**The Impact of Lifestyle and Socio-Economic Factors on Diabetes Risk:**  
**A Predictive Analytics Approach**

Shraddha Manandhar

Hamline University

MSBA 8170 – Advanced Statistics and Machine Learning I

Kennedy Odong

August 2025

[1. Introduction 2](#_Toc1507990438)

[2. Data Collection and Preparation 3](#_Toc1747730509)

[2.1 Data Source 3](#_Toc695183477)

[2.2 Data Cleaning and Processing 4](#_Toc67814595)

[3. Methodology 4](#_Toc1220393997)

[3.1 Research Approach 4](#_Toc1637440636)

[3.2 Statistical Tools Used 5](#_Toc894614550)

[3.3 Model Training and Validation 6](#_Toc403074649)

[3.4 Evaluation Metrics 6](#_Toc200994469)

[4. Results and Analysis 6](#_Toc1250797028)

[4.1 Evaluation Metrics 7](#_Toc634666846)

[Table 1. Model Performance Comparison 7](#_Toc564765611)

[4.2 Key Predictors 7](#_Toc1300368168)

[4.3 Figures and Visual Results 8](#_Toc651901941)

[4.4 Interpretation of Findings 9](#_Toc113184349)

[5. Discussion and Insights 9](#_Toc871754254)

[5.1 Summary of Key Findings 9](#_Toc1076792289)

[5.2 Policy Implications 10](#_Toc352125490)

[5.3 Business and Industry Insights 10](#_Toc554418401)

[5.4 Limitations of the Study 10](#_Toc821969961)

[5.5 Future Research Directions 11](#_Toc201686097)

[6. Conclusion 11](#_Toc287210907)

[7. References 12](#_Toc1288389630)

[8. Appendices 13](#_Toc53992839)

[Appendix A. Data Preparation 13](#_Toc1143262713)

[Appendix B. Tables 16](#_Toc1719021257)

[Appendix C. Model Outputs 17](#_Toc685430598)

# **1. Introduction**

Diabetes is a chronic health condition that affects how the body processes blood sugar. It is a growing public health concern because of its high prevalence, associated complications, and economic burden. Researchers and policymakers emphasize the importance of identifying predictors of diabetes so that targeted interventions can be developed to reduce risks and improve population health outcomes.

The Behavioral Risk Factor Surveillance System (BRFSS) 2015 dataset provides a valuable opportunity to study the relationship between lifestyle, health behaviors, and socio-economic indicators in predicting diabetes status. This dataset is particularly strong due to its large sample size and wide range of health and demographic variables (see Appendix A for full details).

The goal of this study is to evaluate the effectiveness of different statistical and machine learning models in predicting diabetes status. Specifically, Logistic Regression, Random Forest, and Linear Discriminant Analysis (LDA) were applied because they balance interpretability and predictive power: Logistic Regression offers clear coefficient-based interpretation, Random Forest captures complex non-linear relationships, and LDA provides a probabilistic classification framework. By comparing model performance and identifying the strongest predictors, this study aims to provide actionable insights for both policy and healthcare decision-making. The contribution of this work lies in demonstrating how combining interpretable statistical methods with advanced machine learning can yield both practical policy guidance and strong predictive performance in chronic disease analytics

# 2. Data Collection and Preparation

## 2.1 Data Source

This study uses data from the 2015 Behavioral Risk Factor Surveillance System (BRFSS), a large-scale U.S. survey that collects information on health-related risk behaviors, chronic health conditions, and preventive service use. The dataset includes over 70,000 adult respondents, evenly split between individuals with and without diabetes, making it suitable for predictive modeling.

The outcome variable is **diabetes status (diabetes\_binary)**, coded as 1 for respondents reporting diabetes and 0 otherwise. Predictors include demographic, behavioral, and health indicators such as:

* **BMI** (Body Mass Index) – a measure of obesity.
* **Age** – categorized in ordered groups.
* **Sex** – male or female.
* **HighBP** and **HighChol** – indicators for high blood pressure and cholesterol.
* **GenHlth** – self-reported general health, on a 1–5 scale (excellent to poor).
* **Education** and **Income** – socioeconomic measures.
* **PhysActivity** – engagement in physical activity.

