

**Predicting Diabetes: A Data-Driven Statistical Study on Key Risk Factors**

Sameera Viraj Malwaththa Pathirannehelage - (0001119746)

Course: Statistical Data Analysis for Applied Physics

Master’s in applied physics

May 2025

Table of Contents

[1 Introduction 3](#_Toc198038348)

[1.1 Hypotheses and Statistical Tests 3](#_Toc198038349)

[2 Methodology 3](#_Toc198038350)

[2.1 Hypothesis testing: 4](#_Toc198038351)

[2.1.1 Hypothesis 1: Higher BMI is associated with an increased likelihood of diabetes 4](#_Toc198038352)

[2.1.2 Hypothesis 2: Smoking history affects the risk of diabetes 5](#_Toc198038353)

[2.1.3 Hypothesis 3: Older individuals have a higher risk of diabetes 5](#_Toc198038354)

[2.1.4 Hypothesis 4: Diabetes prevalence differs by gender 6](#_Toc198038355)

[2.2 Full Logistic Regression Model 6](#_Toc198038356)

[3 Results and Discussion 7](#_Toc198038357)

[3.1 Exploratory Data Visualization 7](#_Toc198038358)

[3.1.1 Distribution of BMI 7](#_Toc198038359)

[3.1.2 Distribution of age 7](#_Toc198038360)

[3.1.3 Gender 8](#_Toc198038361)

[3.1.4 Smoke history 9](#_Toc198038362)

[3.1.5 Diabetes Status 10](#_Toc198038363)

[3.2 Results of tested hypothesis 10](#_Toc198038364)

[3.2.1 Hypothesis 1: Higher BMI is associated with an increased likelihood of diabetes 10](#_Toc198038365)

[3.2.2 Hypothesis 2: Smoking history affects the risk of diabetes 10](#_Toc198038366)

[3.2.3 Hypothesis 3: Older individuals have a higher risk of diabetes 11](#_Toc198038367)

[3.2.4 Hypothesis 4: Diabetes prevalence differs by gender 11](#_Toc198038368)

[3.3 Full Model: Age, BMI, Smoking History, and Gender as Predictors of Diabetes 11](#_Toc198038369)

[3.4 General discussion 11](#_Toc198038370)

[4 Conclusion 12](#_Toc198038371)

# 

# Introduction

Diabetes is considered a major global health concern, and the identification of its contributing factors is viewed as essential for prevention and early intervention. In this project, a dataset sourced from Kaggle was analyzed, containing information collected from 100,000 individuals. Various attributes were included in the dataset, such as age, gender, location, ethnicity, hypertension status, heart disease history, BMI (Body Mass Index), smoking history, HbA1c level, and blood glucose level. Statistical analyses were carried out using R software, where logistic regression and chi-square tests were applied to examine associations between selected variables and the presence of diabetes. Four hypotheses were tested, focusing on BMI, smoking history, age, and gender, in order to identify which factors were significantly associated with an increased risk of diabetes.

## Hypotheses and Statistical Tests

In this study, four hypotheses were formulated to explore the relationship between specific risk factors and diabetes prevalence. Each hypothesis was tested using appropriate statistical methods to determine its significance. The first hypothesis examined whether a higher Body Mass Index (BMI) is associated with an increased likelihood of diabetes. The second focused on whether smoking history has an effect on diabetes risk. The third hypothesis evaluated the impact of age, testing whether older individuals are more likely to develop diabetes. Lastly, the fourth hypothesis assessed whether diabetes prevalence differs by gender. These hypotheses were chosen based on commonly studied risk factors and were tested using logistic regression for continuous variables and chi-square tests for categorical variables.

# Methodology

Before performing the statistical analysis, the dataset was carefully examined and underwent thorough preprocessing to ensure its quality and reliability. The main objectives of this step were to filter out irrelevant (“garbage”) data, eliminate duplicate entries, and remove records with missing values. Since the dataset was in CSV (Comma-Separated Values) format, it was loaded into the R environment using the read.csv () function. Duplicate and incomplete data were then removed using the following R commands:

A close-up of a computer code

AI-generated content may be incorrect.

