# Predictive Analytics in Public Health: An Assessment of Diabetes Risk Factors Using CDC's Behavioral Risk Factor Surveillance System Data

Nazmus Sakib Sumon

May 2024

**Abstract**

The Centers for Disease Control and Prevention (CDC) conducts the Behavioral Risk Factor Surveillance System (BRFSS) annually, engaging over 400,000 Americans to collect data on health-related risk behaviors, chronic health conditions, and the use of preventive services. Utilizing the 2015 BRFSS data, this report explores significant predictors of diabetes by analyzing 22 variables through various statistical methods. We employed five baseline methods and two advanced ensemble techniques, focusing on parameter optimization through cross-validation. Among the baseline methods, Logistic Regression emerged as the most effective, demonstrating the lowest error rates. Random Forest, among all the methods, provided highly accurate predictions, outperforming other tested methods. This analysis underscores the efficacy of combining traditional and contemporary statistical approaches in predicting health outcomes.

## Introduction

Diabetes, one of the most prevalent chronic illnesses in the United States, affects millions annually and significantly impacts the national economy. This condition impairs an individual's ability to regulate blood glucose levels, potentially reducing life expectancy and quality of life. During digestion, sugars are released into the bloodstream, signaling the pancreas to release insulin, which facilitates the cellular utilization of blood sugars as an energy source. Typically, diabetes arises from inadequate insulin production or inefficient use by the body. Although incurable, its adverse effects can be mitigated through healthy eating, exercise, weight management, and medical care. Predictive models are invaluable, as early diagnosis can lead to effective lifestyle modifications and improved treatment outcomes.

## Data Set Description

The CDC annually surveys critical health-related issues and preventive service usage since 1984. The primary data source is the 2015 Kaggle dataset, supplemented by data from the UCI Irvine Machine Learning Repository. According to CDC data, as of 2018, 88 million Americans have prediabetes, and 34.2 million have diabetes. Many are unaware of their condition, which also imposes a significant economic burden, with diagnosed diabetes costing over $327 billion annually. Our analysis used a cleaned dataset of 253,680 survey responses from the BRFSS 2015 survey. The original dataset contained 330 features from 441,455 respondents. The target variable, 'Diabetes-binary', distinguishes between no diabetes (0) and prediabetes or diabetes (1).

## Exploratory Data Analysis

The dataset comprises 253,680 rows and 22 columns, with no missing values. An analysis of data types indicated discrepancies, prompting conversion of most fields to categorical as per Table 1's specifications.

The histogram analysis revealed a right-skewed distribution for BMI values, with a notable long tail towards higher ranges. A stepwise logistic regression identified significant predictor variables, which were then visualized in bar charts. Significant variables were selected based on p-values less than 0.05 from dummy variables. Variables with more than two categories were excluded from these charts to maintain clarity.

