STUDY ON THE FACTORS AFFECTING THE INCIDENCE OF GESTATIONAL DIABETES IN FEMALE PATIENTS

Dissertation paper / project to be submitted in partial fulfillment of

the requirement of the Degree of

BSc. Statistics Honours.

by

Shreyasi Mondal.

Roll No. : 3-14-21-0420

Registration No.: 041-1212-0276-20

Session : 2021-2024

Under Supervision of

Prof. Dr. Surabhi Dasgupta.

DEPARTMENT OF STATISTICS ST. XAVIER’S COLLEGE (AUTONOMOUS),
KOLKATA

DECLARATION

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

Shreyasi Mondal.

Shreyasi Mondal.

Department of Statistics

St Xavier's College

(Autonomous), Kolkata.

Date: 6th April 2024.

PAPER NAME

DISSERTATION ON GESTATIONAL DIAB ETES STSA_Roll No. 0420_Sem .6.pdf

AUTHOR

shreyasi mondal

WORD COUNT

8840 Words

CHARACTER COUNT

40737 Characters

PAGE COUNT

50 Pages

FILE SIZE

1.3MB

SUBMISSION DATE

Apr 2, 2024 12:25 PM GMT+5:30

REPORT DATE

Apr 2, 2024 12:26 PM GMT+5:30

● **12% Overall Similarity**

The combined total of all matches, including overlapping sources, for each database.

- 12% Internet database
- Crossref database
- 8% Publications database
- Crossref Posted Content database

● **Excluded from Similarity Report**

- Bibliographic material
- Cited material
- Methods and Materials
- Quoted material
- Abstract
- Small Matches (Less then 15 words)

● **16% words excluded by Custom Sections**

CONTENTS

Page No.

(1) Introduction -- 4 - 5

(2) Data Set --5 - 7

(3) Methodology ---7 - 17

(4) Results and Discussion ---18 - 40

(5) Conclusion -- 40

(6) Limitation of the Study and Future Scope --------------------------------------41

(7) References --42 - 43

(8) Acknowledgement --44

(9) Appendix --45 - 50

INTRODUCTION

Future events are necessarily uncertain, so guaranteed accurate information about the future is impossible. Prediction can be useful to assist in making plans about possible developments.

A prediction or forecast is a statement about a future event or about future data. Predictions are often based upon experience or knowledge of forecasts.

* WHAT IS DIABETES?

Diabetes is a chronic disease that directly affects the pancreas, and the body is incapable of producing insulin. Insulin is mainly responsible for maintaining the blood glucose level.

Gestational Diabetes: When a woman has uncontrolled blood sugar levels during her pregnancy it is called gestational diabetes.

Gestational diabetes can occur during pregnancy because of reduced production of insulin. Risk factors include body mass index, previously having gestational diabetes, a family history of type 2 diabetes, and having polycystic ovarian syndrome. Diagnosis is done by blood tests. For those at normal risk, screening is recommended between 24 and 28 weeks' gestation. For those at high risk, testing may occur at the first prenatal visit.

After the delivery, the blood glucose metabolism of most patients with Gestational Diabetes may return to normal, but some patients with Gestational Diabetes may develop T2DM (Diabetes Mellitus in Trimester 2). Moreover Gestational Diabetes can lead to perinatal complications, such as microsomia, dystocia, cesarean section, and hypoglycemia. At the same time, Gestational Diabetes has long-term adverse effects on T2DM in the mother and causes obesity in the offspring.

In these study they developed the risk prediction model of Gestational Diabetes based on the age of pregnant women, BMI in first trimester, parity, BP, blood lipid profile, UA, and inflammatory indices in the first trimester; thereafter, the final predictive factors of the model were determined by regression model. Then they found the sensitivity, specificity, positive predictive value and negative predictive values from the model and plotted an ROC curve.

There were short-term and long-term threats to the health of the mother and offspring due to Gestational Diabetes. In this study, it was found that age, BMI, triglycerides, and glycosylated hemoglobin in the first trimester were the independent risk factors of Gestational Diabetes. These have certain predictive value for it, and the sensitivity and specificity of the combined prediction will be higher.

OBJECTIVE OF OUR STUDY :Our main motivation for this study is mainly the correct

prediction model construction and verification methods to evaluate the risk prediction model of GESTATIONAL DIABETES, in order to predict the risk of this in the second and third trimester during pregnancy. So here our interest of this study is to correctly identify the factors affecting incidence of GESTATIONAL DIABETES.The main objective on this Gestational diabetes framework is to use statistical techniques such as regression analysis, linear regression models, categorical data analysis etc. to predict the probability of a pregnant woman developing this kind of diabetes. The framework can be applied for early disease detection, risk assessment, and the development of effective gestational diabetes prevention and treatment strategies. It can assist healthcare professionals and researchers in identifying high-risk persons and in providing targeted interventions to those most likely to benefit. By providing an accurate and efficient way of identifying individuals at risk of developing diabetes, the framework can help personalize diabetes prevention and treatment strategies and, ultimately, reduce the disease burden on individuals and society, as well as lower healthcare costs associated with diabetes management.

DATA SET

SOURCE OF DATA:

(a) Original owners: National Institute of Diabetes and Digestive and
Kidney Diseases

(b) Donor of database: Vincent Sigillito (vgs@aplcen.apl.jhu.edu)
Research Center, RMI Group Leader
Applied Physics Laboratory
The Johns Hopkins University
Johns Hopkins Road
Laurel, MD 20707
(301) 953-6231

(c) Date received: 9 May 1990

DETAILS OF THE DATA SET:

In our data set we have data on 768 individuals which has 9 attributes namely number of times pregnancies, Plasma glucose concentration a 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), Body mass index (weight in kg/(height in m)^2), Diabetes pedigree function, Age (years), Class variable or Outcome (0 or 1). Here the value of Class variable or Outcome 1 is interpreted as “tested positive for diabetes”.

Suppose we consider variables Y, X1, X2, X3, X4, X5, X6, X7 and X8. Descriptions of the variables are given below:

I. Y: denotes the Outcome variable, which takes the value 0 if a person is “tested negative for diabetes” and 1 if a person is “tested positive for diabetes”.

This is a categorical variable. In our study this is the binary response variable.

II. X1: denotes the number of times the patient was pregnant. In our original data the variable named “Pregnancies”.

III. X2: denotes 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.

IV. X3: denotes Diastolic blood pressure in mm of Hg of the individual. Diabetes damages the small blood vessels in our body, causing the walls of the blood to stiffen. This increases pressure, which leads to high blood pressure.

V. X4: denotes triceps skin fold thickness of the individual measured in mm. Skin thickness (the interface between the epidermal surface and skin fat), which is primarily determined by collagen content, is greater in patients with insulin-dependent diabetes mellitus

(IDDM).

VI. X5: denotes two-hour serum insulin (mu U/ml). Insulin plays an important role in the development of type 2 diabetes. This regulates blood sugar level (glucose) of the body.

VII. X6: denotes body mass index (weight in kg/ (height in m) ^2) of the individual.

Overweight (BMI of 2529.9 BMI) or obesity (3039.9 BMI) or morbid obesity (more than 40 BMIs) enhances the risk of developing diabetes.

VIII. X7: denotes the function that shows the chance of tested positive for diabetes based on his/her family. This measure of genetic influence gives us an idea of the hereditary risk one might have with the onset of diabetes mellitus.

