ST. XAVIER'S COLLEGE (AUTONOMOUS), KOLKATA

DEPARTMENT OF STATISTICS

NAME: SOHINI MANDAL

ROLL NUMBER: 0481

SUPERVISOR'S NAME: PROF. MADHURA DASGUPTA

TITLE: A STATISTICAL STUDY ON DIABETES USING LOGISTIC REGRESSION

SESSION: 2021-2024

DECLARATION:

I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.

Signature

ACKNOWLEDGEMENT

I have had a lot of support and assistance while conducting my study. I express my gratitude and acknowledgement to my supervisor, Prof. Madhura Dasgupta, for her tremendous cooperation and commitment in helping me formulate my topic and sharing her astute observations.

Additionally, I would like to thank my other professors at St. Xavier's College in Kolkata, specifically Prof. Dr. Ayan Chandra, Prof. Debjit Sengupta, Prof. Dr. Surupa Chakraborty, Prof. Dr. Surabhi Dasgupta, Prof. Pallabi Ghosh, Prof. Dr. Durba Bhattacharya, Prof. Dr. Sancharee Basak, and Prof. Rahul Roy. Their assistance in helping me cultivate a research-oriented mindset has allowed me to finish the project.

It would be unfair to overlook the assistance and encouragement I got from my peer group, whose suggestions I properly considered and included into my study.

Finally, I would want to express my gratitude to my parents for their unwavering support and advice during my undergraduate studies.

CONTENTS

ABSTRACT

"Health is wealth"

Foreseeing the future is crucial to managing the results. Though it is crucial to plan for the future in every aspect of life, early disease identification and treatment before issues occur in the healthcare system make it even more crucial. Diabetes is one of the deadly diseases which is posing a great threat to mankind. Elevated blood sugar levels resulting from inadequate insulin synthesis, reduced insulin action, or both, are hallmarks of diabetes mellitus, a chronic metabolic condition. It includes several types, each with unique etiologies and risk factors, including as type 1 (autoimmune destruction of insulin-producing cells), type 2 (cells resist insulin), and gestational diabetes (pregnancy related insulin resistance).

According to the World Health Organization, diabetes affected 8.5% of persons aged 18 and above in 2014. In 2019, diabetes was the direct cause of 1.5 million fatalities, with 48% occurring before the age of 70. Diabetes is responsible for 460000 renal disease fatalities and accounts for approximately 20% of all cardiovascular deaths. Between 2000 and 2019, age-standardized diabetes mortality rates increased by 3%. Diabetes-related mortality has increased by 13% in low- and middle-income nations. The World Health Organization also estimates that by 2040, the number might reach around 642 million, meaning that one in ten people would have diabetes as a result of an unhealthy lifestyle and inactivity.

Frequent urination, unintentional weight loss, weariness, increased thirst, polyphagia (increased hunger) and impaired vision are possible symptoms.

Diabetes has a complex effect on health, resulting in a series of problems that damage different organ systems. Vascular impairment makes cardiovascular issues common, such as heart disease and stroke. Kidney disease, nerve dysfunction, and diabetic retinopathy frequently result in renal failure, neuropathy, and visual impairment, respectively. The major causes of foot problems, such as ulcers and amputations, are nerve damage and inadequate circulation. Skin conditions, mental health issues, and dental issues add even more to the load. If diabetes is not treated, it can seriously impair quality of life and put overall health and wellbeing at risk. It also hastens the aging process and increases the likelihood of long-term effects. Insulin treatment, medication, and lifestyle changes are usually part of management to lower blood sugar and avoid problems. The global health impact of diabetes necessitates comprehensive strategies for management and prevention.

In light of the aforementioned information, it is critical that we utilize historical data to identify the key characteristics of diabetes in order to implement the required preventative actions. Statistical data also aids the development of intuition regarding real-world results and lends credibility to medical research. We will investigate and make a diagnostic prediction about the presence or absence of diabetes in a person using binary logistic regression and other statistical techniques.

INTRODUCTION

Using statistical tools and data processing techniques, statistical prediction is the act of projecting future events based on previous observations. It involves looking for patterns and connections within the data in order to provide well-informed projections regarding future occurrences or trends.

As an illustration:

- Weather Forecasting: Based on historical weather data, satellite images, and atmospheric conditions, meteorologists utilize statistical models to forecast weather patterns. People may better organize their schedules and make wise judgments with the aid of these forecasts.- Financial Forecasting: Economists and financial analysts use statistical models to forecast stock prices, GDP growth, inflation rates, and other economic variables. These estimates provide guidance for policy and investment decisions.

- Healthcare Predictions: Using statistical models, doctors predict the occurrence of diseases, patient outcomes, and treatment efficaciousness. For example, predictive models can help identify individuals who have a higher risk of developing a certain disease based on their medical history and lifestyle choices.

- Sports analytics: Teams in sports use statistical models to predict player performance, game outcomes, and best practices. For instance, in basketball, models may predict a team's chances of winning by taking into account individual player statistics, home court advantage, and past performance. Through the use of historical data and trend analysis, statistical prediction aids in the estimation of future events or trends in each of these situations.

Francis Galton coined the term "regression." The statistical method known as regression analysis is used to examine how one dependent variable (usually represented by the letter "Y") depends on one or more independent variables (usually represented by the letter "X"). The goal is to predict the value of the dependent variable based on the values of the independent factors and to comprehend how the independent variables affect the dependent variable. Regression analysis models the relationship between the variables by fitting an equation to the observed data. This equation provides the best-fitting line or curve for minimizing the discrepancies between observed and predicted data points. There are different types of regression analysis depending on the nature of the dependent and independent variables: Linear Regression, Logistic Regression, Polynomial Regression and Non-linear Regression.

Any regression model where the response variable is categorical is referred to as logistic regression. There are three kinds of logistic regression models.

- **Binary Logistic Regression:** The response variable in binary logistic regression models can only fall into one of two kinds.
  Few instances are given below.
  Example 1: NBA Draft
  Let us say a data scientist in the field of sports wishes to forecast the likelihood that a certain college basketball player would be selected in the NBA draft using the predictor variables (1) points, (2) rebounds, and (3) assists. The data scientist would choose a

binomial logistic regression model since the response variable has only two possible outcomes: drafted or not drafted.

Example 2: Spam Detection

Let us say a company wishes to determine the likelihood that a certain email is spam using the predictor variables (1) word count and (2) country of origin. The company would choose a binomial logistic regression model since the response variable has just two possible outcomes: spam or non-spam.

- **Multinomial Logistic Regression:** There is no inherent ordering among the categories in multinomial logistic regression models, a kind of logistic regression where the response variable might fall into one of three or more categories. Here are a couple examples:

  Example 1: Political Preference

  Assume a political scientist wishes to forecast the likelihood that a person would support one of the four presidential candidates using the predictor variables (1) yearly income and (2) years of schooling. Since there are more than two possible outcomes (there are four potential candidates) for the response variable and there is no natural ordering among the outcomes, the political scientist would use a multinomial logistic regression model.

  Example 2: Sports Preference

  Suppose a sports analyst wants to use the predictor variables (1) TV hours viewed per week and (2) age to predict the probability that an individual will pick either basketball, football, or baseball as their preferred sport. Since there are more than two possible outcomes (there are three sports) for the response variable, the sports analyst would use a multinomial logistic regression model.

- **Ordinal Logistic Regression:** Ordinal logistic regression models are a type of logistic regression in which the response variable can belong to one of three or more categories and there is a natural ordering among the categories. Here are a few examples:

  Example 1: School Ratings

  Suppose an academic advisor wants to use the predictor variables (1) GPA, (2) ACT score, and (3) SAT score to predict the probability that an individual will get into a university that can be categorized into “bad”, “mediocre”, “good”, or “great.”

  Since there are more than two possible outcomes (there are four classifications of school quality) for the response variable and there is a natural ordering among the outcomes, the academic advisor would use an ordinal logistic regression model.

  Example 2: Movie Ratings

  Let us say a movie reviewer wishes to forecast the likelihood that a particular film will receive a rating between 1 and 10 using the predictor variables (1) total run length and (2) genre. The movie critic would choose an ordinal logistic regression model as there are more than two possible outcomes (ten ratings are conceivable) for the response variable and a natural ordering among the outcomes.

