Nguyen, Trang
Gomez, Maria
Farooqui, Numayr
Chithajhallu, Sai Siva Deekshita
Group 1
BDA 610: Advanced Business Statistics

Impact of Smoking on Diabetes Prediction Using Logistic Regression and Propensity Score Methods

1. Introduction

Diabetes is a global epidemic affecting millions of people worldwide. It is a chronic disease when glucose accumulates in the bloodstream due to the pancreas not producing enough insulin or the body's inability to utilize it effectively. The frequent episodes of hyperglycemia are a common effect of uncontrolled diabetes [1]. Over time, they can lead to serious health problems like kidney disease, heart disease, and vision loss.

The primary objective of this study is to investigate the impact of smoking on the prediction of diabetes, as well as other factors that may contribute to the prediction of diabetes. This investigation will involve the use of logistic regression and propensity score analysis. To what extent does smoking behavior impact the prediction of diabetes incidence, and are there other factors that contribute to the prediction of diabetes? How do logistic regression and propensity score analysis help to understand the relationship between predictors and the incidence of diabetes? By analyzing these relationships, we aim to develop a predictive model to assist healthcare professionals in early detection and intervention.

2. Literature Review

Regression analysis is a powerful and efficient method used in clinical studies. Many researchers have used this method to examine whether clinical and lifestyle factors correlate with the likelihood of individuals developing diabetes.

Gray, Picone, Sloan, and Yashkin conducted a study on the prediction of risk factors for diabetes using a dataset consisting of 14,657 patients from the Medicare Current Beneficiary

Survey (MCBS) from 1991 to 2010. The study used a proportional hazards regression model, the Cox regression model, to examine the relationship between BMI and the probability of developing type 2 diabetes mellitus (DM). Their findings showed that obesity and high BMI continuously raised the risk of diabetes and associated consequences. Women with a BMI of 40 or higher had a significantly increased chance of developing insulin dependency (HR = 3.57) compared to those with a BMI of 25 to 27.49 (HR = 1.77). In the BMI range of 27.5 to 29.99, men were more likely to develop diabetic complications. The study also found that women and men with BMIs between 30 and 39.9 had a higher risk of cardiovascular disease associated with obesity and an increased risk of developing diabetes [2].

Another study conducted by Aeschbacher et al. utilized a dataset of 2,142 participants to examine the association between smoking and diabetes using a multivariable logistic regression model. The authors found that individuals diagnosed with type 2 diabetes mellitus had a higher prevalence of current smokers and high blood pressure. Smoking cigarettes directly has a negative impact on the body's ability to use insulin properly and increases insulin resistance [3].

A study in Bangladesh that used information from the 7,535 participants in the 2011 BDHS discovered a significant DM prevalence of 33.3% in the 50-54 age range. Age was a significant factor in increasing the likelihood of developing DM. The study showed that females tended to develop DM stronger than males. Furthermore, hypertension and diabetes were correlated, increasing the risk by 63%. Additionally, being overweight nearly doubles the risk of developing diabetes. The research indicated that high levels of physical activity reduced the risk of DM. This demonstrates the requirement for focused policy initiatives to combat diabetes in Bangladesh [4].

Another researcher found that diabetes and hypertension influenced each other, with insulin resistance and weight gain playing a role. The development of hypertension from normal blood pressure involves a significant and rapid increase in blood pressure values. Diabetic individuals with borderline blood pressure should be closely monitored, and specific antidiabetic drugs that lower blood pressure may benefit them [5].

3. Data Descriptions and Visualization Analysis

3.1 Data Collection

There is only one dataset involved in this study. The dataset is sourced from Kaggle and titled 'Diabetes Prediction Dataset.' The file is downloaded in .xlsx format and has a size of 751 KB. It includes data from 100,000 patients, encompassing medical and demographic details of

individuals diagnosed with diabetes or those at risk. This information is sourced from surveys, medical records, and laboratory tests, capturing medical history, demographic details, and lifestyle factors.

3.2 Data Cleaning and Preprocessing

Before conducting the study, data cleaning and pre-processing are crucial steps in analyzing the data, which involves removing missing values, inconsistencies, and errors to ensure data accuracy. The study's dataset comprises nine variables, including gender, age, hypertension, heart disease, smoking history, BMI, HbA1c level, and blood glucose level. These variables are crucial for the study's analysis. The dependent variable in this study is diabetes, represented as a binary variable. The dataset contains independent variables that consist of both qualitative and quantitative attributes. Quantitative variables are characteristics that can be measured using numeric values and are classified as either discrete or continuous variables. The quantitative variables examined in this study consist of age, hypertension, heart disease, BMI, HbA1c, blood glucose level, and diabetes. On the other hand, qualitative or categorical variables are descriptive attributes that can be categorized as either ordinal or nominal. In this study, nominal categories are gender and smoking history. The Data Dictionary of the dataset is presented in Table 1.

