ISSN : 2350-0743

International Journal of Recent Advances in Multidisciplinary Research

Vol. 12, Issue 05, pp.11198-11200, May, 2025

RESEARCH ARTICLE

A MACHINE LEARNING APPROACH TO EARLY DETECTION OF DIABETES

[1*]Sadhu Vasavi, [1]Simhadhati Lokesh, [1]Sampath Kumar, K.L.P and [2]Krishna Mohan, G.S.S.S.S.V.

[1]B. Tech, Department of ECE, Aditya institute Of Technology and Management
[2]Associate Professor, Ph. D, Department of ECE, Aditya institute Of Technology and Management

ARTICLE INFO

Article History
Received 20th February, 2025
Received in revised form
27th March, 2025
Accepted 26th April, 2025
Published online 30th May, 2025

Keywords:

Health Indicators, Machine Learning, Diabetes Prediction Health Metrics, Random Forest, Data Prepossessing, and Feature Engineering, Risk factors, early detection, predictive modeling.

Corresponding author: Sadhu Vasavi,

ABSTRACT

The goal of this research is to use machine learning and health indicator data to create a prediction model for diabetes diagnosis. The study finds important patterns linked to diabetes by examining a data set that includes a variety of health metrics, including BMI, blood pressure, cholesterol levels, lifestyle factors (such as smoking, physical activity, and food), and socioeconomic characteristics (such as income and education). To guarantee model robustness, the data pipeline include prepossessing procedures including feature scaling, encoding, and managing class imbalances. The model is constructed using sophisticated algorithms like Random Forest and logistic regression which are assessed using measures like accuracy, recall. With its effective, data-driven approach to early diabetes identification, this technology enhances preventive care tactics for at-risk groups and gives medical professionals the ability to make well-informed judgments.

Citation: Sadhu Vasavi, Simhadhati Lokesh, Sampath Kumar, K.L.P and Krishna Mohan, G.S.S.S.S.V. 2025. "A Machine Learning Approach to Early Detection of Diabetes.". *International Journal of Recent Advances in Multidisciplinary Research,* 12, (05), 11198-11200.

INTRODUCTION

The goal of this research is to use machine learning and health indicator data to create a prediction model for diabetes diagnosis. The study finds important patterns linked to diabetes by examining a data set that includes a variety of health metrics, including BMI, blood pressure, cholesterol levels, lifestyle factors (such as smoking, physical activity, and food), and socioeconomic characteristics (such as income and education). The model is constructed using sophisticated algorithms like Random Forest and Gradient Boosting, which are assessed using measures like accuracy, recall.

Literature survey: Several research studies have been conducted in recent years focusing on diabetes prediction using machine learning techniques. This section compares existing work in terms of models used, features considered, deployment mechanisms, and practical applicability—and contrasts it with the approach taken in this project. Modak and Jha (2023) applied multiple machine learning classifiers to predict diabetes and found Random Forest to perform better than SVM and K-Nearest Neighbors.. (2021) dealt with the issue of imbalanced healthcare datasets and proposed a framework for improving classification performance. While our dataset (BRFSS) was already balanced (50-50 class split), we still applied scaling and cleaning techniques to enhance model reliability. Logistic Regression is often chosen for interpretability, as shown in Shah and Patel (2021), but it may underperform in handling feature interactions. Our comparative model evaluation shows this trade-off clearly, with Random Forest offering higher recall and F1-score, which is crucial in identifying actual diabetic cases. In conclusion, our project matches or exceeds existing work in terms of accuracy, feature selection, and usability.

METHODOLOGY

Problem Statement: Diabetes is a rapidly growing chronic condition affecting millions of individuals worldwide, often going undiagnosed until serious complications arise. Traditional diagnostic methods depend on clinical tests and medical supervision, which may not be readily accessible, especially in rural or resource-limited regions. There is a pressing need for a non-invasive, fast, and accessible tool that can predict an individual's risk of developing diabetes using easily available health and lifestyle indicators. This project

aims to address this issue by developing a machine learning-based prediction system integrated with a user-friendly web interface, allowing users to input personal health data and receive real-time risk assessments.

Aim of The Project

Develope a Machine Learning Model to Predict Diabetes: The target of this project is to use machine learning algorithms to forecast a person's risk of developing diabetes based on a number of health indicators such as age, BMI, blood pressure, cholesterol, physical activity, and other lifestyle factors. The objective is to develop a system that can recognize patterns linked to diabetes risk by training the model on past health data.

