Simple Predictive Analysis on Predicting Diabetes from CDC Health Indicators

Author: Runhe Zhang

Date: October 20 2025

Course: Biomedical Data Science

1. Introduction

Diabetes is a chronic disease that affects over 400 million people worldwide and is strongly influenced by both genetic and lifestyle factors. Early detection of individuals at risk is crucial for prevention and improved health outcomes. In this project, I analyze the CDC 2015 Behavioral Risk Factor Surveillance System dataset to identify the most important health and behavioral indicators associated with diabetes.

The dataset contains more than 250,000 U.S. survey responses with 35 health-related variables, including BMI, blood pressure, cholesterol, age, physical activity, and diet habits. The objective is to build simple predictive models including logistic regression and random forest to classify whether a person has diabetes, and to interpret which variables contribute most strongly to the outcome.

2. Methods

2.1 Dataset

The source of dataset is CDC BRFSS 2015 (UCI Machine Learning Repository), having rows about 253,680, 35 columns, target variable of diabetes binary (0 = No, 1 = Yes). Predictors include lifestyle, health, and demographic indicators such as BMI, HighBP, GenHlth, Age, and PhysActivity. After removing the target column, the remaining features were standardized where appropriate for the linear model. Data was split into 80% training and 20% testing, preserving class balance (≈ 85% non-diabetic, 15% diabetic).

2.2 Models

Two supervised classification algorithms were implemented using scikit-learn. The first one is logistic regression balanced, interpretable linear baseline; class weights adjusted for imbalance. The other one is random forest, nonlinear ensemble of 400 trees for capturing complex relationships. Model performance was evaluated with 5-fold cross-validation on training data and test-set metrics: accuracy, recall, precision, F1-score, and ROC-AUC.

3. Results

3.1 Exploratory Data Analysis

The dataset is moderately imbalanced (Figure 1). The mean BMI was 27.9 kg/m², and about 33% of respondents reported high blood pressure.

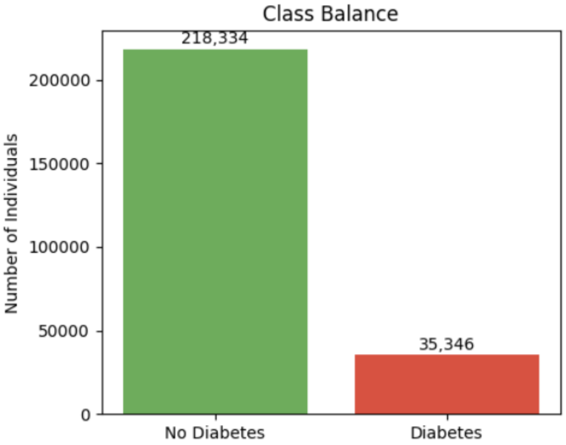
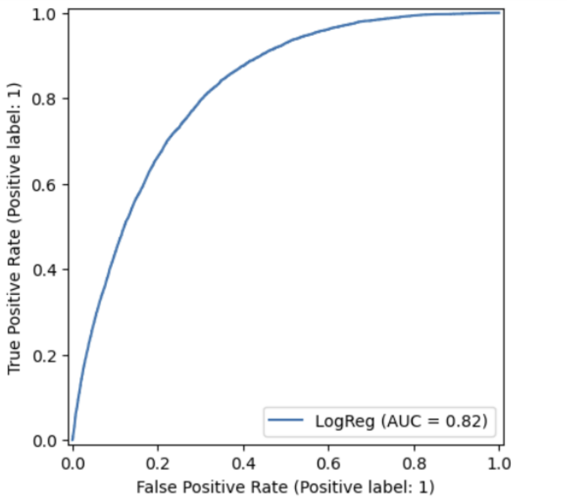


Figure 1. Distribution of Diabetes vs. Non-Diabetes respondents

3.2 Model Performance



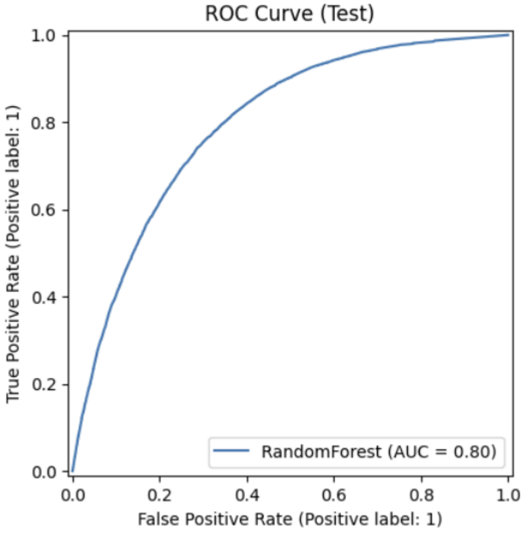


Figure 2.1. ROC curve of the Logistic Regression classifier (AUC = 0.82).
Figure 2.2. ROC curve of the Random Forest classifier (AUC = 0.80).

