Multiclass Diabetes Classification Using XGBoost with SHAP Interpretation on BRFSS 2015 Data

*Prepared by: Supreet Mutsuddi*

Table of Contents

[1. Executive Summary 2](#_Toc199776302)

[2. Introduction 2](#_Toc199776303)

[3. Data Preparation 2](#_Toc199776304)

[3.1 Initial Exploration 2](#_Toc199776305)

[3.2 Feature Relationship Analysis 3](#_Toc199776306)

[3.3 Handling Class Imbalance 3](#_Toc199776307)

[4. Model Building: XGBoost 3](#_Toc199776308)

[4.1 Data Splitting 3](#_Toc199776309)

[4.2 Model Training 3](#_Toc199776310)

[4.3 Model Evaluation 3](#_Toc199776311)

[5. Model Interpretation Using SHAP 4](#_Toc199776312)

[5.1 Why SHAP? 4](#_Toc199776313)

[5.2 SHAP Calculation 4](#_Toc199776314)

[5.3 Feature Importance by Class 4](#_Toc199776315)

[6. Discussion 5](#_Toc199776316)

[7. Conclusion 5](#_Toc199776317)

# 1. Executive Summary

This report presents a machine learning analysis to classify individuals into three categories of diabetes status using the Behavioral Risk Factor Surveillance System (BRFSS) 2015 dataset. An XGBoost model was trained on a balanced version of the dataset, and model interpretability was achieved through SHAP values. The final model demonstrated good classification performance, with features like General Health, BMI, and High Blood Pressure consistently emerging as the most important predictors.

Project GitHub Repository: <https://github.com/Supreet1982/Diabetes_BRFSS>

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# 2. Introduction

Diabetes is a major public health issue, and early prediction can enable timely intervention. This study aims to develop a robust and interpretable predictive model that classifies individuals as:

* 0: No Diabetes
* 1: Prediabetes
* 2: Diabetes

Using the BRFSS 2015 dataset, which includes health indicators collected via surveys, we apply advanced modeling and interpretation techniques.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# 3. Data Preparation

## 3.1 Initial Exploration

The dataset contains thousands of observations with health-related features such as BMI, physical activity, and cholesterol levels. The target variable, Diabetes\_012, showed a significant imbalance:

* Class 0 (No Diabetes): 213,703 observations
* Class 1 (Prediabetes): 35,346 observations
* Class 2 (Diabetes): 4,631 observations

This skewed distribution posed a challenge for model training and required appropriate balancing techniques.

## 3.2 Feature Relationship Analysis

To understand the relationship between input features, pairwise correlations were calculated. Although some variables exhibited moderate to strong correlations, none were removed at this stage. These correlations helped in identifying potential redundancies and understanding the data structure, informing future feature selection and interpretation.

## 3.3 Handling Class Imbalance

Given the imbalance, we used the SMOTE technique followed by downsampling to ensure all three classes had equal representation (n = 35,346 each). This helped prevent the model from being biased toward the majority class.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# 4. Model Building: XGBoost

## 4.1 Data Splitting

The balanced dataset was split into 70% training and 30% test sets using caret::createDataPartition.

## 4.2 Model Training

The model xgb.tuned3 was trained using the xgbTree method in caret with hyperparameter tuning via grid search and 5-fold cross-validation. The tuning grid explored a range of values for eta and nrounds, while holding other parameters constant to reduce complexity.

Model selection was based on accuracy, and the final combination of hyperparameters chosen was:

* nrounds = 150
* eta = 0.1

These values achieved the highest accuracy during cross-validation and were used to train the final model.

## 4.3 Model Evaluation

The final model's performance was evaluated using a confusion matrix. The overall classification accuracy on the test set was **0.8294**. The results indicate that the model achieved high accuracy, particularly in identifying individuals with prediabetes (class 1), while also performing reasonably well across the other two classes.

**Confusion Matrix of Final Model**

|  |  |  |  |
| --- | --- | --- | --- |
| **Prediction \ Reference** | **Class 0** | **Class 1** | **Class 2** |
| Class 0 | 7593 | 114 | 2126 |
| Class 1 | 0 | 10312 | 0 |
| Class 2 | 3010 | 177 | 8477 |

This performance demonstrates the model’s success in managing the class imbalance introduced by the original data and the utility of SMOTE plus downsampling. Class 1 (prediabetes) was classified with perfect precision, though some misclassifications between classes 0 and 2 were observed.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# 5. Model Interpretation Using SHAP

## 5.1 Why SHAP?

SHAP values provide additive feature attributions, enabling both local and global interpretation of the model. They help identify which features contribute most to predictions for each class.

## 5.2 SHAP Calculation

Using the predict() function with predcontrib = TRUE on the final XGBoost model, SHAP values were calculated for each class. The BIAS term was removed to isolate individual feature contributions.

## 5.3 Feature Importance by Class

The top five features for each class, based on mean absolute SHAP values, are summarized below:

|  |  |
| --- | --- |
| Class | Top SHAP Features |
| 0 | GenHlth, BMI, HighBP, Age, PhysActivity |
| 1 | GenHlth, HighBP, HighChol, BMI, Age |
| 2 | GenHlth, HighChol, BMI, HighBP, PhysActivity |

These features consistently contributed the most to the model's predictions. Bar plots generated from the SHAP values helped visualize and confirm the influence of these variables.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# 6. Discussion

The model achieved good classification performance and highlighted health indicators that align with clinical understanding. General Health (GenHlth) was consistently the top feature, indicating self-perception of health is strongly tied to diabetes risk.

Unexpectedly, some known factors like smoking or alcohol consumption had lower influence, suggesting the need for feature engineering or better proxy variables. Class balancing was essential for achieving fair performance.

Limitations include potential overfitting, exclusion of regularization tuning, and lack of external validation.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# 7. Conclusion

The XGBoost model, trained on a balanced dataset and interpreted via SHAP, effectively classified individuals into diabetes status groups. The approach is transparent and suitable for deployment in decision-support dashboards like Power BI or Tableau.

Future work could include regularization, ensemble methods, or external validation on different population data.