International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 43-49 March 2026-Sept 2026 www.csjournals.com

Identification of Critical Lifestyle and Demographic Risk Factors for Diabetes Using Feature Selection and Predictive Modelling

Vishakha[1], Dr. vishal verma[2], Bajrang[3]

[1]MCA Student, [2]Professor, [3]Asst. Prof.
Deptt. of Comp. Sc. and Application, Chaudhary Ranbir Singh University, Jind, Haryana, India

Abstract

Diabetes is among the common long-term diseases globally and causes a major challenge to the health of the population. Early prediction for diabetes can help healthcare professionals well timed treatment and reduce the risk of serious health conditions. This research suggests a ML-based model to identify critical lifestyle as well as demographic risk factors associated with diabetes using feature selection techniques and predictive modelling. The Kaggle Diabetes Health Indicators dataset containing 253,680 records and 22 features was used for analysis. The dataset was pre-processed using train-test splitting, feature scaling, and SMOTE technique to address class imbalance. Feature importance analysis was applied to determine the most most risk indicators element for impacting diabetes prediction. Three ML algorithms, namely Logistic Regression, Random Forest, and Gradient boosting, were used and compared. Results demonstrate that XGBoost achieved the highest prediction accuracy of **84.53%**, Random Forest with **83.93%**, and Logistic Regression achieved **64.19%**. The study recognizes BMI, Age Group, Increase Blood Pressure, High Cholesterol, General Health as most influential predictors of diabetes.

Keywords— Diabetes Prediction, Machine Learning, Feature Selection, Random Forest, XGBoost, BRFSS Data set, Healthcare Analytics.

Introduction

Diabetes mellitus is a long-term metabolic disorder defined by higher blood glucose levels due lack of insulin production. It has turned into one of the fastest-growing global health issues because of its long-term problems, such as heart disease, kidney failure, nerve damage. As reported by international health reports, a large number of people around the world deal with diabetes, and this number is expected to increase strongly.

Early identification of person at high risk of diabetes plays an important role in reducing disease complications and improving healthcare outcomes. Traditional diagnosis generally depends on laboratory investigations. Still , many persons remain undiagnosed during the early stages of the disease, leading to late treatment. Therefore, predictive models are capable of identifying diabetes risk before critical symptoms appear have becoming more important.

ML has resulted as an effective approach for disease prediction because it can identify hidden relationships within large healthcare datasets and provide high -accuracy classification results. Healthcare datasets usually contain numerous variables, many of which may be irrelevant or redundant. The existence of unnecessary features increases computational complexity and may reduce prediction performance.Feature selection techniques address solves this issue by identifying the most useful feature that contribute strongly for prediction accuracy. Selecting relevant features not only improves classification performance but also increases model interpretability.

Literature Review

S.NO	Authors Name	Objective	Method	Key Findings
1	Kakoly et. Al.	To improve diabetes prediction based on relevant features.	ML with Feature Selection	Feature selection reduced computational complexity and improved prediction accuracy by identifying important diabetes risk factors.
2	Bhat et al.	To identify the early-stage diabetes using lifestyle-related factors.	Machine Learning	Lifestyle related factors like diet, physical activity, smoking, overweight significantly influence diabetes risk and support early intervention.
3	Authors in [3]	To make predictions using demographic characteristics.	Machine Learning	Age, gender, education, and income effectively supported early diabetes risk assessment before clinical symptoms appeared
4	Authors in [4]	To compare classification algorithms and feature selection methods	Multiple ML Classifiers with Feature Selection	Feature selection improved accuracy, reduced dimensionality, and enhanced interpretability. Ensemble methods achieved the best performance.
5	Authors in [5]	To identify the most important diabetes risk factors	Inclusive Feature Selection Strategy	BMI, age, blood pressure, cholesterol, and physical activity were identified as the strongest predictors.
6	Authors in [6]	To evaluate ML algorithms for diabetes prediction.	Random Forest, XGBoost, as	Ensemble methods, especially Random Forest and XGBoost, outperformed traditional