IX. X8: denotes age of the individual in years. The risk for diabetes increases with age, making diabetes more common in older adults.

METHODOLOGY

STEP (1) - DATA VISUALISATION

At first we will plot the data and start visualizing the data set. Then we will interpret from the graphs and if we notice any absurd data point we will try to remove it by our next step.

STEP (2) – REGRESSION DIAGNOSTICS

After step (1) we will do Regression diagnostics in order to find outliers. Then if outliers are present we will try to remove that data point.

Outliers are any unusual observation in the context of study which are in discordance with other points.

There are two kinds of outliers:

1. Outlier in y direction (commonly called error outlier or outlier)

2. Outlier in x direction (high leverage points)

The residual is defined as the difference between the observed and fitted value of the study variable. The residuals can be viewed as the observed value of the model errors. So, it can be expected that if there is any departure from the assumption of random errors, then it should reflect in the residuals.

Thus, analyzing residuals help us in finding model inadequacies. In logistic regression, the residual is the difference between the observed probability that

Y = 1 and the predicted value that Y = 1 for any value given of the predictors.

We define, πi = P (Yi = 1)

And

$\widehat{\pi i}$ = P (Yi = 1)

Pearson Residual: $ri = \frac{\pi i - \widehat{\pi i}}{\sqrt{\widehat{\pi i}\,(1-\widehat{\pi i})}}$, for all i = 1(1) n

Standardized residual: ti $= \frac{ri}{\sqrt{1-hi}}$, for all i = 1(1) n

Where, hi is leverage for ith observation.

Rule for Detection of outlier: The ith observation is an outlier if |ti | > 2

Rule for Detection of high leverage points: The ith observation is a high leverage point if

$hi > l0 = 3 *(\frac{p+1}{n})$; where p is the number of predictors and n is the total number of observations

STEP (3) - DIVIDING THE DATA SET

Since our data set has a binary response and some numerical predictors, so at first we did data breakup. The sample size of our data set is 768. Then we used the subset data of size 500. This data set is called the Test Data set. Then we used primarily this Test data set to obtain all the results using all the steps written below. Now the new data set formed by the rest of the observations in our original data set is called the applied data set. The size of this data set is 268. Then we will apply all the results obtained from the Test data set and conclude our final findings.

STEP (4) - REGRESSION ANALYSIS

After the step (1) we will do MULTIVARIATE LINEAR REGRESSION ANALYSIS.

- In our given dataset we have data on 9 variables, so it is a multivariate data. Like Bivariate data (dataset which has 2 variables) , here also the focus study is the association between the variables. This association is studied through regressions.

One of the 9 variables is selected as a variable of interest, called as the response variables. Conventionally the response variable is labeled as y. The response sometimes is a natural choice among all the available variables, another time there can be more than one choice of the response and we select anyone.

In our model the response variable is a natural choice that is, namely, the Outcome variable, which takes the value 0 if a person is "tested negative for diabetes" and 1 if a person is "tested positive for diabetes".

The idea is to study how the remaining variables, namelyx1,x2,x3,x4,x5,x6,x7,x8 jointly influence the response variable y, through the Linear Regression model. The variables x1,x2,x3,x4,x5,x6,x7,x8 are called the regress or predictors of the model.

The response variable y can be modeled on the regress x1,x2,x3,x4,x5,x6,x7,x8. But before modeling the response and the predictors we have to check whether all the available variables are required or not. That is, we are to check whether x1,x2,x3,x4,x5,x6,x7,x8 are correlated or not. This is called Multicollinearity.

STEP (5) - MULTICOLLINEARITY

But before doing a linear regression model we have to check if any predictor variables have Multicollinearity or not.

The term "Multicollinearity" was first coined by Ragnar Frisch, which in statistics means the existence of exact or "perfect" linear relationship among some or all predictors of a multiple regression model.

In multiple regression analysis, the term multicollinearity indicates the linear relationships among the independent variables. Collinearity indicates two variables that are close to perfect linear combinations of one another. Multicollinearity occurs when the regression model includes several variables that are significantly correlated not only with the dependent variable but also to each other. Multicollinearity makes some of the significant variables under study to be statistically insignificant. Multicollinearity increases variance of the regression coefficients making them unstable, which brings problems to interpret the coefficients. That's why a proper detection of this is very important.

Techniques for Detecting Multicollinearity: The primary techniques for detecting the multicollinearity are
i) correlation coefficient, ii) variance inflation factor, iii) eigenvalue method.

In our study we will use the Variance Inflation Techniques (VIF) to detect Multicollinearity of the predictors.

VARIANCE INFLATION TECHNIQUE (VIF) : VIF quantifies how much the variance is inflated. Variance inflation factor is used to measure how much the variance of the estimated regression coefficient is inflated if the independent variables are correlated.

Calculation of VIF is given below.

VIF for the kth predictor is given as:

$$\text{VIFk} = \frac{1}{1-Rk^2}$$

$$= \frac{1}{Tolerance}$$

Where, is Rk^2 the R^2 value obtained by regressing kth predictor on the remaining predictors. Here the tolerance is simply the inverse of the VIF. The lower the tolerance, the more likely is the multicollinearity among the variables.

The value of VIF = 1 indicates that the independent variables are not correlated to each other. If the value of VIF is 1< VIF < 5, it specifies that the variables are moderately correlated to each other.

Selection Rule: We conventionally consider that the value of VIF exceeding 5 as the signs of serious multicollinearity; requiring correction.

Remedy: The easiest remedial measure to tackle severe multicollinearity is "Dropping of one of the redundant collinear variables.”

STEP (6) - LINEAR REGRESSION MODEL :

Then after completing step (3) we can prepare a Linear Regression Model.

A statistical model which is linear in its parameters is said to be a LINEAR MODEL. In our study the statistical model can be denoted as y= f(x1,x2,x3,x4,x5,x6,x7,x8, ε) -------- 1, where y is our variable of interest , and (x1,x2,x3,x4,x5,x6,x7,x8) are the predictors. The model 1 is said to be linear if it is linear in its’ parameters i.e f(x1,x2,x3,x4,x5,x6,x7,x8, ε) is of the form $\beta 0 + \beta 1$x1 + $\beta 2$x2 + ⋯ +$\beta 8$x8 , where $\beta 0$, $\beta 1$,….., $\beta 8$ are the parameters of the model.

Thus the linear model is given by

$$yi = \beta 0 + \beta 1x1i + \beta 2x2i + \cdots \beta 8x8i + \varepsilon i \qquad \text{, for all } i = 1(1)\ \text{n}$$

In our data set we can clearly observe that the response variable y is a binary variable.
Here x1, x2, x3, x4, x5, x6, x7, x8 are the covariates.

We assume,

$$yi\sim \text{Bernoulli } (\pi i) \qquad \text{, for all } i = 1(1)\ \text{n}$$

[Yi’s are non-identical but independent]

$$\text{P(Y=1| x1, x2, x3, x4, x5, x6, x7, x8)} = \frac{exp(\beta 0+\beta 1x1i+\beta 2x2i+\dots\dots\dots+\beta 8x8i)}{1+exp(\beta 0+\beta 1x1i+\beta 2x2i+\dots\dots\dots+\beta 8x8i)}$$

The logistics distribution, whose CDF is the simplified logistic function yields a good link and is given by:

$$\text{E(y)} = \frac{exp(y)}{1+exp(y)}$$

Here the systematic part in E(y) is the linear predictor and is denoted by

$$\eta i = \sum_{j=0}^{n} \beta jxj = xi\beta \qquad \text{, for all j= 1(1) 8}$$

g(μi) = ηi : g(.) is called the link function.

