International Journal of Combined Research & Development (IJCRD)
eISSN:2321-225X; pISSN:2321-2241 Volume: 13; Issue: 7; July- 2024

LEVERAGING MACHINE LEARNING FOR DIABETES PROGNOSIS

Pereira Priyanka Leeya
PG Student
Department of Master of Computer Application
The Oxford College of Engineering
priyanka.pp236@gmail.com

Mrs. Sowmya J
Assistant Professor
Department of Master of Computer Application
The Oxford College of Engineering
sowmyaj@theoxford.edu

ABSTRACT One of the riskiest chronic illnesses, diabetes is brought on by high blood glucose levels and, if left unchecked, can bring a host of consequences. Conventional diagnostic techniques usually require lengthy trips to diagnostic facilities and medical professionals. However, machine learning (ML) provides an answer by making diabetes detectable and treatable early. This paper uses five machine learning (ML) algorithms—Logistic Regression, Decision Tree, Support Vector Machine (SVM), K Nearest Neighbor, and Random Forest—to create a highly accurate diabetes prediction model. For the experiments, Kaggle's Bangladesh Diabetes Dataset (BDD) was used. The Random Forest method outperformed Logistic Regression, as seen by the results, which showed that it achieved the greatest accuracy of 96%. The created model is included into an easy-to-use Flask web application. Due to the disease's extensive impact on numerous organs, traditional diabetes diagnostic procedures rely on physical and chemical tests, which can be difficult. In data science, machine learning is a fast-growing subject that uses past data to improve forecast accuracy. In order to maximize the results of predictions, this paper combines a number of machine learning approaches, including ensemble learning. It also seeks to offer a thorough web portal with details on various forms of diabetes, its causes, preventative measures, and typical drugs. The research makes use of 17 different variables' worth of data from the UCI repository in order to facilitate the creation of this predictive model.

Keywords: Machine learning, models, Diabetes, Dataset

I. INTRODUCTION

Diabetes mellitus is a chronic condition characterized by high blood sugar levels, manifested due to the body system's failure either to produce sufficient insulin or to respond to its presence. The syndrome may result in critical diseases, which include life-threatening conditions like diabetic ketoacidosis, heart complications, and strokes. In 2016, diabetes claimed about 1.6 million lives; approximately 422 million people currently struggle with the disease.

Diabetes broadly has two types: Type 1 and Type 2. Type 1 accounts for only 5–10% of the cases and usually results due to malfunctioning in the pancreas during childhood years. The symptoms occur when 80–90% of the cells producing insulin get damaged. The remaining 90% is Type 2. It results due to continuous hyperglycemia, in older, overweight people. Algorithms involving Random Forest, Decision Trees, KNN, SVM, Logistic Regression, and Deep Neural Networks that employed the principles of machine learning and deep learning techniques were all very accurate in predicting diabetic outcomes.

The genetic basis of diabetes also includes chromosome 6 abnormalities that disturb the body's response to antigens and probable triggers such as viral infections. At the moment, there are 53 million adults ages 20-79 years old who have diabetes; it is estimated that this number will increase to 643 million in 2030 and 783 million in 2045. Type 2 diabetes is hereditary in origin and lifestyle-related, while Type 1 diabetes is an autoimmune disease that occurs when the immune system mistakenly attacks

the cells in the body that produce insulin. A third type of diabetes may be the gestational diabetes which occurs during pregnancy.

It increases the risk of cataracts and blindness, thus affecting eyesight. Control of diabetes through a balanced diet and regular exercise is necessary, but early detection also is an important aspect. Early interventions can prevent complications and reduce healthcare costs while improving quality of life if a diagnosis is made early. Machine learning can be used to find patterns in large data sets, thereby aiding in better diabetes prediction and control.

A program is under development, and this will involve software designed to provide diabetes prediction and treatment with accurate predictions, coupled with an easy-to-use interface for data entry. Personalized treatment plans, efficient management programs, and early detection are the things that are critical in enhancing patient outcomes and lessening the impact of the condition.