DATA DESCRIPTION AND DATA VISUALIZATION

Here, originally, the Pima Indians Diabetes Dataset is taken from the "National Institute of Diabetes and Digestive and Kidney Diseases" repository.

The dataset aims to use binary logistic regression to diagnose and predict diabetes in patients based on certain diagnostic measures. The choices of these examples were made from a bigger database under a number of restrictions. Specifically, all of the patients at this facility are Pima Indian females who are at least 21 years old.

Before proceeding with binary logistic regression, some descriptive studies are done, followed by some categorical data analysis, and checking for multicollinearity among the predictors. Then after fitting the logistic regression model, a list of statistically significant predictor variables which influence the response is obtained. Confusion matrix is made to check the accuracy of this model.

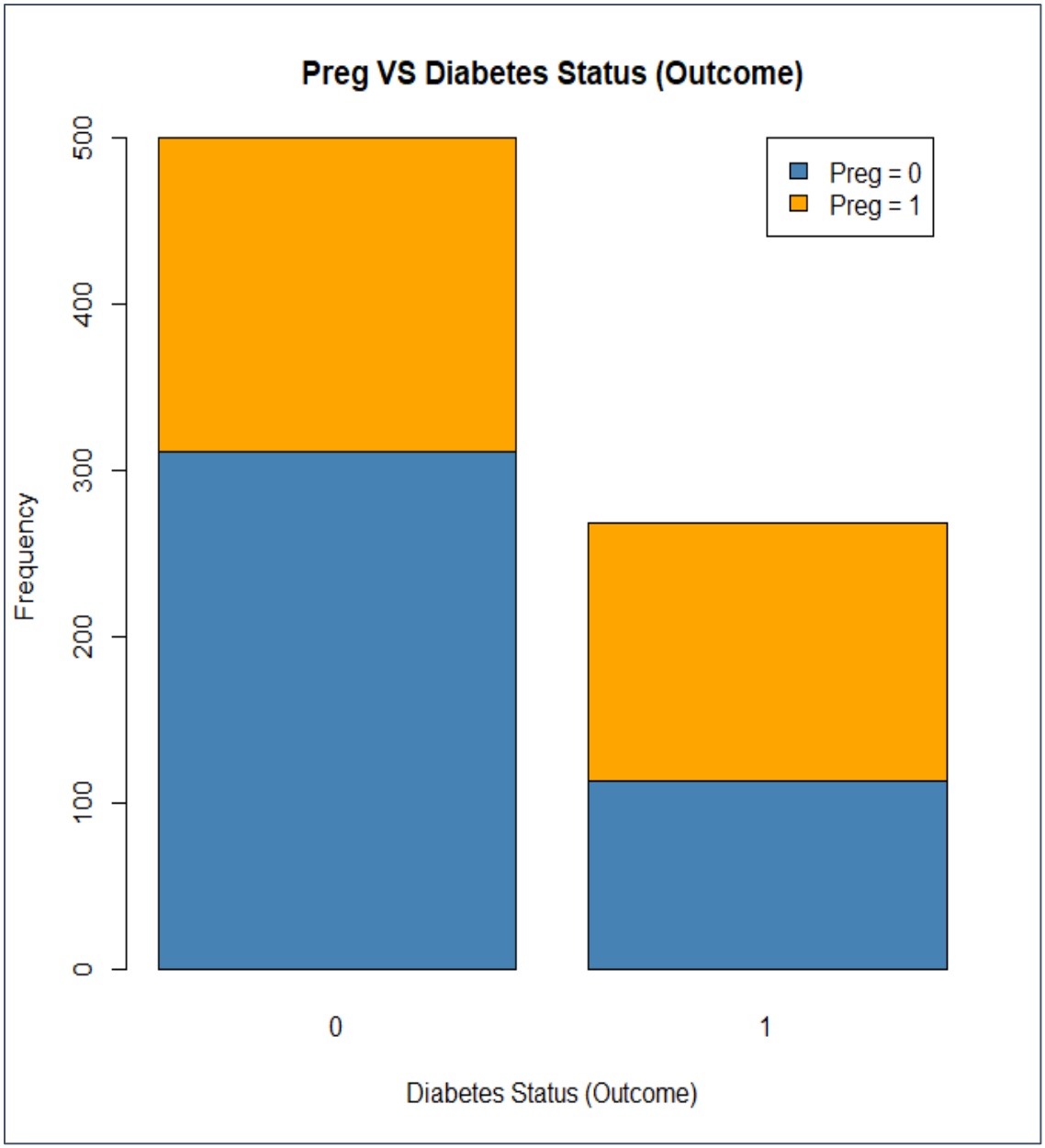
Here, the dataset consists of several (eight) medical independent predictors which are continuous variables except one which is a discrete variable. There is one binary categorical response variable (outcome). 768 patients were tested. Predictors include the number of pregnancies the female has had, their BMI, insulin level, age, glucose level, diastolic blood pressure, skin thickness and diabetes pedigree function.

All the variables are renamed as Y, X_1, X_2, X_3, X_4, X_5, X_6, X_7 and X_8. Description of all the nine variables are given below:

- **Outcome (Y)**
  It is a binary categorical response variable, which takes 0 if non-diabetic and 1 if diabetic.

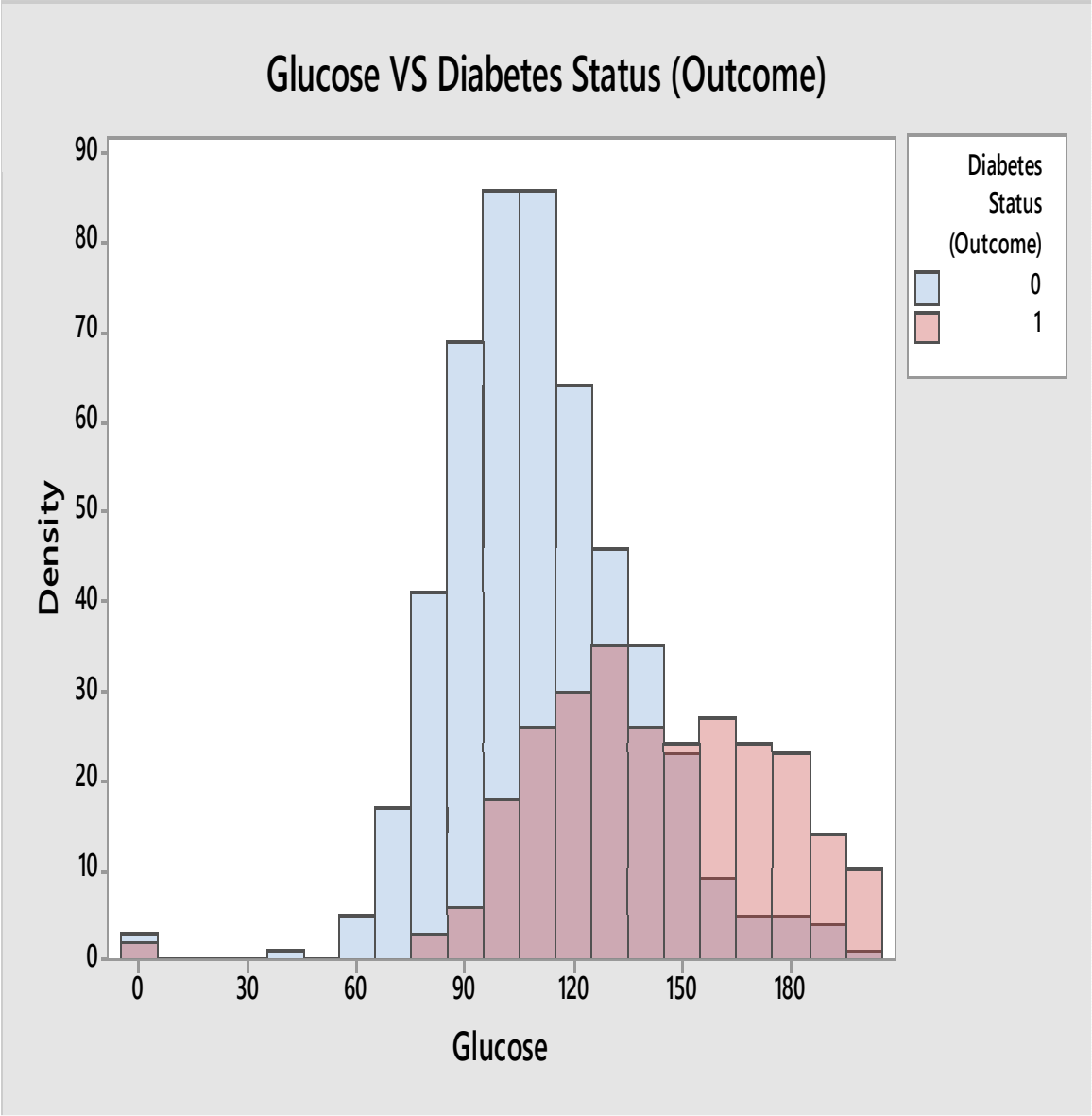
- **Preg (X_1)**
  In the original dataset, we have a variable named "Pregnancies" that indicates how many times the patient was pregnant. A new variable, Preg, is defined. If "Pregnancies" is less than or equal to 3, it takes the value 0, and if not, it takes the value 1. It is a binary categorical variable.

