Diabetes Prediction Analysis

Name: Sunanda Maity

Registration Number: A01-2112-0834-20

Roll Number: 407

Session: 2020 – 2023

Paper Code: *HSTDS6043D*

St. Xavier's College (Autonomous), Kolkata

Supervised by: Dr. Surabhi Dasgupta

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

DEPARTMENT OF STATISTICS

APRIL 5, 2023

Table Of Contents:

- **Introduction:** 3
- **Data Description:** 6
- **Methodology:** 9
  - **Checking Multicollinearity:** 9
  - **Data Splitting And Fitting of the Model:** 13
  - **Fitting of logit model:** 15
- **Results And Discussion:** 17
  - **Graphical Interpretation:** 17
  - **Decision:** 25
  - **Interpretation:** 26
  - **Reconstruction of the Model:** 30
  - **Prediction:** 31
  - **Accuracy Of the Model:** 37
- **Conclusion:** 42
- **Scope Of Further Study:** 43
- **References:** 43
- **Acknowledgement:** 44
- **Appendix:** 45

❖ Introduction:

Diabetes is a chronic disease that affects millions of people around the world. Diabetes is a disease that affects the way our body processes glucose, a type of sugar that is our main source of energy. It affects the body's ability to produce or use insulin, a hormone that regulates blood sugar levels. There are two main types of diabetes:

1. Type 1 diabetes: This type of diabetes is an autoimmune disease that occurs when the body's immune system attacks and destroys the insulin-producing cells in the pancreas, leading to a deficiency of insulin which requires individuals to take insulin injections or use insulin pump to regulate their blood sugar levels. Type 1 diabetes typically develops in children and young adults and requires lifelong insulin therapy.

2. Type 2 diabetes: This type of diabetes occurs when the body becomes resistant to insulin or does not produce enough insulin to meet its needs. Type 2 diabetes is typically associated with lifestyle factors such as obesity, lack of physical activity, and an unhealthy diet. It can often be managed with lifestyle changes, medication, and insulin therapy if necessary. This type of diabetes can be observed among the adults, specially.

Both types of diabetes can cause a range of complications, including heart disease, stroke, kidney disease, nerve damage, and blindness. These complications can be severe and can lead to disability or even death. Symptoms of diabetes may include increased thirst and urination, blurred vision, fatigue, slow healing of wounds, and frequent infections. Individuals with diabetes need to work with their healthcare providers to manage their condition and prevent these complications.

Being conscious of diabetes means being aware of the risk factors, symptoms, and complications associated with the disease. This awareness can help individuals take proactive steps to prevent or manage diabetes. Diabetes can have serious consequences if left untreated or poorly managed. High blood sugar levels over time can damage blood vessels and nerves, leading to a variety of complications such as heart disease, stroke, kidney disease, blindness, and nerve damage.

The best way to manage diabetes is through a combination of medication, healthy lifestyle choices, and regular medical care. People with type 1 diabetes need to take insulin injections to regulate their blood sugar levels. People with type 2 diabetes may also need to take medication to help their bodies use insulin more effectively or to simulate insulin production.

Healthy lifestyle choices are also important in managing diabetes. Eating a healthy, balanced diet that is low in sugar and saturated fat can help to keep blood sugar levels in check. Regular physical activity can also improve insulin sensitivity and maintain a healthy weight. Quitting smoking and reducing alcohol consumption can also help to reduce the risk of complications associated with diabetes.

Regular medical care is also important in managing diabetes. This includes regular blood sugar monitoring, regular check-ups with a healthcare provider, and regular eye and foot exams to check for any signs of complications. Additionally, regular screening for diabetes can help to identify the disease early on and prevent complications.

In addition, being conscious about diabetes because the disease is becoming increasingly common worldwide. According to the World Health Organization (WHO), the number of people with diabetes has risen from 108 million in 1980 to 422 million in 2014. By raising awareness about diabetes and promoting healthy lifestyle habits, we can work to prevent and manage this disease for generations to come.

So, all in all, diabetes prediction is important for several reasons:

1. Early detection can help prevent or delay the onset of complications such as nerve damage, blindness, kidney disease, and heart disease.

2. Early detection of diabetes allows healthcare providers to start appropriate treatment and management strategies. Effective management can help to prevent complications and improve the quality of life for individuals with diabetes.
3. Early detection and management of diabetes can help to reduce healthcare costs associated with complications and hospitalizations. By identifying high-risk individuals, healthcare providers can target interventions for those who need them most, potentially reducing the overall cost of diabetes care.

In this study, we will work on the type two diabetes dataset, which is collected from Kaggle. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. In particular, here all the patients are females at least 21 years old of Pima Indian Heritage. This diabetes dataset contains information about 768 women and there is a total of 8 covariates present, all of which are risk factors for a woman to be diabetic.

Here, we aim to predict the probability of a female patient having diabetes based on their clinical measurements and try to find a relationship between the attributes and the presence of diabetes. So, we will be able to state which of the covariates are significant and how they are having an impact on diabetes. This would be the analysis of the dataset. Here, in this dataset, the target variable is a binary variable indicating whether the patient has diabetes or not. Our task however is to create a model exploring the relationship and evaluate them.

As here the response variable is binary, logistic regression will be appropriate to model this dataset. To fit a logistic regression model to this dataset, we first need to split the dataset into a training set and a testing set so that we can save some data to be able to accurately evaluate the performance of the model. This modelling can help to identify the factors that are strongly associated with the likelihood of diabetes, which will help the public health policies and interventions aimed at reducing the prevalence of diabetes.

❖ Data Description:

In this dataset, we can find data on 768 women who were tested for diabetes. There are 9 columns i.e., there are 8 covariates and one binary response variable.

- **Covariates:**

The covariates in this dataset are "Pregnancies", "Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI", "DiabetesPedigreeFunction", and "Age". Some of them are continuous and some are discrete type of covariate.

No.	Covariates	Description
1.	Pregnancies	The number of previous pregnancies of a particular woman.
2.	Glucose	The blood glucose level of a woman from a glucose tolerance test.
3.	BloodPressure (Bp)	Diastolic blood pressure of each woman (mm Hg)
4.	SkinThickness	Triceps skin fold thickness(mm) of each woman.
5.	Insulin	2-Hour serum insulin (mu U/ml) level of each woman.
6.	BMI	Body mass index (weight in kg/ (height in m) ^2) of a woman.

7.	DiabetesPedigreeFunction	It is a function that scores the likelihood of diabetes based on the family history of each woman.
8.	Age	Age(years) of each woman in the dataset.

- **Response Variable:**

Here, our response variable is "Outcome", which indicates the presence or absence of diabetes. In this dataset,1 indicates the presence of diabetes disease and 0 indicates the absence of diabetes disease.

- **Source of the Data:**

This dataset is secondary data collected from the Kaggle. This dataset was collected and made available by the National Institute of Diabetes and Digestive and Kidney Diseases as part of the Pima Indians Diabetes database.

Methodology:

In this section, we will analyze the dataset by some appropriate methods and try to figure out which are the significant covariates. To do that, we shall divide this segment into sections:

1. **Checking Multicollinearity**:

In this section, we will try to find out whether multicollinearity exists in the dataset. After checking for multicollinearity, we will opt for necessary remedial measures if required.

2. **Data Splitting And Fitting of the Model:**

Here, we first need to split the data into a training set and a testing set.

- Fitting of the model: The training dataset will be used to fit the appropriate model. As mentioned earlier, we will fit the Multiple Logistic Regression model. After fitting the model, we will get the estimate and standard error of each parameter, and from that, we will be able to identify the significant covariates.

- Checking the accuracy of the model: The testing dataset will be used to evaluate the performance of the model used. Also, this dataset will be used to make predictions about whether a new patient is likely to have diabetes based on their clinical features.

Results And Discussion:

In this section, we will discuss the necessary results that we have obtained after data splitting and fitting the Multiple Logistic Regression model to our training dataset. Also after pointing out the significant covariates, we will be able to give a proper interpretation of them, i.e., we will be able to state the effect on the presence of diabetes for a unit change in the value of a particular covariate, keeping the other covariates constant. Also, we can judge how accurately we have been able to classify true diabetic patients.

❖ Methodology:

In this section, we will first visualize the whole data by graphical approach. For continuous predictors, we will use histograms and for discrete explanatory variables, we are going to use dot plot.

➢ Checking Multicollinearity:

Before checking for multicollinearity, let us discuss very briefly what multicollinearity is, what problem arises in the model due to its presence of it, and how to deal with multicollinearity.

What is Multicollinearity?

Multicollinearity is a statistical concept where two or more independent variables or covariates are highly correlated, either positively or negatively. It is essentially considered a sample phenomenon, as even if the covariates are not correlated in the population, they may be correlated in the particular sample in hand. Multicollinearity is a problem as it undermines the statistical significance of an independent variable.

What problem arises in the model if Multicollinearity is present?

Though we get the Best Linear Unbiased Estimate of the parameters under the Ordinary Least Square Estimation method even if multicollinearity is present, the minimum variance of the estimates becomes large (can be infinite in the case of exact multicollinearity, but not possible practically). As a result,

1. The confidence intervals get wider and become less useful.
2. For hypothesis testing, the power of tests will be small because of wider acceptance regions.

So, when we fit a model, we should always check whether multicollinearity is present or not, and if present, we should take proper remedial measures to deal with it.

Detection of Multicollinearity:

One way to detect multicollinearity under a multiple logistic regression setup is to check the Variance Inflation Factor (VIF) values for each covariate. The formula for calculating VIF is given by :

$$VIF_i = \frac{1}{1-R_i^2}$$

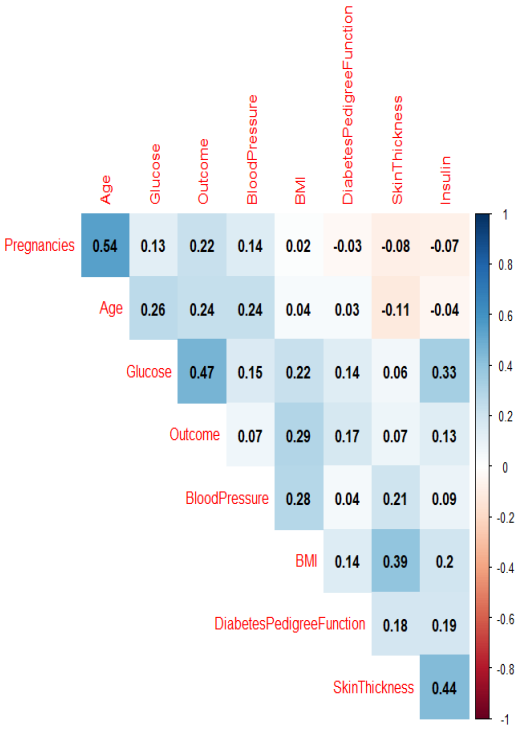
Where i = 1 (1) p, "p" being the number of covariates present in the model
R_i^2 = the multiple correlation coefficient of the ith covariate on all the other covariates.
Here, in our case p =8.