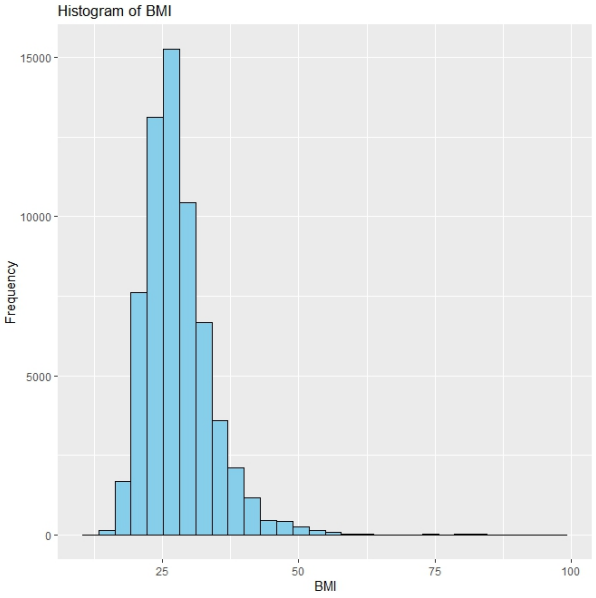


Figure 1: Histogram of BMI values

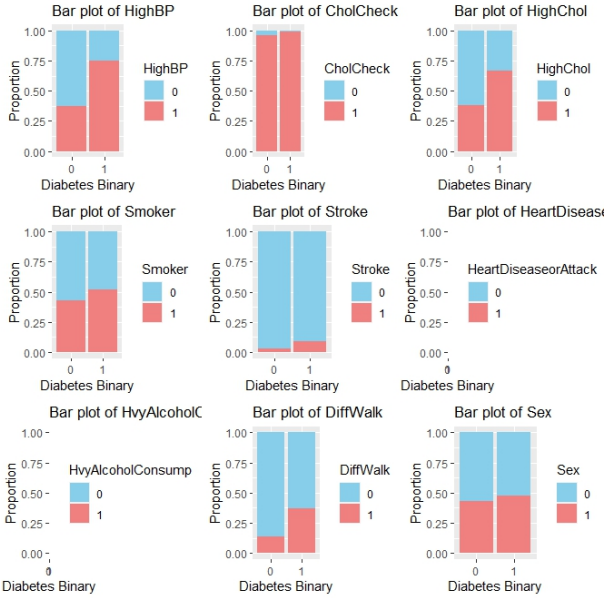


Figure 2: Bar plots of significant variables

Certain categorical variables exhibit distinct behaviors in individuals with and without diabetes, suggesting their potential as significant predictors. Notable among these are High Blood Pressure (HighBP), High Cholesterol (HighChol), Heart Disease or Attack (HeartDiseaseAttack), and Difficulty Walking (DiffWalk).

## Methodology

The initial exploratory data analysis utilized logistic regression. Subsequent analyses incorporated five baseline methods and two ensemble methods to robustly assess predictors of diabetes.

Table 1: Baseline Methods Used in Analysis

| Method Number | Method Name |
|---|---|
| 1 | Linear Discriminant Analysis (LDA) |
| 2 | Quadratic Discriminant Analysis (QDA) |
| 3 | Naïve Bayes |
| 4 | Logistic Regression |
| 5 | K Nearest Neighbor |

Table 2: Ensemble Methods Used in Analysis

| Method Number | Method Name |
|---|---|
| 1 | Random Forest |
| 2 | XG Boosting |

Due to computational limitations, only a quarter of the dataset (63,420 rows and 22 columns) was used, divided into 80% training data and 20% testing data. This subset focused on the 14 variables identified as significant in the logistic regression analysis. Parameters for K Nearest Neighbor were optimized through 5-fold cross-validation, which determined the best k-value for minimizing testing error. Similar parameter tuning processes were applied for Random Forest and XG Boosting, where the mtry parameter for Random Forest and eta, max-depth, and gamma for XG Boosting were adjusted.

## Results

Table 3: Before Cross-Validation Performance

| Method | Training Error | Testing Error | Best Parameter Value |
|---|---|---|---|
| LDA | 0.13790721 | 0.1361002 | - |
| QDA | 0.36286278 | 0.3687549 | - |
| Naïve Bayes | 0.17457058 | 0.1738449 | - |
| Logistic Regression | 0.13492596 | 0.1337510 | - |
| K Nearest Neighbor | 0.12274432 | 0.1427565 | K value = 5 |
| Random Forest | 0.06886476 | 0.1379013 | mtry = 4, ntree = 500 |
| XG Boosting | 0.13879566 | 0.1380580 | Eta = 0.01, Max depth = 3, Gamma = 0 |

Table 4: After Cross-Validation Performance

| Method | Testing Error | Best Parameter Value |
|---|---|---|
| LDA | 0.1400189 | - |
| QDA | 0.360533 | - |
| Naïve Bayes | 0.1721854 | - |
| Logistic Regression | 0.1677369 | - |
| K Nearest Neighbor | 0.1465626 | K value = 5 |
| Random Forest | 0.1361558 | mtry = 4, ntree = 500 |
| XG Boosting | 0.1423841 | Eta = 0.01, Max depth = 3, Gamma = 0 |

## Findings

**Before Cross-Validation Performance:**

Before cross-validation, Logistic Regression exhibited the lowest testing error among the base-line methods with a training error of 0.13492596 and a testing error of 0.1337510. Random Forest outperformed all models in terms of training error (0.06886476) and had a competitive testing error of 0.1379013.