  [*Reason for defining the new variable "Preg"*: In postmenopausal women without a history of gestational diabetes, a statistical study was conducted to determine whether the number of pregnancies during childbearing age was associated with diabetes. The results showed that women who had more than three pregnancies had a significantly higher prevalence of diabetes. The National Health and Nutrition Examination Survey, which was conducted continuously from 1999 to 2014, was their source of data. 9,138 over-40-year-old postmenopausal women without a history of gestational diabetes were chosen for the study. Citation: Lv C, Chen C, Chen Q, Zhai H, Zhao L, Guo Y, Wang N. Multiple pregnancies and the risk of diabetes mellitus in postmenopausal women. Menopause. 2019 Sep;26(9):1010-1015. doi: 10.1097/GME.0000000000001349. PMID: 31453963.]



Comment: From the above subdivided bar diagram, we may say that females who were pregnant for more than 3 times are more prone to developing diabetics than females who were pregnant for less than or equal to 3 times. Opposite result is observed for non-diabetic female individuals.

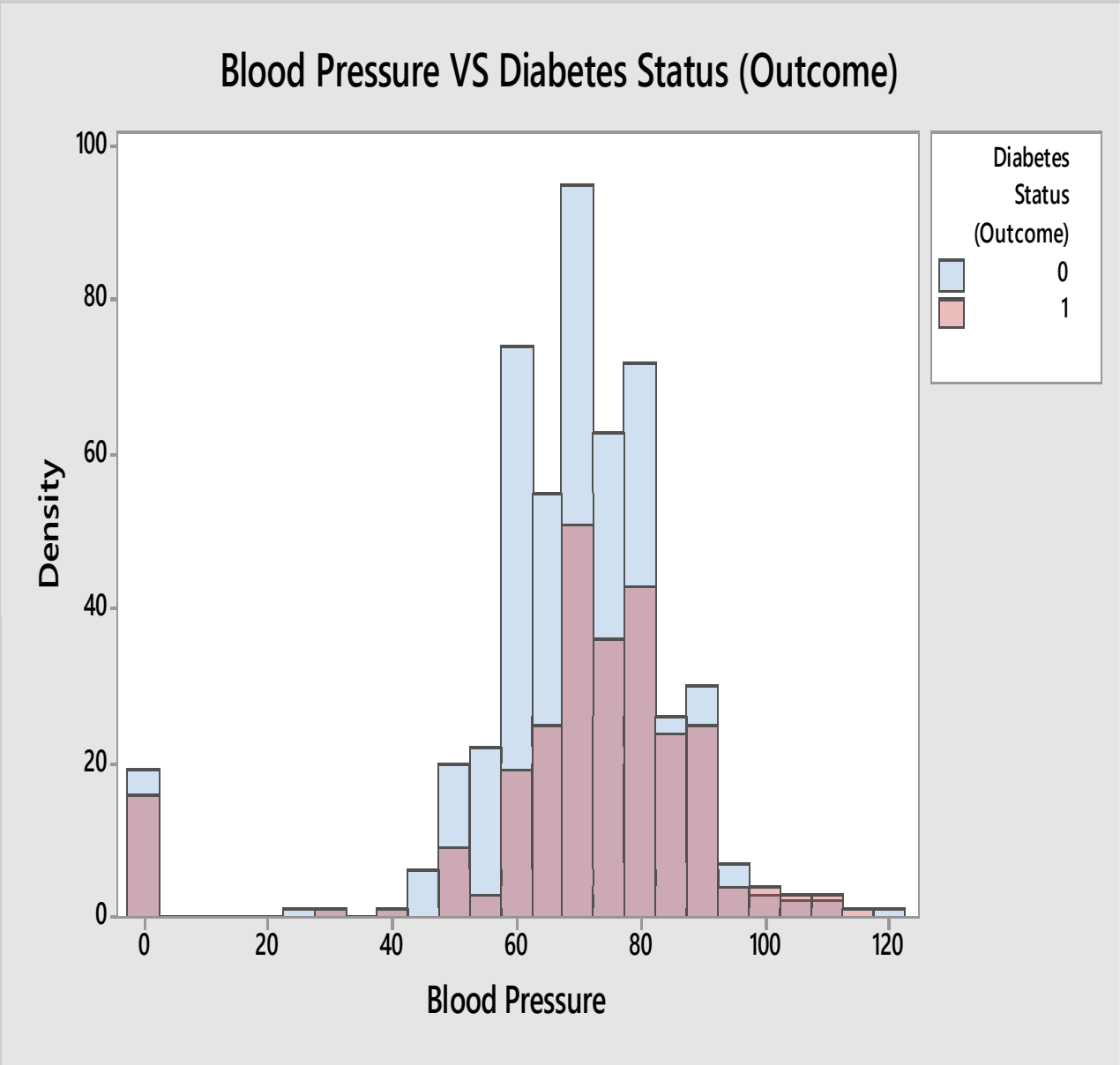
- **Glucose (X_2)**
  Plasma glucose concentration over two hours in an oral glucose tolerance test of the individual measured in mg/dL. Our blood carries glucose to every cell in your body for use in energy. Diabetes is a disease in which the blood sugar level is too high. This covariate is a continuous variable.



Comment: From the graph, it is observed that, on an average, when the level of plasma glucose crosses 145 mg/dL, an individual has a higher chance of being diabetic than she has of not being diabetic. When the glucose concentration, on a average, is between 90 mg/dL and 120 mg/dL, there is a high probability that the person is non-diabetic.