3.3 Exploratory Data Visualizations

3.3.1 Data Visualizations of Continuous Variables

Table 2 presents summary descriptive statistics for four continuous variables in the dataset, including age, BMI, HbA1c level, and blood glucose level. The table provides insights into these variables' central tendencies, variabilities, and distributions. The study examined 100,000 observations of ages ranging from 0.08 to 80 years old, with an average age of around 41.89 years old. In this study, the average BMI is approximately 27.32, ranging from a minimum of 10.01 to a maximum of 95.69. The HbA1c levels range from a minimum of 3.500 to a maximum of 9.000, with a median of 5.800 and an average of about 5.528. The blood glucose levels in this study range from a minimum of 80.0 to a maximum of 300.0, with a median value of 140.0. All these variables have a normal distribution confirmed by the Anderson-Darling Normality Test with p-values less than 2.2e-16, displayed in Table 3. The normal distribution of these variables in the data is visualized in the Q-Q plots in Figure 1.

3.3.2 Data Visualization of Categorical Variables

Bar charts were employed to visualize the frequency of each category in the dataset. According to Table 4 and Figure 2, a total of 4,461 females and 4,039 males are diagnosed with diabetes. However, the gender-based difference is not statistically significant. Among individuals with hypertension, 2,088 have diabetes and hypertension, while 5,397 have hypertension without diabetes. Within the individuals with the heart disease group, 2,675 people do not have diabetes, while 1,267 people have both conditions. In the category of current smokers, 8,338 individuals are without diabetes, while 948 individuals have diabetes. Out of the total number of former smokers, 7,762 individuals do not have diabetes while 1,590 individuals have diabetes. Among the non-smoker group, 31,749 individuals do not have diabetes, and 3,346 individuals have diabetes.

3.4 Data Dimension Reduction.

Table 3 and Figure 3 show a slightly higher prevalence of diabetes in females than males. However, the gender-based difference is not statistically significant. Therefore, gender is not a statistically significant factor in diabetes prevalence, and it is included in data dimension reduction. In addition, Figure 3 demonstrates a positive correlation between diabetes and elevated levels of HbA1c and blood glucose. These indicators are crucial for predicting long-term diabetes complications [6]. Since HbA1c and blood glucose levels are measures of diabetes severity, they have been omitted from the dataset in this study. Furthermore, many research articles reveal that people with diabetes are more likely to have other health conditions that increase their health risks, such as high blood pressure and heart disease [7]. High blood pressure increases the force of blood through the arteries, potentially damaging artery walls and significantly increasing the risk of heart disease [8]. Diabetes is a risk factor for hypertension and heart disease, but not vice versa. As a result, hypertension, and heart disease, along with gender, HbA1c, and blood glucose, are excluded from the dataset in this study.

According to Figure 4, outliers are identified in the BMI boxplot. However, they are retained in this analysis for further investigation. We aim to conduct a study to explore the impact of outliers on this research and their potential connections with the variables, particularly their association with diabetes.

4. Determining Econometric Tasks

In this study, the primary econometric task is to understand and predict the impact of smoking on diabetes incidence, as well as other factors that may contribute to the prediction of diabetes. Our approach to modeling the relationship will involve Logistic Regression and

Propensity Score Methods. Before conducting the study, several articles are reviewed to ensure the rigorous modeling approach aligns with contemporary research. Ram and Chandra, authors of the article titled 'Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches,' state that the Logistic Regression Model is a powerful and efficient method for making appropriate predictions of type 2 diabetes. Their analysis identifies BMI and age as the main predictors of type 2 diabetes. The study's findings illustrate the proposed model exhibiting an accuracy of 78.26% and an associated cross-validation error rate of 22.86%. The model can predict the likelihood of type 2 diabetes with a relatively low error rate [9].

Another study conducted by Sia et al. examined the association between smoking and glycemic control in a population of 3,044 individuals with type 2 diabetes in Taiwan from 2002 to 2017. Glycemic control refers to the optimal glucose concentration in diabetic patients. The study employed the propensity score method, which was calculated using non-parsimonious multivariable logistic regression. In a 1:1 matching ratio, the study matched 757 smokers with 757 non-smokers. The study highlights that individuals who smoke have HbA1c levels that are significantly higher compared to those who do not smoke [10].

According to Figure 5, the correlation matrix identifies linear relationships and multicollinearity among multiple variables in the diabetes dataset. BMI, current smoker, and former smoker have high correlation values with diabetes compared to other variables. In our study, for further research, the econometric task is centered on predicting the onset of diabetes based on these multiple predictors. With the guidance of insights from the literature, our modeling approach will employ the logistic regression technique and propensity score method to optimize our dataset's characteristics and address our research questions.