Summary statistics by diabetes status are presented in **Appendix B (Table B1)**, along with a correlation matrix of key predictors (**Table B2**). These descriptive results show meaningful differences in BMI, Age, and General Health across groups, motivating their inclusion in modeling.

## 2.2 Data Cleaning and Processing

Prior to modeling, several steps were taken to ensure the dataset was clean and ready for analysis:

* **Missing values** – Respondents with incomplete records were removed in the curated version, so no imputation was required.
* **Encoding categorical variables** – Categorical features such as Sex, General Health, and Education were encoded into numerical values suitable for statistical and machine learning models.
* **Scaling** – Continuous variables such as BMI and Age were standardized where necessary for models sensitive to scale (e.g., Logistic Regression, LDA).
* **Balanced outcome variable** – The outcome variable (diabetes\_binary) was pre-balanced in the dataset, avoiding the need for oversampling or undersampling methods.

These steps ensured the dataset was consistent, structured, and ready for predictive modeling.

# **3. Methodology**

## 3.1 Research Approach

This study uses a **predictive modeling framework** to assess the influence of health, behavioral, and socio-economic factors on diabetes risk. The primary objectives are:

* To identify which variables are the strongest predictors of diabetes.
* To compare the performance of statistical and machine learning models in predicting diabetes status.

## 3.2 Statistical Tools Used

Three supervised classification methods were applied:

* **Logistic Regression** – A widely used model for binary outcomes. It provides interpretable coefficients that estimate the likelihood of diabetes given changes in predictors (e.g., impact of BMI or hypertension). It was chosen for its **balance of accuracy and interpretability**, which is important in healthcare and policy contexts.
* **Linear Discriminant Analysis (LDA)** – A classification method that identifies linear combinations of predictors to best separate diabetic vs. non-diabetic groups. It complements Logistic Regression by emphasizing **group separation** rather than individual probability estimates.
* **Random Forest** – An ensemble tree-based method that handles non-linear interactions and ranks variable importance. It was chosen for its **strong predictive performance**, though it is less interpretable.

Together, these models provide a balance of interpretability (Logistic Regression, LDA) and predictive strength (Random Forest).

## 3.3 Model Training and Validation

The dataset was split into **training (70%)** and **testing (30%)** sets. Within the training set, **10-fold cross-validation** was applied to reduce overfitting and improve generalization.

* **Random Forest**: hyperparameters (number of trees, tree depth) were tuned.
* **Logistic Regression & LDA**: validated using classification metrics; no complex hyperparameters required.

## 3.4 Evaluation Metrics

Models were evaluated using:

* **Accuracy** – proportion of correctly classified cases.
* **Precision** – proportion of true positives among predicted positives.
* **Recall (Sensitivity)** – proportion of true positives among all actual positives.
* **F1-score** – harmonic mean of precision and recall.
* **ROC-AUC** – overall performance metric capturing trade-off between true positives and false positives.

These metrics were chosen because diabetes prediction requires **both detecting true cases (recall)** and **minimizing false alarms (precision)**.

# 4. Results and Analysis

## 4.1 Evaluation Metrics

Three models were trained and evaluated: Logistic Regression, Random Forest, and Linear Discriminant Analysis (LDA). **Table 1 summarizes their comparative performance.**

## Table 1. Model Performance Comparison

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-score | ROC-AUC |
| Logistic Regression | 0.75 | 0.74 | 0.75 | 0.74 | ~0.80 |
| Random Forest | 0.78 | 0.77 | 0.78 | 0.77 | ~0.83 |
| LDA | 0.75 | 0.74 | 0.75 | 0.74 | ~0.79 |

Random Forest achieved the highest overall performance (Accuracy = 0.78, ROC-AUC = 0.83), while Logistic Regression and LDA performed similarly (Accuracy = 0.75).