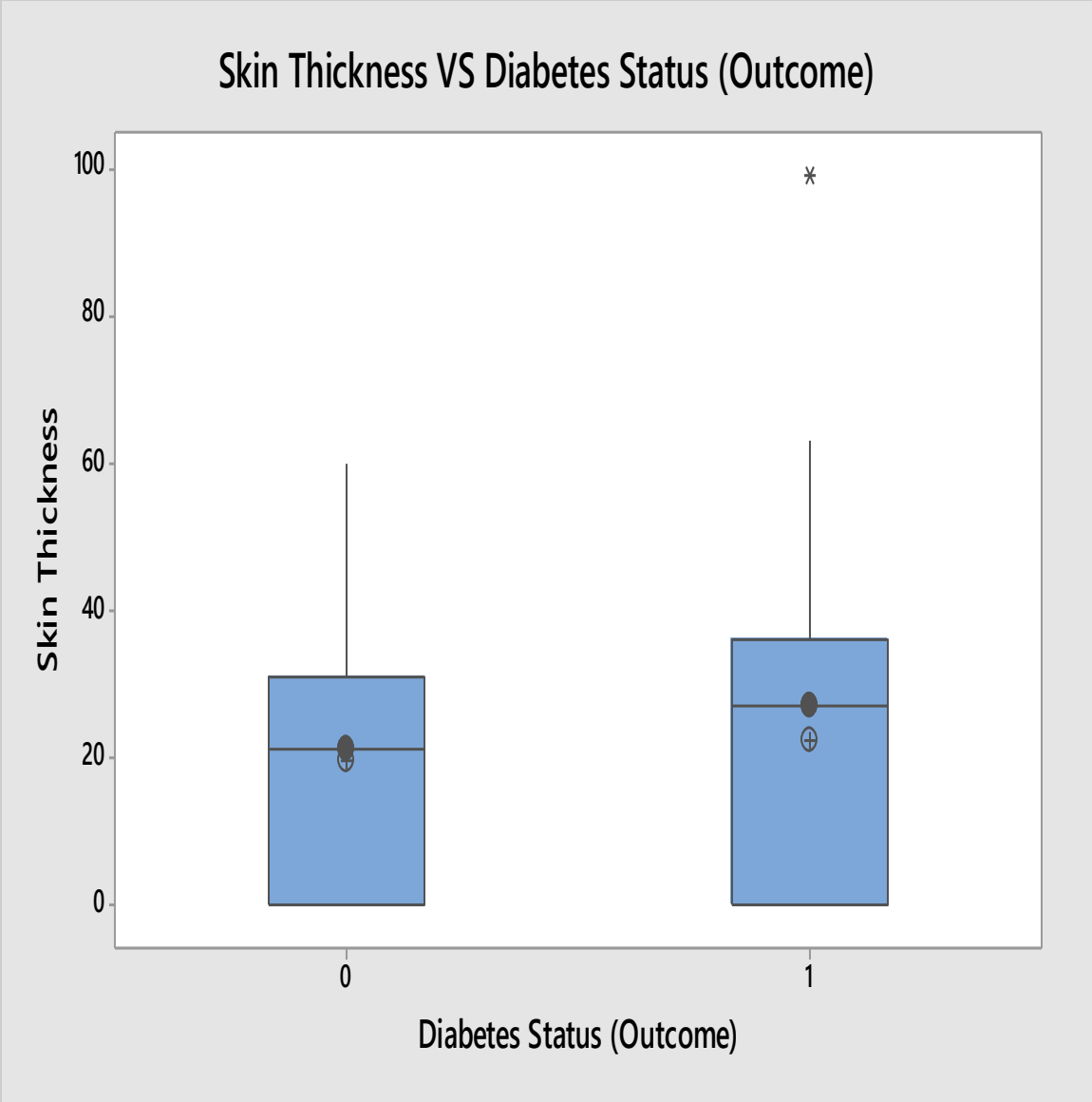
- **BloodPressure (X_3)**

  This predictor displays individual's diastolic blood pressure measured in millimeters of mercury. Diabetes causes damage to our body's tiny blood vessels, stiffening the blood vessel walls in the process. This raises pressure, which causes high blood pressure. It is a continuous variable.



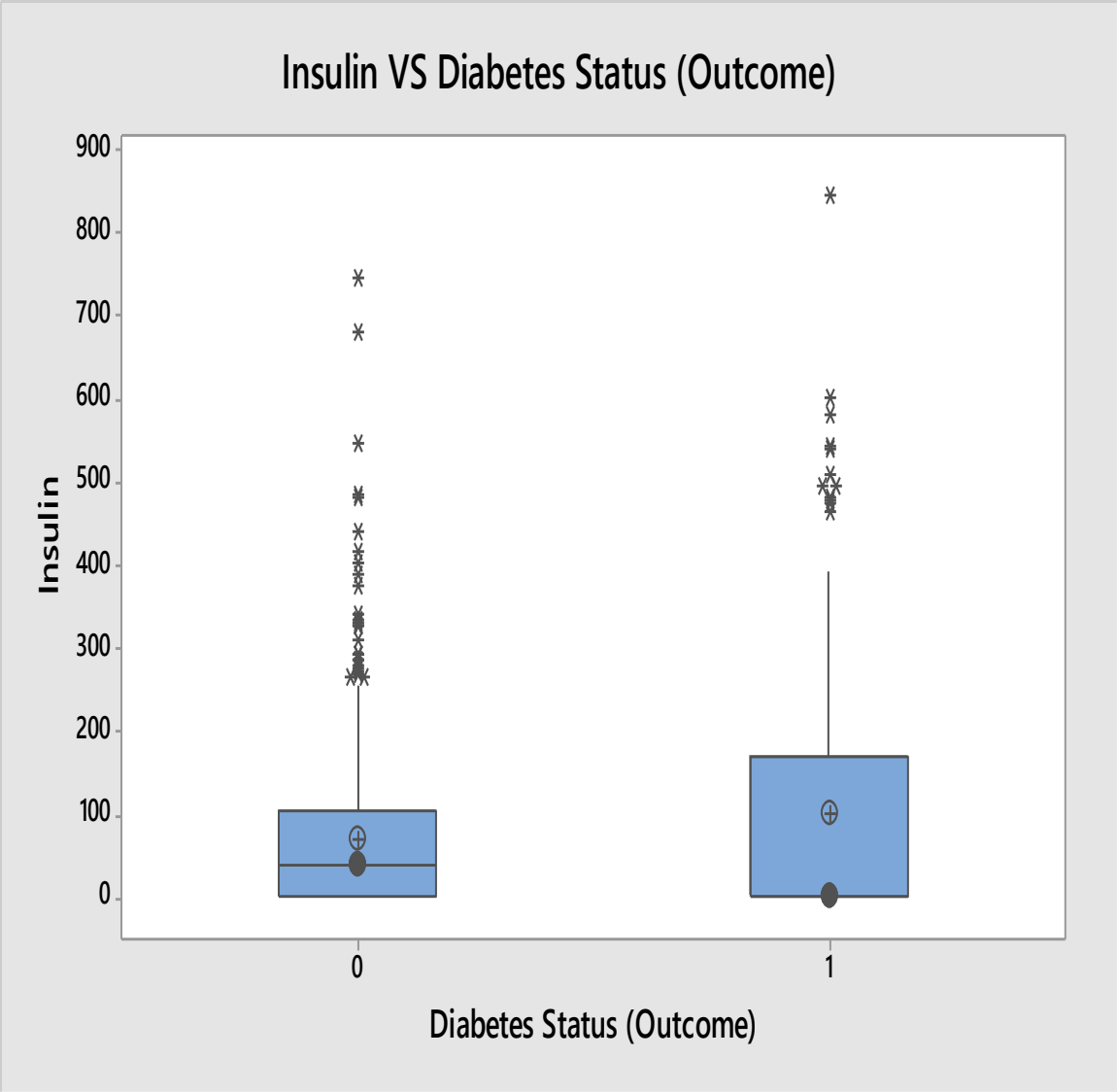
Comment: In the light of the above graph, it can be inferred that on an average, a diabetic person has a lower blood pressure than a non-diabetic one.

- **SkinThickness (X_4)**
  Individual triceps skin fold thickness measured in millimeters. Skin thickness (the interface between the epidermal surface and skin fat), which is mostly governed by collagen content, is larger in individuals with insulin-dependent diabetes mellitus (IDDM) who have being diabetic for more than ten years. This covariate is a continuous variable.



Comment: From the graph, it can be said that, on an average, triceps skin fold thickness of an individual is slightly more if the individual is diabetic. Opposite result is obtained if the person is non-diabetic.

