Proceedings of ICEGEE 2025 Symposium: Sensor Technology and Multimodal Data Analysis
DOI: 10.54254/2753-8818/2025.AU24516

Comparison of Diabetes Risk Prediction Models: A Comparative Analysis of Logistic Regression, Random Forests, and Decision Trees

Jingyi Luo

College of Agronomy and Biotechnology, Southwest University, Chongqing, China
jingyiluo@hes.edu.kg

Abstract. Diabetes is a growing global public health concern, projected to affect over 700 million people by 2045. Accurate risk prediction models are essential for early detection and prevention. This study compares three machine learning models—logistic regression, random forest, and decision tree—using a large-scale health survey dataset from Kaggle. To ensure model robustness, stratified sampling and class balancing were applied, and 14 significant predictors (e.g., body mass index, age, hypertension) were identified using least absolute shrinkage and selection operator (LASSO) regression. Model performance was evaluated using accuracy, recall, F1-score (harmonic mean of precision and recall), and receiver operating characteristic–area under the curve (ROC–AUC). Logistic regression achieved the best balance of accuracy (74%), interpretability, and generalization. Random forest captured complex nonlinear patterns but showed signs of overfitting. Decision trees were less stable and less accurate. Findings reveal key trade-offs in complexity versus interpretability in population-level diabetes risk modeling.

Keywords: Diabetes risk prediction, LASSO regression, feature selection, machine learning

1. Introduction

Diabetes mellitus is a group of metabolic diseases characterized by chronic hyperglycemia, primarily including type 1 (T1DM) and type 2 diabetes mellitus (T2DM), with T2DM accounting for over 90% of cases globally [1]. According to the International Diabetes Federation (IDF) Global Overview (8th edition), the number of people with diabetes is projected to reach 700 million by 2045 [2]. As a high-prevalence country, China faces increasing diabetes rates, imposing significant pressure on its healthcare system and economy [3].

The growing prevalence of diabetes mellitus, particularly type 2 diabetes (T2DM), highlights the urgent need for effective risk prediction tools to support early screening, prevention, and personalized intervention [4]. Accurate identification of high-risk individuals plays a vital role in reducing long-term complications and alleviating the healthcare burden on a population scale. As such, predictive modeling has become a cornerstone in diabetes research and public health management.

DOI: 10.54254/2753-8818/2025.AU24516

Recent research has demonstrated the advantages of combining statistical methods with machine learning techniques. For example, Li constructed a SHAP-enhanced machine learning model for diabetes prediction and found that it significantly improved classification performance and interpretability compared to traditional models [5]. Similarly, Wang et al. applied LASSO regression and random forest to identify key diabetes risk factors and achieved higher prediction accuracy than conventional methods, underscoring the importance of robust feature selection strategies [6].

These studies illustrate how integrating interpretable models such as logistic regression with advanced algorithms like random forests can enhance prediction performance in structured health data [7]. As both accuracy and transparency are essential in medical decision-making, this study aims to contribute to the growing body of research that seeks to balance these priorities in population-level diabetes risk assessment [8].

Based on this background, this study applies biostatistical and machine learning methods to explore key risk factors for diabetes, aiming to provide scientific support for public health strategies and the development of precision medicine and personalized health management.

2. Method

2.1. Data sources and description

The dataset used in this study is from Kaggle, a publicly available health survey dataset that is primarily used to study risk factors for the development of diabetes [9]. The dataset covers several variables that may influence the occurrence of diabetes, including individual health status, lifestyle, and socioeconomic factors [10]. The dataset has a total of 253,680 individual records and contains 22 variables, of which the diabetes binary variable serves as the dependent variable, indicating whether an individual has diabetes or not (0 = not having diabetes, 1 = having diabetes). This dataset was compiled from data from the Behavioral Risk Factor Surveillance System (BRFSS, Behavioral Risk Factor Surveillance System) published by the Centers for Disease Control and Prevention (CDC) in the U.S. The data is highly representative and reliable.