II. LITERATURE REVIEW

Recently, there has been a significant increase in machine learning techniques in diabetes prediction. This is inspired by the fact that some of the research work done on the consideration of different strategies has aided in increased accuracy and maintaining consistent prediction. Among the studies are those done by Ahmed and others in the application of a fused machine learning approach for the purposes of managing diabetes prediction. All being said, the authors combine different algorithms of machine learning. Each of them was armed with their respective strengths with a dataset that includes variables such as age, blood pressure, BMI, and glucose levels related to diabetes. This study found that the fusion of different models like logistic regression, decision trees, and K nearest neighbors with KNN with different preprocessing techniques is unsurprisingly better than the single models and gets maximum prediction accuracy. This strengthens the ongoing evidence in favor of combined machine learning methodologies and demonstrates the feasibility of hybrid models to increase the accuracy of diabetes-related predictions. An optimized model of multivariable regression was designed by Daliya et al. [2] aiming to predict diabetes-related predictions. This research has identified from 1000 diabetes patients the major risk factors such as age, blood pressure, lipid profile, BMI, HbA1c, and blood pressure. The researchers used the multiple linear regression framework, coupled with a stepwise variable selection strategy, to identify the most critical risk factors. It is said that, with 85% accuracy, the intelligent system can make life-or-death decisions much more sensible to healthcare professionals while giving treatment to their patients. This paper throws light on the variables that determine disease progression and on the importance of tailored regression models in the elucidation and anticipation of diabetic trajectory. A very thorough review by Jiajia Song and others, in 2021, considered different machine learning techniques such as artificial neural networks, decision trees, random forests, and support vector machines for diabetes prediction. The review very strongly highlighted the relevance of feature selection and pretreatment of data to improvement in model accuracy. MM: From the work implemented, the authors concluded that the additional multimodal data in their work—genetic, clinical, and lifestyle data—adds to the enhancement of model efficacy in predicting outcomes. In maximizing machine learning applications toward the prediction of diabetes, the paper emphasizes a holistic approach to data integration and preprocessing in the creation of a reliable diabetes predictive model. It also proposes the necessity of further research in this field. In this regard, Hruaping Zhou et al. [4] proposed a much more complex way of allaying concerns over the extraction of temporal and spatial features from the input data by suggesting a deep neural network model that incorporates Long short-term memory (LSTM) networks with Convolutional Neural Networks (CNNs). The method was tested using a public diabetes dataset with a number of state-of-the-art machine learning methods. As per the experimental data, the developed DNN model performs much better than other models since it gives improved values of accuracy, sensitivity, and specificity. This further suggests that elaborative designs of neural networks help capture fine patterns in data and in return give enhanced predictive power

for the diabetes model. Kamrul Hasan, MD, and colleagues using classifiers such as K Nearest Neighbor, Decision Tree, and Support Vector Machine, proposed an ensemble-based method for diabetes prediction [5]. The authors had trained the ensemble model on the dataset of the 'Pima Indians Diabetes' and tested on different parameters of performance evaluation, in which F1 score, sensitivity, accuracy, and specificity were included. The overall performance of the ensemble model was better on all performance metrics than an individual classifier: F1 = 79.03%, sensitivity = 79.21%, and accuracy = 79.17%. This paper will add weight to the utility of integrated machine learning approaches in medical diagnostics, indicating advantages of ensemble methods in merging different classifiers to improve prediction accuracy and reliability. In brief, the research on diabetes prediction through machine learning has improved by leaps and bounds, with very basic ensemble and deep learning models to highly complex algorithms. The optimization of model performances is, in turn, dependent on the integration of many sources of data with sophisticated preprocessing.

III. EXISTING SYSTEM

In most of the studies, diabetes was mainly diagnosed by physical examinations in hospitals and involved several procedures of laboratory testing. Blood sugar levels are usually measured at random, following a fast, or following a glucose solution in these studies. Another common test is the HbA1c, reflecting an average blood sugar over the previous two to three months. The doctors also take into consideration lifestyle, age, weight, family history, and probably the symptoms like being thirsty, passing urine frequently, tiredness, blurred vision, and unexplained weight loss. Urine tests for ketones or glucose are rarely, though sometimes, employed.

However, these traditional methods also have their disadvantages. They are resource-intensive and laborious, needing many visits and dependent on the presence of medical staff and equipment, which is hard to come by in underdeveloped or rural regions of a country. Errors can occur through manual testing. Moreover, the need for lab testing and in-office visits may delay diagnosis and thus treatment, increasing the threat to life. Added to these are the expenses through lost wages and traveling, increasing the burden. It is really hard to successfully treat diabetes between doctor visits when continual blood sugar monitoring is not possible without personal equipment. That is to say, conventional techniques are essential for diagnosing and keeping up with diabetes; they have their flaws. On the contrary, frequent visits, dependence on the healthcare system, and possible human failures, along with financial and logistical difficulties in the case of more conventional approaches, illustrate the potential benefits which, in this case, machine learning can make for improved early diagnosis and individual treatment of diabetes patients.

