Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD28522

A Comparative Study of Logistic Regression and Machine Learning Models for Diabetes Prediction Using the BRFSS Dataset

Yuchen Zhang

ZJUI Institute, Zhejiang University, Hangzhou, China
yuchen.22@intl.zju.edu.cn

Abstract. Diabetes mellitus, particularly type 2 diabetes, is a growing global health challenge with significant clinical and economic consequences. Accurate early prediction of diabetes risk may facilitate prevention and timely intervention. In this study, we conducted a comparative analysis of logistic regression (LR), random forest (RF), and Extreme Gradient Boosting (XGBoost) models for diabetes prediction using the Behavioral Risk Factor Surveillance System (BRFSS) 2015 dataset (N = 253,680). The dataset included demographic, lifestyle, and clinical variables. Models were evaluated on accuracy, area under the receiver operating characteristic curve (ROC AUC), and precision–recall AUC (PR AUC). All models demonstrated strong discriminative performance (ROC AUC = 0.82–0.83), with XGBoost slightly outperforming LR and RF. Feature importance analysis revealed that general health status, high blood pressure, body mass index, cholesterol level, and age were consistent predictors across models. While ML models provided higher predictive performance, logistic regression retained value in interpretability. These findings underscore the potential of integrating machine learning into diabetes risk assessment while maintaining clinical interpretability.

Keywords: Diabetes, Logistic Regression, BRFSS

1. Introduction

Diabetes is a chronic metabolic disorder that affects over 463 million adults worldwide, and this number is likely to rise to 578 million by 2030 [1]. The majority of cases are type 2 diabetes, which develops gradually and is associated with major complications such as cardiovascular disease, kidney failure, neuropathy, and premature mortality [2]. The increasing prevalence of diabetes shows significant clinical and economic burdens, highlighting the urgent need for effective preventive strategies.

For many years, traditional statistical models, particularly logistic regression (LR), have been widely applied for diabetes risk prediction. LR offers simplicity and interpretability, making it attractive in clinical contexts where understanding the influence of each predictor is important [3,4]. However, LR assumes linearity and independence among predictors, which may limit its ability to capture complex relationships in real-world health data. To address these limitations, machine

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.LD28522

learning (ML) methods such as random forest (RF) and Extreme Gradient Boosting (XGBoost) have been increasingly adopted [5–7]. These approaches usually have great performances in handling high-dimensional datasets and nonlinear interactions, often achieving higher predictive accuracy than traditional models.

Several recent studies have demonstrated the value of ML for diabetes prediction. For example, ensemble models have been used to improve the classification of type 2 diabetes using electronic health records, while boosting algorithms have been shown to outperform regression-based approaches in population surveys [8–10]. Despite these advances, questions remain regarding the trade-off between interpretability and predictive performance. A direct comparison between traditional LR and modern ML models using large, representative datasets can therefore provide valuable insights.

The present study aims to evaluate and compare LR, RF, and XGBoost models for predicting diabetes risk in U.S. adults using the Behavioral Risk Factor Surveillance System (BRFSS) dataset. Beyond assessing predictive performance, this research also examines the relative importance of demographic, lifestyle, and clinical risk factors, thereby contributing to the evidence base on diabetes risk prediction and its implications for both clinical practice and public health policy.

2. Methodology

2.1. Data source

This study utilized data from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), a nationwide health-related telephone survey conducted annually by the Centers for Disease Control and Prevention (CDC) [11]. The BRFSS collects information on health behaviors, chronic health conditions, and preventive service use from adults across all 50 U.S. states and territories. The 2015 cycle included responses from over 400,000 adults. After applying exclusion criteria and cleaning procedures, a total of 253,680 participants were included in the final analytic sample. Because the BRFSS is an observational survey based on self-reported data, the study design was retrospective and cross-sectional, and the analysis focused on associations and predictive modeling rather than causal inference.

2.2. Variables

The outcome variable of interest was diabetes status. Participants were asked whether a health professional had ever diagnosed them with diabetes. Individuals who answered “yes” were coded as diabetic, while those who answered “no” were coded as non-diabetic. Cases of gestational diabetes were excluded. Due to the nature of self-reported survey data, the variable did not distinguish between type 1 and type 2 diabetes, and both were included under the same outcome category.