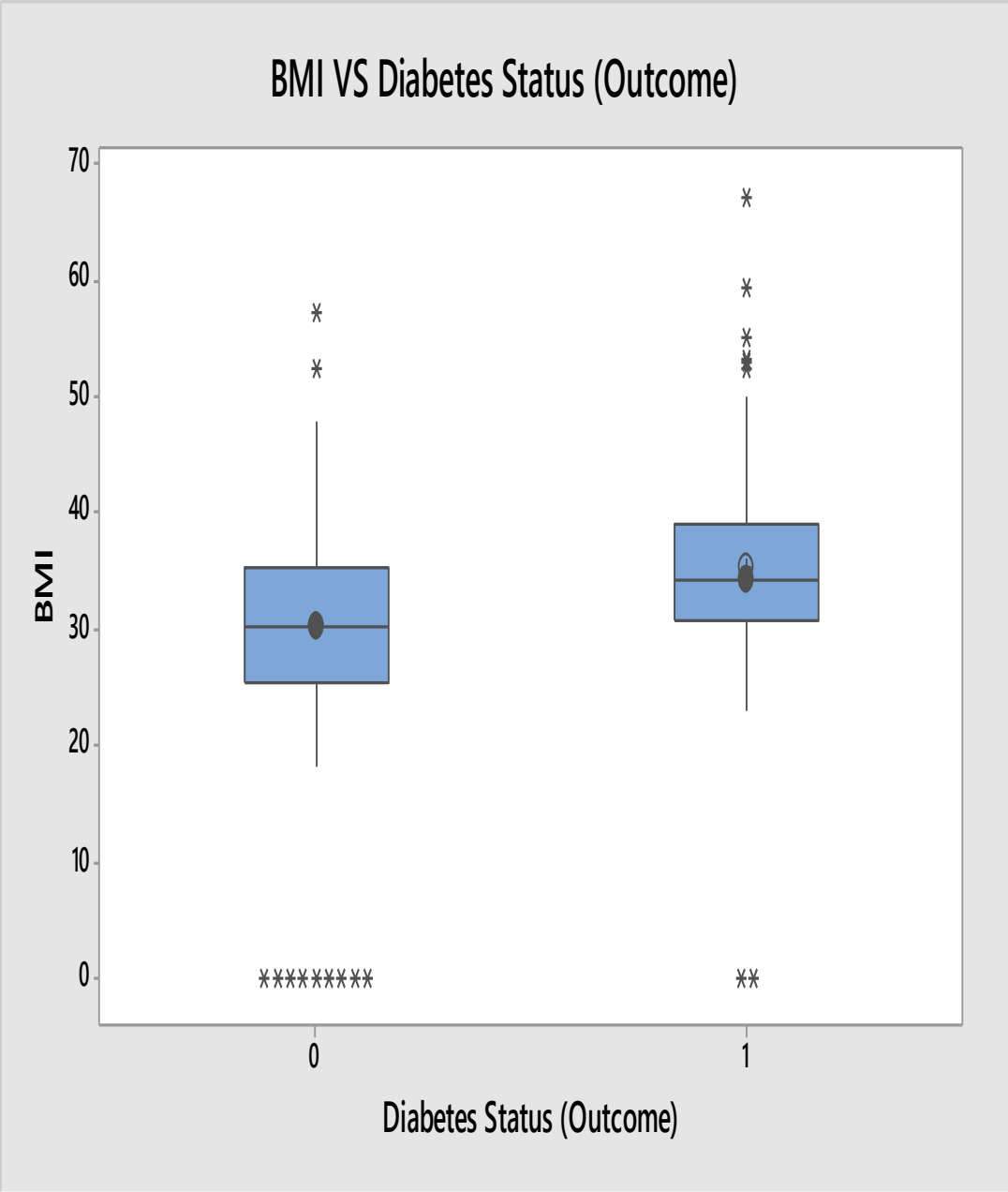
- **Insulin (X_5)**
  Two-hour serum insulin (μU/ml). Insulin has a significant influence in the development of type 2 diabetes. This vital hormone, which the body cannot function without, controls blood sugar levels (glucose). This is a highly complex procedure. This covariate is a continuous variable.



Comment: From the graph, it can be concluded that, for a diabetic person, median of insulin level is less than that of a non-diabetic person. Since, there are very few data points after 500µU/ml, so we have to further check the significance of this factor by another means.

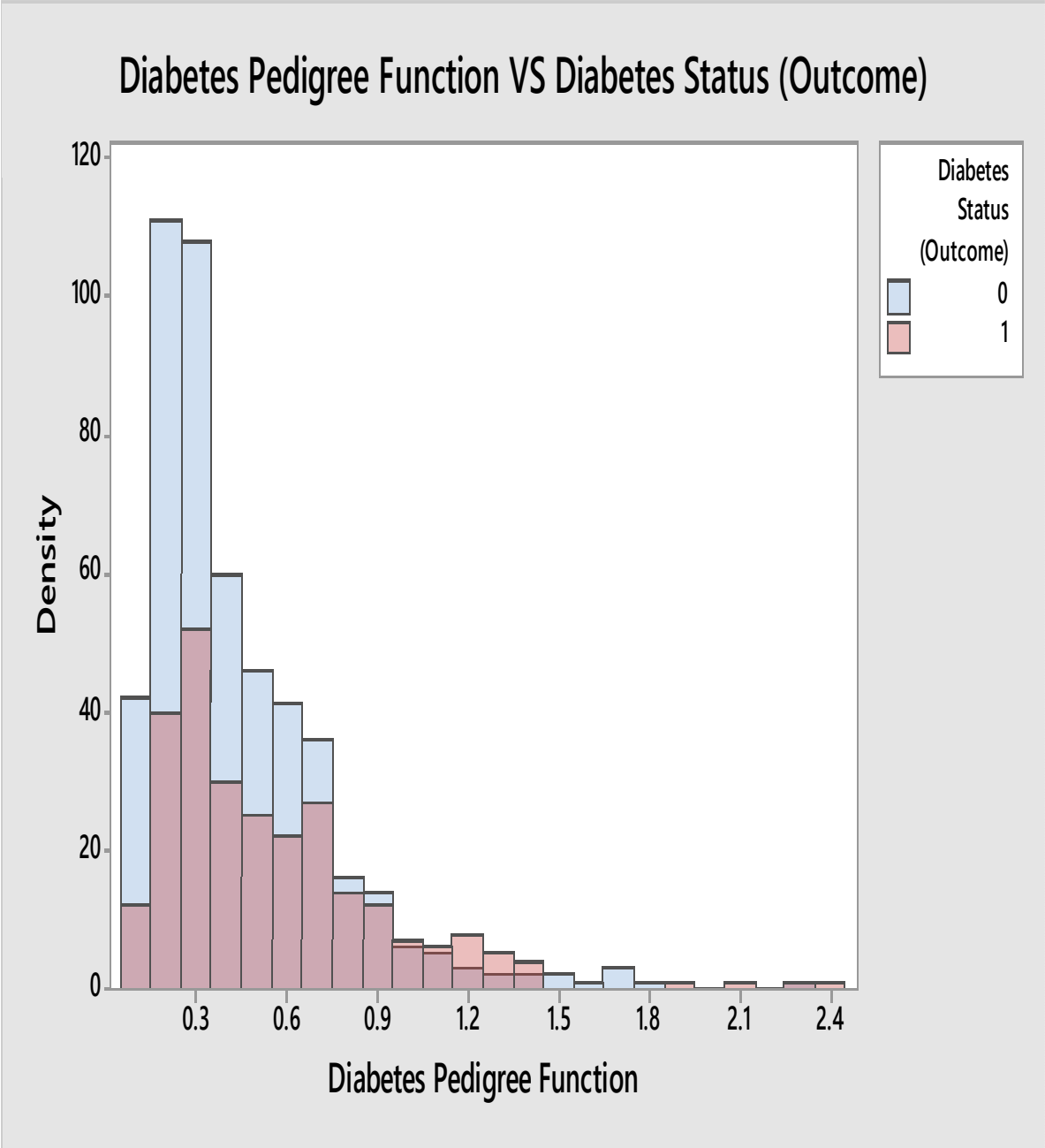
- **BMI (X_6)**

  This independent variable depicts individual's body mass index (weight in kg/height in $m)^2$. Overweight (BMI of 25 to 29.9 BMI) or obesity (30 to 39.9 BMI) or morbid obesity (more than 40 BMIs) enhances the risk of developing type 2 diabetes. The more obese you are, the more your muscles and tissue cells produce insulin hormone. It is a continuous variable.



Comment: From the above boxplots, we can conclude that on an average, BMI for a diabetic individual is higher than that of a non-diabetic individual, though outliers are present.