Process's Analyze health data: A critical phase in the project is data prepossessing, which makes sure the raw data is cleansed and formatted appropriately for model training. In this procedure, missing values will be handled, categorical variables will be encoded, numerical features will be normalized and any possible outliers will be found and eliminated. In order to glean valuable insights from the data set and maybe enhance model performance, feature engineering will also be carried out. In order to better understand the major factors influencing diabetes risk, data exploration and visualization tools will be used to find trends and correlations within the data.

Model Performance: Several metrics, including accuracy, precision, recall, and F1-score, will be used to assess the model's efficacy following training. Cross- validation will be used as part of the evaluation process to make sure the model performs effectively when applied to fresh, untested data. In order to find possible areas for model improvement, an analysis of the false positives and false negatives will also be carried out. The final objective is to guarantee that the model has low error rates for both diabetic and non-diabetic predictions, in addition to being accurate and dependable.

Provide Early Detection and Preventative Insights: The project's goal is to create a tool that can give people and Healthcare professionals early warning signs of possible diabetes development so they can take preventative action. The algorithm will be able to forecast a person's risk of developing diabetes after it has been integrated into an intuitive user interface. Through early identification, the system may be able to lower the prevalence of diabetes, which might greatly enhance public health outcomes and lower medical expenses related to later-stage diabetes management.

Data Acquisition: Finding a data set with significant health markers related to diabetes prediction is part of the data collecting process for this"diabetes_binary_ 5050split_ health_indicators_BRFSS2015. csv," a publically accessible data set, was selected for this research. The BRFSS is a useful tool for health-related prediction models since it gathers health-related data from individuals all around the United States.

Features of the Data set: Numerous variables in the data set are essential for estimating the risk of developing diabetes. Important characteristics include:The target variable that indicates if a person has diabetes (1) or not (0) is called diabetes_binary.High-BP: Phys-activity: Indicates if the person is physically active (1 for yes, 0 for no).Fruits: Indicate if the person often eats fruits (1 for yes, 0 for no).Veggies: Indicate if the person routinely eats veggies 1)for yes, 0 for no).Heavy Alcohol Consumption: Indicates whether the person drinks a lot of alcohol (1 for yes, 0 for no).Any Healthcare: Indicates whether the person has access to medical treatment (1 denotes yes, 0 denotes no).

Pre processing: Before training we should preprocess the machine learning models, several data pre processing steps were carried out to ensure the dataset was clean, consistent, and suitable for analysis. These steps improved the performance and accuracy of the models and prepared the data for real-time prediction use.

Handling Missing Values: We began by checking the dataset for missing or null entries. Upon inspection, we found that the BRFSS 2015 dataset version we used was already cleaned and did not contain any missing values. This allowed us to proceed directly to the next steps without imputation.

Encoding Categorical Variables: Several features in the dataset were categorical in nature, such as “Yes”/“No” responses. These values were converted into binary numerical representations using label encoding, where “Yes” was mapped to 1 and “No” to 0. This transformation enabled the models to process categorical data effectively. dominating the model during training.

Splitting The Dataset: Once the data was encoded and normalized, we split it into two subsets: 80% for training and 20% for testing. This split allowed us to evaluate model performance on unseen data and ensured that our results were generalizable.

Final Data Validation: After preprocessing, we verified the structure, format, and consistency of all features. We also confirmed that the target variable (Diabetes_binary) remained balanced between the two classes (1 = diabetic, 0 = non-diabetic), maintaining the integrity of the dataset for training.

Model Training: After preprocessing the dataset, we moved forward with training machine learning models to classify whether an individual is at risk of diabetes. Our goal was to develop a system that could accurately predict diabetes using only basic health and lifestyle inputs, without requiring complex medical tests.

CHOICE OF ALGORITHMS

We selected two popular and well-established classification algorithms: **Logistic Regression** and **Random Forest Classifier**. Logistic Regression is a simple, linear algorithm that works well for binary classification problems like ours (diabetic or non-diabetic). It helps understand the direct impact of each feature on the outcome.On the other hand, Random Forest is an ensemble learning method that combines the results of multiple decision trees to produce a more accurate and stable prediction. It handles both linear and non-linear data well and automatically deals with interactions between features.