5. Data Methodologies and Techniques

5.1 Logistic Regression Model

Logistic regression is a statistical technique to analyze the relationship between multiple predictor variables and a binary outcome variable. It is considered a nonlinear regression model. This model uses the cumulative logistic distribution function to characterize the relationship between predictor and binary dependent variables. In this study, the model is utilized to analyze the impact of smoking on the prediction of diabetes and other risk factors on the occurrence of diabetes.

Mathematical Model [11]:

$$\Pr(Y = 1 \mid X_1, X_2, \ldots, X_k = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k = = \frac{1}{1 + e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}}$$

Hypothesis Testing

Null Hypothesis: There is no significant relationship between the predictor variables (age, BMI, current smoker, and never smoker) and the probability of having diabetes.

Alternative Hypothesis: There is a significant relationship between at least one of the predictor variables (age, BMI, current smoker, or never smoker) and the probability of having diabetes. Hypothesis testing and testing are conducted as part of the study's research approach to confirm a diagnosis and explore other relevant factors.

5.2 Propensity Score Matching Method (PSM)

Propensity Score Matching (PSM) is a statistical method used to estimate causal treatment effects by equating groups based on a set of observed covariates that might affect the treatment decision. In observational studies, where a random assignment of treatments is impossible, PSM reduces selection bias by matching treated and non-treated units with similar propensity scores. This study's main objective is to investigate the association between smoking and the risk of developing diabetes, and other factors in predicting the incidence of diabetes. Our observational dataset presents a unique challenge. The decision of whether to smoke or not isn't random but may be influenced by various observed and unobserved factors. This non-randomness in the 'treatment' (smoking status) can lead to biases in estimating its effect on the outcome (onset of diabetes).

We turned to Propensity Score Matching (PSM) to tackle this inherent bias. Specifically, we designated smoking status as the treatment variable. Data was classified into two categories: current smoker (1) and never smoker (0). To maintain clarity, we excluded other smoker categories from our analysis.

Propensity Score Matching was employed to match current smokers with non-smokers using their propensity scores. These scores represent the predicted probability of being a current smoker, and they were determined based on observed covariates such as age, BMI, and other factors. This method created a balanced comparison group, effectively simulating a randomized experiment. This helped us to reduce the impact of confounding variables and obtain a more precise estimation of the treatment effect. As a result, we could draw more robust and unbiased

conclusions regarding the significant association between smoking and the risk of developing diabetes.

6. Results

6.1 Predicting factors associated with diabetes incidence using logistic regression analysis

Summary of Logistic Regression Model for Diabetes Incidence is presented in Table 5. Six iterations of the Fisher Scoring optimization algorithm were conducted to estimate the model parameters and achieve the best fit for the observed data in this study. Table 5 shows the statistical significance of factors in predicting diabetes incidence, including age, BMI, and smoking status (current, former, and never), most of which have a p-value (< 2e-16 ***).

Among these variables, current smokers and former smokers stand out as having the strongest associations with diabetes. For every one-unit increase in being a current smoker, there is approximately a 0.64 increase in the probability of having diabetes. Similarly, for every one-unit increase in being a former smoker, there is approximately a 0.50 increase in the probability of having diabetes. Table 5 displays a positive coefficient for non-smokers, implying a statistical correlation with an increased likelihood of contracting diabetes. This suggests that other factors may be contributing to the increased probability of developing diabetes.

Both age and BMI also have statistically significant associations with diabetes, but they are not as strongly associated as the other factors, as indicated in Table 5.

It is concluded that current and former smoking status are significantly and positively associated with the likelihood of having diabetes. As the presence of current or former smoking increases, the likelihood of having diabetes also increases.

6.2 Predicting factors associated with diabetes incidence using Propensity Score Matching (PSM)

The matching process yielded a balanced distribution of covariates between current smokers (treated group) and never smokers (control group). The summary statistics post-matching is presented in Table 6 and Figure 6, which indicate that the means of essential covariates like age and BMI are almost identical between the two groups. This balance strengthens the reliability of our subsequent analyses.

We conducted a logistic regression to determine whether smoking status, the primary predictor, was associated with the presence or absence of diabetes. The model was trained on the matched dataset. The outcome was diabetes, and the main predictor was the treatment variable

(smoking status). The logistic regression analysis results are presented in Table 8 with a coefficient of 0.01299 for the treatment. This positive coefficient suggests that current smokers have marginally elevated odds of being diagnosed with diabetes compared to those who have never smoked. However, this effect is negligible. The p-value associated with the treatment effect is 0.789. This value exceeds the conventional significance threshold of 0.05, denoting that the effect of being a current smoker on the odds of developing diabetes isn't statistically significant after adjusting for covariates.

7. Conclusion

Current smokers and former smokers are the most strongly associated factors in the logistic regression model for predicting diabetes. However, in the propensity score method, being a current smoker doesn't significantly influence the probability of developing diabetes compared to non-smokers.