2.2. Selection and description of indicators

To reduce computational burden while ensuring sample representativeness, this study first performed stratified random sampling on the original dataset, which contains 253,680 individual records. By maintaining consistency in the distribution of key variables such as diabetes status, age, gender, and education level, a moderately sized and representative sample was ultimately obtained.

Given that, in health survey data, the proportion of individuals with diabetes is typically much lower than that of healthy individuals, data balancing was further performed to prevent class imbalance from affecting model training. Referring to the strategy used in the Kaggle dataset employed in this study, the sample was adjusted to achieve a 1:1 ratio between diabetic and non-diabetic individuals. This approach effectively avoided the model's bias toward the majority class during training and improved the robustness and predictive reliability of the classification model.

For feature selection, this study adopted the Least Absolute Shrinkage and Selection Operator (LASSO) regression method. By introducing an L1 regularization term, LASSO automatically eliminates redundant variables or those with severe multicollinearity, thereby enhancing the interpretability and stability of the model. Specifically, after standardizing 21 independent variables, the optimal regularization strength was determined through five-fold cross-validation, and 14 key variables with significant impacts on diabetes prediction were ultimately retained.

DOI: 10.54254/2753-8818/2025.AU24516

2.3. Introduction to the methodology

2.3.1. Data pre-processing

In data preprocessing, initial cleaning was first performed, including removing duplicate values, treating missing values (using mean padding or deleting variables with excessive missing rates), and identifying outliers to improve data quality and model stability.

2.3.2. Feature selection

To filter out the key variables affecting diabetes and to reduce redundant information, this study used several methods for feature selection. First, highly correlated variables were excluded by calculating the correlation matrix between variables to reduce the problem of multicollinearity. Second, analysis of variance (ANOVA) was used to assess the association of continuous variables (e.g., BMI, age) with diabetes, and the significance of categorical variables (e.g., smoking, exercise) with diabetes was assessed using the chi-square test. Subsequently, key variables were further screened using LASSO regression (L1 regularization), which automatically shrinks the coefficients of some variables to zero using a penalty term, thus screening out variables that contribute less to the model.

$$r_{XY} = \frac{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{\sqrt{\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2}\sqrt{\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^2}} \tag{1}$$

where $\bar{X}$ and $\bar{Y}$ are the means of X and Y respectively. All pairwise correlations r_{XY} form the correlation matrix R. Highly correlated variables (e.g., $|r_{XY}|$ above a chosen threshold) were considered to be exhibiting multicollinearity; one of each such pair was removed to avoid redundancy and instability in the model coefficients.

Second, to assess the association of individual features with the outcome, it employed statistical tests. For continuous variables (e.g., BMI, age), a one-way Analysis of Variance (ANOVA) was performed to assess if the mean values differed significantly between the diabetes and non-diabetes groups. ANOVA computes an F-statistic.

2.3.3. Modeling analysis

Various statistical models and machine learning algorithms were used to analyze the risk factors for diabetes. First, Logistic Regression was used as a baseline model to assess the effect of different variables on the prevalence of diabetes. Second, a Decision Tree was constructed to explore the nonlinear relationships between the variables. A Random Forest was used to improve the predictive power and calculate the variables' importance ranking to assist in feature selection.

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n)}} \tag{2}$$

Where $P(Y=1|X)$ represents the probability that the sample belongs to class 1, $X_1, X_2, \ldots, X_n$ are the feature variables, $\beta_0, \beta_1, \ldots, \beta_n$ are the model coefficients, and e is the base of the natural logarithm.

DOI: 10.54254/2753-8818/2025.AU24516

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(X) \tag{3}$$

Where T is the number of decision trees, $f_t(X)$ is the prediction of the t-th tree, and $\hat{y}$ is the final prediction result.

2.3.4. Model evaluation

To evaluate the model's classification performance, the ROC curve (Receiver Operating Characteristic Curve) and the AUC value (Area Under the Curve) were used to measure its differentiation ability. Meanwhile, Accuracy, Precision, Recall, and F1-score were calculated to evaluate the model performance comprehensively.