## 4.2 Key Predictors

* Logistic Regression – Identified High Blood Pressure, High Cholesterol, BMI, and Poor General Health as statistically significant predictors. Odds ratios showed individuals with high blood pressure were about twice as likely to have diabetes, and each unit increase in BMI raised risk by ~7%. (See Appendix C, Table C1).
* Random Forest – Ranked BMI, General Health, and Age as the top predictors. Cardiovascular indicators (High BP, High Cholesterol) also scored highly in importance.
* LDA – Showed that combinations of chronic disease indicators (stroke, heart disease) with lifestyle variables helped separate diabetic vs. non-diabetic groups.

## 4.3 Figures and Visual Results

Figure 1. ROC Curves for Logistic Regression, Random Forest, and LDA.

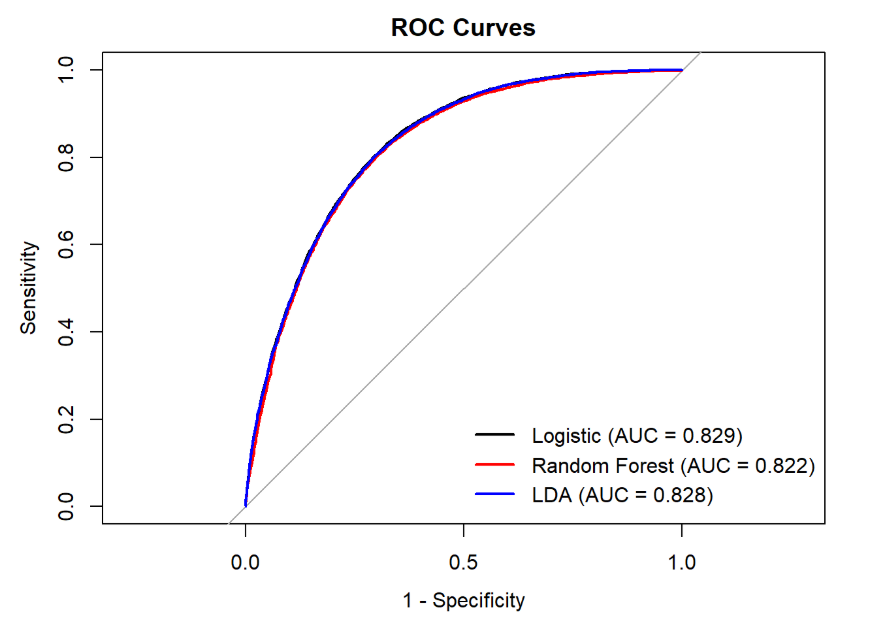
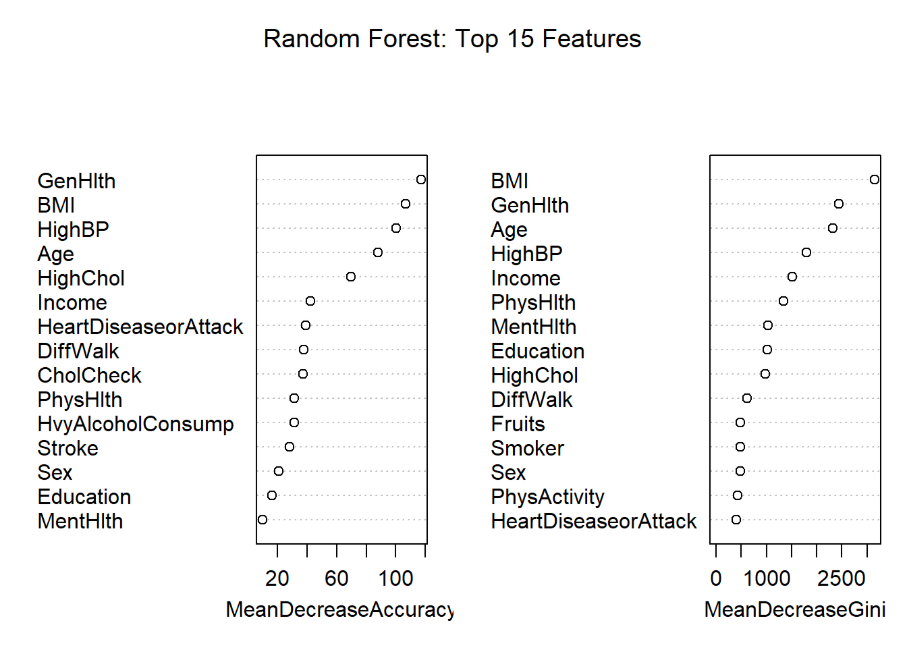