A broad set of predictor variables was included to capture demographic, clinical, and lifestyle-related risk factors. Demographic variables consisted of age, sex, race or ethnicity, education level, and household income. Age was categorized into four groups: 18–34, 35–49, 50–64, and 65 years or older. Clinical risk factors included body mass index (BMI), self-reported history of high blood pressure, high cholesterol, stroke, and heart disease or heart attack. Lifestyle factors included smoking status, alcohol consumption, physical activity, and dietary habits, specifically fruit and vegetable intake. General health status was measured using a five-point self-rated scale ranging from excellent to poor. In addition, the number of days with poor mental health in the past month was included as an indicator of mental health burden. Continuous variables such as BMI and age were

DOI: 10.54254/2755-2721/2025.LD28522

standardized when appropriate, while categorical predictors were converted into dummy variables for analysis.

2.3. Data preprocessing

Data preprocessing was conducted prior to model training. Variables with more than 20% missing data were excluded from the analysis. For predictors with less than 5% missingness, imputation was performed using the median for continuous variables and the mode for categorical variables. Implausible values, such as BMI less than 12 or greater than 70, were treated as invalid and removed. Duplicate records were also excluded. Since the prevalence of diabetes in the dataset was 13.9%, class imbalance was an important consideration. To address this issue, logistic regression was fitted with class weights, and both Random Forest and XGBoost incorporated their built-in mechanisms for handling imbalanced data. Finally, the dataset was randomly split into a training set comprising 70% of the data and a testing set comprising 30%. All preprocessing was carried out using only the training data to prevent information leakage.

2.4. Predictive models

Three predictive models were developed and compared. First, logistic regression served as the baseline statistical approach. It was fitted using a logit link function with diabetes as the binary outcome, and all predictors were entered simultaneously. Multicollinearity was assessed with variance inflation factors, and variables with a variance inflation factor greater than 10 were excluded. Odds ratios and 95% confidence intervals were reported for interpretability. Second, the Random Forest model was constructed as an ensemble of decision trees using bootstrap sampling. At each split, a random subset of predictors was considered to minimize correlation between trees. Hyperparameters, including the number of trees and maximum tree depth, were tuned using cross-validation. Feature importance was evaluated using the mean decrease in Gini impurity. Third, XGBoost, an implementation of gradient boosting, was trained by building sequential decision trees to minimize classification error. Hyperparameters such as learning rate, maximum depth, and number of estimators were optimized through grid search with five-fold cross-validation. Feature importance in XGBoost was assessed using gain values, and Shapley Additive Explanations (SHAP) values were used in sensitivity analyses to provide interpretability.

2.5. Statistical analysis

Model performance was evaluated on the testing dataset. Accuracy was reported as the proportion of correctly classified cases. Discrimination was assessed using the area under the receiver operating characteristic curve (ROC AUC). Because of the imbalanced nature of the data, the area under the precision–recall curve (PR AUC) was also used to provide insight into model sensitivity and positive predictive value. Calibration, or the agreement between predicted probabilities and observed outcomes, was evaluated using calibration curves and Brier scores. Lower Brier scores indicated better calibration. To account for sampling variability, 95% confidence intervals for AUC values and calibration metrics were obtained using 1,000 bootstrap resamples.

Finally, statistical inference differed between models. For logistic regression, statistical significance of predictors was evaluated using Wald tests, with p-values below 0.05 considered significant. In contrast, Random Forest and XGBoost were evaluated primarily through predictive accuracy and feature importance rankings rather than hypothesis testing. Comparisons of model

DOI: 10.54254/2755-2721/2025.LD28522

performance were made descriptively, emphasizing both predictive strength and clinical relevance rather than relying solely on statistical significance. All data preprocessing, model development, and statistical analyses were conducted using Python version 3.10 and R (version 4.5). Data cleaning and model implementation were primarily carried out in Python with the scikit-learn and xgboost libraries, while supplementary statistical testing and visualization were performed in R using packages such as tidyverse and ggplot2.