Remedial Measure:

As a standard practice, we eliminate those covariates from the model which have a VIF value more than 5. To justify this, we can say from the formula of VIF that a higher value of R_i^2 will increase the value of VIF, as high R_i^2 lowers the denominator. Thus, a value of VIF which is greater than 5 for a covariate suggests that the multiple correlation coefficient for that covariate regressed on the rest of the covariates is very high. So, that particular covariate is well explained by the rest of the covariates in the model. As a result, we eliminate that particular covariate and remove multicollinearity.

Multicollinearity in our dataset:

Let us first check the pairwise correlation for all the covariates and see whether any pair is highly correlated. The pairwise correlation heat map is given below:



From the heat map, we can observe that none of the pairs of covariates have a high value of correlation coefficient. So, we proceed to check the VIF values with the help of Multiple Correlation Coefficient.

Now, let us check whether multicollinearity is present in our dataset or not. First, we calculate the VIF values for every covariate present in our dataset with the help of "R-Software".
The values are given by :

No.	**Covariate**	**VIF value**
1.	Pregnancies	1.4084
2.	Glucose	1.2143
3.	BloodPressure	1.1752
4.	SkinThickness	1.5220
5.	Insulin	1.4679
6.	BMI	1.2204
7.	DiabetesPedigreeFunction	1.0343
8.	Age	1.5020

Here none of the covariates' VIF values is greater than 5. So, there is no multicollinearity among the covariates in this dataset. Hence, we can keep all the covariates in our dataset and can proceed to work for our necessary inference.

➢ Data Splitting And Fitting of the Model:

Here, first we will split the dataset into training and testing with a ratio of 7:3. Now, we will fit an appropriate model to the training data. Our response variable here is binary, taking values 0 and 1. This suggests that the appropriate model here would be the "Multiple Logistic Regression" model.

- **Defining the Variables:**

Response Variable:

Y: Let Y be a binary random variable that denotes the presence or absence of diabetes in a randomly selected patient (woman). Here, Y is our response or dependent variable, which takes values:

Y=1 if the patient is diabetic, i.e. the presence of diabetes in a randomly selected woman.

Y=0 if the patient is not diabetic, i.e. the absence of diabetes in a randomly selected woman.

Covariates:

In our dataset, covariates like the no of pregnancies, glucose level, blood pressure level, skin thickness, and insulin are discrete in nature, and covariates like BMI, diabetes pedigree function and age are continuous in nature. Now, we just list out the names of the covariates and what they are denoted by :

Symbol	Corresponding covariate
X1	Pregnancies
X2	Glucose
X3	BloodPressure

X4	SkinThickness
X5	Insulin
X6	BMI
X7	DiabetesPedigreeFunction
X8	Age

- **Random Sample:**

Let $y_1, y_2, y_3, \ldots, y_n$ be the outcome of the women whether they are diabetic or not, chosen randomly from the population. In our dataset, n=768.We also have the corresponding values of covariates $x_1, x_2, \ldots, x_8$ for every value of Y.

- **Splitting of Dataset:**

Before creating the model, we divide our dataset into training and testing sets. So, we have split the dataset where 70% of the dataset will be used as a training set and the remaining 30% as a testing set. We will fit a "Multiple Logistic Regression" model on the training dataset using the logit model.

Logit Model:

$$\Pr(Y=1/\ x1, x2, \ldots\ldots, x8) = \frac{e^{n}}{1+e^{\eta}} + \varepsilon$$

Where,

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_8 x_8$$

Here, $\beta_j' s$ are the unknown parameters corresponding to x_j's which are to be estimated .[j=1(1)8]

- **Assumption:**

We assume that Yi ~ Bernoulli (πi) [i = 1 (1) 768=n] independently but have non-identical distribution. As

πi = E(Yi) = Pr (Y=1/x1, x2, x3,..., x8),

and, Var(Yi) = πi(1- πi). we have:

Using Logit Model,

$$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

Or, $\eta_i = \log\left(\frac{\pi i}{1-\pi i}\right) = g(\pi_i)$, where g(.) is called the logit link

So, the link function under the logit model is given by,

$$\eta = g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

- **Fitting of logit model:**

We will use the method of "Maximum Likelihood" for fitting the model to our training data. The Log-likelihood function to find out the Maximum Likelihood Estimates of the parameters (β0, β1, β2, β3, ..., β8) under logit model is given by:

$$l = \sum_{i=1}^{n} \{yi \ln(\pi_i) + (1 - yi) \ln(1 - \pi_i)\}$$

- **Software:**

Here, we will use the "glm" function in "R - Software" in order to fit the model to the training data. By doing that, we will be able to get hold of the estimated value of the parameters along with their standard errors for logit model, along with a value of Residual Deviance to measure the goodness of fit.

- **Goodness of Fit:**

For measuring the goodness of fit, we will check the value of residual deviance for the Logit model. A lower value of residual deviance indicates a better fit. Residual deviance shows how well the response variable is predicted/classified with the inclusion of independent variables. Whereas, the null deviance shows how well the response variable is classified by a model that includes only the intercept.

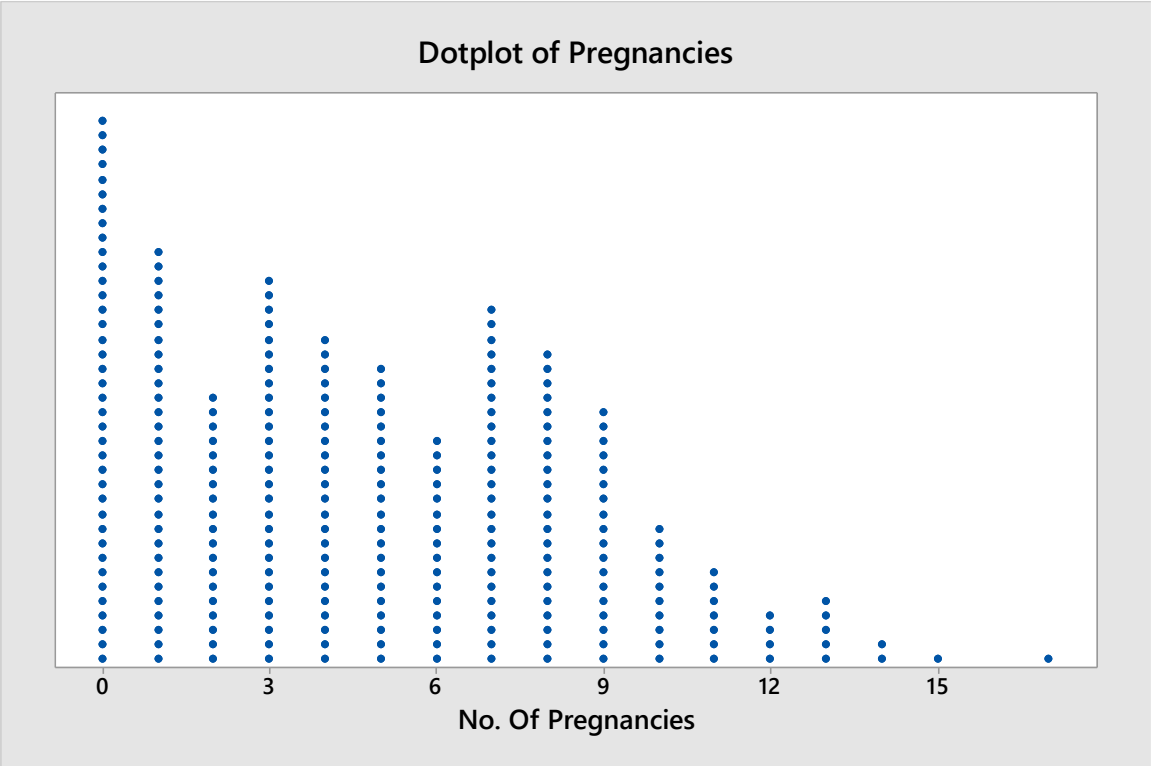
The residual deviance for the logit model applied on the training dataset is coming out as 462.28 and the null deviance is 663.60.

❖ Results And Discussion:

By graphical approach, here we want to visualize how the eight different predictors are affecting the response variable i.e. the status of diabetes disease for each woman. Moreover, we are here more interested in knowing which predictors are responsible for the presence of diabetes. So, here we first extract the data on women who have diabetes and then will make the necessary plots. It will help us to analyze the data appropriately.