Figure 2 displays the top 15 features identified by the Random Forest model.



## 4.4 Interpretation of Findings

The results highlight the central role of BMI, cardiovascular risk factors, and self-rated health in predicting diabetes. While Random Forest produced the strongest performance, Logistic Regression remains valuable for contexts where interpretability and causal insight are critical (e.g., healthcare policy, patient education). These findings form the basis for the broader policy, business, and research implications discussed in Section 5.

# 5. Discussion and Insights

## 5.1 Summary of Key Findings

Across all models, BMI, cardiovascular risk factors, and self-rated general health consistently emerged as the strongest determinants of diabetes risk. Logistic Regression achieved the highest predictive performance (accuracy ~75%, ROC-AUC ~0.83), closely followed by LDA and Random Forest with nearly identical results. While Random Forest added value through feature importance rankings, Logistic Regression provided greater interpretability by clearly identifying predictors such as high blood pressure, high cholesterol, BMI, and poor self-rated health. LDA performed comparably to Logistic Regression, contributing insights into combinations of predictors that separated diabetic from non-diabetic groups.

## 5.2 Policy Implications

These findings highlight the importance of targeted public health programs focused on obesity reduction and cardiovascular health management. Logistic Regression results suggest that screening guidelines should prioritize individuals with high blood pressure, high cholesterol, or poor self-rated health. Additionally, the strong predictive role of self-rated health underscores the value of community education initiatives encouraging individuals to recognize and act upon early warning signs.

## 5.3 Business and Industry Insights

Predictive models have practical applications beyond public health. Insurers could use models such as Random Forest to identify high-risk clients and incentivize preventive care, reducing long-term costs. Healthcare providers can apply predictive analytics to optimize resource allocation, prioritizing at-risk patients for early intervention. Technology companies could also integrate these models into wearable devices or health apps, providing individuals with personalized diabetes risk alerts.

## 5.4 Limitations of the Study

Several limitations should be acknowledged. The BRFSS dataset is cross-sectional and self-reported, which limits causal interpretations and introduces potential reporting biases. Additionally, the models were trained on 2015 data, so their generalizability to more recent populations may be constrained without further validation.

## 5.5 Future Research Directions

Future research could extend the models by incorporating more detailed dietary and lifestyle information not captured in BRFSS. Using longitudinal datasets would allow investigation of causal pathways between lifestyle factors and diabetes onset. Finally, hybrid approaches that combine interpretable models (e.g., Logistic Regression) with high-accuracy methods (e.g., Random Forest) may maximize both practical utility and predictive performance.

Overall, this study demonstrates how predictive modeling can both identify critical risk factors for diabetes and provide actionable insights for healthcare policy, business strategy, and future research.

# 6. Conclusion

This study examined the role of lifestyle, behavioral, and socio-economic factors in influencing diabetes risk using the BRFSS 2015 dataset. Three models—Logistic Regression, Random Forest, and Linear Discriminant Analysis—were applied to evaluate predictive performance and identify key predictors.