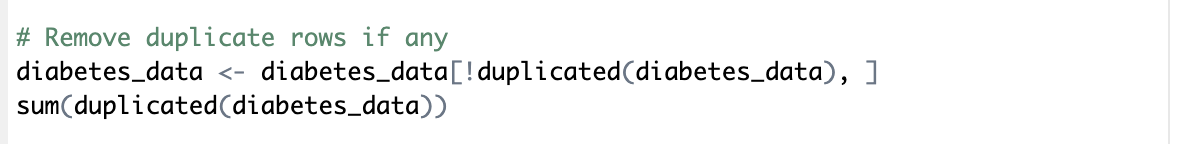


Figure : Data preprocessing, checking and removing duplicates and missing rows

After the dataset was cleansed, the distributions of the available parameters such as BMI, age, gender, smoking history and diabetes cases were examined. Basic comparisons between BMI vs age and HbA1c level vs diabetes status was also performed to gain a brief overview of the dataset. The following commands were used to generates graphs.

A screenshot of a computer program

AI-generated content may be incorrect.

Figure : Commands used in R to check the distributions of selected parameters of the dataset

After examining the behavior of the above-mentioned parameters, the analysis was conducted to test the hypotheses outlined under section 1.1.

## Hypothesis testing:

### Hypothesis 1: Higher BMI is associated with an increased likelihood of diabetes

The first hypothesis was tested to examine the relationship between Body Mass Index (BMI) and the presence of diabetes in the dataset. To evaluate this, a logistic regression model was applied using BMI as the predictor variable and diabetes status as the binary outcome.

* Null Hypothesis (H₀): BMI has no significant effect on the likelihood of diabetes.
* Alternative Hypothesis (H₁): Higher BMI is associated with an increased likelihood of diabetes.

Statistical significance was determined based on the p-value obtained from the regression output, and the effect size was interpreted using the odds ratio. The following commands were used to apply logistic regression model on the BMI and the diabetes status of the dataset.

A close-up of a text

AI-generated content may be incorrect.

Figure : Commands for perform logistic regression test and to evaluate effect of BMI for diabetes

### Hypothesis 2: Smoking history affects the risk of diabetes

The second hypothesis was tested to determine whether there is an association between smoking history and diabetes status. As both variables are categorical, a chi-square test of independence was used to evaluate their relationship.

* Null Hypothesis (H₀): Smoking history has no significant effect on the risk of diabetes.
* Alternative Hypothesis (H₁): Smoking history significantly affects the risk of diabetes.

A contingency table was created to summarize the frequency of diabetes across the different smoking categories, and the chi-square test was applied to assess statistical significance.

This analysis was carried out using a series of R commands, which are shown below.

A computer screen shot of a computer code

AI-generated content may be incorrect.

Figure : Commands used to perform chi-square test

In addition, the following commands were used to measure the strength of significancy of the smoking history to diabetes using Cramér’s V.

A close-up of a text

AI-generated content may be incorrect.

Figure : command for the evaluation of strength of significancy of the smoking history to diabetes

### Hypothesis 3: Older individuals have a higher risk of diabetes

The third hypothesis aimed to evaluate the impact of age on diabetes prevalence. A logistic regression model was applied with age as the predictor variable and diabetes status as the outcome.

* Null Hypothesis (H₀): Age has no significant effect on the likelihood of diabetes.
* Alternative Hypothesis (H₁): Older individuals have a higher risk of developing diabetes.

The significance of the relationship was assessed using the p-value from the regression model, and the effect size was interpreted through the calculated odds ratio.

The logistic regression was performed in R using the commands included below.

A close-up of a white background

AI-generated content may be incorrect.

Figure : Performing logistic regression model for third hypothesis on R

### Hypothesis 4: Diabetes prevalence differs by gender

The fourth hypothesis investigated whether diabetes prevalence varies between males and females. A chi-square test was conducted using gender and diabetes status as categorical variables.