3. Results

3.1. Sample characteristics

The final analytic dataset consisted of 253,680 adults who participated in the 2015 BRFSS survey. Among these, 35,346 individuals (13.93%) reported having diabetes, while 218,334 (86.07%) did not. The relatively low prevalence of diabetes in this dataset underscores the importance of using evaluation metrics that are robust to class imbalance, such as the precision–recall area under the curve (PR AUC). Participants with diabetes were, on average, older and had higher rates of high blood pressure, high cholesterol, and obesity compared with participants without diabetes. In contrast, higher education and income levels were more common among non-diabetic respondents.

3.2. Model performance: discrimination

Discriminative ability was assessed using receiver operating characteristic (ROC) curves. Logistic regression achieved an ROC AUC of 0.82, indicating a strong ability to distinguish between individuals with and without diabetes. Random Forest achieved a nearly identical ROC AUC of 0.82, while XGBoost yielded the highest discrimination with an ROC AUC of 0.83. Although the absolute differences between models were modest, the consistent improvement of XGBoost over both logistic regression and Random Forest suggests incremental gains in predictive accuracy when using gradient boosting methods.

Figure 1 presents the ROC curves for all three models, illustrating that their performance was comparable, with XGBoost showing a slightly superior curve.

DOI: 10.54254/2755-2721/2025.LD28522

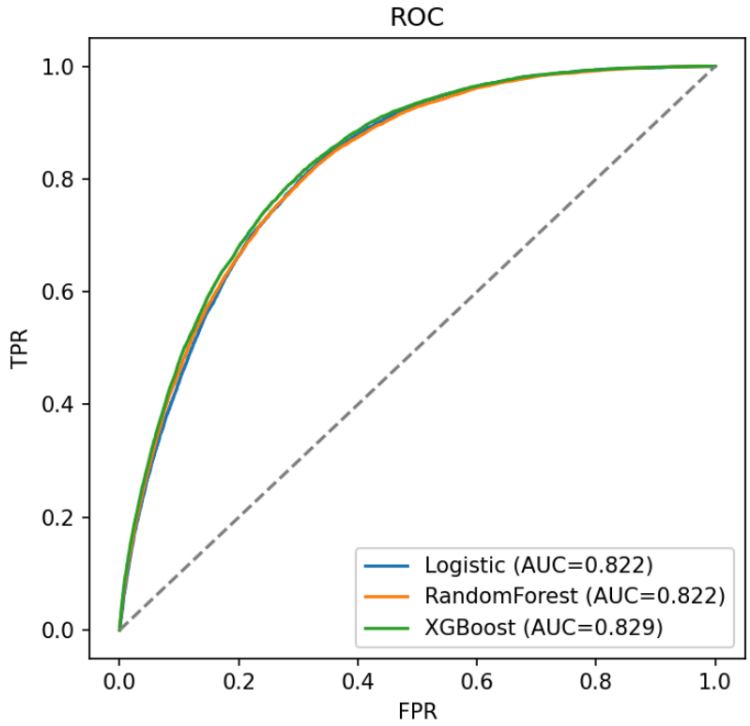

Figure 1. ROC curves

3.3. Model performance: precision–recall analysis

Because of the class imbalance in the dataset, ROC AUC alone does not fully reflect performance in predicting the minority class (diabetes). Therefore, precision–recall (PR) curves were also evaluated. Logistic regression achieved a PR AUC of 0.40, Random Forest achieved 0.42, and XGBoost achieved the highest at 0.43. These values indicate that, while all models performed reasonably well, identifying true positive cases of diabetes remains challenging due to the relatively small proportion of positive cases in the dataset.

Figure 2 shows the PR curves, highlighting that XGBoost consistently outperformed the other two models across most recall thresholds.

