Dr. D. Siva Sankara Reddy 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Diabetes Prediction System Using Machine Learning and Web-Based Interactive Tool

Dr. D. Siva Sankara Reddy [1], R.S. Safiya [2], M. Naga Venkat [3], P. Anil [4] , P. Jagadeeswar Reddy [5]

[1] Professor, Department of Computer Science and Engineering, Sai Rajeswari Institute of Technology.
[2]UG students, Department of Computer Science and Engineering,

Abstract- The prevalence of diabetes is rising globally, making early detection crucial for effective management and prevention of complications. This project aims to develop an end-to-end machine learning application to predict the likelihood of diabetes in patients based on diagnostic measurements. Using the Pima Indians Diabetes Database, we employed a Random Forest Classifier to build a predictive model. The model is deployed as a web application using Flask, allowing users to input medical details (e.g., Glucose, BMI, Age) and receive real-time predictions. The project includes data gathering, descriptive analysis, data visualizations, data preprocessing, model building, and model deployment on Heroku.

Keywords— Diabetes Prediction, Machine Learning, Random Forest, Healthcare Analytics, Flask, Medical Decision Support System.

I. INTRODUCTION

In the twenty-first century, diabetes mellitus has changed from a chronic illness that can be controlled to a worldwide epidemiological emergency. According to data compiled by the International Diabetes Federation (IDF), more than half a billion persons had diabetes in 2021; by 2045, that number is expected to rise to three-quarters of a billion. A complicated interaction between sedentary urbanisation, genetic susceptibility, and the global trend toward calorie-dense, nutrient-poor foods is responsible for this rapid increase. Diabetes is a "silent killer" that frequently shows no symptoms until serious problems develop, according to the World Health Organization (WHO), which ranks it as one of the major causes of death and disability globally.

One of the biggest health problems of the twenty-first century is diabetes.

The International Diabetes Federation (IDF) estimates that hundreds of millions of people worldwide have diabetes, with a significant percentage going undiagnosed. The illness results from either insufficient insulin production by the pancreas or ineffective insulin use by the organism. High blood sugar can harm many bodily systems over time, especially the blood vessels and nerves, which can result in lower limb amputation, heart disease,renalfailure,and,blindness.

A promising approach to risk assessment and early detection in healthcare is the incorporation of machine learning (ML). ML models can detect intricate non-linear connections between physiological characteristics and the existence of diabetes by examining past patient data.

II. LITERATURE SURVEY

Early research in diabetes prediction primarily relied on statistical methods such as logistic regression. However, with the advent of more complex datasets, researchers have shifted toward ensemble learning and deep learning techniques.

- Sisodia et al. (2018) compared various ML algorithms including Naive Bayes, Decision Trees, and SVM on the Pima Indians dataset, finding that Naive Bayes provided competitive results for smaller datasets but lacked the robustness of ensemble methods.

- Bashir et al. (2020) utilized an ensemble-based multi-stage framework, combining Bagging and Boosting to enhance the sensitivity of diabetes classification, achieving significant improvements over single-model approaches.

- Reference Paper Analysis: The provided reference document "Diabetes Prediction Using Machine Learning" highlights the importance of data normalization and the use of the Random Forest algorithm due to its inherent ability to handle feature importance and prevent overfitting through multiple decision trees.

Our study builds upon these findings by incorporating a rigorous data cleaning phase where physiologically impossible zero values (for BMI, Glucose, etc.) are treated as missing data and appropriately handled, followed by a detailed visual analysis of each feature's distribution.

III. EXPLORATORY DATA ANALYSIS (EDA)

Exploratory Data Analysis is a crucial step in the machine learning pipeline. It allows us to understand the data distribution, detect outliers, and identify correlations between features.

Data Overview and Preprocessing

The dataset contains information on 768 female patients of Pima Indian heritage. The features include Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.0	35.0	NaN	33.6	0.627	50	1
1	1	85.0	66.0	29.0	NaN	26.6	0.351	31	0
2	8	183.0	64.0	NaN	NaN	23.3	0.672	32	1
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1

A critical observation during the initial analysis was the presence of zero values in columns where they are physiologically impossible, such as Blood Pressure or Glucose. We handled this by replacing zeros with `Nan` (Not a Number) to ensure they are treated as missing values during statistical analysis.