This suggests that other factors also contribute to diabetes, and further research may be needed to clarify the true relationship between smoking and diabetes. Smoking has been linked to several mechanisms that increase the risk of developing type 2 diabetes, such as insulin resistance, chronic inflammation, changes in fat distribution, direct effects on the pancreas, cellular disruptions, increased belly fat, lifestyle factors, and hormonal disruptions. Each of these mechanisms offers a pathway through which smoking can influence diabetes risk [12].

Appendix

	Variable Name	Description	Data Type
1	gender	Gender of the individual (Female, Male)	Nominal Category
2	age	Age of the individual	Continuous
3	hypertension	Presence of hypertension (0 = no, 1 = yes)	Discrete
4	heart_disease	Presence of heart disease (0 = no, 1 = yes)	Discrete
5	smoking_history	History of smoking (never, current, no info)	Nominal Category
6	bmi	Body Mass Index (BMI)	Continuous
7	HbA1c_level	HbA1c level (average blood sugar over the past 3 months)	Continuous
8	blood_glucose_level	Blood glucose level	Continuous
9	diabetes	Diabetes status (0 = no, 1 = yes)	Discrete

Table 1: Data Dictionary Table of 9 Contributors to Diabetes Prediction Analysis

Descriptive Statistics	Age	BMI	HbA1c_level	Blood_glucose_level
Min	0.08	10.01	3.500	80.0
1st Qu	24.00	23.63	4.800	100.0
Median	43.00	27.32	5.800	140.0
Mean	41.89	27.32	5.528	138.1
3rd Qu	60.00	29.58	6.200	159.0
Max	80.00	95.69	9.000	300.0

Table 2. Descriptive Statistics of 4 Continuous Variables in Datasets

Variables	A	p-value
Age	610.11	< 2.2e-16
BMI	2366.3	< 2.2e-16
HbA1c_level	2533	< 2.2e-16
Blood_glucose_level	2556.6	< 2.2e-16

