International Journal of Research in Engineering Sciences and Management
Vol.1, No.1, July-Dec' 2024
E-ISSN: 3107-4383

Smart Diabetes Prediction System: Utilizing Machine Learning For Early Risk Detection

Dhanya K.R.[1], Dr. R. Suganthi[2]

[1]M.Sc. CSDA, Dr. N.G.P. Arts and Science College, Coimbatore

dhanyafab@gmail.com

[2]MCA, M.Phil, Ph.D., Professor, Dr. N.G.P. Arts and Science College, Coimbatore - suganthi.r@drngpasc.ac.in

Abstract

Diabetes is a growing global health concern, with its rates increasing across all age groups, affecting children, teenagers, adults, and older individuals. This study explores the prediction of diabetes using machine learning algorithms, including K-Nearest Neighbors (KNN), AdaBoost, and Random Forest, with a dataset from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset includes diagnostic measurements like the number of pregnancies, glucose levels, blood pressure, skin thickness, insulin, body mass index (BMI), diabetes pedigree function, and age, with an outcome variable indicating whether diabetes is present. Among the algorithms tested, Random Forest achieved the highest accuracy at 74.03%, compared to 66.23% for KNN and 73.38% for AdaBoost. Based on these results, a predictive model was created using Random Forest, along with a user-friendly interface built with the Streamlit Python library. This interface allows users to register, log in, and fill out a medical form to assess their risk of diabetes. By clicking the "Test Results" button, users receive predictions indicating whether they have diabetes based on their input data. User authentication is managed through MongoDB, ensuring secure storage and validation of credentials. The application is deployed on Render, providing easy access for users on both mobile devices and laptops, making it a convenient tool for evaluating diabetes risk based on personal health information.

Keywords— Diabetes, Machine Learning, Random Forest, K-Nearest Neighbors, AdaBoost, Predictive Model, Health Assessment, Streamlit, MongoDB, User Interface.

1. Introduction

Diabetes is becoming a major health issue around the world, affecting millions of people. Currently, about 537 million individuals have diabetes, which makes up roughly 6.6% of the global population of around 8.1 billion. This means that for every 100 people, around 6 or 7 are living with this condition. Diabetes can occur in people of all ages, including children, teenagers, adults, and old individuals. At its core, diabetes is a problem with how the body manages blood sugar (glucose).

Under normal circumstances, a hormone called insulin, produced by the pancreas, helps move sugar from the blood into the cells, where it is used for energy. However, in diabetes, the body either doesn't make enough insulin or the cells can't use insulin properly, resulting in high blood sugar levels.

There are two main types of diabetes. Type 1 diabetes occurs when the body stops making insulin, requiring insulin injections. Type 2 diabetes, the more common type, is often linked to poor diet, lack of exercise, and obesity, where the body doesn't use insulin effectively, causing high blood sugar.

Gestational diabetes occurs during pregnancy, leading to high blood sugar that can harm both mother and baby. While it usually goes away after childbirth, it increases the mother's risk of Type 2 diabetes later on.

Certain groups of people are more likely to develop diabetes. For example, those with a family history of diabetes, individuals who are overweight or obese, and people over the age of 45 are at a higher risk. Other risk factors include having a sedentary lifestyle or existing health issues like high blood pressure or heart disease.

When diabetes isn't treated well, it can lead to severe health issues. High blood sugar can damage blood vessels, which increases the risk of heart disease and strokes. It can also harm the kidneys, potentially leading to kidney failure, and cause vision problems like diabetic retinopathy, which can result in blindness. Moreover, people with diabetes are at greater risk for gum disease and other dental issues because high blood sugar levels can lead to infections and inflammation.

As diabetes continues to affect more people across all age groups, finding effective ways to predict and manage the condition is becoming increasingly important. New advancements in technology, especially in machine learning, offer promising methods for improving diabetes risk assessments and early detection. This study focuses on developing a machine learning model that helps individuals assess their health status and take steps to manage their diabetes risk.

2. Literature Review

Research on diabetes prediction using machine learning has gained significant attention as healthcare systems strive for innovative solutions for early diagnosis and intervention. Various studies have demonstrated the ability of machine learning techniques to analyze large datasets and identify complex patterns related to diabetes risk factors, ultimately enhancing prediction accuracy. Researchers have explored a range of algorithms, from traditional methods like logistic regression to advanced deep learning approaches, showcasing their effectiveness in tailoring predictions to individual patient profiles.