			well as other ML techniques.	algorithms in prediction accuracy.
7	Authors in [7]	To develop a lifestyle-based diabetes prediction framework.	Machine Learning	Obesity, physical inactivity, unhealthy diet, smoking, and alcohol consumption were effective predictors of diabetes.
8	Authors in [8]	To improve diabetes risk assessment through preprocessing and feature engineering	Data Preprocessing, Feature Selection, and Machine Learning	Proper preprocessing and feature engineering significantly improved classification performance.
9	Authors in [9]	To improve diabetes prediction using ensemble learning.	Framework (Ensemble Learning + Feature Fusion	Ensemble learning with feature fusion achieved better prediction accuracy than individual models
10	Authors in [10]	To develop a correlation-based diabetes prediction algorithm.	Correlation Feature Selection based on ML, Deep Learning	Correlation-based feature selection removed redundant attributes, and deep learning achieved high prediction accuracy

Methodology

Collection of Data

The dataset was obtained from the Kaggle which has 253,680 records with 22 health-related attributes and one target variable. It consists of demographic, lifestyle, and medical features like BMI, Increase Blood Pressure, Raised Cholesterol, Physical Activity, Smoking, Age, Education, Income, General Health.

Preprocessing of Data

The dataset is pre-processed by using EDA to improve data quality. The input attribute is separated from the target variable. The dataset was split into training and testing sets.

Feature Selection

It is used to performed identify the extremely important features for impacting diabetes prediction. Selected features were used for model training, and the feature importance ranking as shown in **Fig. 1.**

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 43-49 March 2026-Sept 2026 www.csjournals.com