3. Results

This study applied logistic regression, random forest, and decision tree models to predict diabetes risk, with a primary focus on evaluating their performance on the test set. By comparing key metrics such as accuracy, recall, and F1-score, it assessed the models' predictive capabilities, generalizability, and classification balance.

3.1. Logistic regression

Table 1: Test set model evaluation results

Tasks	Accuracy	Recall rate	f1-score	Sample size
0	0.76	0.71	0.74	1021
1	0.72	0.77	0.74	979
Accuracy			0.74	2000
Average value	0.74	0.74	0.74	2000
Average (combined)	0.74	0.74	0.74	2000

As shown in Table 1, the logistic regression model achieved the most stable performance, with an overall accuracy of 74% on the test set. It demonstrated a balanced ability to identify both diabetic and non-diabetic individuals, with recall rates of 0.71 and 0.77 respectively, and an F1-score of 0.74 for both classes. This consistency indicates strong generalization without overfitting. Moreover, logistic regression offers clear interpretability through its regression coefficients, making it a practical and reliable tool for public health risk prediction and factor identification.

3.2. Random forest

Table 2: Test set model evaluation results

Tasks	Accuracy	Recall rate	f1-score	Sample size
0	0.74	0.68	0.71	1021
1	0.69	0.75	0.72	979
Accuracy			0.71	2000
Average value	0.72	0.71	0.71	2000
Average (combined)	0.72	0.71	0.71	2000

DOI: 10.54254/2753-8818/2025.AU24516

In comparison, the random forest model achieved a slightly lower accuracy of 71% (Table 2), with a tendency toward classification imbalance. It showed stronger performance in detecting diabetic cases (recall = 0.75) but weaker performance for non-diabetic individuals (recall = 0.68). Although random forests are capable of modeling complex nonlinear relationships and interactions between variables, the results suggest a mild degree of overfitting, likely due to a lack of parameter tuning. Therefore, additional optimization, such as tree depth limitation or regularization, may be required to improve its generalizability.

3.3. Decision tree

Table 3: Test the set model evaluation results

Tasks	Accuracy	Recall rate	f1-score	Sample size
0	0.65	0.63	0.64	1021
1	0.62	0.64	0.63	979
Accuracy			0.64	2000
Average value	0.64	0.64	0.64	2000
Average (combined)	0.64	0.64	0.64	2000

The decision tree model, as shown in Table 3, performed the weakest among the three, with a test accuracy of only 64%. Its recall rates for both classes were below 0.65, and F1-scores did not exceed 0.64. These results suggest poor ability to extract essential features and indicate limited predictive capacity. The poor generalization is likely due to the model's overly complex structure caused by the absence of effective pruning, resulting in overfitting to training data and poor performance on unseen samples.

In summary, logistic regression demonstrated the best stability and interpretability, making it a solid baseline model in structured health data contexts. Random forest showed promise in handling complex feature interactions but requires careful tuning to avoid overfitting. Decision trees, while intuitive and interpretable, lacked sufficient generalization and are better suited for use within ensemble frameworks to enhance predictive robustness.

4. Discussion

4.1. Accuracy

Logistic regression and random forest are both around 74% and 72% on the test set, with very little difference; decision tree is only around 64%, which is significantly behind. This shows that on the data it has processed, linear models like logistic regression have been able to achieve high accuracy rates, and complex models have not brought significant improvement.

4.2. Study limitations

Although this study was based on authoritative BRFSS data and employed multiple mainstream models for comparison, several limitations remain. First, to control for class imbalance and improve training stability, a 1:1 class balancing strategy was applied, which may have affected the representativeness of the original dataset [11]. Second, feature selection was primarily based on survey responses, lacking clinical indicators such as family history or blood glucose levels, which may reduce the model's interpretability and accuracy in clinical applications [12]. Additionally, no

DOI: 10.54254/2753-8818/2025.AU24516

systematic parameter tuning was conducted for the random forest and decision tree models, and more advanced models such as deep learning were not included. Future research could expand model complexity to further improve prediction performance.