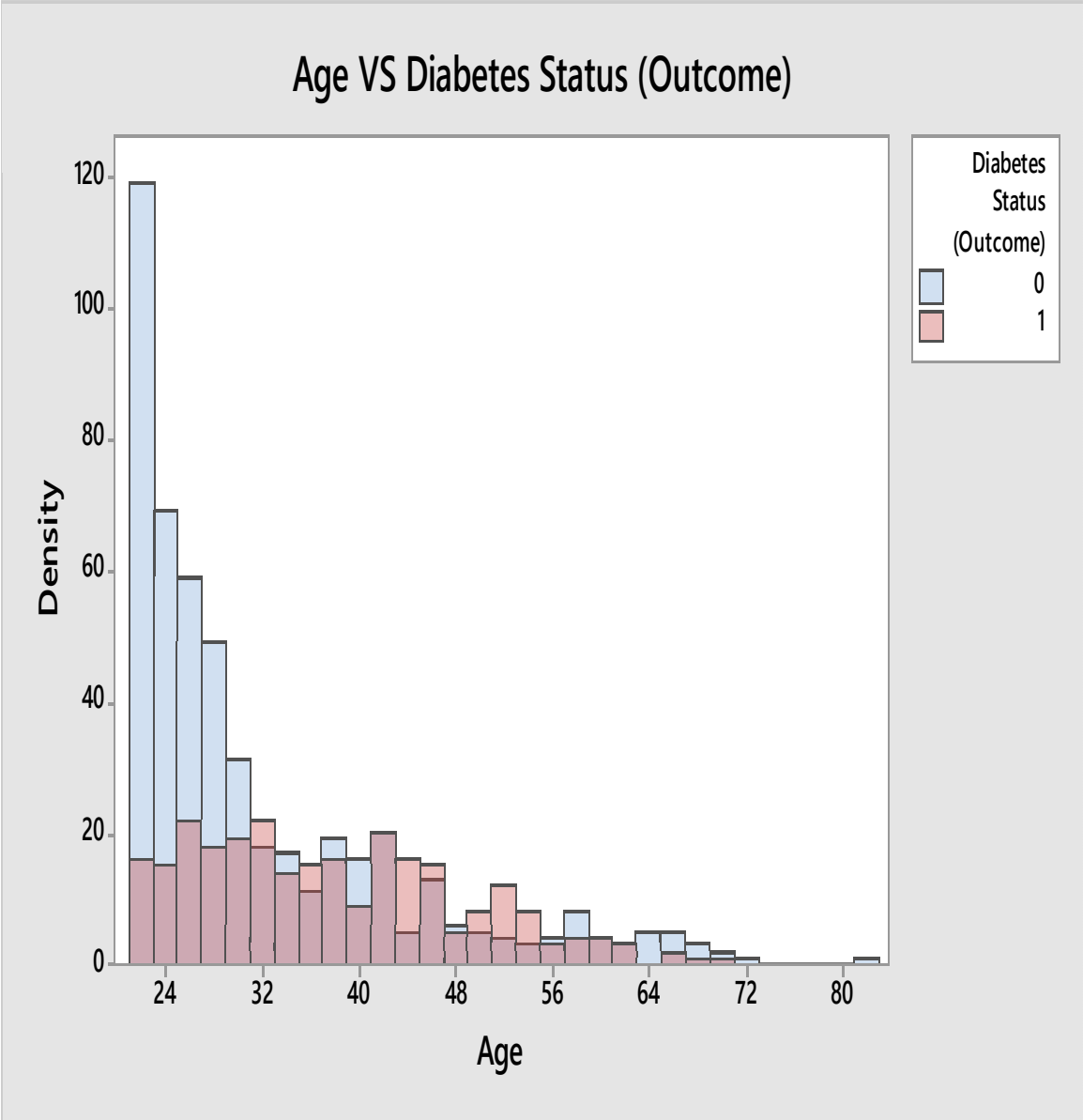
- **DiabetesPedigreeFunction (X_7)**
  A function that determines the risk of diabetes based on family history. This measure of genetic impact provides insight into the hereditary risk associated with the development of diabetes mellitus. It is a continuous independent variable.



Comment: According to the above graph, even though there are many outliers, the average Diabetes Pedigree Function for a diabetic is higher than that of a non-diabetic person.

- **Age (X_8)**

  It depicts the individual's age in years. Older persons are more likely to get diabetes since the risk of the disease rises with age. In fact approximately 25% of adults over the age 60 years have diabetes. It is a continuous independent variable.



Comment: It can be inferred from the above graph that mean age of diabetic persons is higher than that of non-diabetic persons. Therefore, with increasing age, risk of diabetes also increases.

MULTICOLLINEARITY

Finding a correlation between the response and predictor(s) is obvious in regression analysis; however, correlation between predictors is not desirable. Before fitting logistic regression to our data, we may wish to know if the covariates under consideration should be given equal weight in predicting the response. If not, we may remove that variable and focus on the others, which brings us to the concept of association between variables. The idea of multicollinearity comes into play.

It is Ragnar Frisch who coined the word "multicollinearity." Originally it meant the existence of a "perfect," or exact, linear relationship among some or all explanatory variables of a regression model. In multiple regression analysis, the term multicollinearity indicates to the linear relationships among the independent variables. When two variables are almost perfect linear combinations of one another, they are said to be collinear.

Multicollinearity occurs when the factors in a multiple regression model have a high degree of intercorrelation. When an investigator tries to establish how well each predictor can predict or interpret the response variable in a statistical model, multicollinearity can lead to skewed or misleading results. Overall, multicollinearity may lead to less stable probability estimates and a wider confidence interval for the predictors. In other words, results from a multicollinearity model could not be reliable. Multicollinearity may often make it difficult to interpret a regression analysis with multiple factors. Multicollinearity causes the standard errors of each model coefficient to rise, altering the conclusion of the analysis. Some of the relevant variables under investigation become statistically insignificant due to multicollinearity. Multicollinearity makes the regression coefficients more unstable by raising their variance, which makes it more difficult to understand the coefficients. Multiple research works looked at and addressed the issues with multicollinearity in regression models. They also stressed that the main issues with multicollinearity include biased and unequal standard errors as well as unworkable explanations for the data. We have perfect multicollinearity if the correlation between the explanatory variables is ±1.

Mathematically, we represent the relationship as:

$$b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_{ki} = 0 \qquad \forall\, i = 1(1)n$$

where b_j's are constant and X_{ji} is the ith observation on kth covariate. $\forall\, j = 1(1)k$

Nevertheless, because of innate correlations in the system under study, almost multicollinear variables frequently persist after redundancies have been found and eliminated. In this situation, the preceding equation is not true; instead, we have the same equation updated to include an error factor, εi. If the variance of εi's is minimal for a given set of values for the bj, we claim that the variables are almost completely multicollinear bj.

Using Variance Inflation Factors (VIFs) to detect multicollinearity in our dataset:

The Variance Inflation Factor, or VIF, measures how much of the variation is inflated. It is known that if multicollinearity exists, the standard errors and therefore the variances of the computed coefficients would be inflated.

VIF for the k[th] predictor is given by:

$$VIF_k = \frac{1}{1 - R_k^2}$$

where, R_k^2 is the R^2 value obtained by regressing k[th] predictor on the remaining predictors.

Interpretation: A VIF of 1 indicates no association between the k^{th} predictor and the remaining predictor variables, implying that the variance is not inflated.

Selection Rule: We generally regard VIF values greater than 5 to be signals of substantial multicollinearity that must be corrected.

Remedy: The simplest corrective method for severe multicollinearity is "Dropping of one of the redundant collinear variables".

Procedure and Calculations: For each of the eight explanatory variables, we have computed the VIF using the statistical software R.

Predictors	VIF
Preg	1.363458
Glucose	1.207082
BloodPressure	1.176100
SkinThickness	1.533354
Insulin	1.469802
BMI	1.216472
DiabetesPedigreeFunction	1.036229
Age	1.431713

Conclusion: According to the above table, the value of VIF are less than 5. Therefore, we can conclude that there might not exist any multicollinearity among the independent variables and hence they should be included in the further study.

CATEGORICAL DATA ANALYSIS

Data measured using a nominal or ordinal scale gives rise to categorical data. Nominal scale is based on a set of mutually exclusive and exhaustive categories which do not involve any order. Thus, categorical data without any natural ordering is nominal. For example, blood group, hair colour, religion etc. are nominal data. An ordinal scale is based on a set of mutually exclusive and exhaustive categories where the categories are ordered. Interval between the categories need not be necessarily equal. Thus, categorical data having a natural ordering is called ordinal. For example grades obtained by a student in an examination are ordinal. The nominal and ordinal scales taken together are sometimes referred to as categorical scales. Categorical variables are often referred to as qualitative variable to distinguish them from the metric or quantitative variables such as weight, height, age etc. Categorical data analysis is the analysis of categorical data.

Effect of Preg on Diabetes Status (Outcome):

We construct a 2×2 contingency table taking Preg as our binary explanatory variable and response as Diabetes Status (Outcome) of a person and with the help of this contingency table, we can measure the association with Yule's Coefficient of Association, Yule's Coefficient of Colligation and Cramer's Coefficient of association and then we calculate odds ratio for comparison of two levels and from that we draw our inference.

2×2 Contingency Table

Outcome / Preg	**Non-diabetic (0)**	**Diabetic (1)**	**Total**
Preg = 0	311	113	424 = f_{01} (say)
Preg = 1	189	155	344 = f_{02} (say)
Total	500 = f_{10} (say)	268 = f_{20} (say)	768 = n (say)

Where f_{ij} is the cell frequency across the (i, j)th cell of the contingency table.