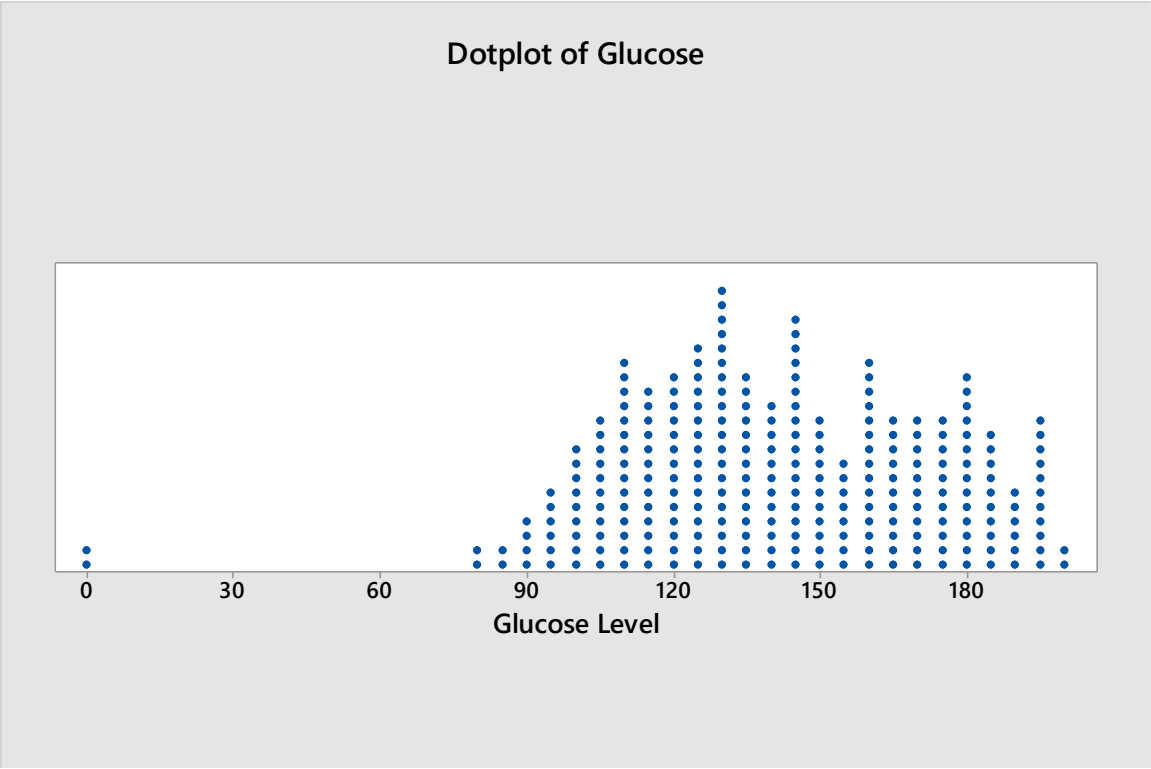
➢ Graphical Interpretation:

1. Pregnancies:



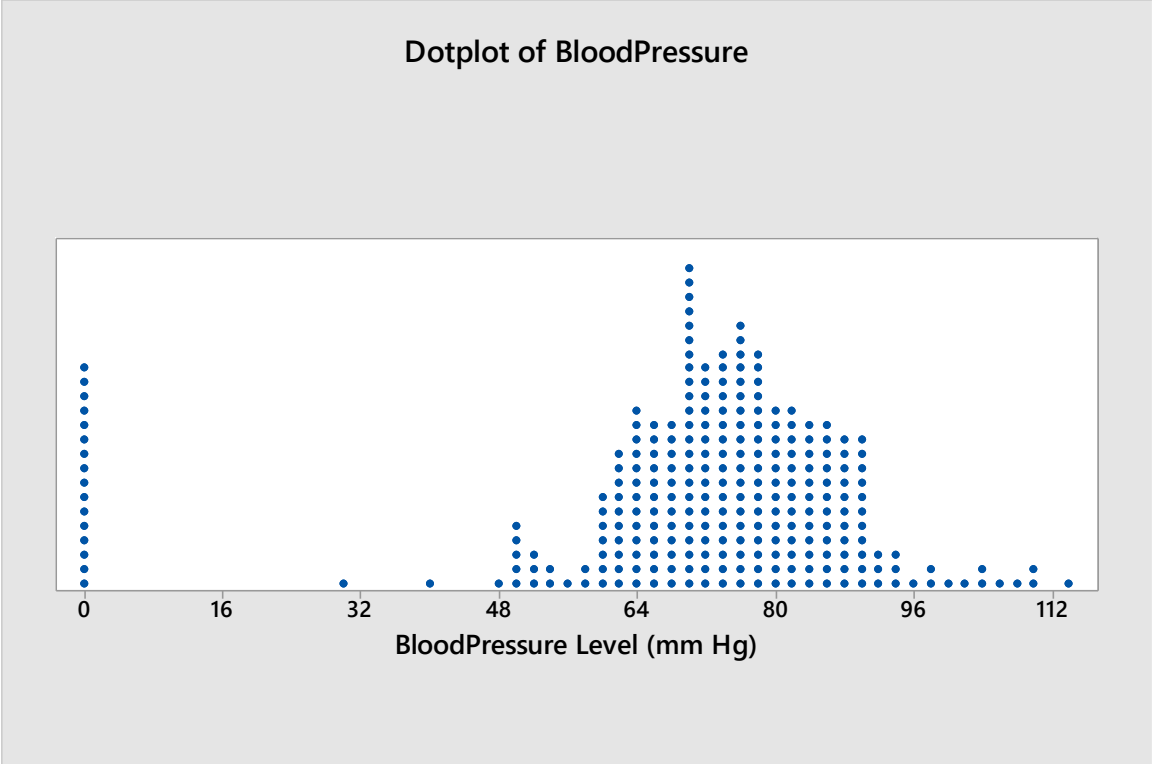
The covariate “pregnancies” is a discrete explanatory variable. So we have used dot plot for it. Also, we are here more interested in knowing which predictors are affecting the presence of diabetes. So, here we have extracted the data on women who have diabetes and make a dot plot to understand how this covariate is affecting the presence of this disease. The graph is positively skewed. So, we can say that more diabetic women have lesser no of pregnancies. Moreover, we can observe that the women having less than 4 pregnancies are more prone to develop the diabetes disease than those who have more than four pregnancies.

2. Glucose:



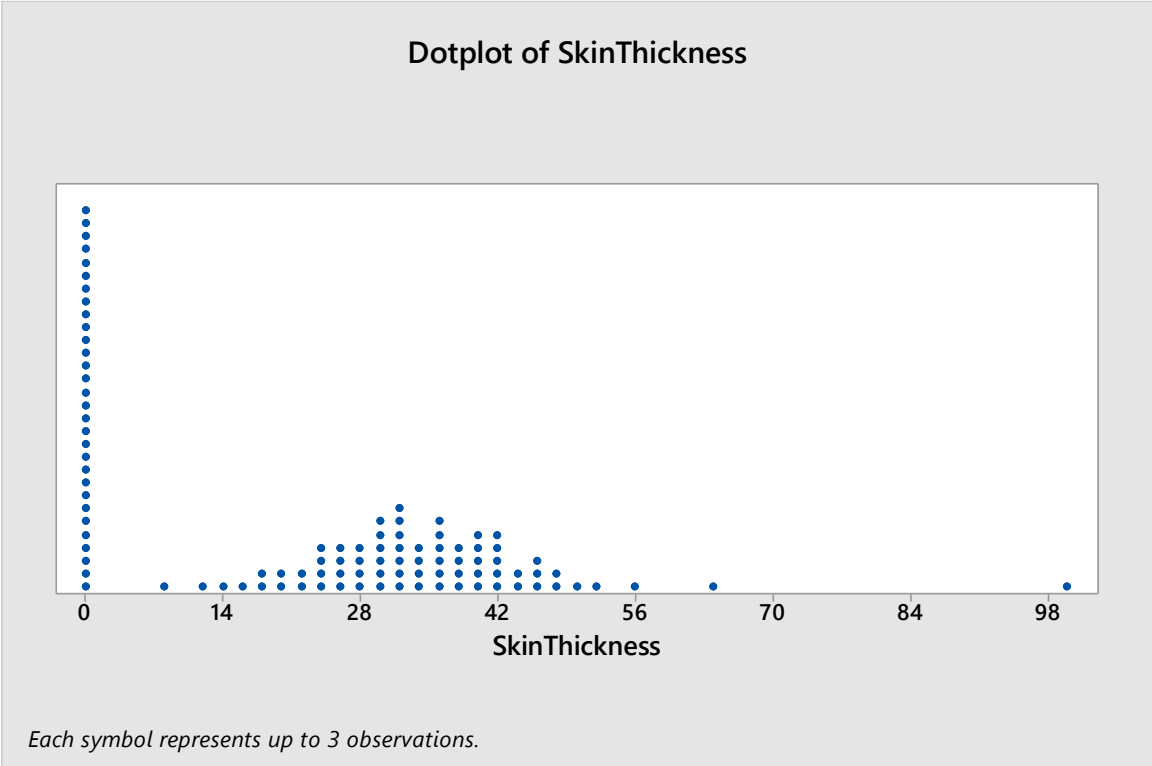
The predictor “Glucose” is also a discrete explanatory variable. So we have made a dot plot of this covariate. The graph is negatively skewed. So, we can say that more diabetic women tend to have a higher level of glucose. In particular, we can observe that women having glucose level between 125 to 180 ,have more tendency to develop the diabetes.