4.3. Model explanatory

Logistic regression provides clear regression coefficients and significance tests, which makes it easy to interpret the influence of each risk factor; decision trees give visualized decision rules, which intuitively explain how to classify the population according to the feature thresholds (which is easy to interpret if the tree is not too deep), and thus are also interpretable to a certain extent; random forests are difficult to directly interpret the predictions of individual predictions due to the large number of trees, but can be explained by the importance of the variables and their significance. interpreted, but general trends can be understood through variable significance and localized interpretation methods. Overall, logistic regression and decision trees are more advantageous in terms of interpretability, while random forests belong to the “black box” model, which is relatively difficult to interpret.

4.4. Suitability

Logistic regression assumes linearly additive effects and is suitable for situations where factors act independently of each other in a linear fashion, and may be underfitted if there are strong nonlinearities in the true relationship; random forests automatically capture nonlinearities and interaction effects, and therefore usually perform better on tasks with complex patterns.

5. Conclusion

This research provides a comparative evaluation of three predictive modeling approaches—logistic regression, random forest, and decision tree—for diabetes risk assessment based on structured survey data. Logistic regression demonstrated the highest overall utility, offering robust accuracy (74%), interpretability through regression coefficients, and stability across test cases. In contrast, random forests, while effective at capturing nonlinear relationships, suffered from mild overfitting due to the lack of hyperparameter optimization. Decision trees, though intuitive, lacked predictive power and generalizability when used independently. These findings suggest that interpretable models like logistic regression remain highly effective for structured health data applications, especially in public health screening scenarios. Future research should incorporate more clinical biomarkers, apply ensemble techniques with regularization, and explore explainability frameworks such as SHAP values to bridge the gap between accuracy and interpretability, thereby enhancing the practical relevance of machine learning in preventive healthcare.

References

[1] International Diabetes Federation. (n.d.). IDF Global Diabetes Overview 8th Edition. https://diabetesatlas.org/upload/resources/previous/files/8/IDF_DA_8e-ZH-final.pdf
[2] Li, R., & Li, D. (2025). A review of the impact of multiple risk factors on individualized management of diabetes mellitus. Advances in Clinical Medicine, 15(11), 170.
[3] Li, Y., Wang, F., & Zhang, W. (2020). Meta-analysis of factors influencing the prevalence of diabetes mellitus in the Chinese population. China Public Health, 36(9), 1378–1384.
[4] Ning, G., Wang, W. Q., & Li, G. W. (2024). Diabetes mellitus in China 1: Epidemiology and risk factors. Translational Medicine.

DOI: 10.54254/2753-8818/2025.AU24516

[5] Wang, L., Zhang, M., & Li, N. (2023). Risk factor analysis of type 2 diabetes mellitus based on LASSO regression and random forest algorithm. Journal of Environment and Health, 40(7), 613–618.
[6] Li, J. S. (2023). Machine learning-based diabetes prediction and SHAP characterization [Master's thesis, Shanghai University of Engineering and Technology].
[7] Zhang, H., Liu, Y., & Wang, Q. (2020). Analysis of risk factors for insulin antibody production in patients with type 2 diabetes mellitus. Advances in Clinical Medicine, 10(5), 361–367.
[8] Wang, M., Li, N., & Zhang, W. (2018). Progress in epidemiologic research on psychosocial risk factors for the development of diabetes mellitus. Chinese Journal of Epidemiology, 39(10), 1343–1347.
[9] Dong, J., Xue, Q., Teng, F., et al. (2024). Progress in the study of diabetes risk factors after transplantation. Organ Transplantation, 15(1), 19–24.
[10] Teboul, A. (2022). Diabetes Health Indicators Dataset. https: //www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset
[11] Centers for Disease Control and Prevention. (n.d.). Behavioral Risk Factor Surveillance System (BRFSS). https: //www.cdc.gov/brfss/
[12] Dong, X., Zhao, D., Li, X. (2022). Effects of family history, sleep, and psychological factors on diabetes prevalence: Based on 2018 China Chronic Disease Surveillance Data. Chinese Journal of Prevention and Control of Chronic Diseases, 30(5), 366–370.