DOI: 10.54254/2755-2721/2025.LD28522

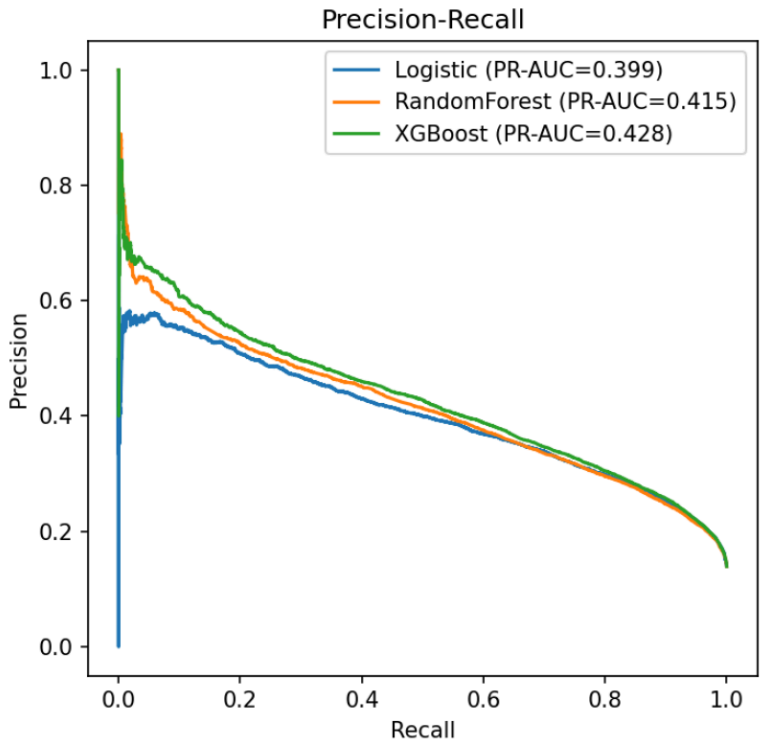

Figure 2. PR curves

3.4. Feature importance and risk factors

Logistic regression results indicated several strong predictors of diabetes. Higher odds of diabetes were observed among individuals with high blood pressure (odds ratio [OR] = 1.46), high cholesterol (OR = 1.33), poor general health (OR = 1.76), higher BMI (OR = 1.50), and older age (OR = 1.46). Conversely, protective associations were observed for younger age groups (18–34 years, OR = 0.089), heavy alcohol consumption (OR = 0.84), higher income (OR = 0.90), and greater educational attainment (OR = 0.97).

Random Forest and XGBoost provided complementary insights through feature importance rankings. Both models consistently highlighted general health, BMI, high blood pressure, cholesterol, and age as the most important predictors of diabetes risk. While logistic regression quantifies associations in terms of odds ratios, machine learning models emphasized relative importance, reinforcing the robustness of these predictors across different analytical frameworks.

Figure 3 and Figure 4 present the feature importance rankings from Random Forest and XGBoost, illustrating the dominance of health-related variables such as BMI and general health status.