**After Cross-Validation Performance:**

The best performing method after cross-validation was Random Forest, with a testing error of 0.1361558. The optimal parameter value for mtry was determined to be 4.

**Performance Tuning:**
Parameter tuning played a crucial role in optimizing the performance for KNN, Random Forest, and XG Boosting. Notably, cross-validation aided in identifying optimal settings, thereby reducing the likelihood of model overfitting and enhancing the predictability of the models.

**Significant Variables:**
Stepwise logistic regression initially identified key predictors for the onset of diabetes. Subsequent analysis confirmed the importance of variables such as High Blood Pressure (HighBP), High Cholesterol (HighChol), Heart Disease or Attack (HeartDiseaseAttack), and Difficulty Walking (DiffWalk). These variables consistently impacted various models, highlighting their predictive significance.

## Future Research

Future studies should consider integrating additional predictors, such as dietary and lifestyle factors, into the models. Furthermore, advanced machine learning techniques like ensemble methods and support vector machines could be evaluated to enhance predictive accuracy. Implementing a larger dataset might also provide deeper insights and a more stable performance evaluation.

## Further Reading

For detailed information on the statistical methods used in this study, please refer to the faraway package documentation on CRAN. For general information on body fat percentage, consult the Wikipedia article: Body Fat Percentage on Wikipedia.

## Appendix

Table 5: Description of Key Variables in the Diabetes Study

| **Variable Name** | **Description** |
|---|---|
| ID | Patient ID |
| Diabetes_binary | 0 = no diabetes, 1 = prediabetes or diabetes |
| HighBP | 0 = no high BP, 1 = high BP |
| HighChol | 0 = no high cholesterol, 1 = high cholesterol |
| CholCheck | 0 = no cholesterol check in 5 years, 1 = cholesterol check in 5 years |
| BMI | Body Mass Index |
| Smoker | Have you smoked at least 100 cigarettes in your entire life? (0 = no, 1 = yes) |
| Stroke | Have you had a stroke? (0 = no, 1 = yes) |
| HeartDiseaseorAttack | Coronary heart disease (CHD) or myocardial infarction (MI) (0 = no, 1 = yes) |
| PhysActivity | Physical activity in past 30 days not including job (0 = no, 1 = yes) |
| Fruits | Consume Fruit 1 or more times per day (0 = no, 1 = yes) |
| Veggies | Consume Vegetables 1 or more times per day (0 = no, 1 = yes) |
| HvyAlcoholConsump | Heavy drinkers (0 = no, 1 = yes) |
| AnyHealthcare | Have any kind of health care coverage (0 = no, 1 = yes) |
| NoDocbcCost | Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? (0 = no, 1 = yes) |
| GenHlth | General health condition (1 = excellent, 2 = very good, 3 = good, 4 = fair, 5 = poor) |
| MentHlth | Number of days in the past 30 days that mental health was not good (range: 1-30 days) |
| PhysHlth | Number of days in the past 30 days that physical health was not good (range: 1-30 days) |
| DiffWalk | Do you have serious difficulty walking or climbing stairs? (0 = no, 1 = yes) |
| Sex | Sex (0 = female, 1 = male) |
| Age | Age category (1 = 18-24, 9 = 60-64, 13 = 80 or older) |
| Education | Education level (1 = Never attended school or only kindergarten, 6 = College graduate) |
| Income | Income scale (1 = less than $10,000, 8 = $75,000 or more) |