Electrical and Automation Engineering

Vol: 5(1), 2026

REST Publisher; ISBN: 978-81-956353-5-1

Website: http://restpublisher.com/book-series/eae/

DOI: https://doi.org/10.46632/eae/5/1/2



Diabetes Prediction System Using Machine Learning

Karnack Dutta, Rashmi Ranjan Dalai, Ankita Nandi, Meghali Mukherjee, *Arpita Santra, Sandhya Pattanayak, Swati Barui

Narula Institute of Technology, Kolkata, West Bengal, India

*Corresponding author: arpita.santra@nit.ac.in

Abstract: *Diabetes is one of the most common chronic diseases affecting a large population worldwide. Early identification of diabetes is important because delayed diagnosis can lead to severe health complications such as heart disease, kidney failure, nerve damage, and vision-related disorders. In recent years, Machine Learning techniques have been widely used in the healthcare sector for disease prediction and medical data analysis. These techniques help in identifying hidden patterns from historical patient data and assist in faster decision-making. This project presents a machine learning-based diabetes prediction system developed using patient medical attributes such as glucose level, blood pressure, insulin level, body mass index (BMI), age, pregnancies, skin thickness, and diabetes pedigree function. The system uses the Pima Indians Diabetes Dataset for training and testing purposes. Exploratory Data Analysis (EDA) was performed to understand dataset characteristics and analyze feature distribution before model training. For prediction, the Random Forest Classifier algorithm was implemented because of its effectiveness in handling classification problems and structured healthcare datasets. The dataset was divided into training and testing sets, and feature scaling techniques were applied during preprocessing. The trained model was evaluated using performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix. The model achieved satisfactory prediction performance for diabetes classification. To provide an interactive user experience, the trained machine learning model was integrated into a Streamlit-based web application. The application allows users to enter patient health information and receive instant prediction results regarding diabetes risk. Additional visualization features were included to improve understanding of model performance and prediction behavior. The proposed system demonstrates the practical application of machine learning in healthcare prediction systems. The project highlights how data-driven approaches can support preliminary disease risk assessment and contribute toward intelligent healthcare solutions.*

Keywords *- Diabetes Prediction, Machine Learning, Random Forest Classifier, Healthcare Analytics, stream lit, Data Analysis, Classification, Artificial Intelligence, Predictive System.*

1. INTRODUCTION

Diabetes has become one of the most common health-related problems in modern society. It is a disease in which the human body cannot properly control the level of glucose present in the blood. This condition may occur either because the body does not produce sufficient insulin or because the produced insulin does not function effectively. Over the last few years, the number of diabetic patients has increased rapidly due to changes in lifestyle, unhealthy eating habits, obesity, stress, lack of physical activity, and genetic factors. If diabetes is not identified at an early stage, it can create several serious complications such as kidney problems, heart diseases, nerve damage, and eyesight-related disorders. Because of these reasons, early prediction and monitoring of diabetes have become very important in the healthcare field. In traditional healthcare systems, diabetes diagnosis mainly depends on laboratory testing and medical examination performed by doctors. Although these methods are reliable, analyzing large amounts of patient information manually can take considerable time and effort. In many situations, especially where healthcare resources are limited, there is a requirement for systems that can assist in predicting diseases quickly and efficiently. Recent developments in Artificial Intelligence and Machine Learning have created new possibilities for solving such healthcare-related problems through data-driven approaches. Machine Learning is a technique that allows computer

Copyright@ REST Publisher

systems to learn patterns automatically from previous data. Instead of giving fixed rules to the computer, machine learning algorithms study historical records and identify relationships between input data and expected outputs. In medical applications, this approach is highly useful because disease prediction often depends on multiple health parameters that may be interconnected in complex ways. By learning from existing patient records, machine learning systems can help predict the likelihood of diseases in new patients. The main purpose of this project is to develop a diabetes prediction system using machine learning techniques. The system uses medical parameters such as glucose level, blood pressure, insulin, body mass index (BMI), age, pregnancies, skin thickness, and diabetes pedigree function to predict whether a person is diabetic or non-diabetic. For this project, the Pima Indians Diabetes Dataset has been used for training and testing the model. The dataset contains medical records of patients along with their diabetes outcomes, making it suitable for supervised machine learning classification. Before training the model, exploratory data analysis was performed to understand the dataset structure, feature distribution, and class balance. Different preprocessing steps were applied to prepare the data for training. In this project, the Random Forest Classifier algorithm has been used because of its good performance in classification problems and its ability to handle structured healthcare data effectively. The trained model learns patterns from historical patient data and later predicts diabetes risk for new input values provided by users. To make the project interactive and easy to use, the machine learning model has been integrated with a Streamlit-based web application. Users can enter medical details through the interface, and the system instantly generates prediction results. The project also includes visualization components such as confusion matrix analysis and feature importance representation to better understand model behavior and prediction performance. This project demonstrates how machine learning can be applied in healthcare for disease prediction and decision support. Although the system is designed mainly for academic and educational purposes, it shows the potential of intelligent prediction systems in assisting early diabetes risk analysis using patient medical data.