3. **BloodPressure(mm Hg):**



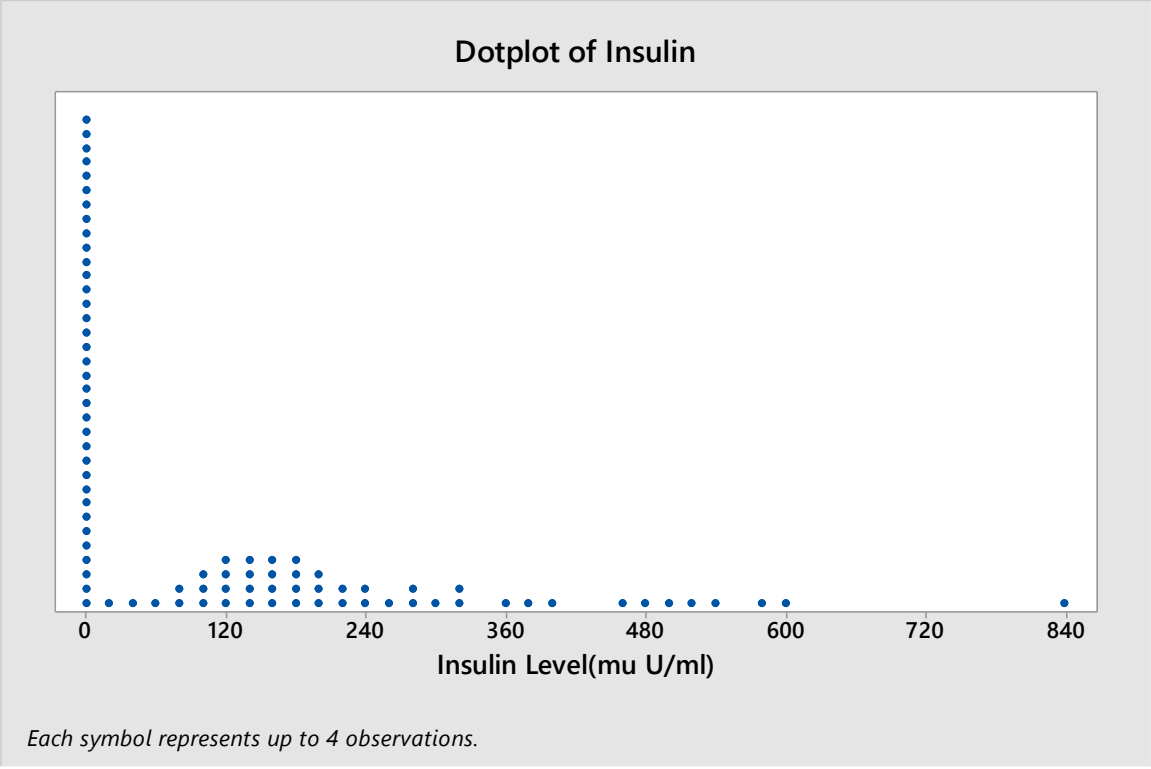
From the plot, it is clear that women having diastolic blood pressure level between 64 mm hg to 80 mm hg, have more tendency to develop the diabetes. Also, blood pressure level 0 mm hg indicates the presence of diabetes. This graph also shows slight negative correlation.

4. **SkinThickness(mm):**



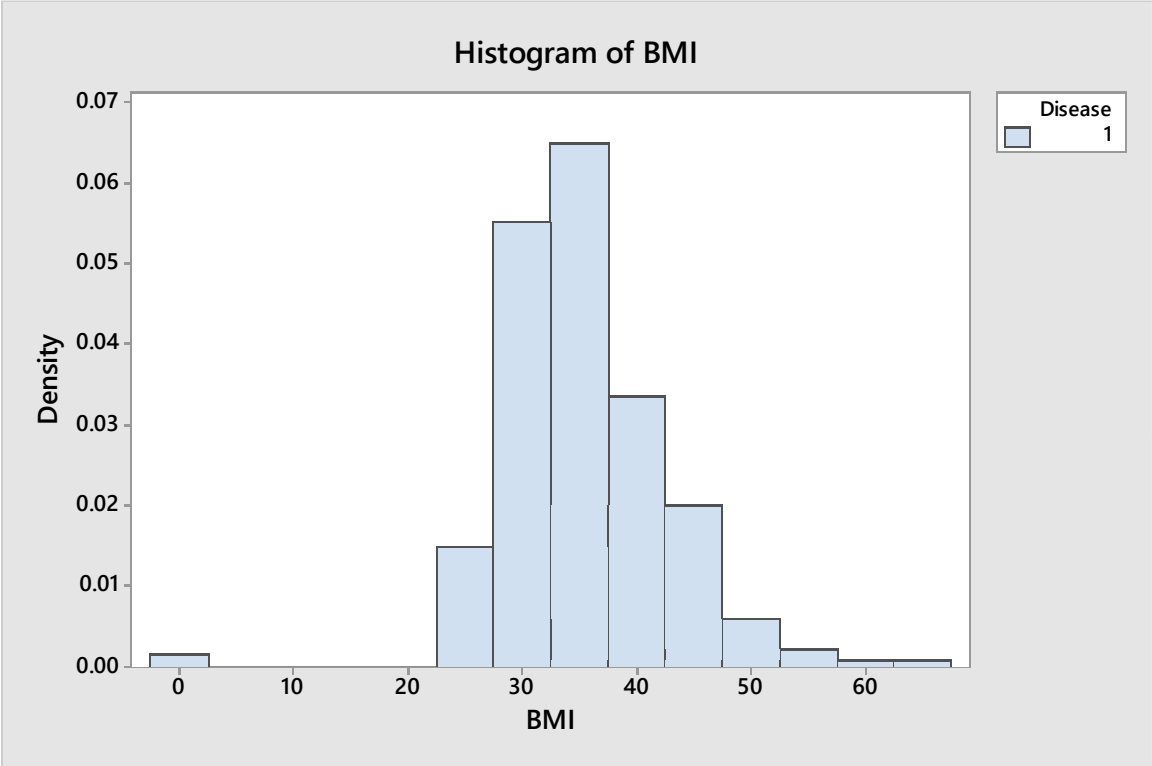
From the graph, we can say that women having skin thickness 0 mm are more prone to develop the diabetes than others.

5. **Insulin(mu U/ml):**



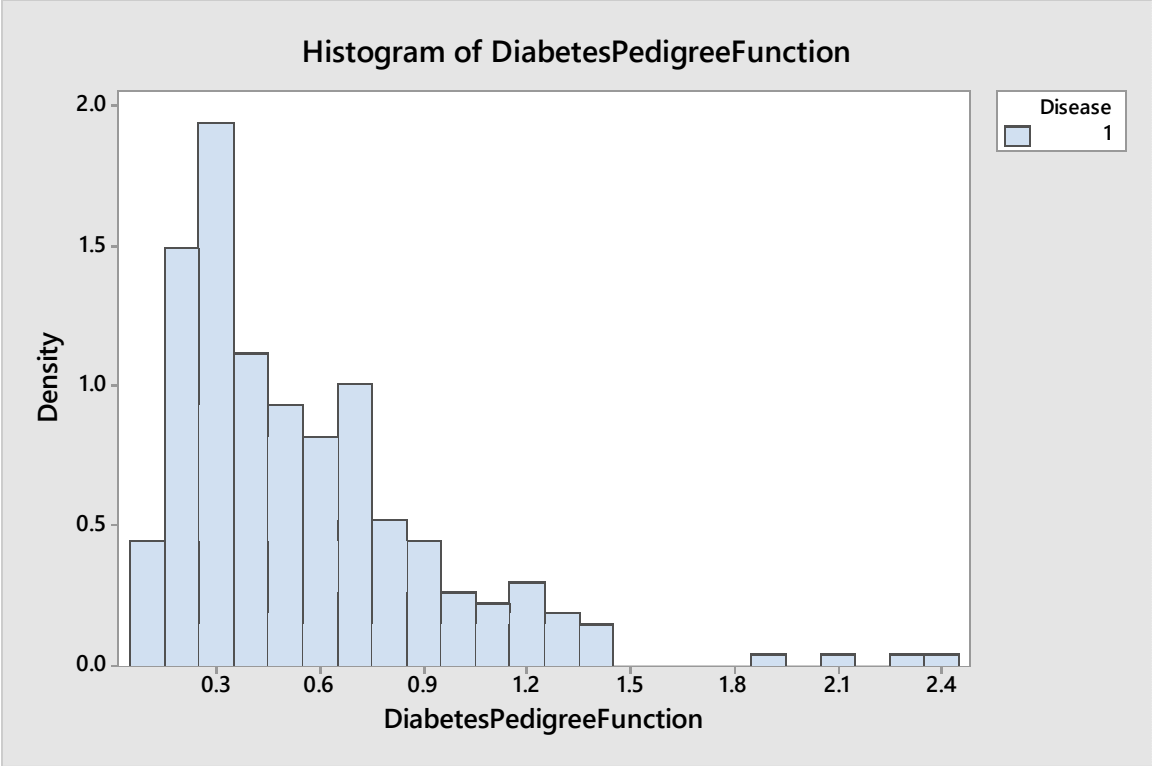
As this covariate is discrete variable, we have used dotplot for it. From the dotplot, we can see that women having insulin level 0 mu U/ml , have high tendency to develop this disease. Higher insulin level does not indicate the presence of diabetes. So, we can graphically state that insulin level is not that important in predicting the diabetes of a woman.

6. BMI (BODY MASS INDEX):



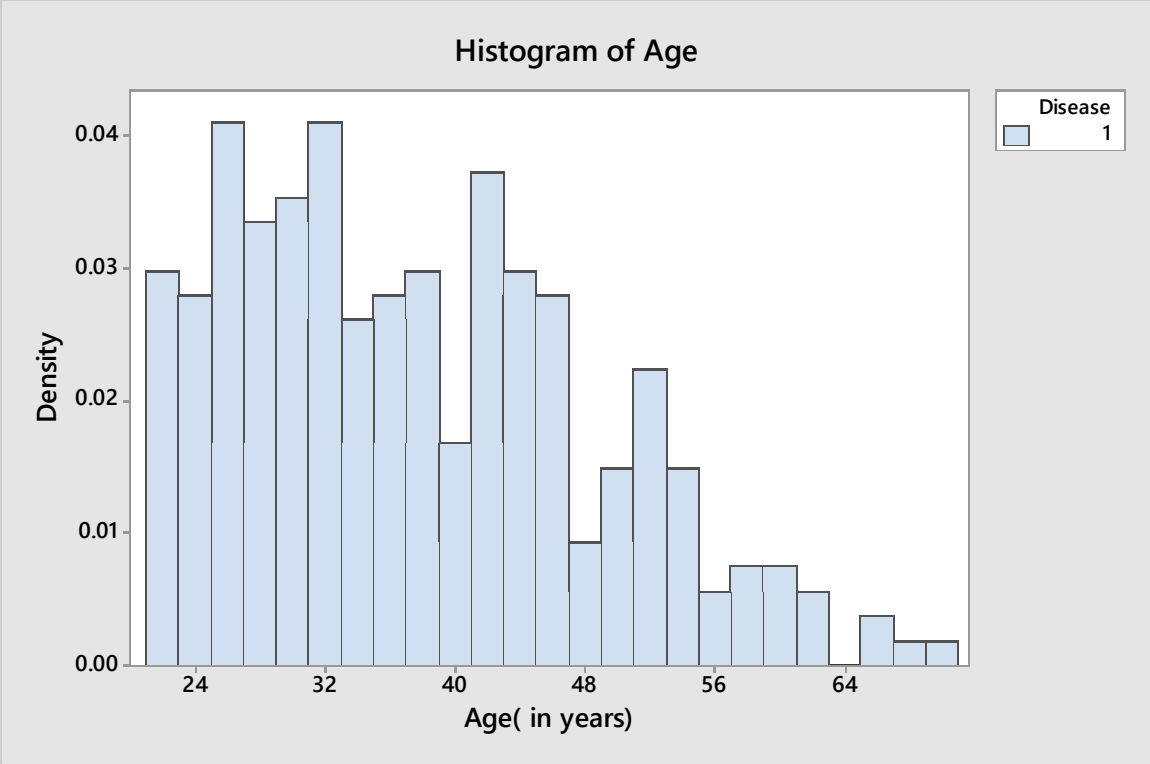
As, BMI is continuous covariate measured in kg, we have used a histogram for it. From histogram, it is clear that women having weight between 25 kg to 45 kg are more prone to develop this diabetes disease.