IV. PROPOSED SYSTEM

The proposed method harnesses the power of ML in predicting the possibility of diabetes at its earlier stages, hence complementing the deficiencies of traditional diagnostic tools. In this research, 17 variables related to diabetes from Kaggle's BDD are used. The steps to the approach start with data preparation, which involves cleaning and normalization of the dataset. Then, feature selection is applied to determine important predictors for diabetes. In this stage, various machine learning methods—like Decision Tree, K Nearest Neighbor, Random Forest, Support Vector Machine, and Logistic Regression—are trained and their results are compared to get the best accurate model. The performance of the models was evaluated using several metrics such as recall, accuracy, and precision. Another important module at the system level is a web-based interface, which utilizes Flask, hence making user interaction easier. This platform will allow patients and doctors to enter the symptoms and get instant forecasts. Early Detection using ML models can avoid serious outcomes by allowing early interventions.

Also, this system will provide recommendations for treatment that is now calibrated due to the patient's specific data—ensuring better results and quality of care. It contains a vast volume of information and resources on diabetes, covering such aspects as its types, causes, wellness advice, and typical treatments. It is designed to provide easy access. This

benefit of accessibility can add to the improvement in patient education and involvement, allowing the general public to self-manage their health. Hence, it gives easy access to more people in remote areas with limited access to healthcare services through its web-based platform. Automated diagnostic procedures that are less dependent on manual analysis reduce human error and bias. This system lowers model bias: representative and diversified training data ensure that models predict accurately in any population. More importantly, quick findings are made since the system quickly processes the symptom data—something that is important for efficient diabetes control.

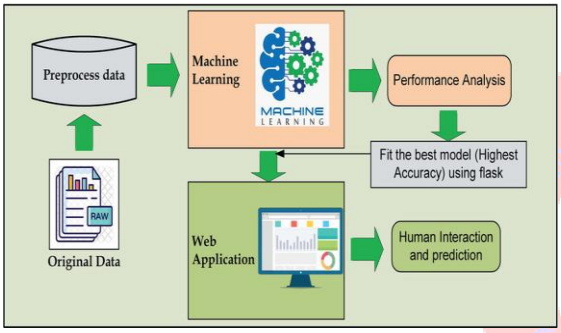

Figure 1. Architecture

In this regard, the machine learning-based system proposed here offers quite a few advantages over conventional techniques by way of early detection, tailored therapy, increased accessibility, cost-effectiveness, and improvement in patient education. Further, this new technology is also fraught with its problems, such as data quality issues, assuring privacy, and ensuring ease of use across user groups. Despite the challenges pointed out in the foregoing, the proposed strategy still has immense potential to enhance diabetes diagnosis and treatment for improved patient outcomes and reduced burden on healthcare systems.

V. IMPLEMENTATION

Data Collection

Data collection is one of the preliminary steps toward developing a predictive model. In this paper, data have been collected from Sylhet Hospital in Bangladesh through direct surveying of diabetes patients after obtaining approval from a doctor. This dataset contains 521 patient records with 16 features and one target variable.

Data Preprocessing

Data preprocessing is crucial to have a good-quality dataset. It includes cleaning, selecting relevant features, and data visualization for a better understanding of data and model performance.

Model Building

Model Building: It will consist of selecting and training appropriate Machine Learning algorithms on the prepared dataset to predict diabetes. Several Algorithms are considered as follows:

a. Logistic Regression:

- Concept: A technique used with binary class variables to predict event probabilities or class membership (0 or 1).
- Sigmoid Function: Projects the predicted values to a probability scale from 0 to 1.
- Equation: This logistic regression model is derived from the equation of linear regression and then transformed by the logistic function.

b. Decision Tree:

- Concept: A flowchart-like structure where every internal node shows a test on a feature, and every leaf node, class label, or class distribution.
- Building Process: recursive partitioning: it divides the data into subsets based on the feature tests.

c. Random Forest:

- Concept: This is an ensemble method whereby predictions from several decision trees are aggregated to improve on accuracy and avoid overfitting.
- Advantages: Efficient with large datasets, high-dimensional data; resistant to missing values.

d. Support Vector Machine (SVM):

- Concept: Find the best hyperplane to separate different classes within a feature space.

- Kernel Functions: To deal with both linear and nonlinear data, SVM does a mapping of data to higher dimensions.

e. K Nearest Neighbor (KNN):

- Concept: It predicts the class of any data point and considers a majority vote of its k nearest neighbors.
- Distance Metrics: Several metrics could be used in order to find how close two data points are to one another.