Table 3. Anderson-Darling Normality Test

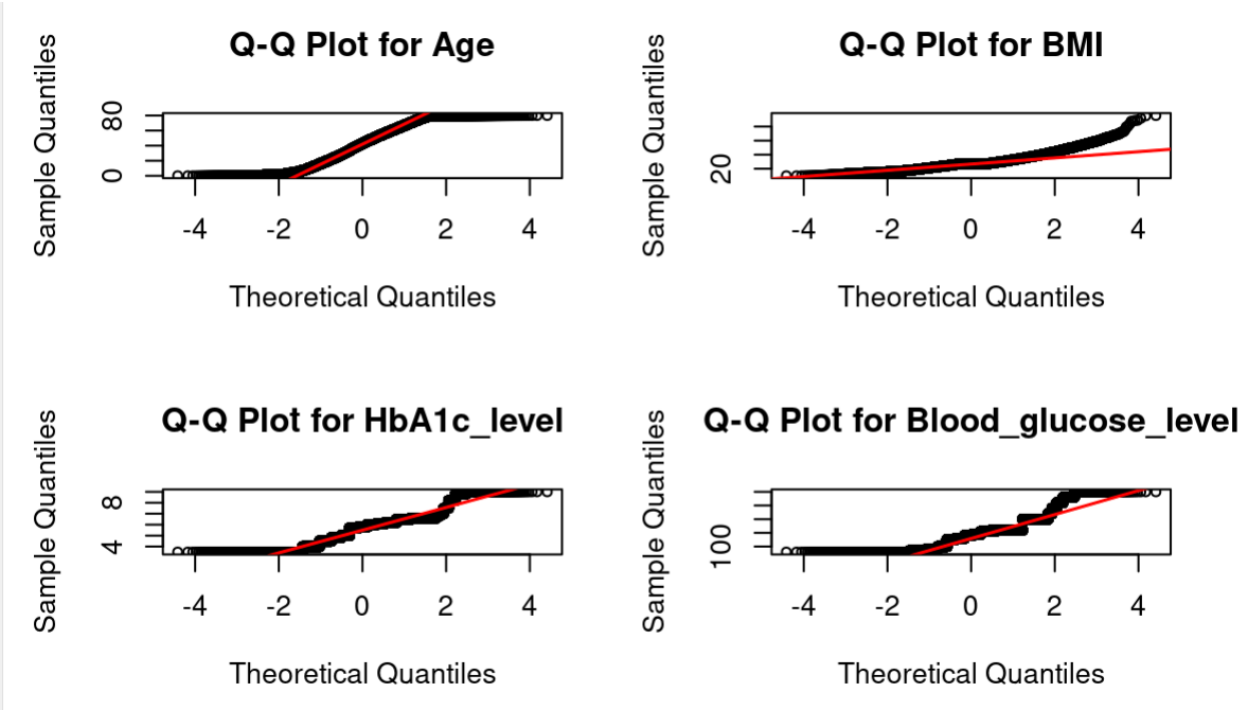

Figure 1. Q-Q Plots of 4 Continuous Variables in Datasets

	No Diabetes (0)	Diabetes (1)
Female	54091	4461
Male	37391	4039
Hypertension (No)	86103	6412
Hypertesion (Yes)	5397	2088
Heart_disease (No)	88825	7233
Heart_disease (Yes)	2675	1267
Current smoker	8338	948
Ever Smoker	3532	472
Former Smoker	7762	1590
Never Smoker	31749	3346
No Info	34361	1454
Not current Smoker	5757	690
Diabetes	91500	8500

Table 4. Statistic Summary of Health and Lifestyle Characteristics by Diabetes Status