Here, μi= ηi , for all i= 1(1) n

Hence this is a canonical link.

i) In case of logistic regression here we use LOGIT link, defined as:

$$\eta\text{i} = ln(\frac{\pi i}{1-\pi i}) \quad \text{, for all i= 1(1) n}$$

Thus, $\pi\text{i} = \frac{1}{1+exp(-\eta i)}$, for all i= 1(1) n

or, $\mu\text{i} = ln(\frac{\pi i}{1-\pi i})$, for all i= 1(1) n

The score equations are given by,

$$\frac{\delta l}{\delta\beta 0} = \sum_{i=1}^{n}(\text{yi} - \pi i) = 0 \quad \text{, for all i= 1(1) n}$$

$$\frac{\delta l}{\delta\beta 1} = \sum_{i=1}^{n}(\text{yi} - \pi i)x1i = 0 \quad \text{, for all i= 1(1) n}$$

$$\frac{\delta l}{\delta\beta 2} = \sum_{i=1}^{n}(\text{yi} - \pi i)x2i = 0 \quad \text{, for all i= 1(1) n}$$

.
.
.

$$\frac{\delta l}{\delta\beta 8} = \sum_{i=1}^{n}(\text{yi} - \pi i)x8i = 0 \quad \text{, for all i= 1(1) n}$$

After solving the score equations the fitted values are given by

$$\widehat{yi} = \widehat{\pi i}$$

$$= \frac{1}{1+exp(-\widehat{\eta i})} \quad \text{, for all i= 1(1) n}$$

Now after partially differentiating the score equations wrt the predictors β0, β1,…, β8 we get the Fisher Information matrix . It is given by,

$$\begin{bmatrix} I11 & \ldots. & I19 \\ : & : & : \\ : & : & : \\ I91 & \ldots. & I99 \end{bmatrix}$$

The standard error of estimates is thus given by:

$$(\beta k) = \sqrt{I^{(k+1)*(k+1)}} \qquad \text{, for all } k = 0(1)8$$

ii) Now in order to fit the logistic regression model if we use PROBIT link,

$$\eta i = \Phi^{-\pi i} \qquad \text{, for all } i = 1(1)\ n$$

Taking $ui = \eta i$, for all i

we get

$$g' (\mu i) = \frac{\delta ui}{\delta \Phi(ui)}$$

$$= \frac{1}{\phi(\eta i)} \qquad \text{, for all } i$$

Where ϕ is the pdf of a N (0, 1) distribution.

The score equations are given by,

$$\sum_{i=1}^{n} \frac{(yi-\pi i)}{\pi i*(1-\pi i)} * \phi(\eta i) = 0 \qquad , \text{ for all } i$$

$$\sum_{i=1}^{n} \frac{(yi-\pi i)}{\pi i*(1-\pi i)} * \phi(\eta i) * xi = 0 \quad , \text{ for all } i$$

Now after partially differentiating the score equations wrt the predictors β0, β1,…, β8 we get the Fisher Information matrix . It is given by,

$$\begin{bmatrix} I11 & \dots & I19 \\ : & : & : \\ : & : & : \\ I91 & \dots & I99 \end{bmatrix}$$

Similarly we know that inverting the Fisher Information matrix we can obtain the Standard errors.

STEP (7) - GOODNESS OF FIT:

After completion of step (4) now we are concerned with which model gives best fit to the Test data set which is called analyzing Goodness of Fit.

Under the Classical linear regression model, with the assumption of normality of the errors, a suitable measure of goodness of fit is R^2 , i.e square of correlation coefficient between x and y.

But in the case of a logistic regression model one of the popularly used measures is given by Deviance Function.

Intuitively, it measures the deviance of the fitted logistic model with respect to a perfect model for P[Y=1 $x1, x2…. x8$].

The model which involves only one parameter for all observation and is of the form

$$yi= \mu+ €i \qquad , \quad \text{for all } i = 1(1)8$$

is called the Null model.

A benchmark for evaluating the magnitude of the deviance is the Null deviance,

$$D0= -2*\text{log-likelihood (Null model)} +2*\text{log-likelihood (Saturated model)}.$$

Now, The model with n parameter µi , for all i = 1(1)8 and is of the form

$$yi= \mu i+ €i \qquad , \quad \text{for all } i = 1(1)8$$

is called the Saturated model.

More precisely, the deviance is defined as the difference of log likelihoods between the fitted model and the saturated model. That is

$$D= 2*\text{log-likelihood (Saturated model)} -2*\text{log-likelihood (Fitted model)}$$

A measure of the goodness of fit might be given with the help of the deviance *R*^2 statistic, which is a generalization of the coefficient of determination *R*^2 used in multiple linear regression, and is given by,

$$R^2 = 1- \frac{D}{D0}$$

$$= 1- \frac{Deviace}{Null\ Deviance}$$

STEP (8) - PRODUCTION OF LOGISTIC REGRESSION MODEL

Then after step (5) we want to find the confusion matrix and the misclassification error.

LOGIT MODEL: We get as an estimate of logit model,

$$log(\frac{pi}{1-pi}) = \widehat{\beta 0} + \widehat{\beta 1^{xi}} \quad \text{, for all } i = 1(1)8$$

Which is given by a transformation,

$$\widehat{Yi} = \widehat{\pi i} = \frac{1}{1+exp(-\widehat{\eta i})} \quad \text{, for all } i =1(1)\ n$$

The problem of predicting y is carried out by choosing a threshold, say p* and assigning 1 as the predicting value if $\widehat{pi}$ > p* and 0 otherwise.

To find the optimal cut off point we use the optimal cut off function.

Here we take some more pi* values which are commonly used values like mean of $\widehat{pi}$ values and median of $\widehat{pi}$ values. Then we arrange the units and their observed values yi's in decreasing order of $\widehat{pi}$ values. Choosing each pi* has a threshold, classifying the predicted values into 1's & 0's. Through the classification, there occurs misclassification between the observed & predicted value of y for pi* .We classify the mismatches in a 2×2 matrix. These matrices are known as CONFUSION matrices.

After that we find the misclassification error for the different values of $\widehat{pi}$ and plot them against all the $\widehat{pi}$ values.
We carry out the above process for the model which fits the data most appropriately decided through Goodness of fit of the 2 models. We then select that value of $\widehat{pi}$ for which the misclassification error is minimum for the model. After that the selected pi* which is that value of pi* for which misclassification error is minimum is taken as the threshold.
Then we report the confusion matrix and misclassification error for the chosen threshold value.

STEP (9) - TESTING OF SIGNIFICANCE OF PREDICTORS

Then after step (6) we want to check which predictor variables are significant in predicting Gestational Diabetes.

Now we have obtained the estimates of the parameter above, now our objective is to test the significance of all the predictors in determining whether a person has Gestational diabetes or not.

We are to test,

$$H_{oj}: \beta 0 = 0 \quad \text{vs} \quad H1: \beta 1 \neq 0 \qquad , \text{ for all } j = 1(1)\ 8$$

The test statistic under *Hoj* is given by:

$$Tj = \frac{\widehat{\beta j}}{standard\ error\ (\widehat{\beta j})} \qquad , \text{ for all } j = 1(1)\ 8$$

Critical Region: We reject *Hoj* at α level of significance, if |*Tj obsv*| > τ α⁄2

Using R software we know,

τ α⁄2= 1.96, at α=0.05

Then we select the predictor variables which are rejected and interpret the findings.