2. DATASET

The development of the proposed diabetes prediction system required the use of a healthcare dataset, machine learning libraries, and software tools for data analysis, model training, and application deployment. The project was implemented using Python because of its simplicity, flexibility, and extensive support for data science and machine learning applications. For this project, the Pima Indians Diabetes Dataset was used as the primary dataset for model training and evaluation. This dataset is widely used in healthcare-related machine learning research and contains medical records of female patients along with diabetes outcomes. The dataset consists of 768 patient records and 9 attributes. Among these attributes, 8 columns represent input medical features and 1 column represents the target output variable named Outcome. The Outcome column contains binary values where 0 indicates non-diabetic patients and 1 indicates diabetic patients. The medical attributes available in the dataset include Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index (BMI), Diabetes Pedigree Function, and Age. These parameters are commonly associated with diabetes diagnosis and risk analysis. The dataset contains both diabetic and non-diabetic cases, enabling the machine learning model to learn patterns from historical patient records and classify new patient data accordingly. Before model training, exploratory data analysis was performed to understand the structure and characteristics of the dataset. Statistical analysis helped identify feature ranges, average values, and data distribution patterns. During analysis, it was observed that some features contained zero values in places where zero is medically unrealistic, such as glucose level, BMI, and blood pressure. These values were treated as part of the available dataset during project implementation. Several Python libraries were used throughout the project development process. Pandas was used for dataset loading, manipulation, and exploratory data analysis. NumPy supported numerical operations and data handling. Scikit-learn was used for preprocessing, train-test splitting, machine learning model implementation, and evaluation metrics. The Random Forest Classifier algorithm was selected for prediction because of its effectiveness in handling classification problems and structured tabular datasets. Pickle was used to save the trained model and preprocessing scaler for future use without retraining the system repeatedly. The web-based prediction interface was developed using Streamlit, which enabled the creation of an interactive and user-friendly application. Through this interface, users can enter medical information and obtain prediction results instantly. The project was developed and executed using Visual Studio Code as the programming environment. The implementation was tested on a standard computer system with Python installed along with the required machine learning libraries.
The materials and technologies used in this project provided an effective environment for implementing a machine learning-based healthcare prediction system capable of analyzing patient health data and generating diabetes prediction results.

Copyright@ REST Publisher

3. METHOD

The methodology of the proposed diabetes prediction system consists of multiple stages including dataset collection, exploratory data analysis, data preprocessing, model training, evaluation, and web application integration. The complete workflow was designed to transform raw healthcare data into a machine learning-based prediction system capable of identifying diabetes risk from patient medical attributes. The first stage of the project involved collecting and loading the dataset into the Python environment. The Pima Indians Diabetes Dataset was imported using the Pandas library and stored in a Data Frame for further processing. After loading the dataset, exploratory data analysis was performed to understand the structure and characteristics of the data. Functions such as head (), describe (), and value counts () were used to inspect sample records, statistical summaries, and class distribution. This analysis helped in understanding feature ranges, identifying anomalies, and examining the balance between diabetic and non-diabetic records. After dataset analysis, the input features and target output variable were separated. The input attributes included medical parameters such as glucose level, blood pressure, insulin, BMI, age, pregnancies, skin thickness, and diabetes pedigree function. The target variable was the Outcome column, which represented whether a patient was diabetic or non-diabetic. Separating features and labels were necessary because the machine learning model must learn relationships between patient attributes and diabetes outcomes. The next step involved preprocessing the dataset before training the machine learning model. Since the dataset features had different numerical ranges, feature scaling was applied using the Standard Scaler technique from the Scikit-learn library. Scaling standardizes the feature values and helps maintain consistency during model training. This preprocessing step transforms the data into a normalized format suitable for machine learning algorithms. After preprocessing, the dataset was divided into training and testing sets using the train-test split method. In this project, 80% of the dataset was used for training and 20% was used for testing. The training dataset was used by the model to learn patterns from historical patient records, while the testing dataset was used to evaluate prediction performance on previously unseen data. This separation helps measure how effectively the model generalizes to new patient inputs. For diabetes prediction, the Random Forest Classifier algorithm was implemented. Random Forest is an ensemble machine learning algorithm that combines multiple decision trees to improve classification accuracy and reduce overfitting. During training, the algorithm analyzed different combinations of patient medical parameters and learned patterns associated with diabetic and non-diabetic cases. Multiple decision trees were generated internally, and the final prediction was determined based on the collective output of these trees. Once the model training process was completed, prediction results were generated for the testing dataset. The trained model was evaluated using various performance metrics including accuracy, precision, recall, F1-score, and confusion matrix analysis. Accuracy measured the overall correctness of predictions, while precision and recall helped evaluate classification quality. The confusion matrix provided detailed insight into correct and incorrect predictions, including false positive and false negative classifications. To avoid retraining the model every time the application runs, the trained Random Forest model and Standard Scaler object were saved using the Pickle library. Model serialization enabled the trained system to be reused efficiently during application execution. Finally, the trained model was integrated into a Streamlet-based web application to create an interactive prediction interface. The application allows users to enter patient medical details through input fields. Once the data is submitted, the application preprocesses the input using the saved scaler and sends the transformed values to the trained machine learning model for prediction. The prediction result is then displayed on the interface, indicating whether the patient is likely diabetic or non-diabetic. The overall methodology demonstrates the complete implementation flow of a machine learning-based healthcare prediction system, beginning from raw dataset analysis and ending with real-time prediction through an interactive web application.