- **Yule's Coefficient Of Association:**

$$Q_{AB} = \frac{f_{11}f_{22} - f_{12}f_{21}}{f_{11}f_{22} + f_{12}f_{21}} = 0.38596$$

Interpretation: Preg and Outcome are somewhat positively associated because $Q_{AB} > 0$, but not near to 1.

- **Yule's Coefficient Of Colligation:**

$$Y_{AB} = \frac{\sqrt{f_{11}f_{22}} - \sqrt{f_{12}f_{21}}}{\sqrt{f_{11}f_{22}} + \sqrt{f_{12}f_{21}}} = 0.20075$$

Interpretation: Preg and Outcome are roughly positively associated because $Y_{AB} > 0$, but not close to 1.

- **Cramer's Coefficient Of Association:**

$$V_{AB} = \frac{f_{11}f_{22} - f_{12}f_{21}}{\sqrt{f_{10}f_{01}f_{20}f_{02}}} = 0.3448$$

Interpretation: Preg and Outcome are roughly positively associated because $V_{AB} > 0$, but not near to 1.

- **Odds Ratio:**

$$OR = \frac{P(Y = 1|X_1 = 1) \Big/ P(Y = 0|X_1 = 1)}{P(Y = 1|X_1 = 0) \Big/ P(Y = 0|X_1 = 0)}$$

An estimate of Odds Ratio based on the given data is given by:

$$\widehat{OR} = \frac{f_{11}f_{22}}{f_{12}f_{21}} = 2.257$$

Interpretation: Odds of occurrence of diabetes when given number of pregnancies is greater than 3 is 2.257 times odds of that when number of pregnancies less than or equal to 3.

FITTING OF BINARY LOGISTIC REGRESSION MODEL

Logistic regression is calculated using the maximum likelihood estimation (MLE) technique, whereas linear regression is computed using the ordinary least squares (OLS) method. OLS can be applied only when the response variable is continuous, independent and identically distributed. We want to fit a binary logistic regression to our data since our response variable (Y) is binary categorical in nature.

Our study variable is Y and X_1, X_2, X_3, X_4, X_5, X_6, X_7 and X_8 are our covariates.

Here, Y is the Outcome which denotes the Diabetes Status of an individual.

$$Y = \begin{cases} 1, \text{if the individual is diabetic,} \\ 0, \text{otherwise.} \end{cases}$$

We have observations on 768 individuals.

The values of Y, that is, $y_1, y_2, \ldots, y_{768}$ are independent or at least uncorrelated.

In binary logistic regression, we model the probability of success on the basis of the given predictors.

So, we model $P(Y=1| x_1, x_2, \ldots, x_8) = p$, where $0<p<1$. In particular, $P(Y_i=1| x_{i1}, x_{i2}, \ldots, x_{i8}) = p_i$, i=1(1)768.

We assume, $Y_i \sim$ Bernoulli (p_i) , i=1(1)768.

We are interested to predict Y for a random set of individuals drawn from the relevant population. This is done with the help of covariates.

From the theory of Generalized Linear Models, we choose the logit link to model p_i's on x_{ij}'s $\forall$ i=1(1)=768, j=1(1)8.

Logit Link: $\log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$, $i = 1(1)n$, where $n = 768$ and $k = 8$.

$$\Rightarrow \frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik}}, i = 1(1)n.$$

$$\Rightarrow p_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik}}}{1+e^{\beta_0+\beta_1 x_{i1}+\cdots+\beta_k x_{ik}}}, i = (1)n.$$

β_0 is the intercept term and $\beta_1, \beta_2, \ldots, \beta_k$ are the regression coefficients corresponding to $x_1, x_2, \ldots, x_k$ respectively. β_i's are the unknown real valued parameters which will be estimated using MLE method. Let $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are the maximum likelihood estimates of the parameters.

Interpretation of Model Parameters:

- β_0 can be interpreted as the logarithm in odds of Y=1 against Y=0 when X_j=0, j=1(1)k.

- Here, X_1 is a dummy covariate taking the values 0 and 1. So, β_1 can be interpreted as the log odds ratio of Y=1 against Y=0 as X_1 changes from 1 to 0, keeping the other covariates fixed.
- $X_2, X_3, \ldots, X_k$ are continuous independent variables. β_j is interpreted as the amount by which the log odds of success changes due to one unit change in X_j, keeping all the other covariates fixed, $\forall$ j=2(1)k.

Model Fitting:

The estimates are obtained using the R software. A snapshot from the R console is given below:

```
Call:
glm(formula = Outcome ~ Preg + Glucose + BloodPressure + SkinThickness +
    Insulin + BMI + DiabetesPedigreeFunction + Age, family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6404  -0.7319  -0.4180   0.7690   2.8385

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -8.318234   0.713878 -11.652  < 2e-16 ***
Preg                      0.616246   0.213881   2.881  0.00396 **
Glucose                   0.034782   0.003691   9.424  < 2e-16 ***
BloodPressure            -0.013420   0.005213  -2.574  0.01004 *
SkinThickness             0.002110   0.006900   0.306  0.75972
Insulin                  -0.001221   0.000904  -1.351  0.17666
BMI                       0.087917   0.014975   5.871 4.34e-09 ***
DiabetesPedigreeFunction  0.945575   0.298237   3.171  0.00152 **
Age                       0.020699   0.009097   2.275  0.02289 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 993.48  on 767  degrees of freedom
Residual deviance: 730.34  on 759  degrees of freedom
AIC: 748.34

Number of Fisher Scoring iterations: 5
```

Hence the fitted regression model is given by:

$$\hat{p}_i = \frac{e^{\hat{\eta}_i}}{1 + e^{\hat{\eta}_i}} \quad , \forall\, i = 1(1)n$$

where $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$, $i = 1(1)n$, where $n = 768$ and $k = 8$.

η_i is known as the linear predictor, $\forall\, i = 1(1)n$.

$$\hat{\eta} = -8.318 + 0.6162\, Preg + 0.0348\, Glucose - 0.0134\, BloodPressure + 0.0021\, SkinThickness - 0.0012\, Insulin + 0.0879\, BMI + 0.9456\, DiabetesPedigreeFunction + 0.0207\, Age$$

TEST OF SIGNIFICANCE OF PREDICTORS – WALD TEST

After obtaining the estimates for the parameters, our current objective is to determine which factors are significant in predicting an individual's risk of developing diabetes. The test of significance for individual regression coefficients in logistic regression is called the Wald test.

To Test:

$H_{0j}: \beta_j = 0$ against $H_{1j}: \beta_j \neq 0$ $\forall\, j = 1(1)8$

Test Statistic:

Under H_{0j},

$$Z_j = \frac{\hat{\beta}_j}{S.E.(\hat{\beta}_j)} \sim N(0,1) \quad \forall\, j = 1(1)8$$

where $S.E.$ stands for Standard Error.

Critical Region:

We reject H_{0j} at level of significance α iff $|\text{observed } Z_j| > \tau_{\alpha/2}$, where τ_α is the upper α point of N(0, 1) distribution and $\tau_{\alpha/2} = 1.96$, for α = 0.05.

Decision Table:

Predictors	Z-statistic	P-value	Decision
Preg	2.881	0.00396	Reject
Glucose	9.424	< 2e-16	Reject
BloodPressure	-2.574	0.01004	Reject
SkinThickness	0.306	0.75972	Accept
Insulin	-1.351	0.17666	Accept
BMI	5.871	4.34e-09	Reject
DiabetesPedigreeFunction	3.171	0.00152	Reject
Age	2.275	0.02289	Reject

In the table Z-statistic indicates $|\text{observed } Z_j|, j = 1(1)8$. The null hypothesis is rejected iff the p-values are less than our α.

Conclusion:

- In the light of the given data, pregnancy has a significant influence in determining a person's risk of developing diabetes, with category 1 pregnant females more likely to have the disease than category 0 females.
- In the light of the given data, an individual's blood glucose level is a highly important predictor of their diabetes status; the higher the blood glucose level, the greater the likelihood of developing diabetes.
- According to the given data, individual blood pressure is a comparatively less important factor in determining whether or not a person has diabetes; the lower the blood pressure, the greater the likelihood of diagnosis.
- In accordance with the given data, the skin thickness of an individual does not significantly influence the likelihood that they will have diabetes.
- In the light of the given data, the skin thickness of an individual does not significantly influence the likelihood that they will have diabetes.
- In accordance with the given data, the individual's BMI has a highly significant influence in determining whether or not they have diabetes; the higher the BMI, the greater the likelihood of diabetes.
- In accordance with the given data, individual DiabetesPedigreeFunction plays a significant role in determining diabetes status, and the higher the DiabetesPedigreeFunction, the more probable the person is to develop diabetes.
- According to the given data, an individual's age can indicate whether or not they have diabetes, with older individuals having a higher likelihood of having the disease. However, age is a comparatively less important factor in this prediction.