STEP (10) - NOTING THE RESULTS FOUND FROM THE APPLIED DATA SET

Then we note all the results serially found after implementing all the steps written above.

STEP (11) - APPLYING ALL THESE RESULTS ON THE TEST DATA SET

After that we will apply all these results on our Applied data set

RESULTS AND DISCUSSION

(1) Now at first we will plot the data and interpret our findings.

1. Histogram of Age variable

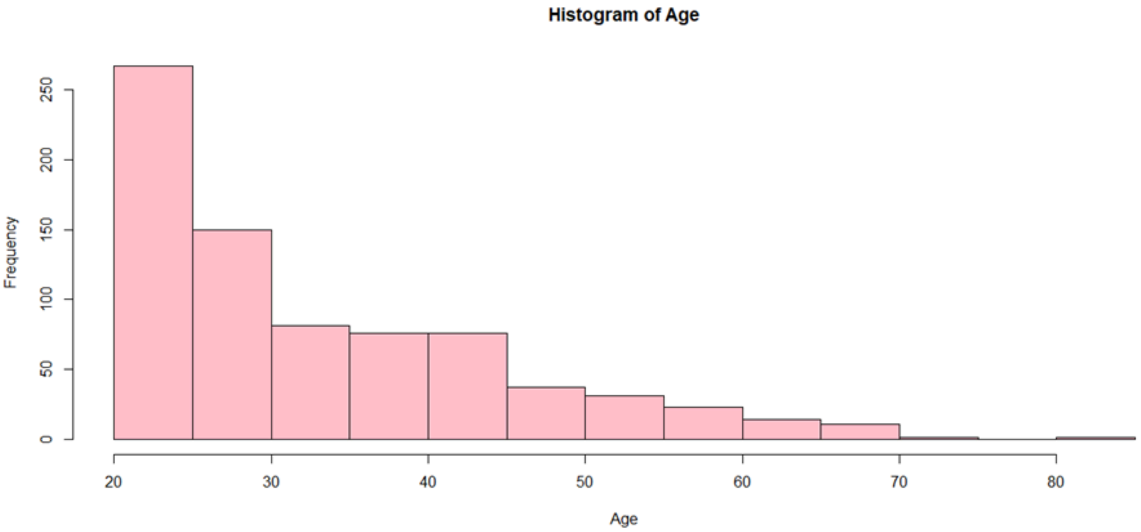


Fig - 1

Comment: From the graph we can clearly notice that most of the patients have ages between 21-30.

2. Boxplot between BMI vs Different Ages

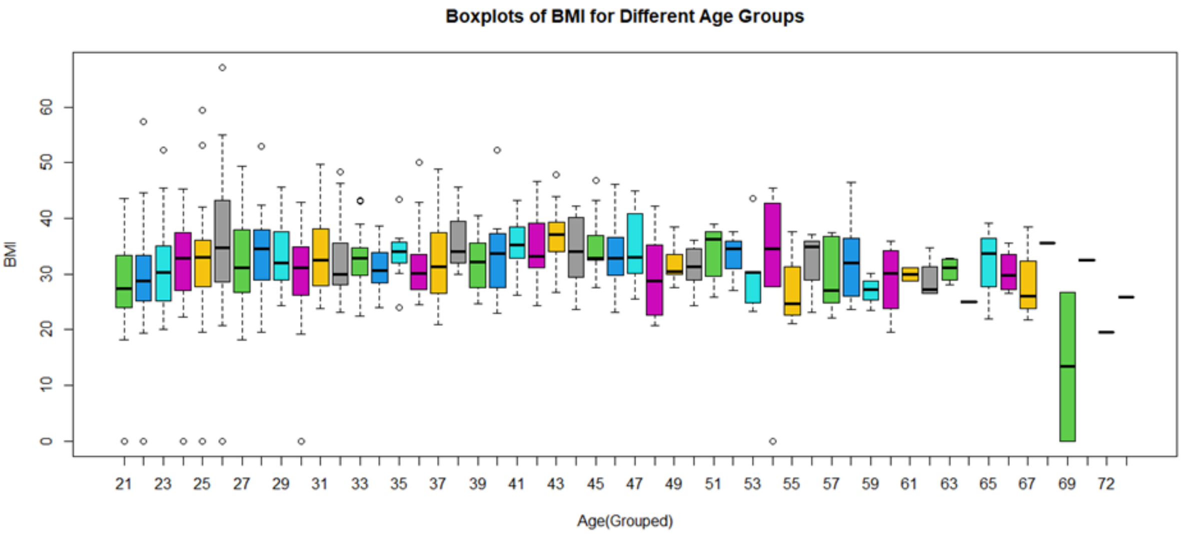


Fig - 2

Comment: We can clearly observe from the graph that BMI for all ages is more or less similar. Also, it can be seen that as the age decreases, the no. of overweight people decreases. But for certain ages we can see that 0 BMI is recorded, which is not possible, indicating error in the dataset.

3. Boxplot between BMI vs Different Pregnancy groups

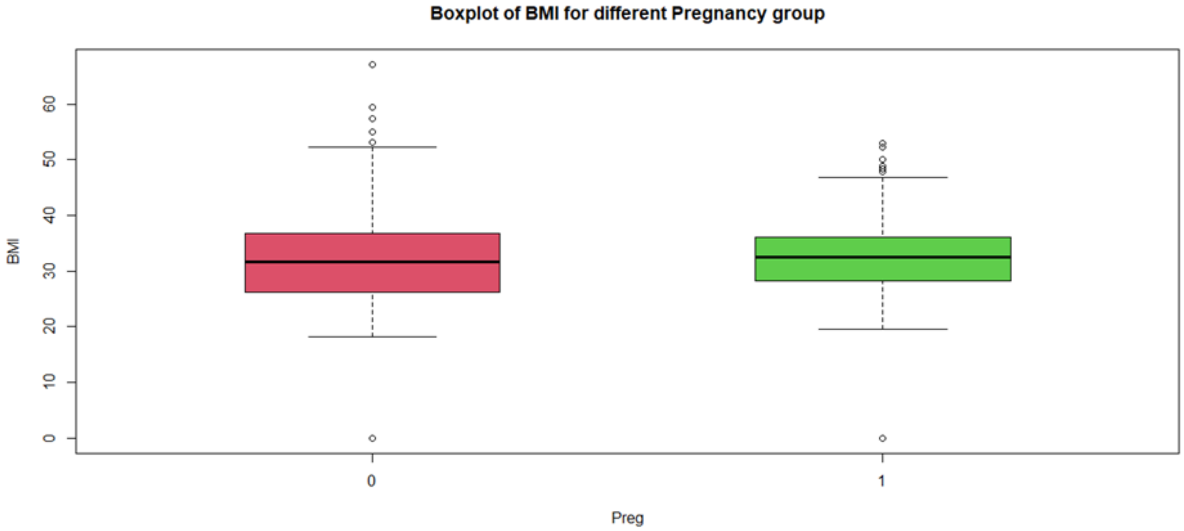


Fig- 3

Comment: We can notice that BMI is almost the same for Different Pregnancy groups. But again BMI is coming to be 0 for some cases which is not possible.