Md. Tanvir Islam et al. [2] investigated "Typical and Non-Typical Diabetes Disease Prediction using Random Forest Algorithm." Their use of the Random Forest method, which combines multiple decision trees, improved prediction accuracy and helped identify key risk factors for diabetes. However, they highlighted challenges such as data quality issues, class imbalance, and the complexity of interpreting results, which can make it difficult for clinicians to apply the findings effectively. In a complementary study, Muhammad Exell Febrian et al. [3] explored "Diabetes Prediction Using Supervised Machine Learning."

They compared two popular algorithms, k-Nearest Neighbors (KNN) and Naive Bayes, finding that Naive Bayes outperformed KNN in this context.

Expanding the conversation on machine learning methods, Ambika Choudhury and Deepak Gupta [4] presented a comprehensive review in their paper "A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques," discussing various methods, including artificial neural networks and support vector machines, utilizing the PIMA Indian Diabetes dataset. Their review emphasized how machine learning can enhance diagnosis and early detection of diabetes, although they acknowledged that some methods require substantial computational resources, which can restrict their accessibility in clinical environments.

Isfafuzzaman Tasin et al. [5] explored diabetes prediction in their research titled "Diabetes Prediction Using Machine Learning and Explainable AI Techniques," developing an automatic prediction system using various algorithms, particularly emphasizing the XGBoost method for its high accuracy. They also integrated explainable AI techniques to enhance model transparency, making predictions easier to understand for users; however, their reliance on a private dataset raised concerns about the generalizability of their findings.

Finally, Victor Chang et al. [6] assessed different machine learning models in their study "An Assessment of Machine Learning Models and Algorithms for Early Prediction and Diagnosis of Diabetes Using Health Indicators," finding that the Random Forest model was the most effective, achieving high accuracy rates. However, their research highlighted the importance of continuously evaluating models as healthcare data and methodologies evolve.

3. Proposed System

A. Data Collection

The dataset used for this study comes from the National Institute of Diabetes and Digestive and Kidney Diseases[1]. It aims to predict if a patient has diabetes based on several health measurements. The dataset includes information like the number of pregnancies, glucose levels after a test, blood pressure, skin thickness, insulin levels, body mass index (BMI), family history of diabetes, and age. The main goal is to determine if a patient is diabetic (1) or not (0). This dataset was selected due to its completeness, reliability, and relevance to the research. A review of alternative datasets indicated issues such as missing data or low quality, making this labeled dataset, which contains a total of 768 entries, the most suitable option for the study. Below, Table 1 shows the outcomes in the dataset:

Table 1: Outcomes in the dataset

Outcome	Number of Entries
No.of.Diabetes(0)	500
Diabetes(1)	268
Total	768

B. Data Preprocessing

In the data preprocessing stage, the dataset was first examined for missing values using the function `data.isnull().sum()`, which counts the missing entries in each column. If any missing values were identified, they were filled with a specified value, such as 0, by applying `data.fillna(0)`. Next, the data types of various features were converted to ensure they were in the correct formats for analysis. These preprocessing steps were crucial for preparing the dataset for accurate analysis and modeling.

C. Feature Selection

Next, feature selection is carried out to identify the most important health measurements that significantly contribute to predicting diabetes. The key features chosen in this study include glucose levels, blood pressure, and insulin levels. High glucose levels indicate issues with sugar management, while abnormal blood pressure can point to insulin resistance. Insulin levels are also critical, as they regulate blood sugar. Although other factors like BMI, age, and family history of diabetes offer additional insights, focusing on glucose, blood pressure, and insulin provides the most reliable basis for predicting diabetes. Statistical methods were applied to determine the significance of these features, ensuring a more accurate model for assessing diabetes risk.

D. Dataset Splitting

For this study, the dataset was split into two parts: 80% for training and 20% for testing. This approach ensures that the model is trained on a significant portion of the data while leaving a separate set to evaluate its performance.

E. Model Training

In the model training phase, three machine learning algorithms were assessed for their effectiveness in predicting diabetes using essential features from the dataset. K-Nearest Neighbors (KNN) classifies instances by analyzing the proximity of nearby data points, selecting the most similar features, such as glucose and blood pressure, to determine the likely outcome based on the majority class of its neighbors. AdaBoost works by creating a series of weak classifiers, each focusing on previously misclassified instances, effectively weighing features like insulin levels and BMI to improve accuracy on challenging cases. In contrast, Random Forest constructs multiple decision trees that evaluate the importance of each feature by examining how they split the data and contribute to accurate classifications. This ensemble method allows it to consider interactions between features, such as how glucose and insulin levels together impact the likelihood of diabetes.