DOI: 10.54254/2755-2721/2025.LD28522

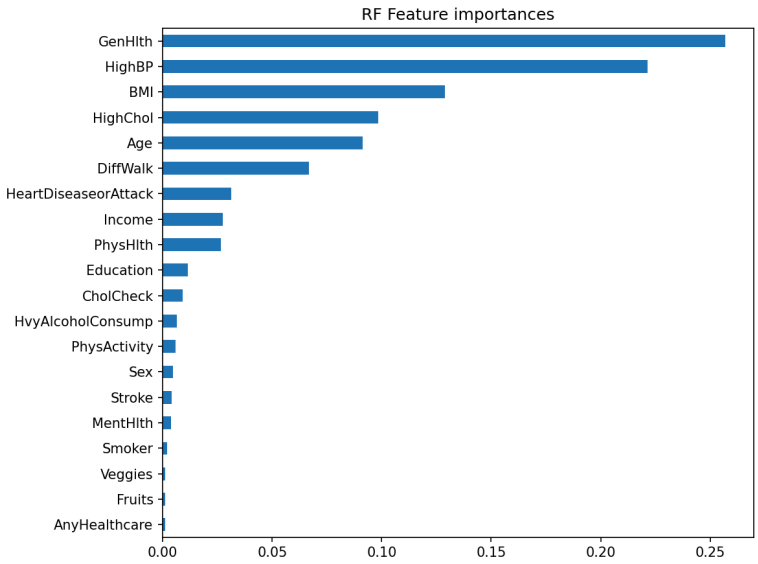

Figure 3. Random forest feature importance plots

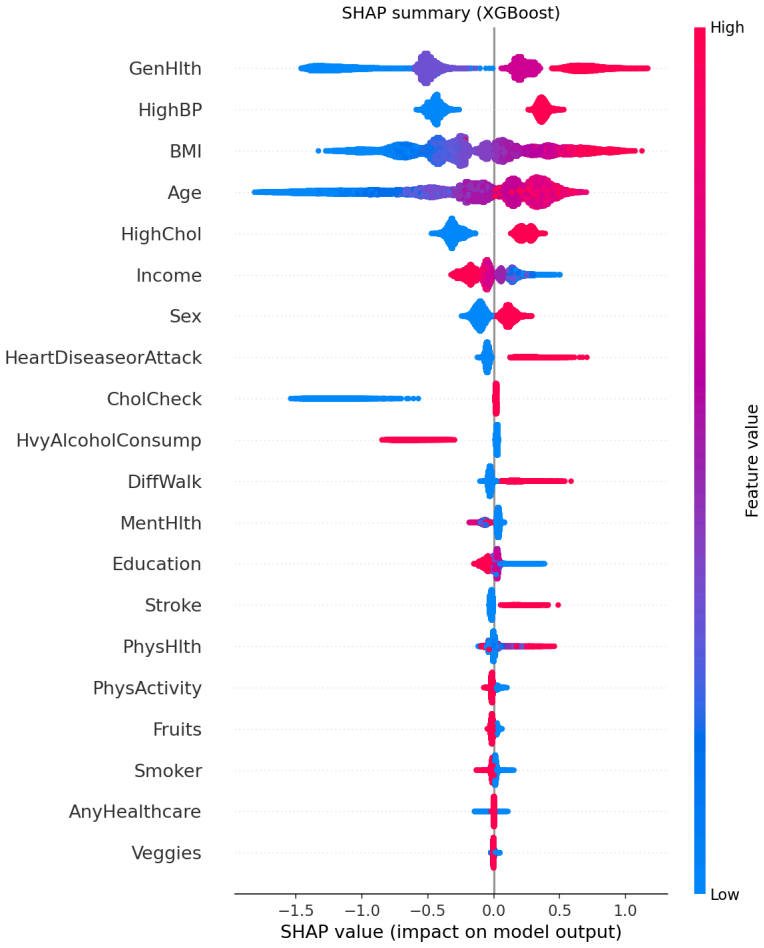

Figure 4. SHAP summary in XGBoost

DOI: 10.54254/2755-2721/2025.LD28522

3.5. Model calibration

Calibration analysis was conducted to evaluate the extent to which predicted probabilities matched observed outcomes. Logistic regression, Random Forest, and XGBoost all demonstrated reasonable calibration, with Brier scores of 0.100, 0.098, and 0.097, respectively. The calibration curve for XGBoost was closest to the diagonal reference line, suggesting that it provided the most reliable probability estimates. Logistic regression tended to underestimate risk at higher probability thresholds, while Random Forest slightly overestimated risk in the upper range.

Figure 5 shows the calibration curves, where the predicted probabilities for XGBoost aligned most closely with actual diabetes prevalence, followed by Random Forest and then logistic regression.

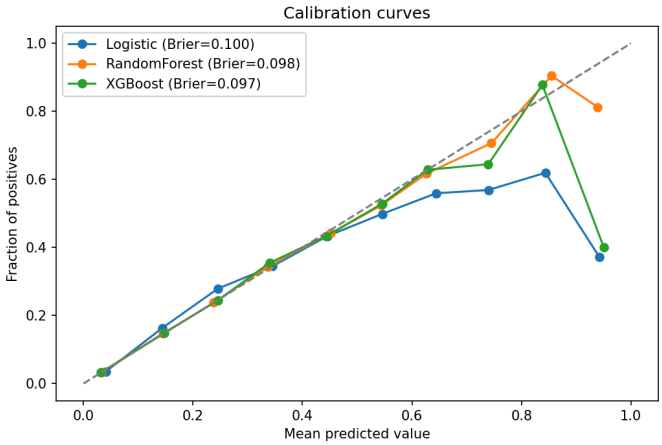

Figure 5. Calibration curves

3.6. Summary of findings

In summary, all three models achieved strong predictive performance in discriminating diabetes status in a large, nationally representative sample. XGBoost consistently outperformed both logistic regression and Random Forest, though the differences were relatively small. As shown in Table 1, logistic regression provided interpretable odds ratios, identifying classical risk factors such as age, BMI, high blood pressure, and general health. Machine learning models reinforced these findings through importance rankings and offered slightly improved calibration and discrimination.