7. DiabetesPedigreeFunction:



From the histogram, it can be seen that women having this covariate value between 0.25 to 0.8 have tendency to develop this disease.

8. Age(years):



As, age is a continuous covariate, we have used a histogram to understand its influence on women who have diabetes, graphically. From the histogram, we can observe that the graph is positively skewed and women having age between 24 years to 47 years have high risk of developing the type two diabetes .

Output:

After fitting the model on the training dataset, we got the values of the estimates of the parameters along with their standard errors. The values are:

No.	Parameters ($\boldsymbol{\beta}$)	Corresponding covariate	Estimate ($\widehat{\boldsymbol{\beta}}$)	Standard Error(S.E.($\widehat{\boldsymbol{\beta}}$))	Z=$\frac{Estimate}{S.E}$
-	Intercept (β_0)	-	9.4127520	0.9687635	-9.716
1.	β_1	Pregnancies	0.1761778	0.0409764	4.299
2.	β_2	Glucose	0.0382632	0.0048569	7.878
3.	β_3	BloodPressure	-0.010837	0.0068521	-1.582
4.	β_4	SkinThickness	-0.003181	0.0089047	-0.357
5.	β_5	Insulin	-0.000641	0.0012095	-0.530
6.	β_6	BMI	0.0998158	0.0191977	5.199
7.	β_7	DiabetesPedigreeFunction	1.1937903	0.3899254	3.062
8.	β_8	Age	0.0103346	0.0118757	0.870

So, under logit setup, the model is given by:

$$\pi = \Pr(Y=1/x_1, x_2, \ldots, x_8) = \frac{e^{\eta}}{1+e^{\eta}}$$

Where, η =9.4127520+0.1761778*x_1 + 0.0382632*x_2 − 0.0108379 ∗ x_3 − 0.0031814*x_4 − 0.0006411 ∗ x_5 + 0.0998158 ∗ x_6 + 1.1937903*x_7+0.0103346 * x_8

Testing of Hypothesis:

Here, Z= $\frac{Estimate}{S.E}$ is the test statistic for the testing problem:

Hoxj : The covariate xj is not significant. i.e., $\beta j = 0$, using Logit model

Against

H1xj : The covariate xj is significant. i.e., $\beta j \neq 0$, using Logit model

Here, [j = 1(1)8]

The critical point here is $\tau_{0.025} = 1.96$, i.e., we reject H0 in favour of H1 if |Zobs|> 1.96 at 5% level of significance.

Here, Z~ AN(0,1)

➢ Decision:

Let us execute the testing problem for each covariate and state our decision.

No.	Corresponding Covariate	Zobs Value	Decision
1.	Pregnancies(x1)	4.299	|Zobs|>1.96 So, we reject Ho in favour of H1

2.	Glucose(x2)	7.878	|Zobs|>1.96 So, we reject Ho in favour of H1
3.	BloodPressure(x3)	-1.582	|Zobs|<1.96 So, we accept Ho.
4.	SkinThickness(x4)	-0.357	|Zobs|<1.96 So, we accept Ho.
5.	Insulin(x5)	-0.530	|Zobs|<1.96 So, we accept Ho.
6.	BMI (x6)	5.199	|Zobs|>1.96 So, we reject Ho in favour of H1
7.	DiabetesPedigreeFunction (x7)	3.062	|Zobs|>1.96 So, we reject Ho in favour of H1
8.	Age (x8)	0.870	|Zobs|<1.96 So, we accept Ho.

Note:

We have not tested for the intercept parameter as that is not important for our analysis.

➢ Interpretation:

Let, $Odds_1$= The odds of a woman to be diabetic

$$= \frac{\mathbf{Pr}(\boldsymbol{Y=1/x1,x2,\ldots..,x8})}{\mathbf{1-Pr}(\boldsymbol{Y=1/x1,x2,\ldots.x8})} = \boldsymbol{e^{\beta_0+\beta_1*x1+\beta_2*x2+\beta_3*x3+\cdots+\beta_8*x8}}$$

Also let, we are trying to interpret the covariate x1 WLG. So, we increase the value of x1 by 1, keeping the value of all the other covariates unchanged.

So, $Odds_2$= the odds of a woman to be diabetic patient, after one unit increment of the value of the covariate x1

$$= \frac{\Pr(Y=1/((x1+1),x2,\ldots..,x8)}{1-\Pr(Y=1/((x1+1),x2,\ldots..,x8)}$$

$$= e^{\beta_0+\beta_1(x1+1)+\beta_2 x2+\cdots+\beta_8 x8}$$

So, Odds Ratio = $\frac{Odds2}{Odds1} = \frac{e^{\beta_0+\beta_1(x1+1)+\beta_2 x2+\cdots+\beta_8 x8}}{e^{\beta_0+\beta_1 x1+\beta_2 x2+\cdots+\beta_8 x8}} = e^{\beta_1}$ = OR (say)

Or, $Odds_2 = e^{\beta_1} * Odds_1$

Or, Log(OR) = β_1 → So, β_1 is log odd's ratio

Now, if the sign of β_1 is positive, the odds of a woman to be diabetic will increase with the value of x1. And if it is negative, the odds will decrease with the increase of the value of x1.

Also, we can interpret that for one unit increment in value of x1, the log odd's ratio increases by the amount β_1.

Covariate Interpretation:

X1 : Pregnanacies

For the covariate, "Pregnancies", we have seen that |Zobs|> $\tau_{0.025}$ = 1.96 and as a result, we reject Hox1 in favour of H1x1 at 5% level of significance. So, we can say that the covariate "Pregnancies" is a significant covariate to determine the diabetic status of a woman.

Further, as $\hat{\beta}_1$=4.299 under logit model, we can say that for one unit increment in number of pregnancies, the odds of a woman being diabetic increases by $e^{\beta 1}$ times, i.e.,$e^{4.299}$= 73.6261 times under the logit model, keeping the value of all the other covariates unchanged.

In other words, for one unit increment in x1, the log odd's ratio increases by the amount 4.299.

X2 : Glucose

For the covariate, "Glucose", we have seen that |Zobs|> $\tau_{0.025}$ = 1.96 and as a result, we reject Hox2 in favour of H1x2 at 5% level of significance. So, we can say that the covariate "Glucose" is a significant covariate to determine the diabetic status of a woman.

Further, as $\hat{\beta}_2$=7.878 under logit model, we can say that for one unit increment in glucose level, the odds of a woman being diabetic increases by $e^{\beta 2}$ times, i.e.,$e^{7.878}$= 2638.5900 times under the logit model, keeping the value of all the other covariates fixed.

In other words, for one unit increment in x2, the log odd's ratio will be incremented by the amount 7.878.

X3:BloodPressure

For the covariate, "BloodPressure", we have seen that |Zobs|< $\tau_{0.025}$ = 1.96 and as a result, we accept Hox3 at 5% level of significance. So, we can say that the covariate "BloodPressure" is not a significant covariate to determine the diabetic status of a woman.

Further, as $\hat{\beta}_3$=-1.582 under logit model, we can say that for one unit increment in blood pressure level, the odds of a woman being diabetic decreases by $e^{\beta 3}$ times, i.e.,$e^{-1.582}$= 0.2055 times under the logit model, keeping the value of all the other covariates fixed.

In other words, for one unit increment in x3, the log odd's ratio will be decremented by the amount 1.582.

X4:SkinThickness

For the covariate, "SkinThickness", we have seen that |Zobs|< $\tau_{0.025}$ = 1.96 and as a result, we accept Hox4 at 5% level of significance. So, we can say that the covariate "SkinThickness" is not a significant covariate to determine the diabetic status of a woman.

Further, as $\hat{\beta}_4$= -0.357 under logit model, we can say that for one unit increment in skin thickness level, the odds of a woman being diabetic decreases by $e^{\beta 4}$ times, i.e.,$e^{-0.357}$= 0.6997 times under the logit model, keeping the value of all the other covariates fixed.

In other words, for one unit increment in x4, the log odd's ratio will be decremented by the amount 0.357.

X5:Insulin

For the covariate, "Insulin", we have seen that |Zobs|< $\tau_{0.025}$ = 1.96 and as a result, we accept Hox5 at 5% level of significance. So, we can say that the covariate "Insulin" is not a significant covariate to determine the diabetic status of a woman.

Further, as $\hat{\beta}_5$= -0.530 under logit model, we can say that for one unit increment in Insulin level, the odds of a woman to be diabetic decreases by $e^{\beta 5}$ times, i.e.,$e^{-0.530}$=0.5886 times under the logit model, keeping the value of all the other covariates fixed.

In other words, for one unit increment in x5, the log odd's ratio will be decremented by the amount 0.530.

X6:BMI

For the covariate, "BMI", we have seen that |Zobs|> $\tau_{0.025}$ = 1.96 and as a result, we reject Hox6 in favour of H1x6 at 5% level of significance. So, we can say that the covariate "BMI" is a significant covariate to determine the diabetic status of a woman.

Further, as $\hat{\beta}_6$= 5.199 under logit model, we can say that for one unit increment in BMI, the odds of a woman being diabetic increases by $e^{\beta 6}$ times, i.e.,$e^{5.199}$=181.0910 times under the logit model, keeping the value of all the other covariates unaltered. Also, we can say that for one unit increment in BMI, the log odds of a woman being diabetic will be incremented by 5.199 amount.