4. Box Plot between Diabetes Pedigree Function vs Different Outcome

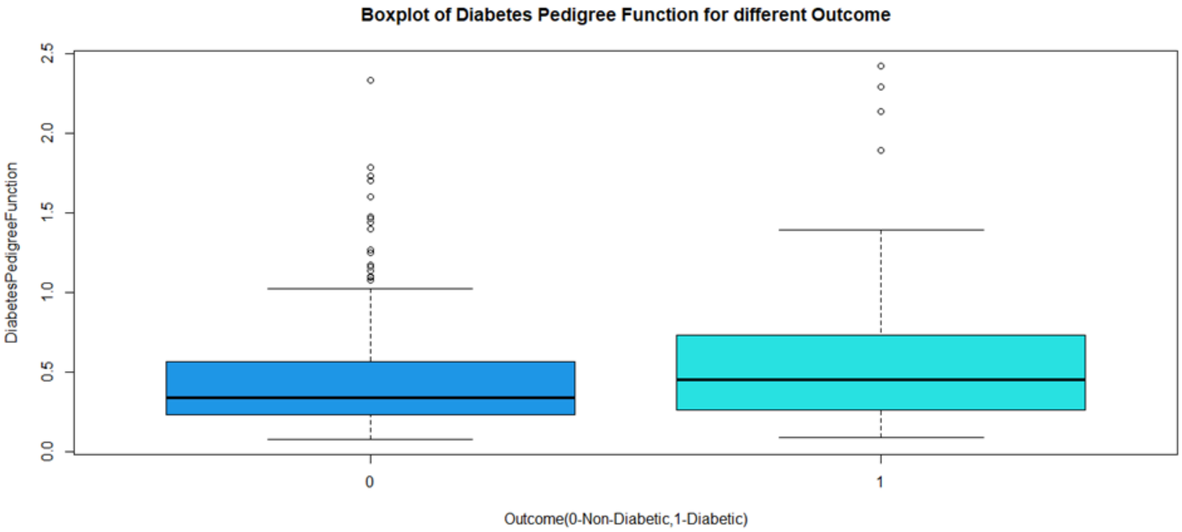


Fig- 4

Comment: On an average Diabetes pedigree function for diabetic individual is greater than that of non-diabetic individuals but we can notice that numerous outliers are present.

(2) After plotting the data set we can clearly notice that outliers may be present here. So we will now perform our next step which is step (2) Regression Diagnostics.

Here we have,

l0 = 0.0352

Fits and Diagnostics for Unusual Observation:

Obs	Outcome	Fit	Resid	Std Resid	
5	1.0000	0.8332	0.1668	0.43	X
7	1.0000	0.0574	0.9426	2.36	R
9	1.0000	0.6654	0.3346	0.85	X
10	1.0000	0.0027	0.9973	2.54	R
14	1.0000	0.6084	0.3916	1.02	X
46	1.0000	0.9644	0.0356	0.09	X
50	0.0000	0.0197	-0.0197	-0.05	X
59	0.0000	0.7332	-0.7332	-1.87	X
110	1.0000	0.1023	0.8977	2.25	R
125	1.0000	0.1801	0.8199	2.06	R
194	1.0000	1.0545	-0.0545	-0.14	X
198	1.0000	0.1535	0.8465	2.12	R
229	0.0000	1.0135	-1.0135	-2.64	R X
248	0.0000	0.6112	-0.6112	-1.57	X
285	1.0000	0.1722	0.8278	2.08	R
328	0.0000	0.8398	-0.8398	-2.11	R
333	1.0000	0.9548	0.0452	0.12	X
350	1.0000	-0.2416	1.2416	3.15	R
358	1.0000	0.9097	0.0903	0.23	X
371	1.0000	0.8531	0.1469	0.38	X
372	0.0000	-0.0071	0.0071	0.02	X
401	1.0000	0.1704	0.8296	2.08	R
446	0.0000	1.2456	- 0.2456	-0.64	X
449	1.0000	0.1845	0.8155	2.04	R
454	0.0000	0.4626	-0.4626	-1.19	X
488	0.0000	0.8841	-0.8841	-2.24	R
490	0.0000	0.8753	-0.8753	-2.21	R
503	1.0000	-0.1515	1.1515	2.92	R
538	0.0000	-0.0851	0.0851	0.22	X

580	1.0000	0.9124	0.0876	0.23	X
585	1.0000	0.3807	0.6193	1.58	X
623	0.0000	1.0073	-1.0073	-2.55	R
660	1.0000	0.1964	0.8036	2.02	R
685	0.0000	0.1381	- 0.1381	-0.35	X
707	1.0000	0.1499	0.8501	2.17	R X
745	0.0000	0.9076	-0.9076	-2.29	R

Here, R Large residual
X Unusual Outcomes

We define, Cook's Distance for the observation i:

$$Di = \sum_{j=1}^{n} \frac{\widehat{yj} - \widehat{yj(i)}}{ps^{2}}$$, for all i =1(1)n

Where, $\widehat{y(i)}$ fitted response value obtained when excluding observation i

and s2 = $\frac{\pi i - \widehat{\pi i}}{np}$, for all i =1(1)n

Detection Rule: If Di is greater than 0.5, then the i th data point may be a potential influential Point.

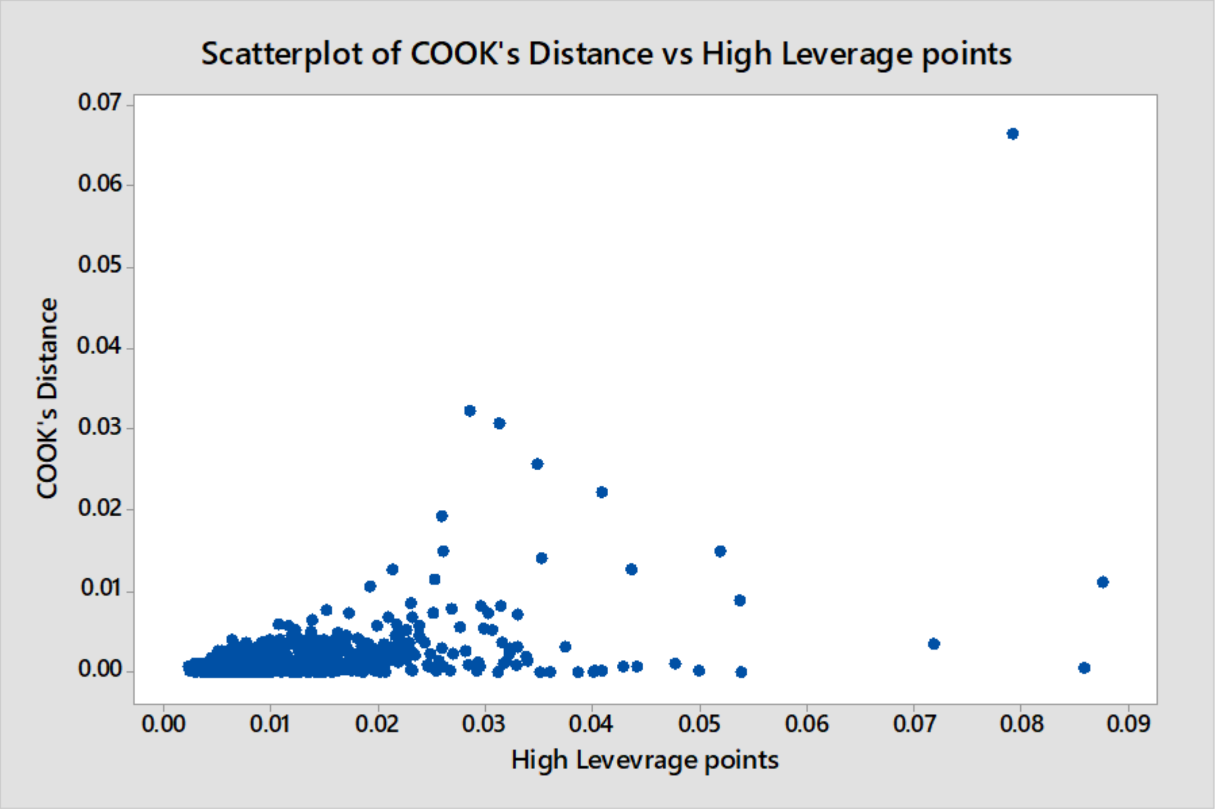


Fig - 5

Comment: Points lying above may be potential influential points. Magenta point seems to be a relatively potential influential point.