Table 1. Some logistic regression results for diabetes risk factors

	coef	std_err	z	p-value	OR	[0.025	0.975]
GenHlth	0.5646	0.010	54.44	<0.0001	1.76	0.5443	0.585
BMI	0.4059	0.007	57.01	<0.0001	1.50	0.392	0.420
Age	0.3809	0.010	37.25	<0.0001	1.46	0.360	0.401
HighChol	0.2885	0.008	35.87	<0.0001	1.33	0.273	0.304
Income	-0.1046	0.009	-11.82	<0.0001	0.90	-0.122	-0.087
PhysActivity	-0.023	0.007	-3.12	0.0018	0.98	-0.376	-0.008
HighBP	0.375	0.009	42.90	<0.0001	1.45	0.358	0.392

DOI: 10.54254/2755-2721/2025.LD28522

4. Conclusion

This study conducted a comparative analysis of logistic regression, random forest, and XGBoost models for predicting diabetes risk using a large, nationally representative dataset from the Behavioral Risk Factor Surveillance System. All three models demonstrated strong discriminative performance, with area under the ROC curve values exceeding 0.80. Among them, XGBoost showed slightly higher performance than logistic regression and random forest, although the improvement was modest.

The observed findings make sense in the context of established knowledge. Our models consistently identified general health status, body mass index, high blood pressure and age as the strongest predictors of diabetes. These factors are well-documented in epidemiological studies and clinical practice, reinforcing the robustness of our analysis.

Despite the strengths of this study, several limitations should be acknowledged. The analysis relied on self-reported data, which may introduce reporting bias. The dataset was limited to U.S. adults, restricting the generalizability of findings to other populations. In addition, the class imbalance, with only 14% of participants reporting diabetes, contributed to relatively low precision–recall performance.

Future research should explore the external validation of models in diverse populations, the use of longitudinal datasets to support causal inference, and the incorporation of interpretability tools such as SHAP values to bridge the gap between predictive accuracy and clinical understanding. Overall, this study highlights the potential of integrating machine learning methods into diabetes surveillance while preserving the interpretability of traditional approaches, ultimately supporting more effective prevention and early intervention strategies.

References

[1] World Health Organization. (2024). Diabetes. https: //www.who.int/news-room/fact-sheets/detail/diabetes
[2] Cho, N. H., et al. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence. Diabetes Research and Clinical Practice, 138, 271–281.
[3] Joshi, R. D. (2021). Predicting type 2 diabetes using logistic regression and decision tree algorithms. Journal of Diabetes Science and Technology, 15(4), 755–763.
[4] Lynam, A. L., & Sutherland, S. (2020). Logistic regression vs machine learning in diabetes classification. npj Digital Medicine, 3, 1–8.
[5] Shin, J. (2022). Development of diabetes prediction models using health screening test parameters. Diabetes & Metabolism Journal, 46(6), 752–763.
[6] Xie, Z., et al. (2019). Building risk prediction models for type 2 diabetes using machine learning. Preventing Chronic Disease, 16, E109.
[7] Li, W. (2024). GA-XGBoost for diabetes prediction. npj Digital Medicine, 7, 1–10.
[8] Kavakiotis, I., et al. (2017). Machine learning and data mining approaches in diabetes research. Computational and Structural Biotechnology Journal, 15, 104–116.
[9] Liang, W., et al. (2019). Development of a machine learning model for diabetes prediction in community settings. Scientific Reports, 9, 1–9.
[10] Wang, Y., et al. (2021). Ensemble machine learning models for predicting type 2 diabetes in adults. Journal of Biomedical Informatics, 115, 103694.
[11] Teboul, A. (n.d.). Diabetes Health Indicators Dataset [Data set]. Kaggle. Retrieved August 31, 2025, from https: //www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset