

Figure 12: Linear Discriminant Analysis Classification Report

3.4 Prediction

Best performing model is used for diabetes prediction or diabetes risk prediction. So Multilayer Perceptron is used for model building.

4. RESULTS

For measuring performance of machine learning we have used Precision, Recall, F1 Score and Accuracy.

Precision = Tp / (Tp + Fp)

Recall = Tp / (Tp + Fn)

F1 Score = 2 * Precision * Recall / (Precision + Recall)

Accuracy = (Tp + Tn) / (Tp + Tn + Fp + Fn)

Where Tp = True Positive, Fp = False Positive, Fn = False Negative, Tn = True Negative

Table 1: Performance of Machine Learning Models

	Precision		Recall		F1 Score		Accuracy
	0	1	0	1	0	1	
Multilayer Perceptron Classifier	0.98	1.0	1.0	0.98	0.99	0.99	0.99
K Neighbors Classifier	0.91	1.0	1.0	0.94	0.95	0.97	0.96
Gaussian Naïve Bayes	0.87	0.91	0.85	0.92	0.86	0.91	0.89
Linear Discriminant Analysis	0.77	0.95	0.93	0.83	0.84	0.88	0.87

From Table 1, among given machine learning algorithms, Multilayer Perceptron has highest accuracy of 99% whereas Linear Discriminant Analysis has lowest accuracy of 87%. So Multilayer Perceptron is used further in order to predict diabetes or diabetes risk.

5. CONCLUSIONS

A good diabetes forecasting model will aid doctors in making precise decisions and ensuring that patients receive prompt medical attention. We used exploratory data analysis on our dataset to determine which characteristics are strongly linked to diabetes. We analysed the performance of four machine learning algorithms. Multilayer perceptron gave highest accuracy of 99% so it used further to predict diabetes or diabetes risk. We can test algorithms on other unstructured diabetes datasets in the future to forecast diabetes.

REFERENCES

[1] Luhar, Shammi, Dimple Kondal, Rebecca Jones, Ranjit M. Anjana, Shivani A. Patel, Sanjay Kinra, Lynda Clarke et al. "Lifetime risk of diabetes in metropolitan cities in India." Diabetologia 64, no. 3 (2021): 521-529.

[2] World Health Organization, 2016. World Health Organization Global Report on Diabetes. Geneva: World Health Organization.

[3] Ma, Juncheng. "Machine Learning in Predicting Diabetes in the Early Stage." In 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), pp. 167-172. IEEE, 2020.

[4] El_Jerjawi, Nesreen Samer, and Samy S. Abu-Naser. "Diabetes prediction using artificial neural network." International Journal of Advanced Science and Technology 121 (2018).

[5] Tafa, Zhilbert, Nerxhivane Pervetica, and Bertran Karahoda. "An intelligent system for diabetes prediction." In 2015 4th Mediterranean Conference on Embedded Computing (MECO), pp. 378-382. IEEE, 2015.

[6] Ahmed, Tarig Mohamed. "Developing a predicted model for diabetes type 2 treatment plans by using data mining." Journal of Theoretical and Applied Information Technology 90, no. 2 (2016): 181.

[7] Anand, A., & Shakti, D. (2015, September). Prediction of diabetes based on personal lifestyle indicators. In 2015 1st International Conference on Next Generation Computing Technologies (NGCT) (pp. 673-676). IEEE.

[8] Singh, Jaswinder, and Rajdeep Banerjee. "A study on single and multi-layer perceptron neural network." In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 35-40. IEEE, 2019.

[9] Ray, Susmita. "A quick review of machine learning algorithms." In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon), pp. 35-39. IEEE, 2019.

[10] Agarwal, S., B. Jha, T. Kumar, M. Kumar, and P. Ranjan. "Hybrid of naive bayes and Gaussian naive bayes for classification: a map reduce approach." International Journal of Innovative Technology and Exploring Engineering 8, no. 6S3 (2019): 266-268.

[11] Tharwat, Alaa, Tarek Gaber, Abdelhameed Ibrahim, and Aboul Ella Hassanien. "Linear discriminant analysis: A detailed tutorial." AI communications 30, no. 2 (2017): 169-190.

[12] Sokolova, Marina, and Guy Lapalme. "A systematic analysis of performance measures for classification tasks." Information processing & management 45, no. 4 (2009): 427-437.