* Null Hypothesis (H₀): Diabetes prevalence is the same across genders.
* Alternative Hypothesis (H₁): Diabetes prevalence differs by gender.

A two-way contingency table was used to organize the data, and the chi-square test was then applied to examine the association.

The R code used to carry out this test is provided in the section below.

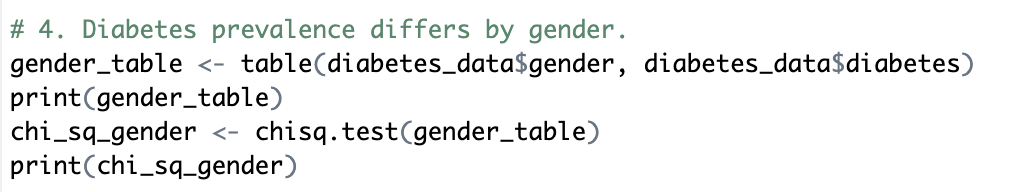


Figure : R commands used for chi-square test

After the test, the strength of the association between gender and diabetes was measured using Cramér’s V value with the following command.

A black text on a white background

AI-generated content may be incorrect.

Figure : Command utilized for calculate Cremèr's V

## Full Logistic Regression Model

Following the individual hypothesis tests, a multiple logistic regression model was developed to assess the combined effects of age, BMI, smoking history, and gender on diabetes status. The purpose of this full model was to determine how these factors jointly influence the likelihood of developing diabetes.

In this model, diabetes status (0 = No, 1 = Yes) was the dependent variable. Age and BMI were treated as continuous predictors, while smoking history and gender were treated as categorical variables. The model was fitted using the glm() function in R with a binomial family link.

The model’s output was summarized to evaluate the significance of each predictor, and odds ratios were calculated by exponentiating the regression coefficients to aid interpretation. Predicted probabilities of diabetes were generated for each individual based on the fitted model. Using a classification threshold of 0.5, the predicted probabilities were converted into binary predicted classes. Model performance was evaluated by constructing a confusion matrix comparing predicted and actual diabetes status, and by calculating the overall prediction accuracy.

This approach provided insight into the relative importance of multiple predictors in explaining diabetes risk.

# Results and Discussion

This chapter presents the findings of the statistical analyses conducted to explore factors associated with diabetes.

## Exploratory Data Visualization

To gain an initial understanding of the dataset, the distribution of each variable was explored through appropriate plots. This step helped identify patterns, outliers, and the general behavior of each parameter.

### Distribution of BMI

A graph of a frequency

AI-generated content may be incorrect.

Figure : Histogram of BMI data available in the dataset

According to the histogram Figure 9 the BMI values ranged broadly across the population, with most individuals falling between approximately 25 and 35. This suggests that a large portion of the sample population is either overweight or obese, which is relevant for diabetes risk analysis.

### Distribution of age

The following figure shows the histogram of the ages of persons involved in the study.