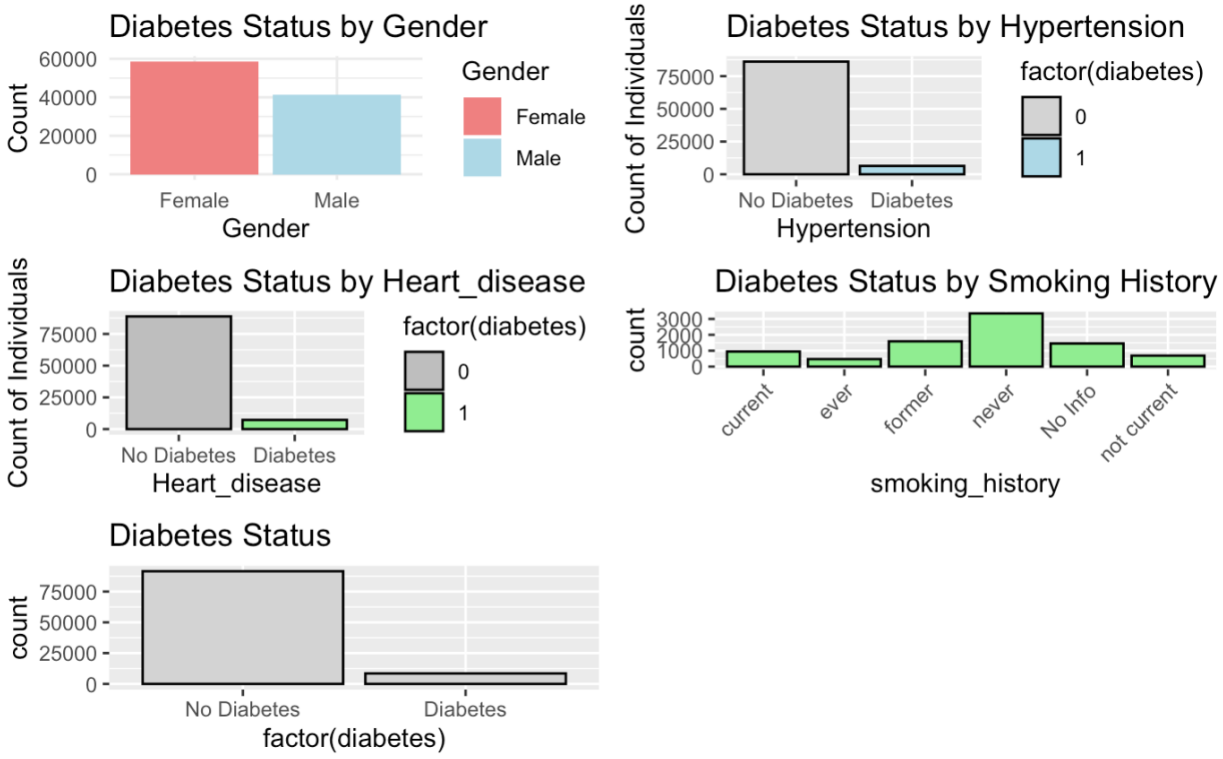

Figure 2. Visualizing distributions of Categorical Variables

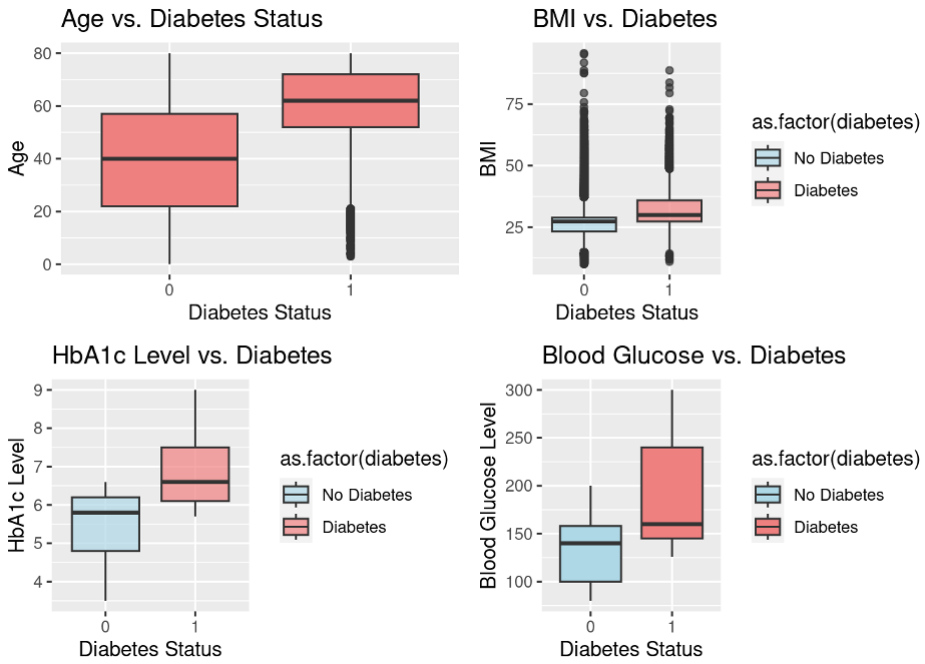

Figure 3. Boxplot Association between Predictors and Response

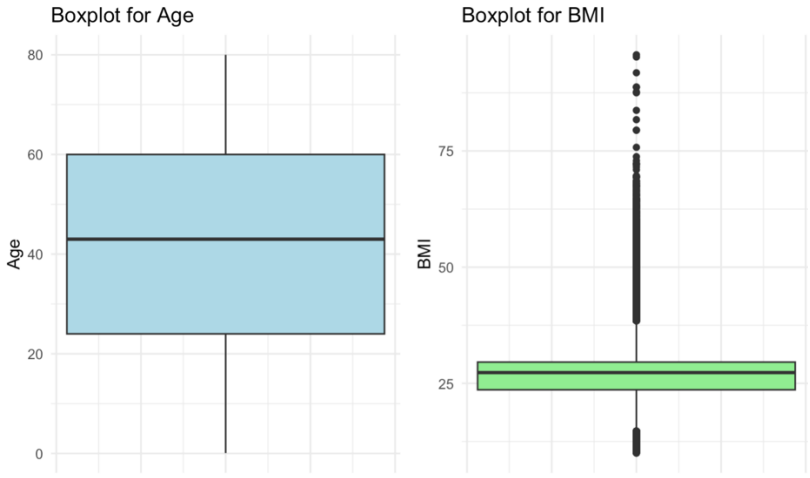

Figure 4. Box Plots of Distribution Age and BMI

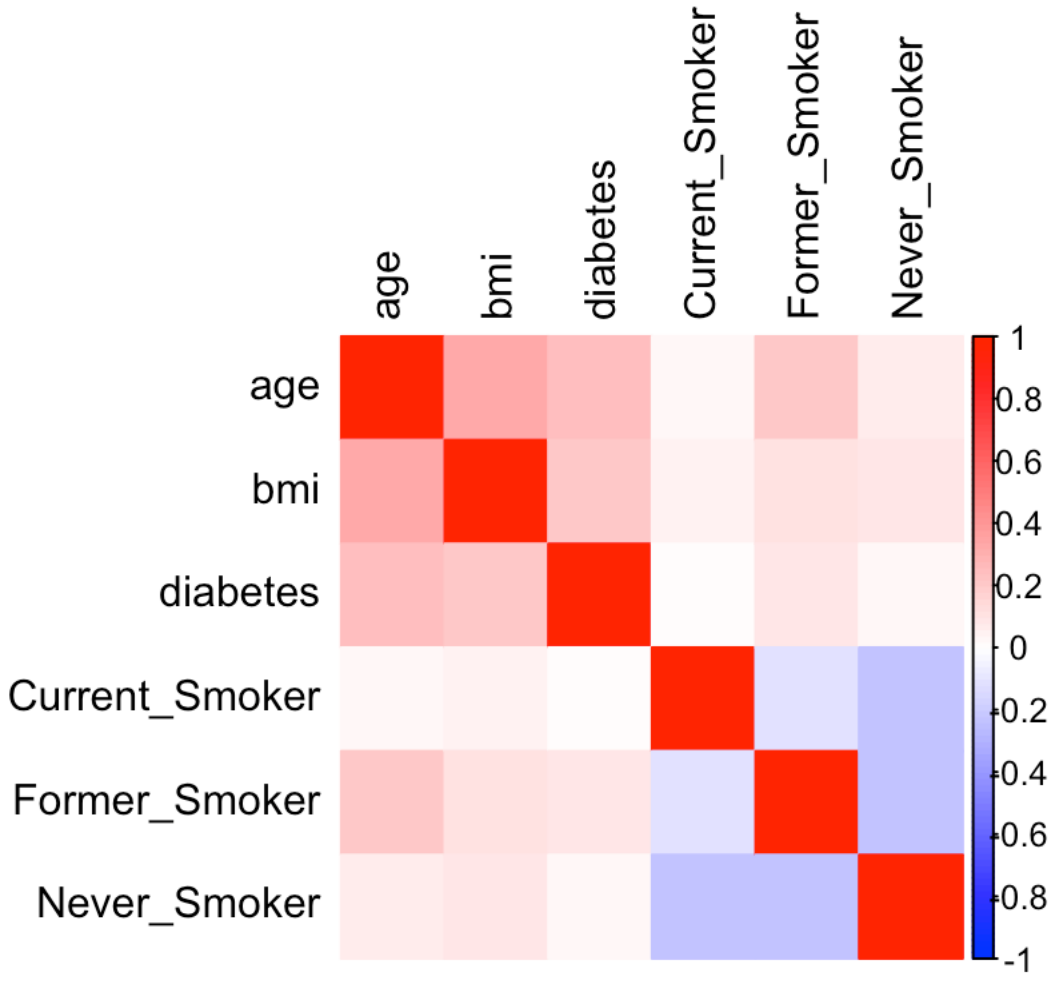

Figure 5. Heatmap Correlation between Predictors and Diabetes

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.18359	0.079304	-103.2	<2e-16 ***
Age	0.053231	0.000773	68.9	<2e-16 ***
BMI	0.092916	0.001707	54.4	<2e-16 ***
Current_Smoker	0.641885	0.042253	15.2	<2e-16 ***
Former_Smoker	0.504912	0.036254	13.9	<2e-16 ***
Never_Smoker	0.355244	0.028881	12.3	<2e-16 ***

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 58163  on 99998  degrees of freedom
Residual deviance: 47461  on 99993  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 47473