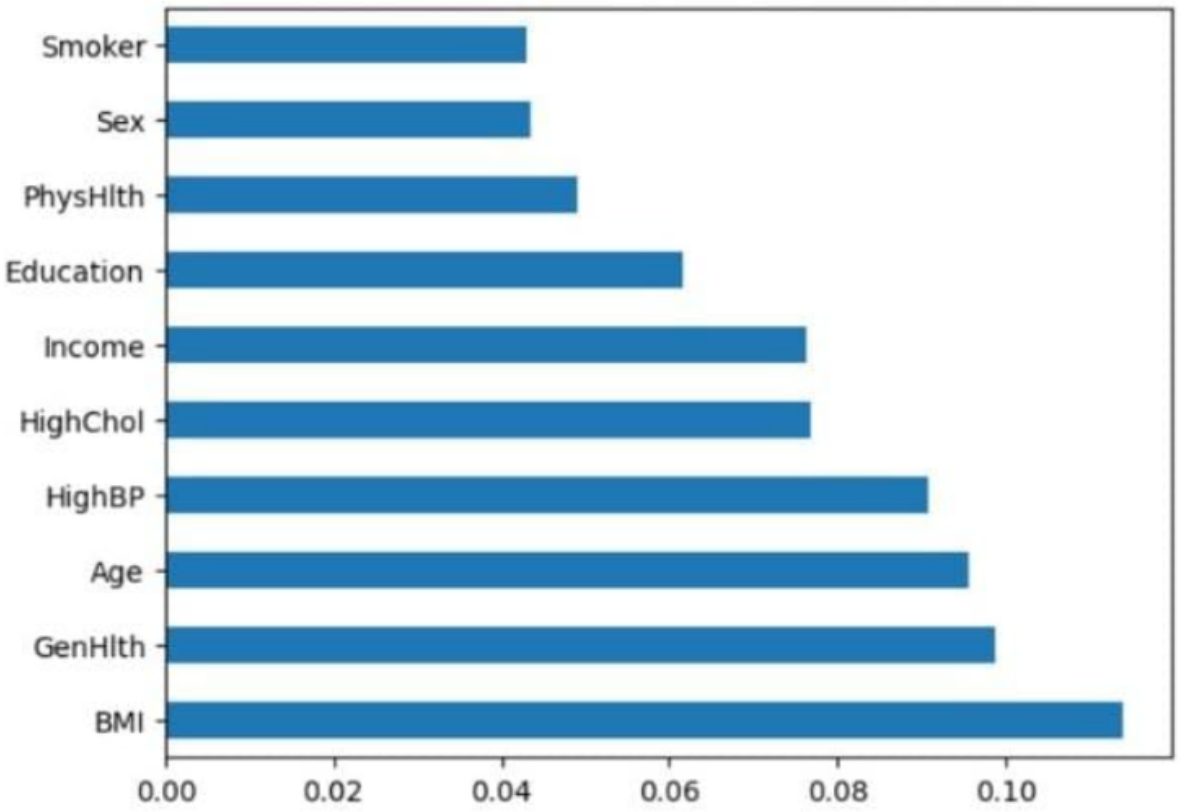


Fig 1. Top 10 Important Features

Class Imbalance Correction

The dataset contained an imbalanced distribution of diabetes classes. SMOTE was used to balancing these classes.

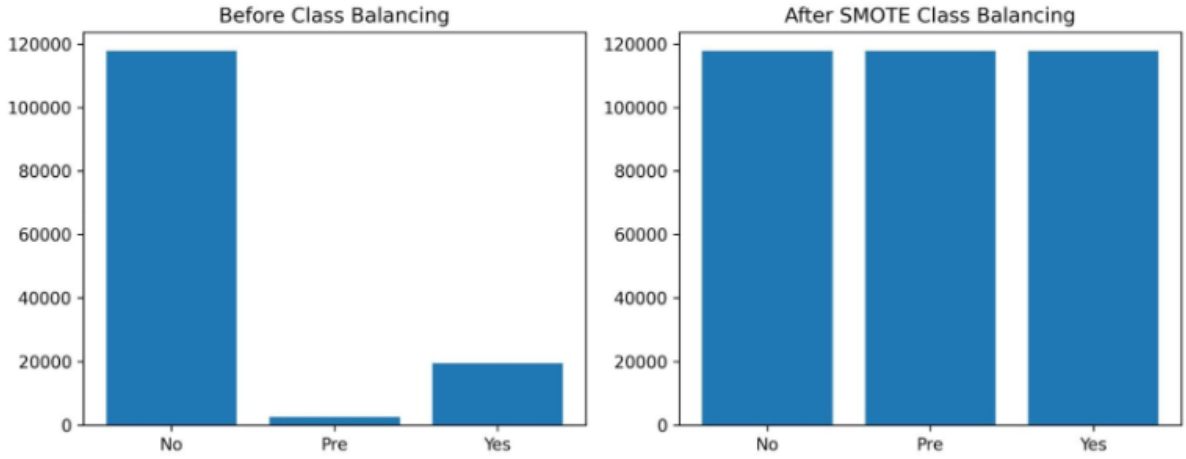


Fig 2. Class Imbalance Correction

ML Models Used:

There are three types of ML models were implemented:

1) Logistic Regression

2) Random Forest

3) XGBoost

Evaluation Metrics

This is used to check the performance of model. Different type of techniques is used:

Accuracy

Precision

Recall

F1-Score

Results

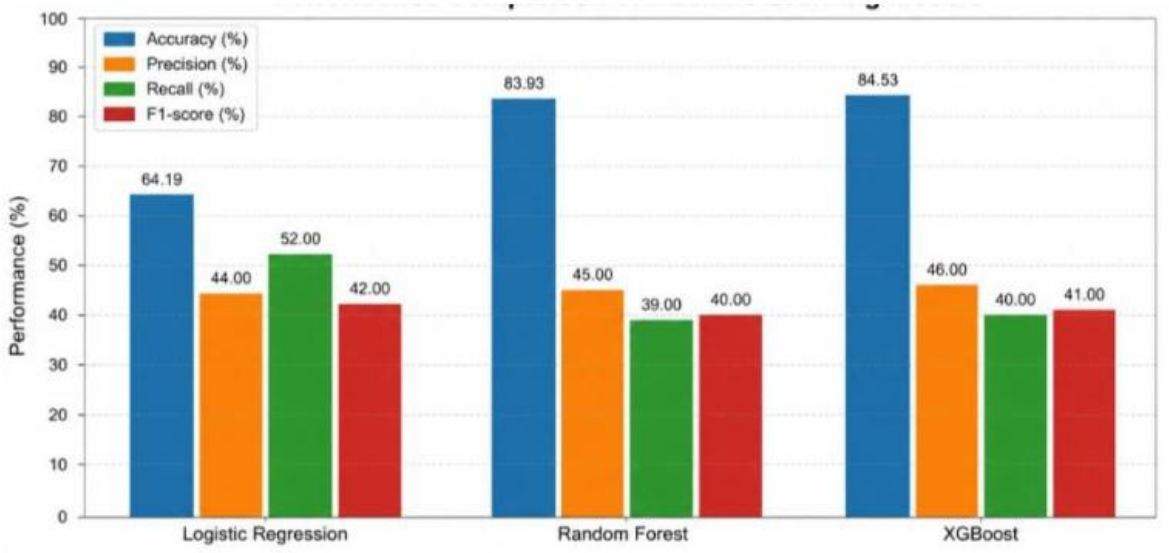
Three ML models, LR, RF, and Gradient Boost, were successfully implemented to this dataset used to predict diabetes risk. This dataset was prepared using data cleaning, feature scaling, train-test splitting, and SMOTE for class balancing.

The performance of the models was analyzed using Accuracy, Precision, Recall, and F1-score. By comparing three machine learning algorithms, XGBoost predict the highest accuracy (84.53%) , Random Forest predicts(83.93%) while Logistic Regression achieved (64.19%).

Performance Comparison:

Model	Accuracy (%)	Precision	Recall	F1-Score
Random Forest	83.93	0.45	0.39	0.40
Logistic Regression	64.19	0.44	0.52	0.42
XGBoost	84.53	0.46	0.40	0.41

The observe results indicate that **XGBoost** is the most efficient model for diabetes prediction in this study.



Discussion

By comparing these ML algorithms: LR, RF and XGBoost in all evaluation metrics XGBoost recorded the highest prediction accuracy (84.53%).

The results shows that XGBoost and Random Forest, are more capable in handling complex relationships among lifestyle, demographic, and health-related factors. The use of feature selection can increase the efficiency of the prediction models. These outcomes suggest that both lifestyle habits and demographic characteristics have a strong impact on diabetes risk.

Future Scope

This latest research uses ML algorithms such as (Logistic Regression, Random Forest, and XGBoost. In future studies , various type of deep learning techniques, like ANN, CNN, can help in order to improve prediction performance and identify more complex patterns in healthcare data.

References

[1] J. Kakoly, M. R. Hoque, and N. Hasan, “Data-driven diabetes risk factor prediction using machine learning algorithms with feature selection technique.”

[2] S. S. Bhat, V. Selvam, and G. A. Ansari, “Predicting lifestyle of early diabetes mellitus using machine learning technique.”

[3]“Machine Learning-Based Approach for Predicting Diabetes Employing Socio-Demographic Characteristics,” 2023.

[4] “Analyzing Classification and Feature Selection Strategies for Diabetes Prediction Across Diverse Diabetes Datasets,” 2024.

[5] "Critical Factor Analysis for Prediction of Diabetes Mellitus Using an Inclusive Feature Selection Strategy," 2024.

[6] "Predictive Modeling for the Development of Diabetes Mellitus Using Key Factors in Various Machine Learning Approaches," 2024.

[7] "Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type," 2022.

[8] "A Risk Assessment and Prediction Framework for Diabetes Mellitus Using Machine Learning Algorithms," 2023.

[9] "KFPredict: An Ensemble Learning Prediction Framework for Diabetes Based on Fusion of Key Features," 2023.

[10] "An Effective Correlation-Based Data Modeling Framework for Automatic Diabetes Prediction Using Machine and Deep Learning Techniques," 2023.