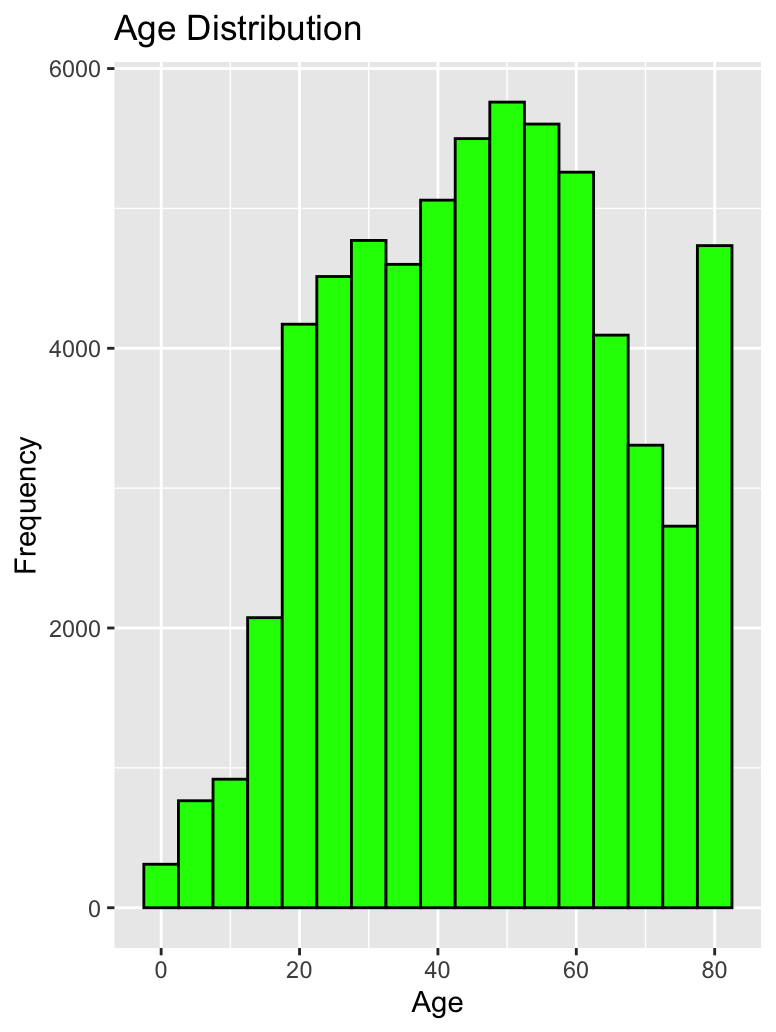


Figure : Histogram of ages of participants

According to the age histogram the distribution was skewed toward middle-aged and older adults, with a noticeable concentration of individuals in the 40–70 age range. This reflects the typical demographic most at risk for developing diabetes.

### Gender

According to the gender distribution of the participants, approximately 60% were female and 40% were male. This reflects a relatively balanced contribution from both genders, ensuring fair representation for gender-based comparisons

A graph of a bar graph

AI-generated content may be incorrect.

Figure : Gender distribution

### Smoke history

Another parameter was the smoking history of the participants, which was recorded under five categories: current, former, ever, never, and not current. The data distribution for each category is shown in the following bar graph.

A graph with different colored bars

AI-generated content may be incorrect.

Figure : Graph of smoking history of participants

Based on the graph, the majority of individuals were classified as “never” smokers, followed by “former” and “current” smokers. Less common categories such as “ever” and “not current” were also present in the dataset.2.1 Hypothesis 1: Higher BMI is associated with an increased likelihood of diabetes

### Diabetes Status

A graph of a number of people

AI-generated content may be incorrect.

Figure : Distribution of diabetes status in the cleaned dataset (0 = Non-diabetic, 1 = Diabetic)

The bar chart shows (Figure 13) the distribution of diabetes status among individuals in the cleaned dataset. A slightly higher number of individuals are non-diabetic (coded as 0), while a significant portion of the population is diabetic (coded as 1). This balanced distribution allows for reliable comparative analysis and supports the application of both statistical tests and predictive modeling in the study.

## Results of tested hypothesis

### Hypothesis 1: Higher BMI is associated with an increased likelihood of diabetes

A simple logistic regression was conducted to evaluate the association between BMI and diabetes status. The model demonstrated a notable improvement in fit, with the residual deviance (42,113) being lower than the null deviance (44,418). The BMI variable was statistically significant (p < 0.001), indicating a strong association with diabetes. The odds ratio for BMI was 1.086, suggesting that for every one-unit increase in BMI, the odds of having diabetes increased by approximately 8.6%. These results support the hypothesis that higher BMI is associated with an increased likelihood of developing diabetes.

### Hypothesis 2: Smoking history affects the risk of diabetes

A chi-square test of independence was conducted to examine the relationship between smoking history and diabetes status. The test produced a chi-square statistic of 430.57 with 4 degrees of freedom and a p-value less than 2.2e-16, indicating a statistically significant association between smoking history and diabetes. This result supports the hypothesis that smoking history affects the risk of diabetes. However, the effect size measured by Cramér's V was 0.082, which indicates a small practical effect. While smoking history is statistically linked to diabetes, the strength of this association in practical terms is relatively weak.