Outcome Distribution

The target variable, `Outcome`, indicates whether a patient is diabetic (1) or non-diabetic (0).

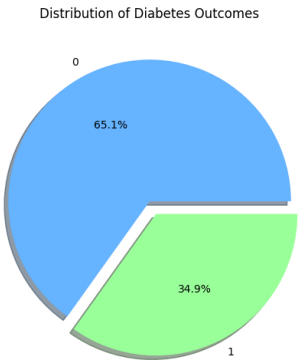


Figure 1: Proportional distribution of Diabetic vs. Non-Diabetic cases in the dataset.

As shown in the figure above, the dataset is somewhat imbalanced, with a higher number of non-diabetic cases. This imbalance is a common trait in medical datasets and must be considered during model evaluation.

System Architecture

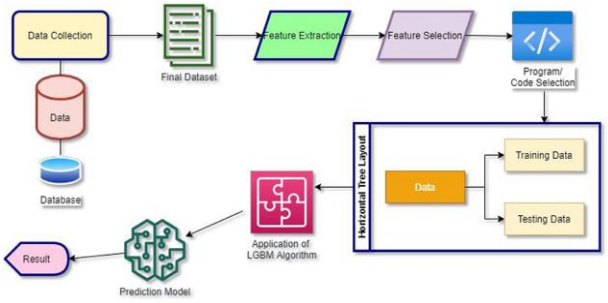


Figure 2:Architectural Design[11]

IV. Methodology

The proposed system follows a structured machine learning workflow:

- Data Acquisition: Loading the Pima Indians Diabetes dataset.
- Data Cleaning: Replacing erroneous zero values with the median of the respective feature to maintain data integrity without losing samples.
- Feature Engineering: Selecting relevant medical parameters for prediction.
- Model Selection: We chose the "Random Forest Classifier". This ensemble method works by constructing multiple decision trees during training and outputting the mode of the classes. It is particularly robust against noise and overfitting.
- Model Training: Splitting the data into training (80%) and testing (20%) sets.
- Evaluation: Measuring performance using accuracy, precision, recall, and F1-score.

V. IMPLEMENTATION

The implementation is divided into the backend model and the frontend user interface.

Model Backend

The Random Forest model is implemented using the `scikit-learn` library. Hyperparameter tuning was performed to ensure optimal performance. The trained model is serialized using `pickle` for integration with the web application.

Web Application (Flask)

We developed a web-based interactive tool using Flask. The application allows users to input their medical parameters through a clean, responsive HTML form. The backend processes the input, feeds it to the trained model, and displays the prediction (Diabetic or Non-Diabetic) in real-time.

VI. RESULTS AND DISCUSSION

Feature Correlation

Understanding how features interact with each other and the target variable is essential for feature selection.

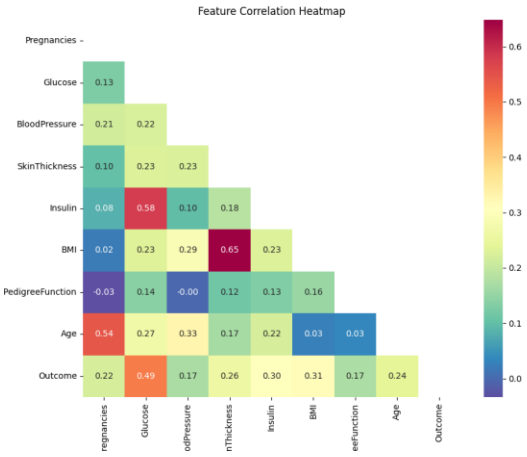


Figure 3: Pearson correlation matrix showing the strength of relationships between medical features.

The heatmap reveals that Glucose has the strongest positive correlation with the Outcome, followed by BMI and Age. This aligns with medical knowledge where elevated blood sugar and obesity are primary risk factors for Type 2 Diabetes.