Data Splitting: After pre process the data set the some amount od data will undergo training

and some amount will undergo for testing The training data was used to help the models "learn" patterns in the health features associated with diabetes. We used Python's Scikit-learn library to implement and train the models. At first, we used default settings for both models to get a baseline performance. Later, we experimented with simple parameter tuning to improve performance—such as increasing the number of trees in Random Forest and adjusting regularization in Logistic Regression

Testing

Evaluation on Testing Data: After training the models using 80% of the dataset, we tested them on the remaining 20% to assess performance on unseen data. We measured accuracy**,** precision**,** recall**,** and F1-score**.** The Logistic Regression model achieved an accuracy of 72.3%**.** The Random Forest model showed improved accuracy at 73.2%**,** along with better recall, making it more effective at identifying individuals at risk. These results indicated that the Random Forest model generalized well and avoided over fitting, making it suitable for deployment.

Realistic Predictions On Sample Inputs: To validate the behavior of the model in practice, we used the trained Random Forest model to predict outcomes for various realistic user inputs: Inputs with multiple risk factors (e.g., high blood pressure, high BMI, stroke history) were consistently predicted as "At Risk." Inputs with healthy profiles (e.g., normal BMI, active lifestyle, no chronic conditions) returned "Not at Risk." This confirmed that the model responded appropriately to different risk combinations.

Integration Testing With Stream lit: The trained Random Forest model was integrated into the Stream lit app. We tested how the app handled real-time predictions when the user submitted health details: The backend successfully transformed inputs into model-ready format. Predictions were generated instantly and displayed clearly. The interface functioned smoothly without crashes or delays.

End-To-End System Validation: We tested the entire workflow—from launching the app in VS Code, entering user data, triggering the prediction, and receiving the result. This validated that:

The model was properly loaded from the .pkl file. The application worked consistently across multiple test cases. The interface and model interaction were fully operational and accurate.

RESULT AND DISCUSSION

The basic prediction model, created with Sci kit-learn, is housed in the machine learning layer. Algorithms like Random Forest and logistic regression Classifier are trained, assessed, and deployed in this layer. In order to improve deployment efficiency, it also enables model serialization using Job lib. Stream lit powers the presentation layer, which acts as the user's front-end interface.

Through an interactive web interface, it enables users to enter health data including blood pressure, BMI, and glucose levels. These inputs are processed by the Stream-lit-managed backend, which then communicates with the machine learning model to produce predictions in real time. By integrating these layers, a unified system that meets the demands of both individuals and healthcare professionals is guaranteed to be dependable and easy to use.

CONCLUSION

The diabetes prediction system shows promise as a dependable and effective tool for early diagnosis by effectively utilizing machine learning techniques to identify those at high risk of developing diabetes. Users may enter health factors and get immediate forecasts with ease thanks to the system's integration of strong data prepossessing, streamlined algorithms, and an interactive Stream lit-based user interface. Even though the results are encouraging, issues like generalization and data set bias point to areas that could use further development. All things considered, the system is a big step toward easily available, tech-driven solutions for proactive diabetic care and illness prevention.

REFERENCES

Patel A. R. and B. Desai, "Smart Healthcare Prediction System using Random Forest Classifier, *"International Journal of Innovative Research in Computer and Communication Engineering***,** vol. 8, no. 4, pp. 892–899, 2020.

Hussain A. and S. Jabin, "A Machine Learning Model for Early Detection of Diabetes, *"International Journal of Engineering Research & Technology* **(**IJERT**)**, vol. 9, no. 6, pp. 450–454, 2020.

Shrivastava P. and N. Sinha, "Effective Diabetes Prediction Using XGBoost and Feature Engineering, "*International Journal of Computer Applications*, vol. 182, no. 12, pp. 20–24, 2019.

Roy K. and R. Sinha, "Prediction of Diabetes Using Logistic Regression and Decision Tree, "*International Journal of Scientific & Technology Research*, vol. 9, no. 2, pp. 4870–4873, 2020.

Agarwal R. and D. Sharma, "Web-Based Application for Diabetes Risk Prediction Using Streamlit, "*International Research Journal of Engineering and Technology* (IRJET)**,** vol. 7, no. 10, pp. 304–309, 2021.

Ahmed F. and S. Gupta, "Ensemble Learning Models for Predicting Diabetes Risk Using Health Data, "*Journal of Computer Science and Applications*, vol. 11, no. 1, pp. 55–60, 2020.

Rao, B. and M. Rani, "Machine Learning-Based Health Prediction System Using Streamlit Framework, "*International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 9, no. 5, pp. 175–180, 2020.

Jain A. and K. Choudhary, "Early Stage Diabetes Risk Assessment Using SVM and Gradient Boosting, "*International Journal of Recent Technology and Engineering* (IJRTE)**,** vol. 8, no. 4, pp. 2306–2310, 2019.

****