MEASUREMENT OF MODEL ACCURACY

The fitted values are then obtained as:

$$\hat{p}_i = \frac{e^{\hat{\eta}_i}}{1 + e^{\hat{\eta}_i}} \quad , \forall\, i = 1(1)n$$

where $\hat{\eta} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$, $i = 1(1)n$, where $n = 768$ and $k = 8$.

Following the fitting of the regression equation, we use the model to predict the outcome of our response, that is, whether Y=1 or Y=0.

The problem of predicting Y is carried out by choosing a **threshold**, say p^* and assigning 1 as predicted value if $\hat{p}_i > p^*$ and 0 otherwise, $\forall\, i = 1(1)n$. Thus, we get the predicted value of the response.

We arrange the units / and their observed values y_i's in decreasing order of $\hat{p}_i$'s. Choosing each $\hat{p}_i$ as a threshold, classify the predicted values into 1's and 0's. Through the classification, there occurs mismatches between the observed and the predicted value of Y. For each $\hat{p}_i$, $i = 1(1)n$, we classify the mismatches in a 2×2 matrix. These matrices are known as **Confusion Matrices**. Creating a confusion matrix is a common way to evaluate the quality of a logistic regression model.

A correct classification occurs when y=0 and Y=0 or y=1 and Y=1. A mis-classification occurs when y=1 and Y=0 or y=0 and Y=1.

Let us define:

True Positive Rate (TPR):

Proportion of rightly predicted Y=1 values among the actual Y values. It is also known as **Sensitivity** of a test. Higher TPR indicates better classification.

$$TPR = P(Y = 1 | y = 1)$$

False Positive Rate (FPR):

Proportion of wrongly predicted Y=1 values among those actual Y =0 values. Further, 1-FPR is referred to as the **Specificity** of a test. Lower FPR or higher specificity indicates better classification.

$$FPR = P(Y = 1 | y = 0)$$

Accuracy:

The percentage of total correct classifications made by the model.

The threshold p^* is chosen as that $\hat{p}_i\ \forall\, i = 1(1)n$, for which $Sensitivity \times Specificity$ is the maximum and the classification for the threshold is considered as the predicted value Y.

Result:

Threshold, that is, $p^* = 0.4173252$.

Confusion matrix for our fitted model when $p^* = 0.4173252$ is given by,

Predicted ($\widehat{Y}$) / Observed (y)	**0**	**1**	**Total**
0	417	83	500
1	98	170	268
Total	515	253	768

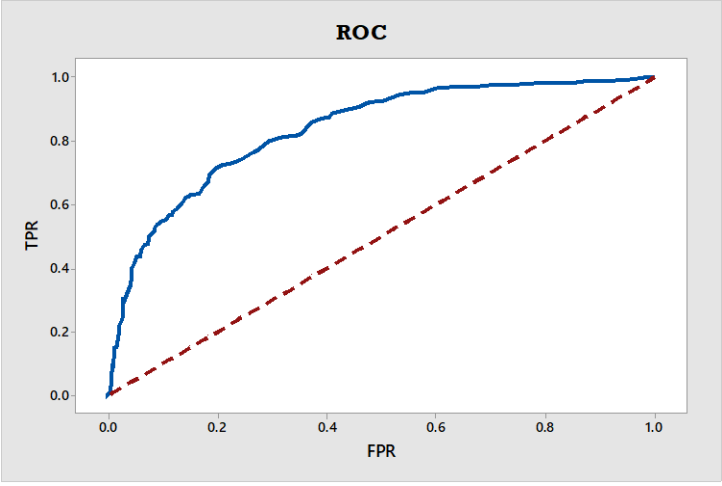
Sensitivity= 0.7126866

Specificity= 0.804

Accuracy rate= 76.43229%

Receive Operated Characteristic (ROC):

The values of TPR and FPR are plotted against each other on a 2-D graph. It is called Receive Operated Characteristic (ROC). The diagonal line in the graph indicates TPR=FPR. This is called the chance line. In practice, steeper the ROC curve, better is the classification. ROC curve becomes steeper when the classification rule takes high values of TPR against low values of FPR, that is, high values of TPR(1-FPR). We obtain the ROC curve using Minitab software:



Conclusion:

The predicted model is moderately accurate.

FINAL CONCLUSION

Finally, it is time to compile the results of the analysis and formulate a final conclusion. As previously stated, we worked with data on diabetes status of an individual provided by the Pima Indians Diabetes Dataset. With the use of eight predictors, the dataset aims to diagnose and determine if a person has diabetes. We have fitted a binary logistic regression model based on all eight covariates. Prior to undertaking binary logistic regression, preliminary descriptive studies are conducted, followed by an examination for multicollinearity among the predictors and categorical data analysis. Subsequently, upon fitting the logistic regression model, a list of statistically significant predictor variables influencing the response is obtained. Finally, a confusion matrix is generated to assess the accuracy of this model.

- By the method of data visualization, following conclusions were drawn:
    - We may say, females who were pregnant for more than 3 times are more prone to developing diabetics than females who were pregnant for less than or equal to 3 times.
    - On an average, increment of plasma glucose concentration may indicate the presence of diabetes in an individual.
    - On an average, a diabetic person has a lower diastolic blood pressure than a non-diabetic one.
    - On an average, triceps skin fold thickness of an individual is slightly more if the individual is diabetic.
    - For a diabetic person, median of insulin level is less than that of a non-diabetic person, though there were very few data points after 500μU/ml.
    - On an average, BMI for a diabetic individual is higher than that of a non-diabetic individual, though outliers are present.
    - The average Diabetes Pedigree Function for a diabetic is higher than that of a non-diabetic person.
    - With increasing age, the likelihood of being diabetic increases.
- We found that the VIF values of the predictors are less than 5. Therefore, we can conclude that there might not exist any multicollinearity among the independent variables.
- From further categorical data analysis, we concluded that females with higher number of pregnancies have higher chance of developing diabetes.
- After model fitting, we tested the significance of predictors, where we found that SkinThickness and Insulin are insignificant predictors, whereas, Glucose and BMI are very strong significant predictors. Preg and DiabetesPedigreeFunction also have a significant role in predicting the presence of diabetes. Age and BloodPressure are relatively less significant factors.

- Then we measured the accuracy of the model using confusion matrix where we found that the model is moderately accurate.

Thus, our study may able to help us to know which factors are mostly affecting the presence of diabetes in an individual so that preventive measures can be taken and early treatment can be done before the disease creates major health complications. Further statistical analysis is required to acquire more information because we used sample data, which cannot offer the most accurate results. Another limitation of our study is that we are training and testing the model on the same dataset because the objective of testing a model is to estimate how well it perform predictions on data that the model did not see. As a result, the model is very likely to overfit. It is true that our model might not always yield the proper results in every analysis, but this study lays the groundwork for enhancing the robustness of future research.

REFERENCES

- "National Institute of Diabetes and Digestive and Kidney Diseases" dataset repository.
- https://www.who.int/news-room/fact-sheets/detail/diabetes
- https://www.statology.org/types-of-logistic-regression/
- Lv, C., Chen, C., Chen, Q., Zhai, H., Zhao, L., Guo, Y., & Wang, N. (2019). Multiple pregnancies and the risk of diabetes mellitus in postmenopausal women. Menopause (New York, N.Y.), 26(9), 1010–1015. https://doi.org/10.1097/GME.0000000000001349
- Basic Econometrics by Gujarati and Porter.
- Shrestha, N. (2020). Detecting Multicollinearity in Regression Analysis. American Journal of Applied Mathematics and Statistics, 8(2), 39-42. http://pubs.sciepub.com/ajams/8/2/1
- https://learn.g2.com/logistic-regression
- https://online.stat.psu.edu/stat462/node/207/
- https://datascience.stackexchange.com/questions/86632/why-is-it-wrong-to-train-and-test-a-model-on-the-same-dataset