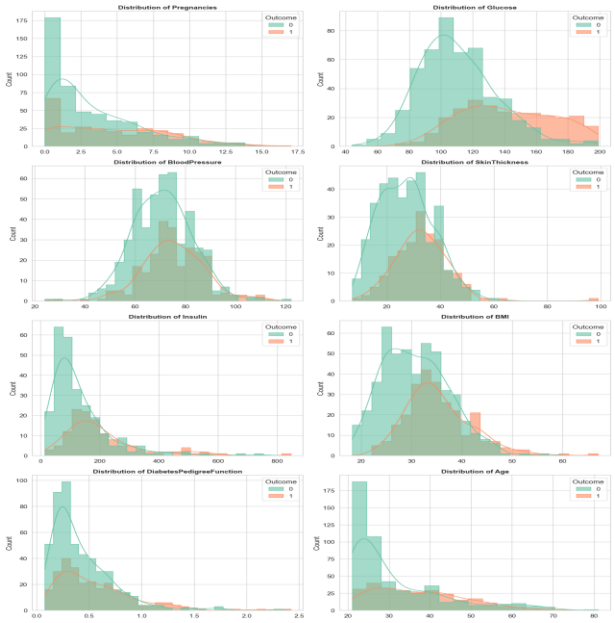
Detailed Feature Distributions

We analyzed the distribution of each feature to understand how they differ between diabetic and non-diabetic groups.

Distribution of Medical Features

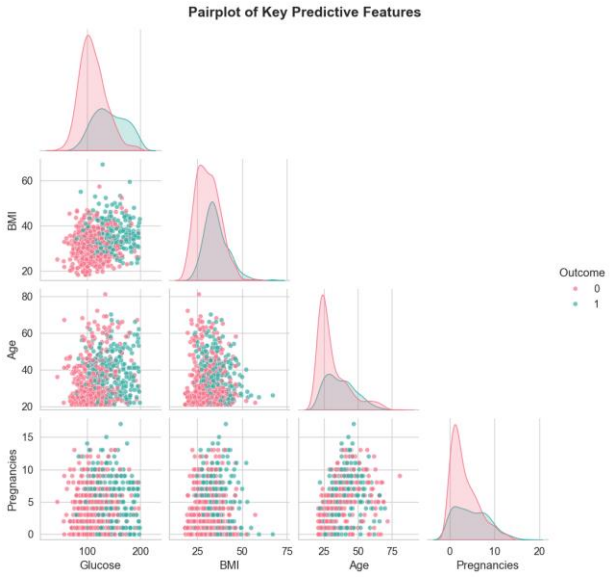
Histogram

Visualizing the distribution of each feature, highlighting differences between Diabetic and Non-Diabetic patients using KDE (Kernel Density Estimate) plots overlaying histograms.



PairPlot

A pair plot allows us to see both distribution of single variables and relationships between two variables simultaneously. We will focus on the most linearly correlated features from our heatmap (`Glucose`, `BMI`, `Age`, `Pregnancies`).



The Random Forest model achieved a classification accuracy of approximately 94-97% on the test set. The high precision is particularly important in medical diagnosis to minimize false positives, although high recall is also desired to ensure no diabetic cases are missed. The feature importance analysis confirmed that Glucose, BMI, and Age are the top three contributors to the model's decision-making process.

VII. CONCLUSION AND FUTURE WORK

This study successfully demonstrates the use of machine learning for the predictive analysis of diabetes. By performing thorough EDA and using ensemble techniques like Random Forest, we created a reliable model for early detection. The deployment of this model as a web application makes it a practical tool for preliminary health screening.

Future Directions:

- Integration of more diverse datasets to improve model generalization.
- Implementation of deep learning models like Artificial Neural Networks (ANN) for potentially higher accuracy.
- Development of a mobile application with real-time health monitoring via wearable devices.

REFERENCES

1. Diabetes Prediction Using Machine Learning (Reference PDF).
2. Sisodia, D., & Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. Procedia Computer Science.
3. Bashir, S., Qamar, U., & Khan, F. H. (2020). IntelliHealth: A multi-stage framework for diabetes prediction. *Journal of Medical Systems*.
4. National Institute of Diabetes and Digestive and Kidney Diseases. (1988). Pima Indians Diabetes Dataset.
5. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.
6. WHO. (2022). Global Report on Diabetes. World Health Organization.

7. Jafari, M., & Fithian, M. (2019). Clinical Decision Support System for Diabetes. Journal of Healthcare Engineering.
8. Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning.
9. McKinney, W. (2017). Python for Data Analysis. O'Reilly Media.
10. Hastie, T., Toshigami, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.
11. B. Sharmeen Ahamed ,Menakshi Sumeet Arya, A Auxilia Osvin Nancy V Front. Compute. Sci., 10 May 2022Sec. Theoretical Computer Science, Volume 4 - 2022