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.732	0.311	0.761	0.441	0.820
Random Forest	0.857	0.460	0.157	0.235	0.796

Figure 3. ROC Curve comparison between Logistic Regression and Random Forest models. Both models achieved fair discrimination; Random Forest slightly outperformed Logistic Regression on AUC and recall.

3.3 Key Predictors

Across both models, several predictors consistently emerged as the most influential factors associated with diabetes. The Body Mass Index (BMI) was the strongest indicator, followed by General Health (GenHlth), which reflects an individual's self-assessed overall well-being. Age was another major contributor, showing the expected trend of higher diabetes prevalence with increasing age. Functional limitations such as Difficulty Walking (DiffWalk) also appeared as an important predictor, suggesting that mobility impairment may be linked to poorer metabolic health. Finally, High Blood Pressure (HighBP) was among the top variables, aligning with the well-established association between hypertension and diabetes risk.

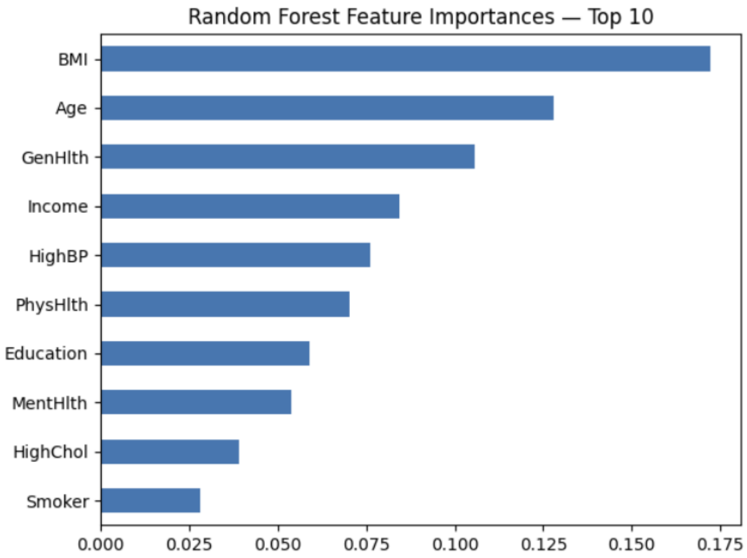


Figure.4 shows the top 10 feature importances from the Random Forest model.

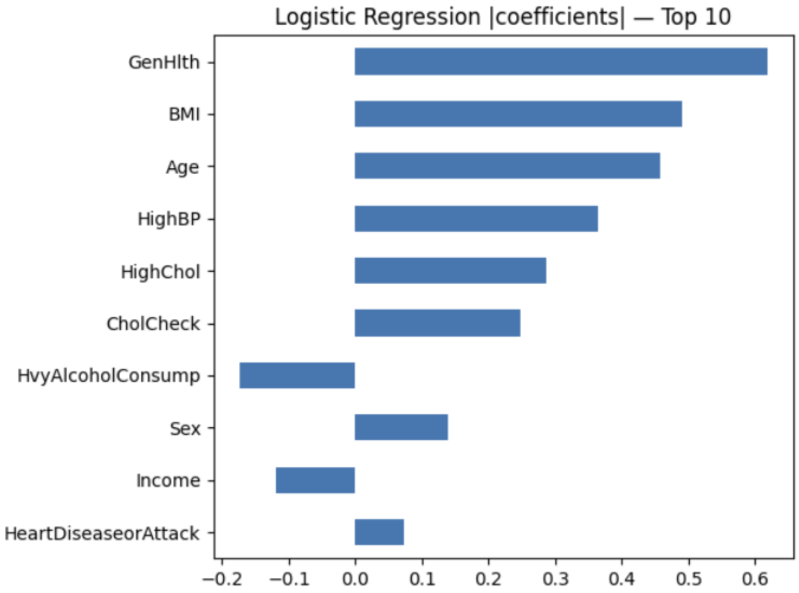


Figure.5 presents the absolute logistic regression coefficients.

This highlights a similar pattern of dominant predictors across both approaches.