Remedy: For 229th observation (point in the above graph) the regression quantiles change markedly, for the rest of the suspected points change is not substantial. Hence, 229th observation is an influential point, thus we delete them from our dataset before further analysis.

Now we will show all the calculations of the steps written earlier and will report all our findings from these calculations. Here will also interpret each and every result obtained after doing the required calculation.

CALCULATION:

(1) Applying step (4) which is Multicollinearity we got,

Using the software R we apply on the 1st (Test data set) dataset the “VIF” function and calculate the values of VIF for all the 8 explanatory variables.

The snapshot pasted below elaborates the process of obtaining VIF and its value obtained:

Table - 1

Predictor Variables	VIF
Pregnancies (x1)	1.3434
Glucose (x2)	1.2638
Blood Pressure(x3)	1.1315
Skin Thickness(x4)	1.5647
Insulin (x5)	1.5455
BMI(x6)	1.1978
Diabetes Pedigree Function (x7)	1.0335
Age (x8)	1.4821

INTERPRETATION: According to the detection rule we can observe that the value of 8 explanatory variables is nearly 1 i.e less than 5. Hence, all the independent variables can be considered uncorrelated and should be included in the further study.

(2) Now after applying step (4) which is Linear Regression Model on the Test data set we get,
Using the software R the values of the fitted coefficients in case of logit model are given by:

Table - 2

Coefficients	Estimate	Standard Error
(Intercept) (β0)	- 7.7143	0.08647
Pregnancies (β1)	0.1158	0.03834
Glucose (β2)	0.0313	0.0045
Blood Pressure (β3)	-0.0098	0.0062
Skin Thickness (β4)	0.0025	0.0085
Insulin (β5)	-0.0011	0.00108
BMI (β6)	0.0914	0.0178
Diabetes Pedigree Function (β7)	0.9157	0.3513
Age (β8)	0.0042	0.0112

The fitted regression equation is given by:

$$Y= - 7.7143 + 0.1159\, x1 + 0.0313\, x2 - 0.0098\, x3 + 0.0025\, x4 -0.0011\, x5 + 0.0914\, x6 + 0.9157\, x7 + 0.0042\, x8$$

INTERPRETATION:

Here,

$\widehat{\beta 0}$ = -7.7143 : When all the other predictors are absent or takes the value 0 the odds of getting Gestational Diabetes decreases by $e^{-7.7143}$ times multiplicatively.

$\widehat{\beta 1}$ = 0.1159 : Keeping other predictors / covariates fixed the odds of having Gestational Diabetes increases by $e^{0.1159}$ times when Pregnancies (x1) increases by one unit.

$\widehat{\beta 2}$ = 0.0313 : Keeping other predictors / covariates fixed the odds of having Gestational Diabetes increases by $e^{0.0313}$ times when the variable Glucose (x2) increases by one unit.

$\widehat{\beta 3}$ = -0.0098 : Keeping other predictors / covariates fixed the odds of having Gestational Diabetes decreases by $e^{-0.0098}$ times when the variable Blood Pressure (x3) increases by one unit.

$\widehat{\beta 4}$ = 0.0025 : Keeping other predictors / covariates fixed the odds of having Gestational Diabetes increases by $e^{0.0025}$ times when the variable Skin Thickness (x4) by one unit.

$\widehat{\beta 5}$= -0.0011 : Keeping other predictors / covariates fixed the odds of having Gestational Diabetes decreases by $e^{-0.0011}$ times when the variable Insulin (x5) increases by one unit.

$\widehat{\beta 6}$ = 0.0914 : Keeping other predictors / covariates fixed the odds of having Gestational Diabetes increases by $e^{0.0914}$ times when the variable BMI (x6) by one unit.

$\widehat{\beta 7}$ = 0.9157 : Keeping other predictors / covariates fixed the odds of having Gestational Diabetes increases by $e^{0.9157}$ times when the variable Diabetes Pedigree (x7) by one unit.

$\widehat{\beta 8}$ = 0.0042 : Keeping other predictors / covariates fixed the odds of having Gestational Diabetes increases by $e^{0.0042}$ times when the variable Age (x8) by one unit.

Using the software R the values of the fitted coefficients in case of **probit model** are given by:

Table - 3

Coefficients	Estimate	Standard Error
(Intercept) (β0)	-4.5471	0.4772
Pregnancies (β1)	0.0691	0.02245
Glucose (β2)	0.0181	0.0025
Blood Pressure (β3)	0.0054	0.0036
Skin Thickness (β4)	0.0011	0.0053
Insulin (β5)	-0.0007	0.0006
BMI (β6)	0.0538	0.0102
Diabetes Pedigree Function (β7)	0.4968	0.2015
Age (β8)	0.0038	0.0066

The fitted regression equation is given by,

$$\begin{aligned} Y = &- 4.5471 + 0.0691\ x1 + 0.0181x2 - 0.0054\ x3 + 0.0011\ x4 \\ &- 0.0007\ x5 + 0.0538\ x6 + 0.4968\ x7 + 0.0038\ x8 \end{aligned}$$

Interpretation:

Here,

The value of $\widehat{\beta j}$, □ j = 1(1) 8 can be interpreted as the change in the probability of having Gestational Diabetes is same when the j th predictor variable changes from 0.2 to 0.3 or from 0.3 to 0.4 and so on. In short the Linear Probability Model assumes that the marginal effect of xij on P(y=1) is constant.

i.e, $(\frac{dP(y=1)}{dxij}) = \widehat{\beta j}$, for all j = 1(1) 8
i = 1(1) n

(3) Now after applying step (5) which is Goodness of Fit we have,

Using R we can easily find the value of Residual Deviance function and Null Deviance function for the LOGIT model. The values are given by 220.21, 336.36

$$R^2 LOGIT = 1- \frac{220.21}{336.36}$$

$$= 0.345$$

Now the value of Residual Deviance function and Null Deviance function for the PROBIT model is given by 493.79 and 655.68.

$$R^2 \text{ PROBIT} = 1- \frac{493.79}{655.68}$$

$$= 0.2469$$

Interpretation:

- Since in the LOGIT model the R^2 value is greater than as in the PROBIT model. Hence the LOGIT model gives a better fit.

(4) Now after applying step (6) which is obtaining the threshold value and finding the confusion matrix and misclassification error we have,

LOGIT MODEL: After using the optimal cut off function in the test data in case of logit model we have the $\hat{p}$ as 0.6065791

p* = 0.6065

≈ 0.61

Here we take a various combinations of p* as 0.24 which is the median of predicted values, 0.36 which is the mean of the predicted values, 0.5 which is a commonly used point and the optimal cut off point 0.61.

