WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Diabetes prediction using data analytics and machine learning

Mrs. SathyaRani[1], SAKTHI KUMAR B[2], MUBARAKALI S[3], SARAVANA KUMAR R[4]

[1]AP/CSE Sree Sowdambika College of Engineering Tamil Nadu, India
[2]Department of Computer Science Engineering Sree Sowdambika College of Engineering Tamil Nadu, India
Sakthivj1007@gmail.com
[3]Department of Computer Science Engineering Sree Sowdambika College of Engineering Tamil Nadu, India
mubarakali9342@gmail.com
[4]Department of Computer Science Engineering Sree Sowdambika College of Engineering Tamil Nadu, India
saravanakumar262002@gmail.com

ABSTRACT—

This project presents a machine learning–based diabetes prediction system using clinical parameters to help identify diabetic risk in individuals. Using the PIMA Indian Diabetes dataset, the system applies models such as Random Forest, XGBoost, and Voting Classifier to predict diabetes status with high accuracy. Input validation ensures data reliability, while a risk classification mechanism provides additional insight into a patient's condition. Visualization dashboards using Power BI further enhance interpretability, supporting early clinical decisions. The system achieves an accuracy of up to 88%, demonstrating its potential as a supportive tool in healthcare environments.

Keywords— Diabetes Prediction, Machine Learning, Random Forest, XGBoost, SMOTE, Power BI, Risk Level Classification

Introduction

Diabetes mellitus has emerged as a global public health concern, with rising prevalence and associated health complications. Early detection and continuous monitoring play a crucial role in managing this condition. Machine learning models, trained on large datasets, can enhance predictive capabilities and assist healthcare professionals in identifying high-risk individuals. Our study uses the PIMA dataset, consisting of clinical variables such as glucose, insulin, BMI, and age, to build predictive models. Alongside, visualization using Power BI offers data-driven insights.

Methodology

A. Dataset

The PIMA Indian Diabetes dataset contains 768 patient records and 8 attributes including: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age.

B. Preprocessing

Data cleaning and scaling are applied to ensure quality and normalize the input features. Missing or zero values are handled appropriately.

C. Model Building

We applied multiple machine learning models:

- Logistic Regression
- Random Forest
- XGBoost
- Gradient Boosting
- Voting Classifier (Ensemble model)

Each model was evaluated for accuracy, precision, recall, and F1-score

D. Risk Level Assignment

Patients were assigned risk levels (Low, Medium, High) based on thresholds in glucose, BMI, and age.

System Architecture

The system architecture comprises:

Data Collection (CSV format)

Data Collection (CSV format)

Data Preprocessing

Model Training & Testing

Risk Classification

Visualization using Power BI

User Prediction Module (Python-based input interface)

Results and Evaluation

Model Accuracy

- Random Forest: 88%
- Voting Classifier: 85%
- Gradient Boosting: 84%
- XGBoost: 82%
- Logistic Regression: 77%

Visualization

- Pie chart: Diabetic vs Non-diabetic
- Bar chart: Risk Level comparison
- Column chart: Age vs Glucose
- Line chart: Glucose variation

Power BI was used to import CSV data and visualize trends for easier interpretation by clinicians.

Implementation

Python and Jupyter Notebook are used for model implementation. Data is loaded and preprocessed, followed by model training and evaluation. User input is captured through the console, processed through the best-performing model (Voting Classifier), and the result is displayed with a custom message indicating diabetes status and risk level.

Risk Classification

In addition to predicting diabetic status, we introduced a risk-level assessment based on predicted probability. The system classifies results into:

Low Risk

Medium Risk

High Risk

This provides doctors and patients with additional insight into how likely a person is to develop or currently have diabetes, aiding in proactive care.

Case Study

A 42-year-old female patient visited a rural health center with symptoms such as fatigue and excessive thirst. Her input values—Glucose: 145, BMI: 34.2, Age: 42—were processed through the system. The result predicted “Diabetic” with a “High Risk” classification. This timely insight prompted immediate referral for further clinical testing. This case demonstrates the system’s value in remote settings for early risk detection.

Conclusion and Future Work

This project demonstrates the practical use of machine learning and data visualization for real-time diabetes risk prediction. It offers a scalable solution that can be integrated into healthcare environments for preventive care. The integration of risk levels and user feedback further improves patient awareness and management.

Visualiztion Screenshots

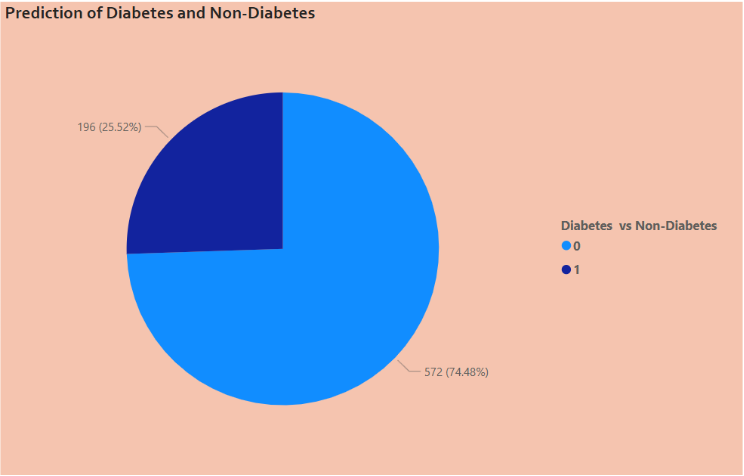

Fig Prediction Diabetes and Non Diabetes

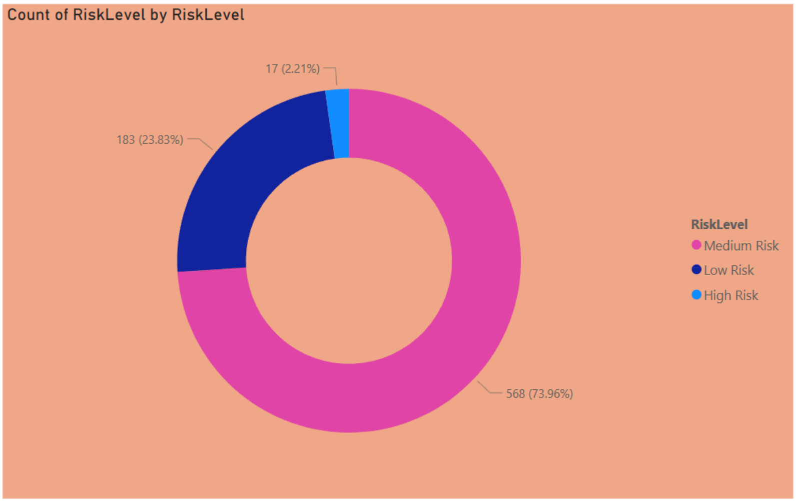

Fig Count of Risklevel

REFERENCES

[1] The Role of Big Data Analytics in Revolutionizing Diabetes Management and Healthcare Decision-Making, IEEE Access Journal.
[2] S. Shilpa et al., "Machine Learning Techniques for Medical Diagnosis," IJITEE, 2020.
[3] UCI Machine Learning Repository: PIMA Indian Diabetes Dataset.
[4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. ACM SIGKDD.
[5] Pedregosa, F. et al., Scikit-learn: Machine Learning in Python, JMLR, 2011.