F. Model Selection

In this study, model selection can be done by evaluating the performance of these algorithms through the Confusion Matrix, using key metrics such as Accuracy, Precision, Recall, and F1 Score.

The Confusion Matrix provides the model's predictions by categorizing them into four types: True Positives (TP), where diabetic patients are correctly classified; False Positives (FP), where non-diabetic patients are incorrectly classified as diabetic; False Negatives (FN), where diabetic patients are incorrectly classified as non-diabetic; and True Negatives (TN), where non-diabetic patients are correctly classified.

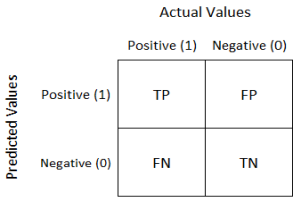

Fig 1: TP,FP,FN,TN Arrangement

Accuracy is the ratio of correctly predicted instances (TP + TN) to the total number of predictions made (TP + TN + FP + FN). It reflects the model's overall ability to correctly classify diabetic and non-diabetic patients, calculated as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Precision measures the proportion of true positive predictions out of all positive predictions made by the model. It emphasizes the model's ability to minimize false positives (i.e., predicting someone has diabetes when they do not). Precision is given by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall, also known as sensitivity, measures the proportion of actual diabetic cases correctly identified by the model. It highlights the model's ability to minimize false negatives, where diabetic patients are misclassified as non-diabetic. The formula for recall is:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1 Score is the harmonic mean of precision and recall, providing a balanced measure when there is an uneven class distribution. It is calculated as:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

G. Streamlit Interface

The Streamlit library was used to build an easy-to-use web application for the Smart Diabetes Prediction System. This interface allows users to register and log in securely using MongoDB Atlas. After logging in, users can enter important health information like the number of pregnancies, glucose level, blood pressure, insulin level, body mass index (BMI) etc. Once they fill out the medical form, users can click the "Predict" button to receive instant

predictions about their diabetes risk based on the best-performing model. Additionally, the application includes a logout feature, ensuring users can safely exit their session.

4. Experiment and Results

The performance metrics for the evaluated algorithms are summarized in the Table 2 and Figure 2 below:

Table 2: Performance Metrics

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.740260	0.627119	0.672727	0.649123
KNN	0.662338	0.524590	0.581818	0.551724
AdaBoost	0.733766	0.625000	0.636364	0.630631

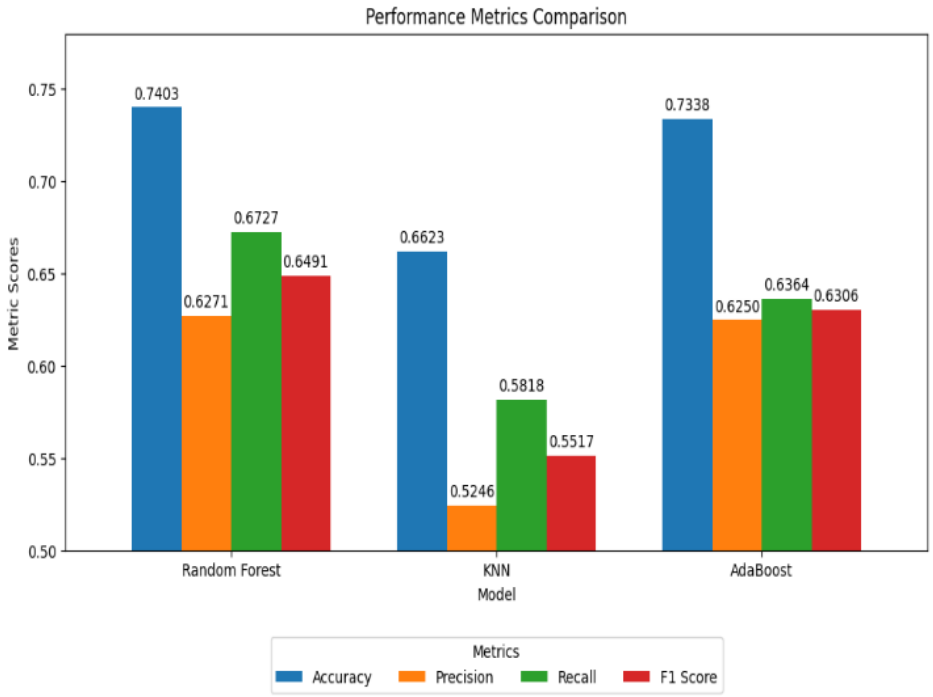

Fig 2: Performance Metrics Comparison

Among the three algorithms tested, Random Forest was the most effective, achieving an accuracy of 74.03%. It demonstrated a precision of 0.63 and a recall of 0.67 for positive cases, indicating its strong capability to identify individuals with diabetes. The confusion matrix for Random Forest is shown in fig.3.