X7: DiabetesPedigreeFunction

For the covariate, "DiabetesPedigreeFunction", we have seen that |Zobs|> $\tau_{0.025}$ = 1.96 and as a result, we reject Hox7 in favour of H1x7 at 5% level of significance. So, we can say that the covariate "DiabetesPedigreeFunction" is a significant covariate to determine the diabetic status of a woman.

Further, as $\hat{\beta}_7$ = 3.062 under logit model, we can say that for one unit increment in DiabetesPedigreeFunction, the odds of a woman being diabetic increases by $e^{\beta 7}$ times, i.e.,$e^{3.062}$= 21.3702 times under the logit model, keeping the value of all the other covariates unaffected. Also, we can say that for one unit increment in diabetes pedigree function, the log odds of a woman being diabetic will be incremented by 3.062 amount .

X8: Age

For the covariate, "Age", we have seen that |Zobs|< $\tau_{0.025}$ = 1.96 and as a result, we accept Hox8 at 5% level of significance. So, we can say that the covariate "Age" is not a significant covariate to determine the diabetic status of a woman.

Further, as $\hat{\beta}_8$= 0.870 under logit model, we can conclude that for one unit increment in age, the odds of a woman being diabetic increases by $e^{\beta 8}$ times, i.e.,$e^{0.870}$= 2.3869 times under the logit model, keeping the value of all the other covariates fixed. In other words, for one unit increment in age, the log odds of a woman to be diabetic will be incremented by 0.870 amount. Although, age has no significant role in determining the diabetes of a woman.

So, covariates x1,x2,x6 and x7 are significant predictors in determining the diabetes disease of a woman.

➢ Reconstruction of the Model:

As, we have pointed out the significant covariates, we now fit again logit model on the training data, but this time only using the significant covariates. The reconstructed model is given by:

Logit Model:

$$\pi = \Pr(Y=1/\ x_1, x_2, x_6, x_7) = \frac{e^{\eta}}{1+e^{\eta}}$$

where, η = -9.516192 + 0.188030*x1 + 0.037514*x2 + 0.088841*x6 + 1.147900*x7

➢ Prediction:

From the analysis of the data, we have obtained which of the covariates are significant among all the covariates, along with the estimated values of the parameters in the model. We can clearly see that the fitted values obtained from the models are fractional values between 0 and 1. But, the outcome of a woman to be diabetic is either "Yes" or "No" i.e., either 1 or 0. So, if we are given the values for all the covariates, we can use the model in order to get a value of the outcome of a woman i.e. whether she is diabetic or not, but that would be a value between 0 and 1, from which we cannot properly say a woman is diabetic or not. Under this section, we will try to deal with this problem.

Also, here we will use the reconstructed models containing only the significant covariates.

- **Optimum Threshold:**

  "Optimum Threshold" is a value between 0 and 1, which is obtained through some proper method. If with the help of the values of all the covariates, we get a value (fitted value) for the outcome of a woman (whether she is diabetic or not) that is greater than the Optimum Threshold, we will call the woman as "diabetic" (i.e. "1" in the dataset). Otherwise, we will call her as "non-diabetic" (i.e. "0" in the dataset)

How to get the predicted value of Y after fitting the Logistic Regression Model?

Step1: Since the estimation of model parameters (reconstructed) have been done, we have got the estimated probabilities $\hat{\pi}_i$.

Step2: Now, we arrange the $\hat{\pi}_i$ values in decreasing order of magnitude, keeping the corresponding observed Y values intact. Also, we only consider the unique $\hat{\pi}_i$ values. (as there may be a repetition of $\hat{\pi}_i$ values). This unique $\hat{\pi}_i$ values will be used as threshold values.

Here, we got such 90 unique cut points or threshold values.

Step3: By taking the unique $\hat{\pi}_i\,'s$, as the cutpoint, we are to classify the response(Y) as "0" and "1". i=1(1)90

[Note: We can also take a sequence of threshold values between 0 and 1]

Let, $\hat{\mathbf{Y}} = \mathbf{1}$ if Y_{fitted} > Optimum Threshold

$= \mathbf{0}$ Otherwise

Where, Y_{fitted} is the fitted value of Y obtained from the model.

So, we can think Optimum Threshold as a cut-point in order to classify the outcome of a woman. A thumb rule is to take the Optimum Threshold as 0.5, but it is preferable to obtain the actual Optimum Threshold, especially in cases related to health.

In order to obtain the Optimum Threshold, let us state some important definitions.

- **Confusion Matrix/Misclassification Matrix:**

  Confusion matrix is a technique for summarizing the performance of a classification algorithm. The number of correct and incorrect classifications are summarized with count values and broken down by each class. The confusion matrix shows the ways in which the classification model is confused when it makes prediction.(In other words, there will be some mismatches between actual Y and predicted Y)

Let us choose a "Threshold" value from 90 such values. Also, let us have a testing dataset in hand and we have already obtained an appropriate Multiple Logistic Regression model on the training dataset. In the dataset, we will have the actual values of the response variable (0 or 1). Now, let us define,

$\hat{\mathbf{Y}} = \mathbf{1}$ if Y_{fitted} > Selected Threshold

$= \mathbf{0}$ Otherwise

Where, $\boldsymbol{Y_{fitted}}$ is the fitted value of Y obtained from the model.

So, here, we will observe the (Y, Ŷ) values, that is the actual value of the response, along with its corresponding predicted value. We will find 4 different observations, namely (1,1), (0,1), (1,0) and (0,0). Let us represent the frequencies in a matrix:

Predicted / Actual	$\widehat{\boldsymbol{Y}} = \mathbf{0}$	$\widehat{\boldsymbol{Y}} = \mathbf{1}$	**Total**
Y = 0	f_{22}	f_{21}	f_{20}
Y=1	f_{12}	f_{11}	f_{10}
Total	f_{02}	f_{01}	n

This is called the Confusion Matrix or the Misclassification Matrix for the selected threshold, where:

f_{21} : No. of women for whom Y = 0 and Ŷ = 1 (Actually they are non-diabetic but classified as diabetic patients)

Similarly, we can define f_{12}, f_{11}, and f_{22}

f_{10} : No of women for whom Y = 1, regardless the value of Ŷ (Total no of women who are actually diabetic patients)

Similarly, we can define f_{20}.

f_{01}: No of women for whom $\hat{Y}$=1, regardless the value of Y (Total no of women who are classified as diabetic)

Similarly, we can define f_{02}.

n: Total no. of woman

- **True Positive Rate (TPR) and Sensitivity:**

From the Confusion matrix, we can see that f_{11} is the number of actual "diabetic" women, who are also classified to be as "diabetic" patients with the help of the model and a selected Threshold. Also, f_{10} is the total number of women in the dataset who have actually diabetes, regardless the predicted value.

The True Positive Rate (TPR) is given by –

$$\Pr(\hat{Y} = 1 / Y = 1) = \frac{f_{11}}{f_{10}}$$

From the value of TPR, we can have an idea about how accurately the model along with that particular selected threshold correctly predicts/classifies a woman as "diabetic", who are actually "diabetic" patients. It is also known as the "Sensitivity" of the threshold value. So, we can clearly see that a high value of Sensitivity is better for a threshold. (Correct Classification)

- **False Positive Rate (FPR) and Specificity:**

From the Confusion matrix, we can see that f_{21} is the number of women in the dataset who are actually non-diabetic woman but they are predicted as diabetic patient with the help of the model and a selected Threshold. Also, f_{20} is the total

number of women in the dataset who are actually non-diabetic women, regardless the predicted value.

The False Positive Rate (TPR) is given by –

$$\Pr(\hat{Y} = 1 / Y = 0) = \frac{f_{21}}{f_{20}}$$

From the value of FPR, we can have an idea about how often the model along with that particular selected threshold makes a mistake and predicts/classifies a woman as diabetic woman, who are actually non-diabetic women (Misclassification). The value of (1-FPR) is known as the "Specificity" of the model. Here, it is obvious that a high FPR value for a threshold is not desirable, i.e., a high value of Specificity is better for a threshold.

- **Method of Finding the Optimum Threshold:**

In order to find the Optimum Threshold, let us take those 90 threshold values between 0 and 1. Now, for each selected value of threshold, we calculate the corresponding Sensitivity and Specificity values.

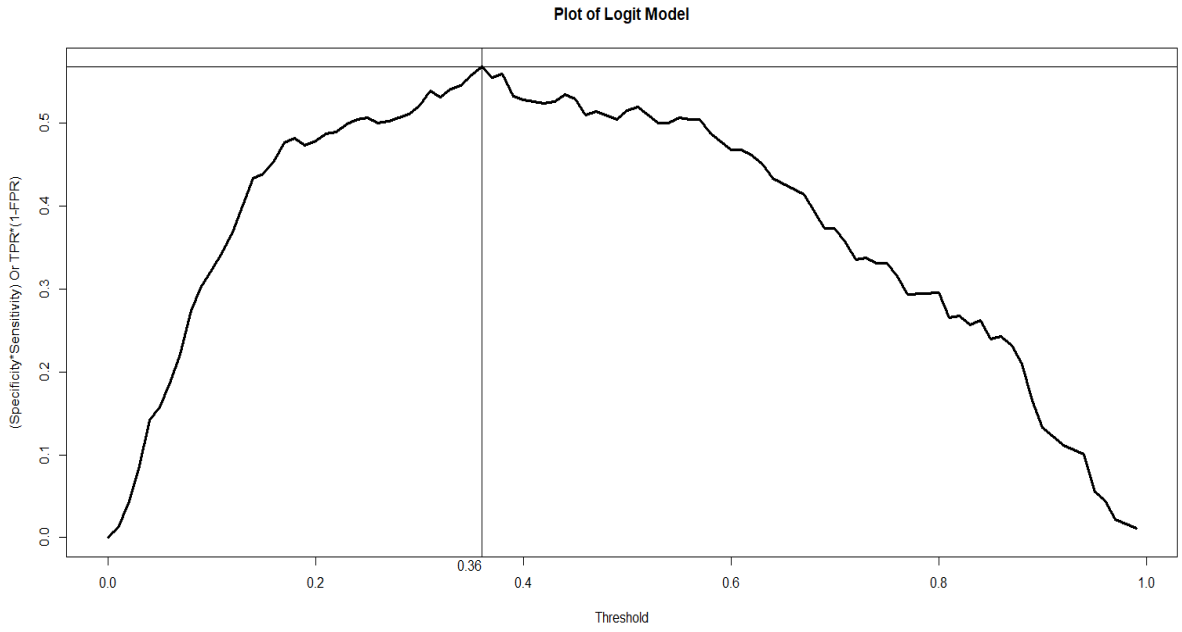
We will select that value of the threshold as "Optimum Threshold" for which the value of (Specificity*Sensitivity)* i.e (1-FPR)*TPR is highest.