Evaluating the Model

It is an important phase where checks with regards to the performance and accuracy of the model are carried out for the predictive model. There are many metrics and tools that help evaluate model performance.

a. Confusion Matrix:

Components:

True Positives (TP)

True Negatives (TN)

False Positives (FP)

False Negatives (FN)

Helps in understanding the classification performance and identifying errors.

b. Accuracy:

The proportion of correctly classified instances out of the total number of instances.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN)$$

c. Precision:

The proportion of correctly predicted positive cases out of all predicted positives.

$$\text{Precision} = TP / (TP + FP)$$

d. Recall:

The ratio of correctly identified positive cases to all actual positive cases.

$$\text{Recall} = TP / (TP + FN)$$

Connecting to Web Interface

Such integration will allow users to interface with the developed machine learning model through a web interface. In this respect, Flask is used as a lightweight Python web framework.

VI. RESULTS

Data Collection

This dataset was downloaded from the Kaggle website. There are 521 samples; each has 16 parameters associated with diabetes symptoms, such as age, sex, genital thrush, polyuria, and polydipsia. The target variable in this research is represented by the column called "class" defining whether a patient has diabetes or not. Each sample is saved as one row in a CSV file.

Preparing the Dataset

Data preparation involved several steps:

- **Identifying Missing Values:** Checking for any missing or null values in the dataset using the `isna()` function returning a Boolean indicating the presence of NA values.
- **Data Encoding:** In this, categorical variables like Yes/No are converted into numerical values to avoid miscommunication during model training.

Exploratory Data Analysis

The given dataset will do an EDA to know the distribution.

Data Correlation

We computed the correlation of different features with the target variable to see which among them drives the greatest influence. The correlation matrix showed strong relationships between target variables and symptoms like polyuria, polydipsia, and sudden weight loss.

Data Normalization

The reason for data normalization in this case was to rescale features to a common range, usually with a mean of 0 and standard deviation of 1. This stage is very critical in ensuring that the magnitude of

features does not disproportionately impact the process of model training.

Model Building

We trained five different algorithms on the dataset: Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, and K-nearest Neighbors. We then calculated the accuracy, precision, and recall for every model.

Algorithm	Accuracy	Precision	Recall
Random Forest	95.51	0.971154	0.961905
KNN	83.97	0.788462	0.964706
Decision Tree	95.51	0.955128	0.955128
Support Vector Classifier	91.67	0.916667	0.916667
Logistic Regression	94.87	0.948718	0.948718

Table 1: Results

Out of all the tested algorithms, Random Forest and the Decision Tree methods gave the highest accuracy. On the other hand, Random Forest was able to attain the highest precision and recall among these six algorithms; thus, it is the optimal algorithm for this project on early-stage diabetes prediction.

VII. CONCLUSION

The objective of the research is to develop a machine learning model that will be helpful in predicting diabetes at an early stage using the symptoms dataset. Five machine learning techniques have been tested, specifically: Random Forest, K-Nearest Neighbors, Decision Tree, Support Vector Machine, and Logistic Regression. In this case, the used dataset consisted of 16 features and 521 instances, and it is sourced from Kaggle. We handled missing values, encoded categorical data, and normalized features as part of our thorough data pretreatment. Some very important information about the distribution of symptoms across demographic groups and their relation to diabetes was revealed by exploratory data analysis. On the models, it was found that the Random Forest Algorithm performed better than the rest, with an accuracy of 95.51 percent, precision of 0.971154, and recall of 0.9611905. Since they are ensembles of decision trees themselves, the Random Forest model is intrinsically resistant to overfitting and robust with respect to handling feature interactions. In summary, based on the information provided, the Random Forest Algorithm is the best model for the prediction of early stages of diabetes. This is a reliable tool for detecting those at risk of diabetes because it has high accuracy, precision, and memory. This enables early intervention and control of the condition. Dataset growing, addition of more features, and more sophisticated machine learning methods can then be focused on in future study to improve the forecast accuracy.

REFERENCES

[1] Ahmed, "Prediction of diabetics empowered with fused Machine-learning", 2022 International Research Journal of Modernization in Engineering Technology and Science, Student, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India.

[2] Daliya, "An Optimized Multivariable Regression Model for Predictive Analysis of Diabetic Disease Progression", 2021 Department of Electronics and Communication Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru, India.

[3] Jiajia Song, Chao Wang, and Wenzhuo Zhao, “Literature review on machine-learning for diabetes prediction”, 2021.

[4] Hruaping Zhou, Raushan Myrzashova and Ruiz Heng, “An enhanced deep neural network (DNN) model for predicting diabetes.”

[5] MD. Kamrul Hasan and MD. Ashraful Alam, "Diabetes Prediction Using Ensembling of Different Machine-learning Classifiers", 2020.