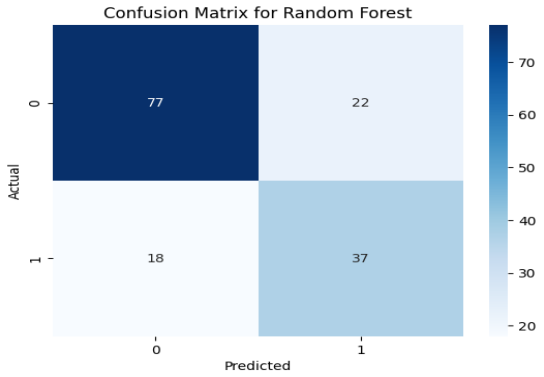

Fig 3: Confusion matrix for Random Forest

In contrast, K-Nearest Neighbors (KNN) had an accuracy of 66.23%, with a precision of 0.52 and a recall of 0.58, indicating its limitations in accurately detecting diabetes. The confusion matrix for KNN is shown in fig.4.

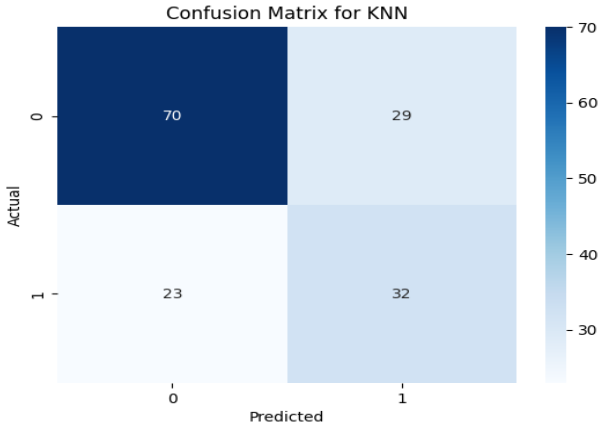

Fig.4: Confusion matrix for KNN

AdaBoost achieved an accuracy of 73.38%, with a precision of 0.62 and a recall of 0.64. While better than KNN, it still lagged behind Random Forest. The confusion matrix for AdaBoost is shown in fig.5.

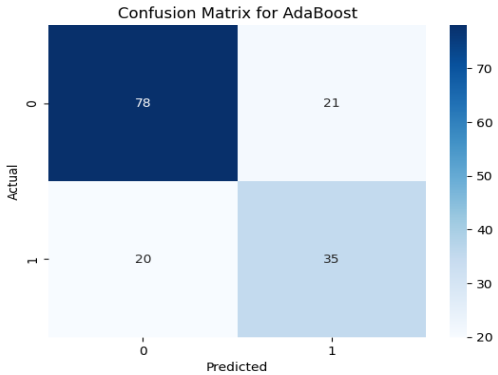

Fig.5: Confusion matrix for AdaBoost

The next step involved integrating the selected random forest model into a user-friendly application using the Streamlit framework. The User Registration feature allows users to create an account, with authentication handled via MongoDB. Once registered, users can securely log in through the Login page.

Below, Fig.6 displays the registration page, and Fig.7 illustrates the login page.

Fig.6: Signup (Registration Page)

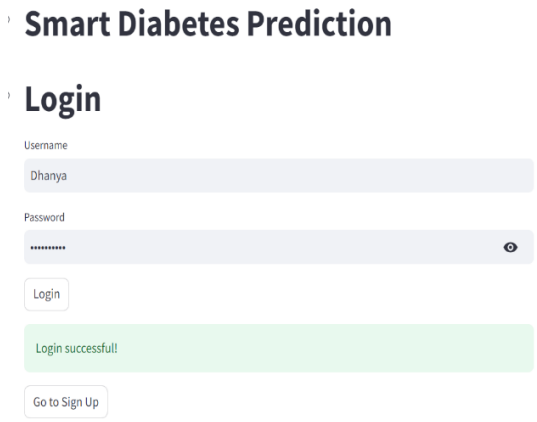

Fig.7: Login Page

After successfully logging in, users are presented with an input form. Fig.8 displays this medical form.