After that using the R software we find the misclassification error for all the above $\widehat{pi}$ values as 0.2684, 0.1869, 0.1103, 0.0291

Then we plot them against the various $\widehat{pi}$ values. The graph is given below,

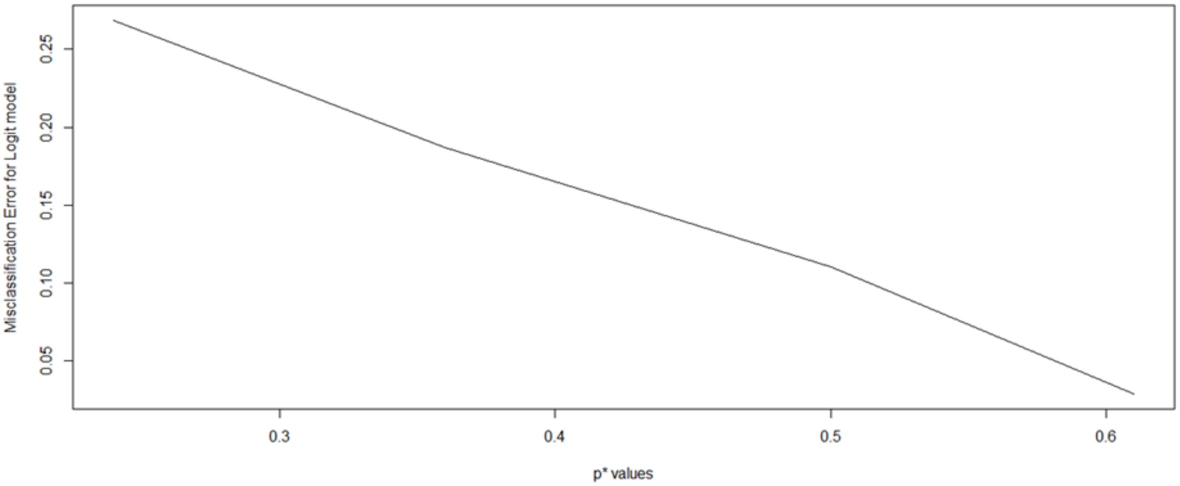


Fig - 6

Figure : Graph of Misclassification Error vs different p* values

Interpretation:

From the above graph we can clearly conclude that the misclassification error for $\hat{p}$ = 0.61 is minimum. Hence we find the confusion matrix for $\hat{p}$ = 0.61 in the case of the logit model.

The confusion matrix is given by,

Table - 4

Predicted (y) Actual (Y)	0	1
0	284	34
1	86	96

A correct classification occurs when

- $y = 0$ & $Y = 0$ or $y = 1$ & $Y = 1$

A misclassification occurs when

- $y = 1$ & $Y = 0$ or $y = 0$ & $Y = 1$

Now we define,

True Positive Rate (TPR) = $P(Y=1 | y = 1)$ and

False Positive Rate (FPR) = $P(Y=1 | y = 0)$

Total Probability of Misclassification = $P(Y=1 | y = 0) + P(Y=0 | y = 1)$

$= 1 - \text{TPR} + \text{FPR}$

Here in case of a LOGIT model we have,

TPR = 0.7385
FPR = 0.2324

Total Probability of Misclassification = 0.02911

Interpretation: Since in case of Logit model the Misclassification error is 0.02911 which is Very small then we can say that the Logit model gives a good fit of the data

set.

(5) Now after applying step (7) which is testing for significant predictors we have, The values of test statistics are given by,

Table - 5

Coefficient	Z value
Pregnancies (β1)	3.023
Glucose (β2)	6.970
Blood Pressure (β3)	-2.572
Skin Thickness (β4)	-0.291
Insulin (β5)	-1.005
BMI (β6)	5.124
Diabetes Pedigree Function(β7)	2.606
Age (β8)	2.379

In the table Z-value indicates | Tj obsv |

Decision Table:

Table - 6

Coefficients	Decision
Pregnancies (β1)	Reject
Glucose (β2)	Reject
Blood Pressure (β3)	Reject
Skin Thickness (β4)	Accept
Insulin (β5)	Accept
BMI (β6)	Reject
Diabetes Pedigree Function (β7)	Reject
Age (β8)	Reject

Interpretation:

1) Pregnancies (β1) : Pregnancies i.e no. of Pregnancies of an individual denoted by the variables x1 play a **significant** role in predicting Diabetes of a person.

2) Glucose (β2) : Glucose content denoted by the variable x2 plays a **significant** role predicting Diabetes status of the person, and higher the glucose in the the body of the individual is more likely to have Gestational diabetes.

3) Blood Pressure (β3) : Blood Pressure of an individual denoted by the variable x3 plays a relatively **less significant** role in predicting Diabetes status of the person .

4) Skin Thickness (β4) : Skin Thickness of an individual denoted by the variable x4 does **not play a significant** role in predicting Diabetes status of a person.

5) Insulin (β5) : Insulin of an individual denoted by the variable x5 does **not play a Significan**t role in predicting Diabetes status of a person.

6) BMI (β6) : BMI of an individual denoted by the variable x6 plays **a very significant** role in predicting Diabetes status of a person.

7) Diabetes Pedigree Function (β7) : Diabetes Pedigree Function of an individual denoted by the variable denoted by the variable x7 plays a **significant** role in predicting Diabetes status of a person.

8) Age (β8) : Age of an individual denoted by the variable x8 plays a **significant** role In predicting Diabetes status of a person.

(6) Results obtained from the Test data set

1) There is no MULTICOLLINEARITY present between all the predictor variables. That is the value of Variance Inflation Function (VIF) is less than 5.

2) In the Test data we concluded that for LOGIT MODEL Keeping other predictors / covariates fixed the odds of having Gestational Diabetes increases by $e^{\beta j}$ times when Pregnancies (x1), Glucose (x2), Skin Thickness (x4), BMI (x6), Diabetes Pedigree Function (x7), Age (x8) respectively increase by one unit.

3) The LOGIT model gives us a better fit as compared to the PROBIT model, because the value of R^2 is greater in case of the LOGIT model.

4) At $\hat{p}$ = 0.61, which is the optimal cut off in case of LOGIT model gives us the minimum value of the MISCLASSIFICATION error . That is when we take the threshold value as 0.61, we get the better fit.

5) After testing of significance of the predictor variables we can conclude that the variables Pregnancies (x1), Glucose (x2), Blood Pressure (x3), BMI (x6), Diabetes Pedigree (x7), Age (x8).

(7) Application of these results in point (6) in the Applied data set

Now we apply the results obtained from the Test data set on the Applied data set. After that we note our final conclusions from that. Our Applied data set is of size 268 which is obtained from sub setting the actual data set of order 768. The predictors and response remain the same as the Test data set.

We first check for multicollinearity of the predictor variables and then we will use the Logit and Probit model in our data set. Then using the Deviance Function we will verify the above result (2). After that we will select an appropriate $p\hat{}$ value and make a Confusion matrix and Misclassification error. Then we will test for significance of the available predictor variables.