4. ANALYSIS AND DISCUSSION

The analysis and discussion phase of the project focuses on evaluating the performance of the machine learning model and understanding how different medical parameters influence diabetes prediction. After preprocessing and training the dataset using the Random Forest Classifier algorithm, the model was tested using unseen patient records from the testing dataset. Various evaluation metrics and visualization techniques were used to examine the effectiveness of the prediction system. The diabetes dataset used in this project contains 768 patient records with 8 input medical features and 1 target output variable. During exploratory analysis, it was observed that the dataset contains both diabetic and non-diabetic patient records. Out of the total records, 500 instances belong to non-diabetic patients while 268 instances belong to diabetic patients. This indicates that the dataset is slightly imbalanced because the number of non-diabetic cases is higher than diabetic cases.

Copyright@ REST Publisher

TABLE 1. Dataset Information

Parameter	Value
Dataset Name	Pima Indians Diabetes Dataset
Total Records	768
Total Features	8
Target Variable	Outcome
Non-Diabetic Cases	500
Diabetic Cases	268
Training Data	80%
Testing Data	20%
Algorithm Used	Random Forest Classifier

```
PS C:\Users\ranja\OneDrive\Desktop\diabetes-predictor> python explore.py
   Pregnancies  Glucose  BloodPressure  SkinThickness  Insulin   BMI  DiabetesPedigreeFunction  Age  Outcome
0            6      148             72             35        0  33.6                     0.627   50        1
1            1       85             66             29        0  26.6                     0.351   31        0
2            8      183             64              0        0  23.3                     0.672   32        1
3            1       89             66             23       94  28.1                     0.167   21        0
4            0      137             40             35      168  43.1                     2.288   33        1
```

FIGURE 1. Sample records from the diabetes dataset

During statistical analysis, several observations were identified from the dataset. The average glucose level was found to be approximately 120.89, while the average BMI value was approximately 31.99. Certain features such as glucose level, BMI, blood pressure, and insulin contained zero values, which are medically unrealistic in real-life conditions. These values may represent missing or unavailable measurements within the dataset. Despite these irregularities, the dataset was retained in its original form for model training and prediction. The dataset was divided into training and testing sets using the train-test split technique. Approximately 80% of the data was used for training the model, while the remaining 20% was used for testing. Feature scaling was applied using Standard Scaler to normalize feature ranges before model training. Scaling helped maintain consistency among features with different numerical ranges. The Random Forest Classifier algorithm was selected because of its ability to handle structured tabular data efficiently and generate reliable classification results. During training, the algorithm created multiple decision trees internally and combined their outputs to generate final predictions. The trained model learned relationships between medical attributes and diabetes outcomes from historical patient records. After training, the model was evaluated using performance metrics including accuracy, precision, recall, F1-score, and confusion matrix analysis. The model achieved an overall prediction accuracy of approximately 72.7%, indicating that the system was able to classify a significant portion of patient records correctly.