Number of Fisher Scoring iterations: 6
```

Table 5. Summary of logistic regression model for diabetes incidence

Variable	Means_Treated	Means_Control	Std_Mean_Diff
distance	0.193	0.075	0.9129
age	60.9466	40.1152	1.4317
bmi	31.9884	26.8872	0.6749
never0	0.6064	0.653	-0.0955
never1	0.3936	0.347	0.0955
current0	0.8885	0.9089	-0.0648
current1	0.1115	0.0911	0.0648

Table 6. Propensity Score Matching Analysis

Sample	Control	Treated
All	91500	8500
Matched	8500	8500
Unmatched	83000	0
Discarded	0	0

Table 7. Sample Size After Propensity Score Matching

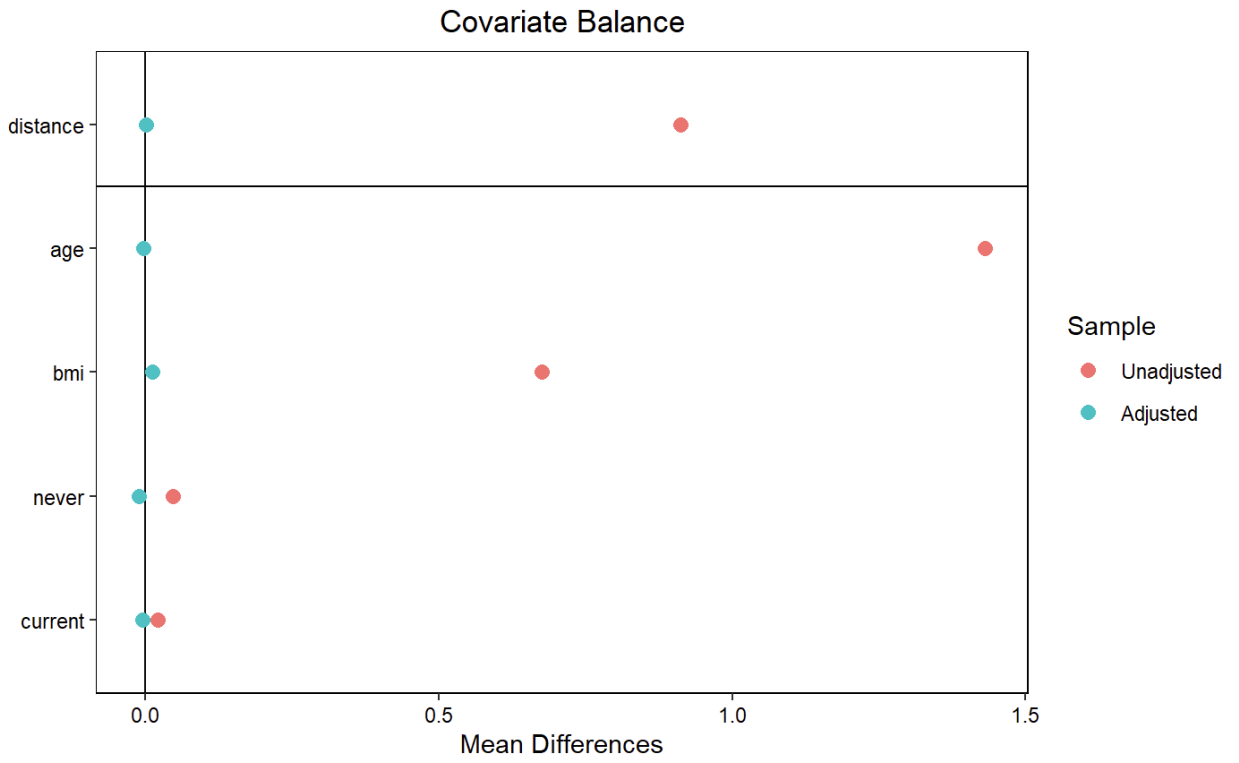

Figure 6. Covariate Balance Propensity Score Matching (PSM)

Variable	Estimate	Standard Error	Z-Value	P-Value
(Intercept)	-2.18721	0.03445	-63.489	< 2e-16
treatment	0.01299	0.04859	0.267	0.789

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 8. Regression Analysis Result: Diabetes Prediction

References

1. Hyperglycemia (high blood glucose). American Diabetes Association. https://www.diabetes.org/healthy-living/medication-treatments/blood-glucose-testing-and-control/hyperglycemia. Accessed July 6, 2022

2. Gray N, Picone G, Sloan F, Yashkin A. Relation between BMI and diabetes mellitus and its complications among US older adults. South Med J. 2015 Jan;108(1):29-36. doi: 10.14423/SMJ.0000000000000214. PMID: 25580754; PMCID: PMC4457375.

3. Aeschbacher S, Schoen T, Clair C, Schillinger P, Schönenberger S, Risch M, Risch L, Conen D. Association of smoking and nicotine dependence with pre-diabetes in young and healthy adults. Swiss Med Wkly. 2014 Oct 8;144:w14019. doi: 10.4414/smw.2014.14019. PMID: 25295968.

4. Rahman, M. A., Zaman, M. M., & Rahman, M. Prevalence and risk factors of type 2 diabetes in Bangladesh: A systematic review. Scientific Reports. 2020;10:1-10. doi: 10.1038/s41598-020-66084-9.

5. Tsimihodimos V, Gonzalez-Villalpando C, Meigs JB, Ferrannini E. Hypertension and Diabetes Mellitus: Co-prediction and Time Trajectories. Hypertension. 2018 Mar;71(3):422-428. doi: 10.1161/HYPERTENSIONAHA.117.10546. Epub 2018 Jan 15. PMID: 29335249; PMCID: PMC5877818.

6. Sherwani SI, Khan HA, Ekhzaimy A, Masood A, Sakharkar MK. Significance of HbA1c Test in Diagnosis and Prognosis of Diabetic Patients. Biomark Insights. 2016 Jul 3;11:95-104. doi: 10.4137/BMI.S38440. PMID: 27398023; PMCID: PMC4933534.

7. Dansinger, M. (2023, May 23). *Diabetes and High Blood Pressure*. WebMD – Better information. Better health. https://www.webmd.com/diabetes/high-blood-pressure

8. Oparil S, Acelajado MC, Bakris GL, Berlowitz DR, Cífková R, Dominiczak AF, Grassi G, Jordan J, Poulter NR, Rodgers A, Whelton PK. Hypertension. Nat Rev Dis Primers. 2018 Mar 22;4:18014. doi: 10.1038/nrdp.2018.14. PMID: 29565029; PMCID: PMC6477925.

9. Joshi RD, Dhakal CK. Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. Int J Environ Res Public Health. 2021 Jul 9;18(14):7346. doi: 10.3390/ijerph18147346. PMID: 34299797; PMCID: PMC8306487.

10. Sia HK, Kor CT, Tu ST, Liao PY, Wang JY. Association between smoking and glycemic control in men with newly diagnosed type 2 diabetes: a retrospective matched cohort study. Ann

Med. 2022 Dec;54(1):1385-1394. doi: 10.1080/07853890.2022.2075559. PMID: 35576130; PMCID: PMC9126565.

11. Stock, J. H., & Watson, M. W. (2018). Regression with a Binary Dependent Variable. In Introduction to Econometrics (4th ed., pp. 355–365). essay, Pearson.

12. Dilworth L, Facey A, Omoruyi F. Diabetes Mellitus and Its Metabolic Complications: The Role of Adipose Tissues. *International Journal of Molecular Sciences*. 2021; 22(14):7644. https://doi.org/10.3390/ijms22147644