Fig.8: Medical Form

Based on the user's input, which includes medical details, the system utilizes the Random Forest model to determine the likelihood of the user having diabetes. The result is then displayed, indicating whether the user is diabetic or non-diabetic. Below, Fig.9 illustrates an example of a diabetic result, while Fig.10 shows a non-diabetic result.

Fig.9: The person is Diabetic

Fig.10: The person is Non- Diabetic

Next, after the prediction results, if the user clicks the logout button, the system logs them out of the session and redirects them to the login page, ensuring secure access control.

The application is deployed on Render as a dynamic web page, making it accessible on both mobile devices and desktop systems.

5. Conclusion and Future Scope

The Smart Diabetes Prediction System effectively uses machine learning to predict diabetes risk through the Random Forest algorithm. The user-friendly interface, along with secure registration and login via MongoDB, allows easy access to results for users. This demonstrates how technology can assist in managing health conditions like diabetes. The future scope of this prediction system includes integrating advanced features and adopting a microservices architecture for better scalability and maintenance. This approach would allow the application to be divided into smaller, independent services, making updates easier. Incorporating more advanced machine learning techniques, such as deep learning, could enhance prediction accuracy. Real-time data analytics and integration with health APIs would improve user input processing and provide richer datasets. Additionally, developing a dedicated mobile application, implementing cloud services for data management, and enhancing data security through blockchain technology could create a more comprehensive health management platform.

References

[1] Dataset from: National Institute of Diabetes and Digestive and Kidney Diseases.

[2] Islam MT, Raihan M, Farzana F, Aktar N, Ghosh P, Kabiraj S. (2020). Typical and non-typical diabetes disease prediction using random forest algorithm. In: Abstracts of the 11th international conference on computing, communication and networking technologies, IEEE, Kharagpur, 1-3 July 2020.

[3] Febrian ME, Ferdinan FX, Sendani GP, Suryanigrum KM, Yunanda R. (2022). Diabetes prediction using supervised machine learning.

[4] Choudhury A, Gupta D. (2019). A survey on medical diagnosis of diabetes using machine learning techniques. In: Kalita J, Balas VE, Borah S, Pradhan R (eds) Recent Developments in Machine Learning and Data Analytics. Advances in Intelligent Systems and Computing, vol 740. Springer, Singapore, pp 67–78. https://doi.org/10.1007/978-981-13-1280-9_6.

[5] Tasin I, Ullah Nabil T, Islam S, Khan R. (2022). Diabetes prediction using machine learning and explainable AI techniques.

[6] Chang V, Ganatra MA, Hall K, Golightly L, Xu QA. (2022). An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators.

[7] Kamble MT, Patil ST. (2016). "Diabetes detection using deep learning approach." International Journal for Innovative Research in Science & Technology 2(12):342-9.

[8] Vijayan VV, Anjali C. (2015). "Prediction and diagnosis of diabetes mellitus—A machine learning approach." In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) Dec, IEEE 10: 122-127.

[9] Sneha N, Gangil T. (2019). "Analysis of diabetes mellitus for early prediction using optimal features selection." Journal of Big Data.

[10] Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM. (2020). "Classification and prediction of diabetes disease using machine learning paradigm." Health Information Science and Systems.

[11] Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. (2018). "Predicting diabetes mellitus with machine learning techniques." Frontiers in Genetics.

[12] Early-stage diabetes risk prediction dataset. (2020). UCI Machine Learning Repository.

[13] Gauri D. Kalyankar, Shivananda R. Poojara, and Nagaraj V. Dharwadkar. (2017). "Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop." International Conference On I-SMAC, 978-1-5090-3243-3.

[14] B.M. Patil, R.C. Joshi, and Durga Toshniwal. (2010). "Association Rule for Classification of Type-2 Diabetic Patients." ICMLC '10 Proceedings of the 2010 Second International Conference on Machine Learning and Computing, February 09 - 11, 2010.

[15] P. Suresh Kumar and S. Pranavi. (2017). "Performance Analysis of Machine Learning Algorithms on Diabetes Dataset using Big Data Analytics." International Conference on Infocom Technologies and Unmanned Systems, 978-1-5386-0514-1, Dec. 18-20.