3.4 Confusion Matrices

The Random Forest correctly identified most diabetic cases (true positives) but also misclassified some non-diabetics (false positives), reflecting a trade-off between sensitivity and specificity.

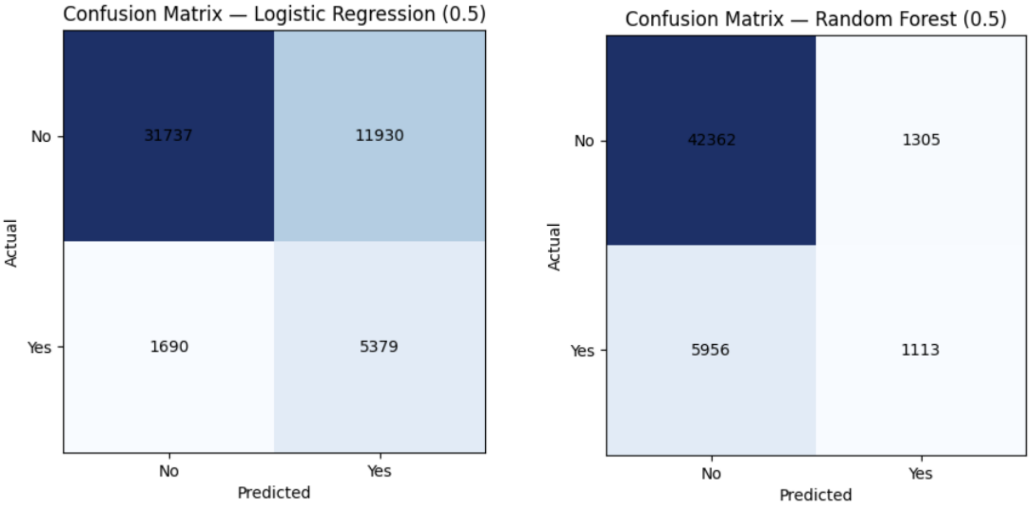


Figure 6. *Confusion Matrices at 0.5 threshold for both models.*

4. Discussion

The results show that both models effectively distinguished between diabetic and non-diabetic individuals, each with different strengths. Logistic Regression achieved an accuracy of 0.73 and a ROC-AUC of 0.82, performing well at identifying diabetic cases (recall = 0.76) but with lower precision (0.31), indicating some false positives. In contrast, the Random Forest reached higher overall accuracy (0.86) but much lower recall (0.16), missing many diabetic cases despite a comparable ROC-AUC of 0.80. As seen in the ROC curves and confusion matrices (Figures 1–2, 6), Logistic Regression was more sensitive, while Random Forest was more conservative.

Feature analysis (Figures 4–5) revealed consistent predictors across both models, with BMI, General Health, and Age as the most influential, followed by High Blood Pressure, Physical Health, and Difficulty Walking. These variables align with known clinical risk factors linking obesity, poor health, and limited mobility to diabetes. Overall, while Random Forest provided slightly higher accuracy, Logistic Regression offered a better balance between interpretability and sensitivity, making it more suitable for public health screening applications.

Limitations and Future Work

This study is limited by the self-reported nature of the BRFSS dataset, which may contain inaccuracies or biases. The strong class imbalance (fewer diabetic cases) also impacts precision and recall. Future analyses could apply oversampling (e.g., SMOTE), threshold tuning, and cross-year validation using newer BRFSS datasets to improve robustness. Incorporating additional features such as dietary patterns or genetic factors could further enhance predictive performance.

5. Conclusion

Using publicly available health survey data, this project demonstrated that simple machine learning models can effectively identify patterns associated with diabetes risk. The Logistic Regression model, with its clear interpretability and strong recall, provides a practical baseline for early screening, while the Random Forest adds nonlinear flexibility but may require threshold adjustments to achieve higher sensitivity.

Key predictors such as BMI, General Health, Age, and High Blood Pressure were consistently identified as significant, reaffirming known lifestyle and physiological risk factors. These findings highlight the potential of low-cost, survey-based analytics for public health applications—supporting early identification, prevention, and resource allocation for populations most at risk of developing diabetes.

6. References

Centers for Disease Control and Prevention (CDC). *Behavioral Risk Factor Surveillance System (BRFSS) 2015 Survey Data.*
Available: https://www.cdc.gov/brfss/

Dua, D. & Graff, C. (2019). *UCI Machine Learning Repository.* Irvine, CA: University of California, School of Information and Computer Science.

Pedregosa, F. et al. (2011). *Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research*, 12, 2825–2830.

Saito, T. & Rehmsmeier, M. (2015). *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS ONE*, 10(3): e0118432.