- **Computation:**

For the logit model, let us first select an array of those 90 values between 0 and 1 and calculate the corresponding values of (Specificity*Sensitivity). In the table below we are reporting some of them to get an overall idea about the corresponding (**Specificity*Sensitivity**) value against a threshold.

Threshold Value	Under Logit (Specificity*Sensitivity) TPR*(1-FPR)
0.05	0.15716800
0.10	0.32218763
0.15	0.43860837
0.20	0.47840801
0.25	0.50683633
0.30	0.52010288
0.35	0.55841343
0.36	0.56734804
0.45	0.52856369
0.50	0.51536483
0.55	0.50683633
0.60	0.46764586
0.67	0.41349668
0.70	0.37315554
0.75	0.33139299
0.80	0.29565453
0.85	0.23974550
0.90	0.13239475
0.95	0.05584134
0.99	0.01123596

From the graph provided below, we can observe the value of (Specificity*Sensitivity) plotted against corresponding values of threshold.



Using R-Software, we have obtained the optimum threshold.

Logit Model threshold:0.36

So, for threshold value 0.36 , TPR*(1-FPR) has come out as maximum. Therefore, the predicted value of Y will be finally reported using 0.36 as the cut point.

- **Accuracy Of the Model:**

In order to evaluate the accuracy of the model, let us first find the Confusion Matrix for the Optimum Threshold of respective models.

Using Logit Model,

Predicted / Actual	$\widehat{Y}$ = 0	$\widehat{Y}$ = 1	Total
Y = 0	f_{22}=127	f_{21}=39	f_{20}=166
Y=1	f_{12}=23	f_{11}=66	f_{10}=89
Total	f_{02}=150	f_{01}=105	n= 255

Now, let us define:

$$\text{Accuracy} = \frac{No.of\ women\ predicted\ correctly\ by\ the\ model\ and\ optimum\ threshold}{Total\ no.of\ women}$$

Accuracy is also calculated as (1-Misclassification Probability) where Misclassification Probability is given by:

Misclassification Probability =

$$\frac{No.of\ women\ classified\ incorrectly\ by\ the\ model\ and\ optimum\ threshold}{Total\ no.of\ women}$$

Here,

$$\text{Accuracy}_{\text{Logit}} = \frac{f_{11}+f_{22}}{n} = \frac{66+127}{255} = \frac{193}{255} = 0.7568$$

Also, TPR = 66/(23+66) = 0.741573

FPR = 39/(127+39)= 0.2349398

So, we can say that the Logit Regression Model along with the optimum threshold (0.36) can predict/classify 75.68% of total diabetic woman patients accurately.

It is evident that the overall prediction accuracy of the logistic model (using only significant covariates) is moderately high.

Just for experiment, let us consider all the covariates and using the first fitted model on the training data, let us classify the response based on the testing dataset. Also, using the previous model (taking into account all the covariates), we can construct the confusion matrix and can examine how the accuracy level will change including all the covariates in the model. Moreover, we can make a comparative analysis of accuracy for two different thresholds,

- One using only the significant covariates
- And, another one using all the covariates

So, using all the covariates, we have constructed a confusion matrix, given below:

Predicted / Actual	$\widehat{Y}$ = 0	$\widehat{Y}$ = 1	Total
Y = 0	f_{22}=124	f_{21}=42	f_{20}=166
Y=1	f_{12}=23	f_{11}=66	f_{10}=89
Total	f_{02}=147	f_{01}=108	n= 255

In this case, we have got the optimum threshold value (including all the covariates) 0.3464487.

Here, $\text{Accuracy}_{\text{Logit}} = \frac{f_{11}+f_{22}}{n} = \frac{66+124}{255} = \frac{190}{255} =$ 0.745098 → Accuracy level using all the covariates

In this case, TPR= 66/(66+23) = 0.741573

FPR= 42/(42+124)= 0.253012

So, we can say that the Logit Regression Model along with the optimum threshold (0.34) can classify 74.50 % of total diabetic woman patients accurately.

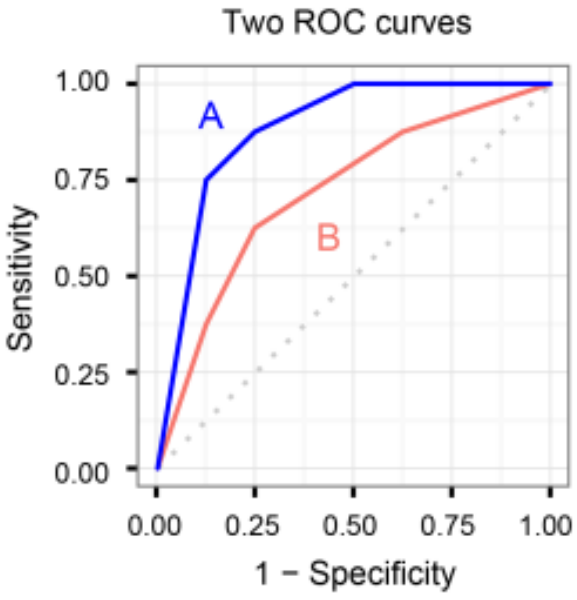
Clearly, it can be observed that when we include all the covariates, the accuracy level has decreased slightly. Whereas, when we have only taken into account the significant covariates in our model, then the accuracy level was 75.68 %, which is quite higher . Also, FPR value for this case is higher than the previous one (when only used significant covariates).

So, we can make a conclusion on the fact that, it is reasonable to use only the significant covariates in our model to find the optimum threshold value and proceed to make inference on necessary results.

- **Receiver Operating Characteristic (ROC) Curve:**

The term ROC stands for **Receiver Operating Characteristic** curve. FPR values are plotted in the x-axis while TPR values are plotted in the y-axis. A 45 degree line is drawn, which represents the baseline for ay model. If a model shows a 45^0 ROC curve, then the model yields TPR=FPR at every point and it is termed as "Useless Model". This 45 degree line is called as "chance line". The performance of a diagnostic test or model is directly proportional to the distance of its ROC curve from the 45^0 diagonal line. In other words, further the ROC curve from the diagonal line better is the performance. A diagnostic model will be called "perfect model" if its ROC curve is y axis.

In order to obtain the Optimum Threshold, we have to come across a large number of Confusion Matrices (one for each selected value of Threshold). As we have a large number of selected values of threshold, it is hard to represent how each of them are behaving as a threshold, i.e., how accurate are each one of them in predicting/classifying the diabetic status of each woman. So, a way to present that is by the ROC Curve, where for each selected threshold, we note down the value of TPR (Sensitivity) and FPR (1-Specificity). Here is an example of a ROC Curve:



We previously noted, that a high value of TPR and a low value of FPR is desirable for a threshold. So, in the example, a low value of (1-Specificity) and a

high value of Sensitivity suggests us that the corresponding threshold for one of the points which are at the top left side of the curve is the Optimum Threshold. ROC Curve is another way to identify the Optimum Threshold, but it is most useful for comparing two different models.

In the above example, we can see two curves for two different thresholds A and B. Clearly, we can say that the model A is better here, because it has more thresholds which has a high value of TPR and low value of FPR. In other words, we can say that the curve A has more area under it than B. So, from ROC Curves, the model with higher AUC (Area Under the Curve) is a better model.

- **AUC Measure:**

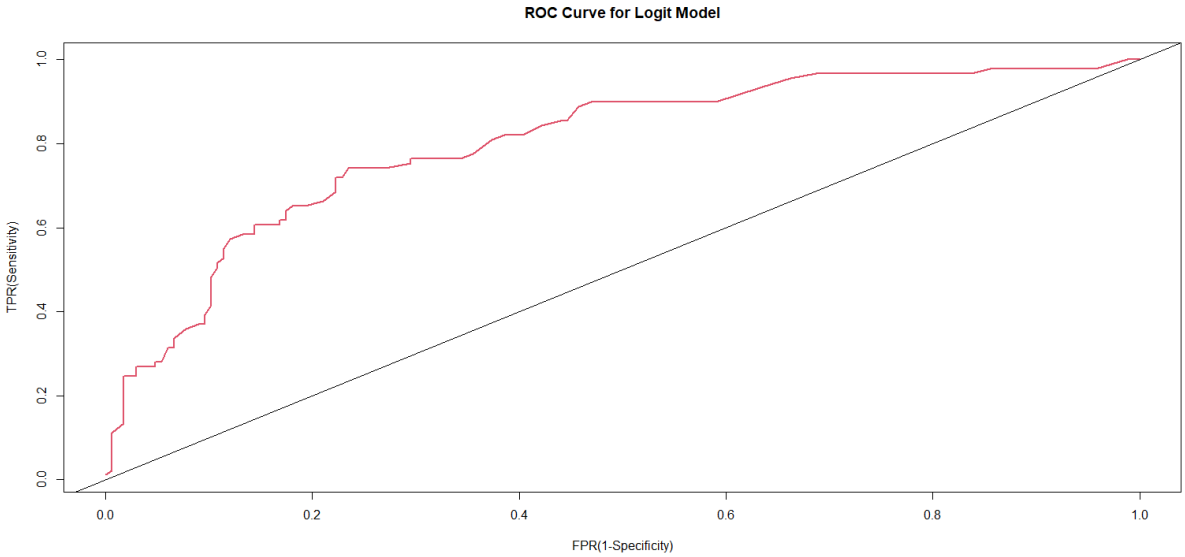
The term AUC stands for "Area under the ROC Curve". It measures the total 2D area under the ROC curve. It varies from 0 to 1. If all the classifications are wrong, then the AUC measure is 0. And, if all the predictions on the test data are correct, then AUC is 1.