TABLE 2. Model Performance Metrics

Metric	**Value**
Accuracy	72.7%
Precision (Class 0)	0.79
Recall (Class 0)	0.79
F1-Score (Class 0)	0.79
Precision (Class 1)	0.62
Recall (Class 1)	0.62
F1-Score (Class 1)	0.62

Copyright@ REST Publisher

```
PS C:\Users\ranja\OneDrive\Desktop\diabetes-predictor> python model.py
Accuracy: 0.7272727272727273
              precision    recall  f1-score   support

           0       0.79      0.79      0.79        99
           1       0.62      0.62      0.62        55

    accuracy                           0.73       154
   macro avg       0.70      0.70      0.70       154
weighted avg       0.73      0.73      0.73       154

Model saved!
```

FIGURE 2. Model evaluation metrics generated after training

The confusion matrix was used to analyze the detailed prediction behavior of the model. Unlike accuracy, which provides only overall correctness, the confusion matrix shows the exact distribution of correct and incorrect predictions. This helped identify the number of diabetic and non-diabetic patients classified correctly and incorrectly by the model.

The confusion matrix generated by the system contained the following values:

- True Negatives (TN): 78
- False Positives (FP): 21
- False Negatives (FN): 21
- True Positives (TP): 34

The model correctly identified 78 non-diabetic patients and 34 diabetic patients. However, 21 healthy patients were incorrectly predicted as diabetic, and 21 diabetic patients were incorrectly classified as non-diabetic. In healthcare applications, false negative predictions are particularly important because they represent diabetic patients who were not detected correctly by the system.

TABLE 3. Confusion Matrix Analysis

Actual Class	Predicted non-diabetic	Predicted Diabetic
Non-Diabetic	78	21
Diabetic	21	34

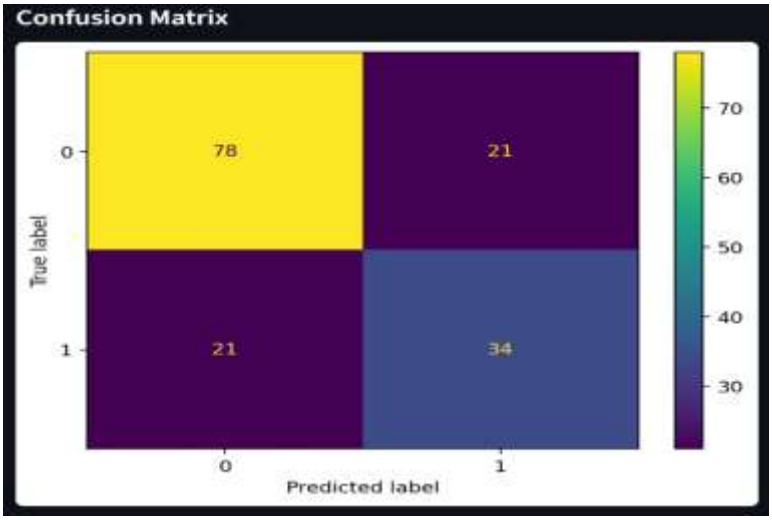


FIGURE 3. Confusion matrix showing classification performance of the Random Forest model

Feature importance analysis was also performed to identify which medical attributes contributed most strongly toward diabetes prediction. According to the trained Random Forest model, glucose level had the highest influence on prediction, followed by BMI and age. This observation is medically meaningful because glucose concentration and obesity-related factors are strongly associated with diabetes risk.

Copyright@ REST Publisher

TABLE 4. Feature Importance Analysis

Feature	Importance Score
Glucose	0.2588
BMI	0.1699
Age	0.1409
Diabetes Pedigree Function	0.1237
Blood Pressure	0.0881
Pregnancies	0.0765
Insulin	0.0761
Skin Thickness	0.0656