### Hypothesis 3: Older individuals have a higher risk of diabetes

The analysis showed a statistically significant relationship between age and diabetes status (p < 0.001). The model’s deviance decreased from 44,418 to 39,726, indicating a better fit when age was included as a predictor. The odds ratio for age was 1.050, meaning that each additional year of age increased the odds of having diabetes by approximately 5%. These results support the hypothesis that older individuals are more likely to develop diabetes.

### Hypothesis 4: Diabetes prevalence differs by gender

The results showed a statistically significant difference in diabetes prevalence between males and females (p < 0.001). Among females, 3,707 out of 38,850 individuals had diabetes, while among males, 3,339 out of 25,318 individuals were diabetic. The association was statistically significant, but the effect size, measured by Cramér’s V (0.057), was small. This suggests that while gender is associated with diabetes status, the strength of the relationship is limited in practical terms.

## Full Model: Age, BMI, Smoking History, and Gender as Predictors of Diabetes

The full logistic regression model assessed the combined influence of age, BMI, smoking history, and gender on diabetes status. All predictors were statistically significant (p < 0.05), indicating meaningful contributions to the model. Age and BMI were both positively associated with diabetes; for each additional year of age, the odds of having diabetes increased by 5.6% (OR = 1.056), and for each unit increase in BMI, the odds increased by 9.8% (OR = 1.098).

Compared to current smokers, all other smoking history categories showed lower odds of diabetes, with “not current” smokers having the lowest odds (OR = 0.75). Gender was also a significant factor: males had higher odds of diabetes than females (OR = 1.45).

The model showed a strong overall fit, with the residual deviance reduced to 37,148 (from a null deviance of 44,418), and an AIC of 37,164. The classification accuracy was 89.1%, indicating that the model correctly predicted diabetes status in the majority of cases.

## General discussion

The findings of this study align well with existing medical literature, reinforcing the well-established link between higher body mass index, older age, and the risk of developing diabetes. These factors likely reflect underlying metabolic and physiological changes that accumulate with age and excess body weight. The statistical significance and strong predictive power of these variables in both individual and multivariable models highlight their reliability as indicators of diabetes risk in population-level data.

Although smoking history and gender were also found to be significantly associated with diabetes, their effect sizes were comparatively small. This suggests that while these variables contribute to risk, they do so more modestly than age and BMI. Interestingly, the full logistic regression model revealed that even after controlling for age and BMI, smoking status and gender retained independent predictive value. This underscores the multifactorial nature of diabetes and the need for holistic approaches to prevention, which consider lifestyle, biological, and demographic factors together.

# Conclusion

This study investigated the association between several individual characteristics—namely age, body mass index (BMI), smoking history, and gender—and the risk of diabetes using a cleaned health dataset. The analysis began with exploratory data visualization to understand the distribution and behavior of each variable, followed by hypothesis testing and a full logistic regression model to assess both individual and combined effects.

The hypothesis tests revealed that higher BMI and older age were significantly associated with increased diabetes risk. Smoking history and gender were also found to be associated with diabetes status, though their practical effect sizes were relatively small. The full logistic regression model confirmed that all four variables were statistically significant predictors of diabetes. Notably, the model showed that each year of age and each unit increase in BMI raised the odds of diabetes, while males and current smokers were more likely to have diabetes compared to their counterparts. The model achieved a high classification accuracy of 89.1%, demonstrating strong predictive power.

Overall, the findings highlight the importance of age and BMI as primary risk factors for diabetes, while also recognizing the influence of smoking behavior and gender. These insights can inform public health strategies aimed at early identification and prevention of diabetes in at-risk populations.

# Appendix

* The dataset can be found at: <https://www.kaggle.com/datasets/ziya07/diabetes-clinical-dataset100k-rows/data>
* The R code is available at: [https://github.com/MPSVIRAJ/r-diabetes-insights#](https://github.com/MPSVIRAJ/r-diabetes-insights)