There are two reasons behind AUC being an appropriate measure of accuracy.There are:

- AUC is scale-invarient. It measures how well predictions are ranked, rather than their absolute values.

- AUC is classification-threshold-invarient. It measures the quality of the model's predictions irrespective of what classification threshold is chosen.

Here, is the ROC Curve for the Logit model used in our study (using 0.36 as cutoff point) :



From R-Software, we calculate the AUC value for Logit Model to be 0.7533 . Therefore we can conclude that AUC value is moderately high, i.e moderately close to 1. So, we can state that our classification scheme is satisfactory.

❖ Conclusion:

In the project, firstly we found out a relationship between the diabetes disease and all the covariates that remained after checking for multicollinearity. From the model, we pointed out the significant covariates (i.e Pregnancies, Glucose, BMI and DiabetesPedigreeFunction), which are the most influential in affecting the diabetes status of each woman. That was our primary goal. After that, we went on to obtain the optimum threshold (0.36) using Logit model which helped us in predicting/classifying whether a particular woman has diabetes or not, given the values of all the covariates.

As mentioned in the "introduction" section, our ultimate aim is to help the public health policies at reducing the prevalence of diabetes. In order to do that, the obtained model and respective optimum threshold values will be very much helpful in not only

identifying the significant covariates that are responsible for developing the diabetes among women, but also to predict/classify that a particular woman is how much probable to develop the diabetes disease provided the values of covariates. As the model accuracy is quite high (approx. 75.68 %), so we can conclude that our ultimate goal has been achieved quite well.

❖ Scope Of Further Study:

Further, aspect of this project may include developing the model using probit link and using that probit model, one can identify the significant covariates in that case. In that case, one can also examine how the optimum threshold value will change when we use probit model and accuracy level will change or remain same. And ultimately one can make a comparison between Logit and Probit model and state which optimum threshold is better.

❖ References:

1.https://www.kaggle.com/datasets/mathchi/diabetes-data-set (for the dataset)

2.www.wikipedia.com (for various definitions and facts)

3.**Goon A.M., Gupta M.K., Dasgupta, B. (2005),** Fundamentals of Statistics, Vol II, World Press, Calcutta.

4. **Agresti, A. (2007),** An Introduction to Categorical data analysis. Wiley.

❖ Acknowledgement:

I would like to thank my supervisor and dissertation guide Dr. Surabhi Dasgupta for her guidance throughout the duration of completion of my project. I would also like to thank all my respected professors of the St. Xavier's College Statistics faculty, who have inculcated in me, a strong research mindset, curiosity and intrinsic capability to pursue this subject.

Lastly, I would like to extend my gratitude to St. Xavier's College for the opportunity to present a dissertation project paper on a topic of my choice, as well as for imbibing in me a drive to polish my research mindset.
Also, I would like to thank to my parents for their constant support and blessings.

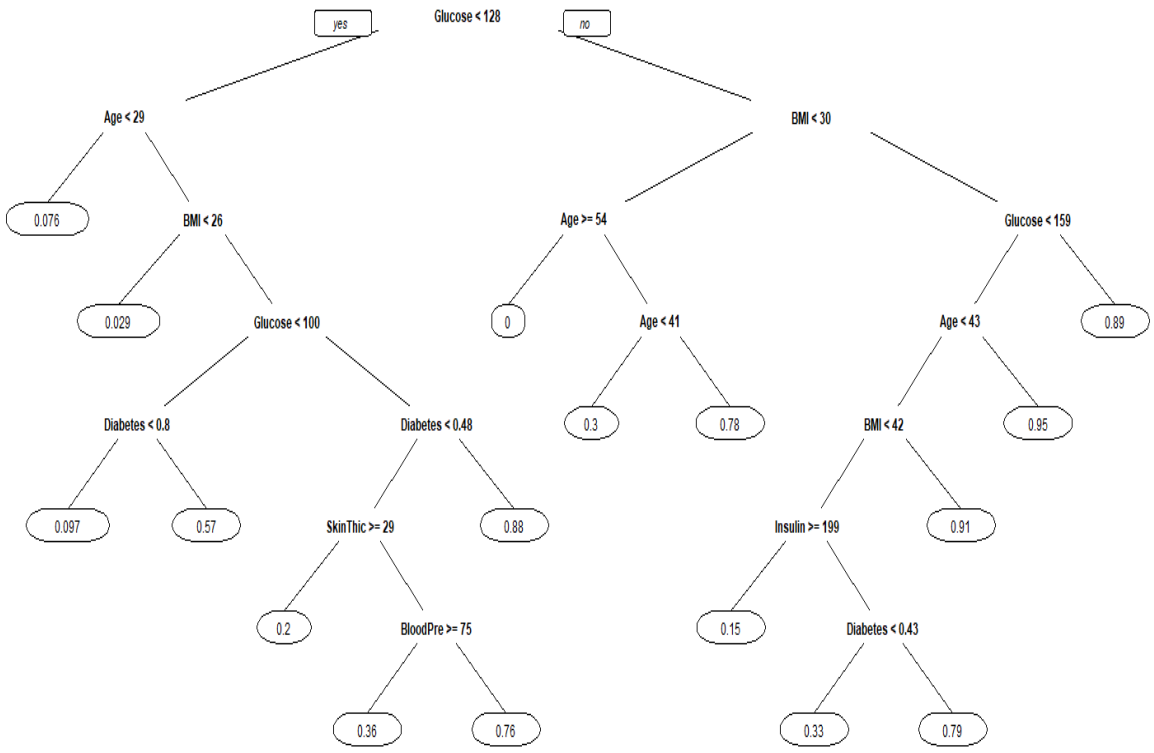
❖ Appendix:

Data Snippet:

To get an idea of the values of the covariates, the first few rows of the dataset are given below:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0

Decision Tree:



R-Code:

```
rm(list=ls())
set.seed(123)

#Calling the necessary libraries
library(caTools)
library(corrplot)
library(car)
library(pROC)

#Calling the dataset in R
data=read.csv("diabetes.csv",header=T)
attach(data)
names(data)
covariates=c( "Pregnancies","Glucose","BloodPressure" ,"SkinThickness" , "Insulin" ,"BMI"
,"DiabetesPedigreeFunction" ,"Age" )
response="Outcome"

colnames(data)=c(covariates,response)
head(data) #Data Snippet
#-----------------------------------------------------------------------------------------------------------------

#Checking Multicollinearity
#1.By checking pair-wise correlation coefficients
z1=cor(data)
corrplot(z1,method="color",type="upper",order="hclust",addCoef.col="black",t1.col="black",t1.srt=
60,
diag=FALSE)
l=1
vif1=0
```

```
vifa=0
vifb=0
verdict=0
corr1=0
for(i in 1:7)
{
for(j in (i+1):8)
{
vif1[l]=1/(1-(cor(data[,i],data[,j])^2))
corr1[l]=(cor(data[,i],data[,j])^2)
vifa[l]=i
vifb[l]=j
if(vif1[l]>4)
{
verdict[l]="correlated"
}else
{
verdict[l]="not correlated"
}
l=l+1
}
}
vif1
# 2. By checking Multiple Correlation Co-efficient
glm_log=glm(Outcome~.,data,family="binomial")
vif(glm_log)

#-------------------------------------------------------------------------------------------------------------
#Splitting The Dataset in 7:3 ratio

split=sample.split(data,SplitRatio=0.7);split
```

```
training=subset(data,split=="TRUE");training

testing=subset(data,split=="FALSE");testing

#-------------------------------------------------------------------------------------------------------------

#Fitting Of the training dataset

#With all the covariates

model=glm(Outcome~.,training,family="binomial");model

summary(model)

#Prediction(with all the covariates)

res=predict(model,testing,type="response");res

#-------------------------------------------------------------------------------------------------------------

#Fitting the model with only the significant covariates

glm.mod=glm(Outcome~Pregnancies+Glucose+BMI+DiabetesPedigreeFunction,training,family=bino
mial(link="logit"))

summary(glm.mod)

res1=predict(glm.mod,testing,type="response");res1  #Prediction

#Selecting the Optimum Threshold Value

d1=data.frame(y=testing$Outcome,res1);d1

d1=d1[order(d1$res1,decreasing=T),];d1

nrow(d1)

cutpoint=unique(round(d1[,2],2));cutpoint

n=length(cutpoint);n

obs_Y=d1$y

obs_Y
```

```
pi=d1$res1;pi

TPR=0

FPR=0

for(i in 1:n)

{

Y_hat=ifelse(pi>=cutpoint[i],1,0)

t=as.vector(table(obs_Y,Y_hat))

TPR[i]=t[4]/(t[4]+t[2])

FPR[i]=t[3]/(t[3]+t[1])

}

TPR[90]=1

FPR[90]=1

TPR

FPR

w=TPR*(1-FPR);w            #(Sensitivity*Specificity)

cutoff=cutpoint[which(w==max(w))];cutoff    #optimum threshold/cutoff point


#-------------------------------------------------------------------------------------------------------------------

#Threshold Graph

plot(cutpoint,w,type="l",lwd=3, main="Plot of Logit
Model",xlab="Threshold",ylab="(Specificity*Sensitivity) Or TPR*(1-FPR)")

abline(h=max(w),v=0.36)

mtext(0.36,side=1,adj=0.36)

#-------------------------------------------------------------------------------------------------------------------

#Accuracy

table(Actualvalue=testing$Outcome,Predictedvalue=res1>0.36)

accuracy=(66+127)/(127+39+23+66);accuracy

#-------------------------------------------------------------------------------------------------------------------

#ROC Curve

matplot(FPR,TPR,type="l",main="ROC Curve for Logit Model",col=2:4,xlab="FPR(1-Specificity)",
ylab="TPR(Sensitivity)")

abline(a=0,b=1)
```

```
#Area Under The Curve(AUC)


Y_hat=ifelse(pi>=0.36,1,0)
auc(obs_Y,Y_hat)  #based on optimum cutoff point
#-----------------------------------------------------------------------------------------------------------------------
#Decision Tree
library(rpart.plot)
library(data.tree)

#Training the decision tree classifier
tree=rpart(Outcome~.,data=training)
tree
prp(tree)

#----------------------------------------------------------END-------------------------------------------------------
```