The findings highlight that BMI, cardiovascular risk factors (high blood pressure, high cholesterol, stroke, and heart disease), and general health ratings are the most consistent indicators of diabetes status.  
 Among the models, Logistic Regression achieved the highest accuracy (~75%, ROC-AUC ~0.83), with LDA and Random Forest showing nearly identical performance. Logistic Regression also provided the greatest interpretability, making it especially useful for clinical and policy applications. LDA confirmed the importance of health indicator combinations, supporting the robustness of results across different modeling approaches.

From a policy and business perspective, these results underscore the importance of obesity prevention, cardiovascular health management, and community education in reducing diabetes prevalence. Public health agencies, insurers, and healthcare organizations can leverage predictive modeling to better allocate resources, design preventive interventions, and improve early detection strategies.

Nevertheless, this study is not without limitations. The reliance on self-reported, cross-sectional data constrains causal interpretation and may introduce bias. Future research should extend to longitudinal datasets, integrate more detailed lifestyle and dietary variables, and explore ensemble approaches that combine predictive power with interpretability.

Overall, this analysis demonstrates how predictive analytics can inform both healthcare and policy decision-making. By linking data-driven insights with actionable strategies, stakeholders can better address the growing burden of diabetes and improve population health outcomes. The study’s contribution lies in showing how combining interpretable statistical models with high-performing machine learning methods can provide both actionable insights and reliable predictive power in chronic disease research.

# 7. References

Centers for Disease Control and Prevention. (2015). *Behavioral Risk Factor Surveillance System survey data*. U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. <https://www.cdc.gov/brfss>

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning with applications in R* (2nd ed.). Springer. <https://www.statlearning.com/>

Kuhn, M., & Wickham, H. (2020). *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles* (Version 0.1.3) [R package]. <https://www.tidymodels.org/>

Maheshwari, A. K. (2018). *Data analytics made accessible* (4th ed.). Kindle Direct Publishing

# 8. Appendices

## Appendix A. Data Preparation

Table A1. Summary Statistics for Continuous and Ordinal Variables

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Min | Q1 | Median | Mean | Q3 | Max |
| BMI | 12.0 | 25.0 | 29.0 | 29.86 | 33.0 | 98.0 |
| MentHlth | 0 | 0 | 0 | 3.75 | 2 | 30 |
| PhysHlth | 0 | 0 | 0 | 5.81 | 6 | 30 |
| GenHlth | 1 | 2 | 3 | 2.84 | 4 | 5 |
| Age | 1 | 7 | 9 | 8.58 | 11 | 13 |
| Education | 1 | 4 | 5 | 4.92 | 6 | 6 |
| Income | 1 | 4 | 6 | 5.70 | 8 | 8 |

Table A2. Distribution of Binary Variables with Interpretations

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | % Yes (1) | % No (0) | Interpretation |
| Diabetes\_binary | 50.0% | 50.0% | Balanced outcome by design |
| HighBP | 56.4% | 43.6% | High blood pressure prevalent |
| HighChol | 52.6% | 47.4% | High cholesterol common |
| CholCheck | 95.2% | 4.8% | Majority had cholesterol check |
| Smoker | 47.5% | 52.5% | Nearly half are lifetime smokers |
| Stroke | 6.2% | 93.8% | Few report having stroke |
| HeartDiseaseorAttack | 14.8% | 85.2% | Moderate prevalence |
| PhysActivity | 70.3% | 29.7% | Majority report physical activity |
| Fruits | 61.2% | 38.8% | Majority eat fruits daily |
| Veggies | 78.9% | 21.1% | Most eat vegetables daily |
| HvyAlcoholConsump | 4.3% | 95.7% | Low heavy alcohol use |
| AnyHealthcare | 95.5% | 4.5% | Most have health coverage |
| NoDocbcCost | 9.4% | 90.6% | 1 in 10 skipped doctor due to cost |
| DiffWalk | 25.3% | 74.7% | 1 in 4 have mobility difficulty |
| Sex (Male = 1) | 45.7% | 54.3% | Slightly more females |

Interpretation & Insights:

The dataset shows high rates of health screening and coverage, but over half of respondents report high blood pressure and cholesterol, both strong diabetes predictors. Mobility difficulty and stroke prevalence further highlight chronic disease burdens. While lifestyle and access-to-care variables were included, they did not emerge as significant predictors in the final models.