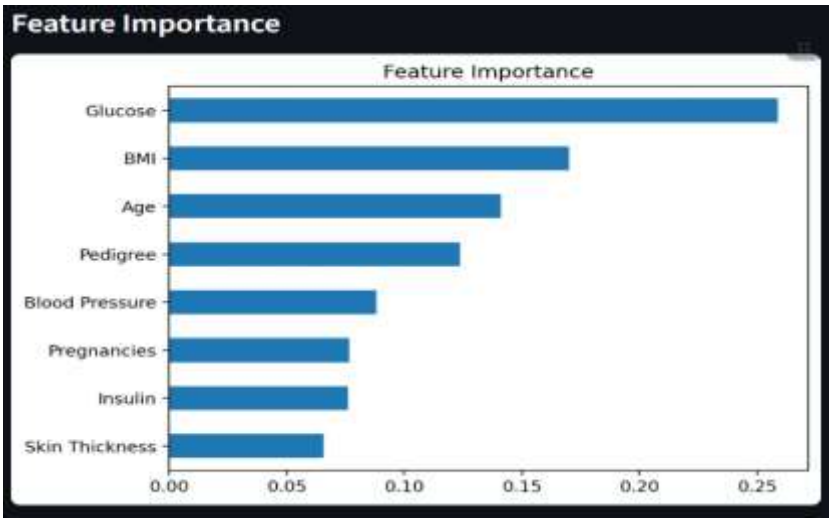


FIGURE 4. Relative importance of medical features in diabetes prediction

The developed streamlet application provided an interactive interface for prediction. Users can enter patient medical details such as glucose level, blood pressure, insulin value, BMI, and age through input fields. The application preprocesses the entered values using the saved Standard Scaler object and sends the transformed data to the trained machine learning model for prediction. The final result is then displayed as diabetic or non-diabetic prediction. The integration of machine learning with a web-based interface improved usability and demonstrated how predictive healthcare systems can be implemented in practical environments. The project successfully combined data analysis, machine learning, and visualization techniques to create a functional diabetes prediction system capable of performing real-time classification based on patient medical information.



FIGURE 5. User interface of the diabetes prediction web application

5. CONCLUSION

The diabetes prediction system developed in this project demonstrates the practical application of machine learning techniques in the healthcare domain. The project successfully utilized patient medical data to predict whether an individual is likely to be diabetic or non-diabetic based on various health-related parameters such as glucose level, blood pressure, insulin, BMI, age, pregnancies, and diabetes pedigree function. By analyzing historical patient records, the machine learning model was able to identify patterns associated with diabetes and generate prediction results for new patient inputs. The implementation process included dataset analysis, preprocessing, feature scaling, model training, evaluation, and deployment through a web-based interface. Exploratory data analysis helped in understanding dataset characteristics and identifying important feature distributions. The Random Forest Classifier algorithm was selected for prediction because of its strong performance in classification tasks and its capability to handle structured medical datasets efficiently. The trained model achieved satisfactory prediction accuracy and demonstrated the ability to classify diabetic and non-diabetic cases effectively. Evaluation techniques such as accuracy score, precision, recall, F1-score, and confusion matrix analysis provided detailed insight into the model's performance and prediction behavior. Feature importance analysis further showed that glucose level, BMI, and age had significant influence on diabetes prediction. The integration of the trained machine learning model with a stream lit-based web application made the system interactive and user-friendly. Users can enter patient medical information through the application interface and receive instant prediction results. This demonstrates how machine learning models can be combined with modern web technologies to create intelligent healthcare support systems. Although the developed system is intended mainly for educational and research purposes, it highlights the potential of machine learning in predictive healthcare applications. In future work, the project can be improved by using larger and more diverse medical datasets, applying advanced machine learning or deep learning algorithms, improving prediction accuracy, and deploying the system as an online healthcare application for broader accessibility. Overall, the project successfully demonstrates the use of machine learning for healthcare prediction and shows how intelligent systems can assist in early diabetes risk assessment using patient medical data.

REFERENCES

[1]. O. World Health Organization, Global Report on Diabetes, Geneva, Switzerland, 2016.
[2]. J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed., Morgan Kaufmann Publishers, 2011.
[3]. T. Mitchell, Machine Learning, McGraw-Hill Education, New York, USA, 1997.
[4]. L. Breiman, "Random Forests," Machine Learning Journal, vol. 45, no. 1, pp. 5–32, 2001.
[5]. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
[6]. W. McKinney, "Data Structures for Statistical Computing in Python," Proceedings of the 9th Python in Science Conference, pp. 51–56, 2010.
[7]. S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed., Packt Publishing, 2019.
[8]. I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning, MIT Press, Cambridge, USA, 2016.
[9]. UCI Machine Learning Repository, "Pima Indians Diabetes Dataset," University of California, Irvine. Available: UCI Machine Learning Repository
[10]. J. Brownlee, Machine Learning Mastery with Python, Machine Learning Mastery Publication, 2017.
[11]. A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd ed., O'Reilly Media, 2019.
[12]. Streamlit Documentation, "Streamlit: The Fastest Way to Build Data Apps in Python." Available: Streamlit Official Website
[13]. Python Software Foundation, "Python Programming Language Documentation." Available: Python Official Website
[14]. NumPy Developers, "NumPy Documentation." Available: NumPy Official Documentation
[15]. Pandas Documentation, "Python Data Analysis Library." Available: Pandas Official Documentation

Copyright@ REST Publisher