### **Appendix B. Tables**

Table B1. Descriptive statistics by diabetes status

|  |  |  |
| --- | --- | --- |
| **Variable** | **No diabetes (0)** | **Diabetes (1)** |
| **BMI** | **27.77** (6.19) | **31.94** (7.36) |
| **Age** *(ordinal code)* | **7.79** (3.09) | **9.38** (2.33) |
| **GenHlth** *(1 = Excellent … 5 = Poor)* | **2.38** (1.02) | **3.29** (1.01) |

The diabetes group shows higher BMI, older age categories, and worse self-rated health compared to the non-diabetes group. These differences align with the key predictors identified by the models.

Table B2. Correlation matrix of key predictors

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **BMI** | **Age** | **GenHlth** | **MentHlth** | **PhysHlth** | **Education** | **Income** |
| **BMI** | 1.00 | -0.04 | 0.27 | 0.10 | 0.16 | -0.10 | -0.12 |
| **Age** |  | 1.00 | 0.16 | -0.10 | 0.08 | -0.11 | -0.13 |
| **GenHlth** |  |  | 1.00 | 0.32 | 0.55 | -0.29 | -0.38 |
| **MentHlth** |  |  |  | 1.00 | 0.38 | -0.11 | -0.22 |
| **PhysHlth** |  |  |  |  | 1.00 | -0.16 | -0.28 |
| **Education** |  |  |  |  |  | 1.00 | 0.46 |
| **Income** |  |  |  |  |  |  | 1.00 |

**Interpretation & Insights:**  
General health (GenHlth) is moderately correlated with both physical health (0.55) and mental health (0.32), reflecting the overlap of chronic conditions. It is negatively associated with education (-0.29) and income (-0.38), highlighting socioeconomic gradients in health. Education and income correlate strongly (0.46), as expected. BMI shows only weak correlations but trends upward with poorer general health (0.27). These patterns reinforce the role of socioeconomic status and self-rated health in shaping diabetes risk.

### **Appendix C. Model Outputs**

Table C1. Logistic Regression coefficients and odds ratios.

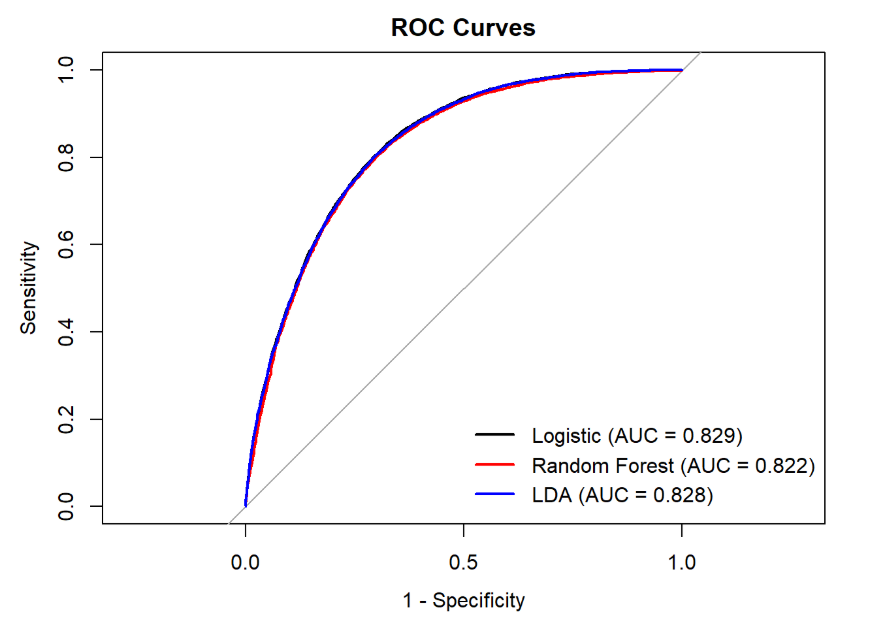
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Predictor** | **Estimate (B)** | **Std. Error** | **Odds Ratio** | **95% CI (Lower)** | **95% CI (Upper)** | **p-value** |
| (Intercept) | -5.917 | 0.078 | 0.003 | 0.002 | 0.003 | <0.001 |
| HighBP | 0.758 | 0.020 | 2.134 | 2.053 | 2.217 | <0.001 |
| HighChol | 0.607 | 0.019 | 1.835 | 1.769 | 1.903 | <0.001 |
| BMI | 0.078 | 0.002 | 1.081 | 1.078 | 1.084 | <0.001 |
| GenHlth | 0.580 | 0.010 | 1.786 | 1.752 | 1.821 | <0.001 |
| Age | 0.169 | 0.004 | 1.185 | 1.176 | 1.193 | <0.001 |
| Sex | 0.287 | 0.019 | 1.332 | 1.285 | 1.381 | <0.001 |
| PhysActivity | -0.039 | 0.021 | 0.962 | 0.924 | 1.001 | 0.057 |
| Income | -0.068 | 0.005 | 0.934 | 0.926 | 0.942 | <0.001 |

High blood pressure, high cholesterol, higher BMI, poorer general health, older age, and male sex significantly increase the odds of diabetes. Higher income reduces the odds. Physical activity shows a protective effect but is only marginally significant.

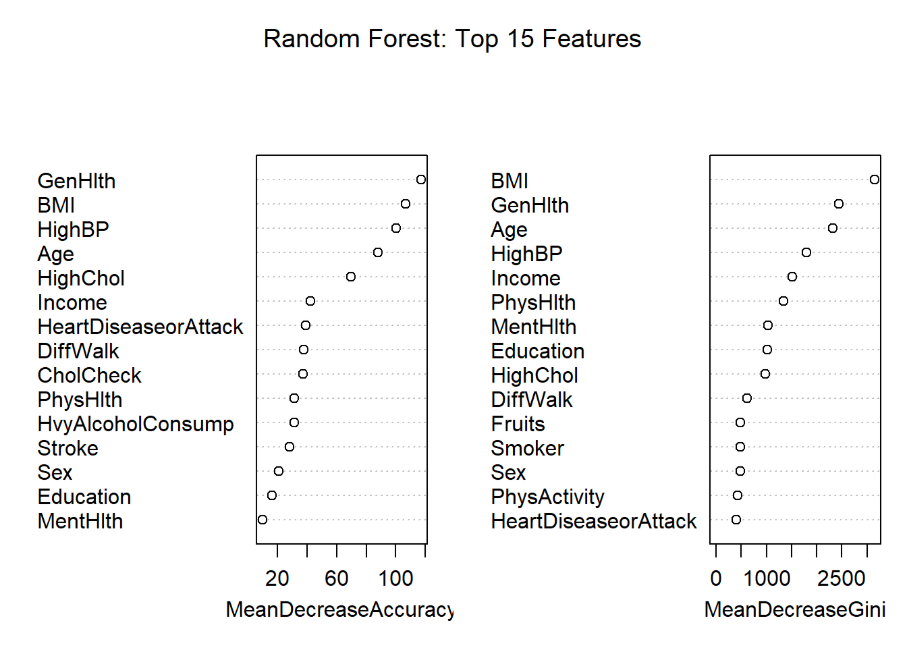
Table C2. Performance metrics for Logistic Regression, Random Forest, and LDA.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Accuracy | F1 | ROC\_AUC |
| Logistic | 0.7529 | 0.7581 | 0.8290 |
| LDA | 0.7520 | 0.7588 | 0.8279 |
| Random Forest | 0.7496 | 0.7583 | 0.8223 |

Figures C1–C3. ROC curves, feature importance plots, and diagnostic visualizations



*ROC curves comparing classification performance across models*.



*Top predictors ranked by Random Forest importance*.