Calculation:

1) Here we used R software to find the values of VIF for all the 8 explanatory variables. The snapshot pasted below elaborates the process of obtaining VIF and its value obtained:

Table - 7

Predictor Variables	VIF
Pregnancies (x1)	1.6299
Glucose (x2)	1.2711
Blood Pressure (x3)	1.4376
Skin Thickness (x4)	1.5154
Insulin (x5)	1.3548
BMI (x6)	1.3162
Diabetes Pedigree Function (x7)	1.0889
Age (x8)	1.7359

CONCLUSION: According to the detection rule we can observe that the value of VIF of all 8 explanatory variables is nearly 1 i.e less than 5. Hence, all the independent variables can be considered uncorrelated and should be included in the further study.

2) Now we will use all the 8 explanatory variables and construct a linear statistical model using the Logit model and Probit model.

The fitted coefficients using LOGIT MODEL using R software are given below,

Table - 8

Coefficients	Estimate
(Intercept) (β0)	-2.77
Pregnancies (β1)	0.5062
Glucose (β2)	0.6243
Blood Pressure (β3)	-0.3608
Skin Thickness (β4)	0.1557
Insulin (β5)	-0.4544
BMI (β6)	0.2277
Diabetes Pedigree Function (β7)	0.239
Age (β8)	0.5852

INTERPRETATION : The result (2) obtained from test data is satisfied in case of the

Applied data as here also for the LOGIT MODEL the log odds of having Gestational Diabetes is increasing for one unit increase in the predictor variables Pregnancies (x1), Glucose (x2), Skin Thickness (x4), BMI (x6), Diabetes Pedigree Function (x7), Age (x8) respectively.

3) In result (3) we notice that LOGIT MODEL gives us the better fit,

Here using R we can easily find the value of Residual Deviance function and Null Deviance function for the LOGIT model. The values are given by 220.21, 336.36

R^2 LOGIT = 1- $\frac{220.21}{336.36}$

= 0.345

4) Here applying the result (4) we will find the confusion matrix, hence we will find the Misclassification error using the threshold point p ̂= 0.61, which is given in result (4).

CALCULATION :

The confusion matrix is given by at $\hat{p}$ = 0.61

Table - 9

Predicted (y) Actual (Y)	0	1
0	173	9
1	42	44

Here,

TPR = 0.8302
FPR = 0.1954

Total Probability of Misclassification = - 0.0255

We know that a classification is best for which model the value of Total Probability of misclassification i.e (1- TPR + FPR) is minimum. Here the Total Probability of Misclassification is very less, so we can say that **LOGIT MODEL** is giving a very **GOOD FIT.**

5) Now we will use our result (5) that the predictor variables Pregnancies (x1), Glucose (x2), Blood Pressure (x3), BMI (x6), Diabetes Pedigree (x7), Age (x8) are significant in order to predict Diabetes in the Applied data set.

CALCULATION:

The values of test statistics are given by,

Table - 10

Coefficients	Z value
Pregnancies (β1)	2.260
Glucose (β2)	6.417
Blood Pressure (β3)	-2.563
Skin Thickness (β4)	-0.250
Insulin (β5)	0.506
BMI(β6)	3.521
Diabetes Pedigree Function (β7)	2.062
Age (β8)	2.387

Since the value of |z| for the variables Pregnancies (x1), Glucose (x2), Blood Pressure (x3), BMI (x6), Diabetes Pedigree (x7), Age (x8) are greater than 1.96 (τ $\alpha/2$= 1.96), then we can say that these variables are **significant** in predicting **Gestational Diabetes.**

CONCLUSION

* According to our data set of size 768 we came to the conclusion that a Logit model fits the data set appropriately. Hence the Logit model is the correct prediction model in our study.

* There were short-term and long-term threats to the health of the mother and offspring due to Gestational Diabetes. In this study, it was found that the predictor variables Pregnancies (x1), Glucose (x2), Blood Pressure (x3), BMI (x6), Diabetes Pedigree Function (x7), Age (x8) are the significant predictors inn predicting this diseases. That is the predictors written above are the independent risk factors of the disease Gestational Diabetes. These have certain predictive value for GDM, and the sensitivity and specificity of the combined prediction will be higher.

* According to our study we can also conclude that a higher level of all these significant predictors indicates that a female patient in our data set is most likely to have Gestational Diabetes.

LIMITATION OF OUR STUDY & FUTURE SCOPE

Limitation of Our Study:

There are some limitations of this study.

* This study is based on a pre-collected data set so we got to know from the description of the data set that the data collection was conducted in a single center and the sample size was small.

* Since this kind of study related to medical diseases requires a huge amount of medical data, we know that it is really hard to get medical data as it is highly confidential. That’s why our data set doesn’t have a huge size but a size of 768 only, which is small as compared to other medical studies.

* Because of the small size of our data set the results obtained from this study may not be as generic as compared to the studies based on huge data points. This is a major drawback of our study.

Future Scope:

* In future studies, multi-centre and a large sample size are needed to accommodate more patients for further study. If we consider large sample sizes then the results and models obtained from that study will be more generic and can then be easily applied on every female patient in order to detect the disease Gestational Diabetes in their 1st trimester only. Early and accurate detection of disease will effectively reduce the risk in the mothers and their offsprings as well.

REFERENCES

1. Original owners: National Institute of Diabetes and Digestive and Kidney Diseases

2. Collected data through Kaggle

   URL : https://www.kaggle.com/datasets/mathchi/diabetes-data-set

3. Prediction of Diabetes Complications Using Computational Intelligence Techniques
   By Turki Alghambadi
   Faculty of Computer and Information Systems, Islamic University of Madinah,

   Madinah 42351, Saudi Arabia

   URL: https://doi.org/10.3390/app13053030

4. American Journal of Applied Mathematics and Statistics, 2020, Vol. 8, No. 2, 39-42
   Published by Science and Education Publishing DOI:10.12691/ajams-8-2-1
   Available online at

   URL: https://pubs.sciepub.com/ajams/8/2/1

5. Analysis of Categorical Data with R: 113 (Chapman & Hall/CRC Texts in Statistical Science)

6. Fundamentals of Statistics, Volume 1:A.M. Goon, M.K. Gupta, B. Dasgupta.

7. Fundamentals of Statistics, Volume 2:A.M. Goon, M.K. Gupta, B. Dasgupta

8. Diabetes – Wikipedia

   URL: https://en.wikipedia.org/wiki/Diabete

9. EPIDEMIOLOGY/HEALTH SERVICES/PSYCHOSOCIAL RESEARCH| JULY 01 2003
   Population Health Significance of Gestational Diabetes
   N. Wah Cheung, PHD; Karen Byth, PHD

   URL:
   https://diabetesjournals.org/care/article/26/7/2005/26732/Population-Health-Significance-of-Gestational

10. Risk prediction model of gestational diabetes mellitus based on nomogram in a Chinese population cohort study

- Xiaomei Zhang,
- Xin Zhao,
- Lili Huo,
- Ning Yuan,
- Jianbin Sun,
- Jing Du,
- Min Nan &
- Linong Ji

Scientific Reports volume 10, Article number: 21223 (2020)

URL: https://www.nature.com/articles/s41598-020-78164-x

ACKNOWLEDGEMENT

I would like to take this opportunity to thank everyone who has helped me and supported me throughout my dissertation work.

First and foremost, a big thanks to my supervisor Prof. Dr. Surabhi Dasgupta, for giving me her valuable time, her immense support and involvement, when it comes to the formulation of my topic and sharing her resourceful insights. Thanks a lot ma'am for your constant guidance and pointing me out in the right direction.

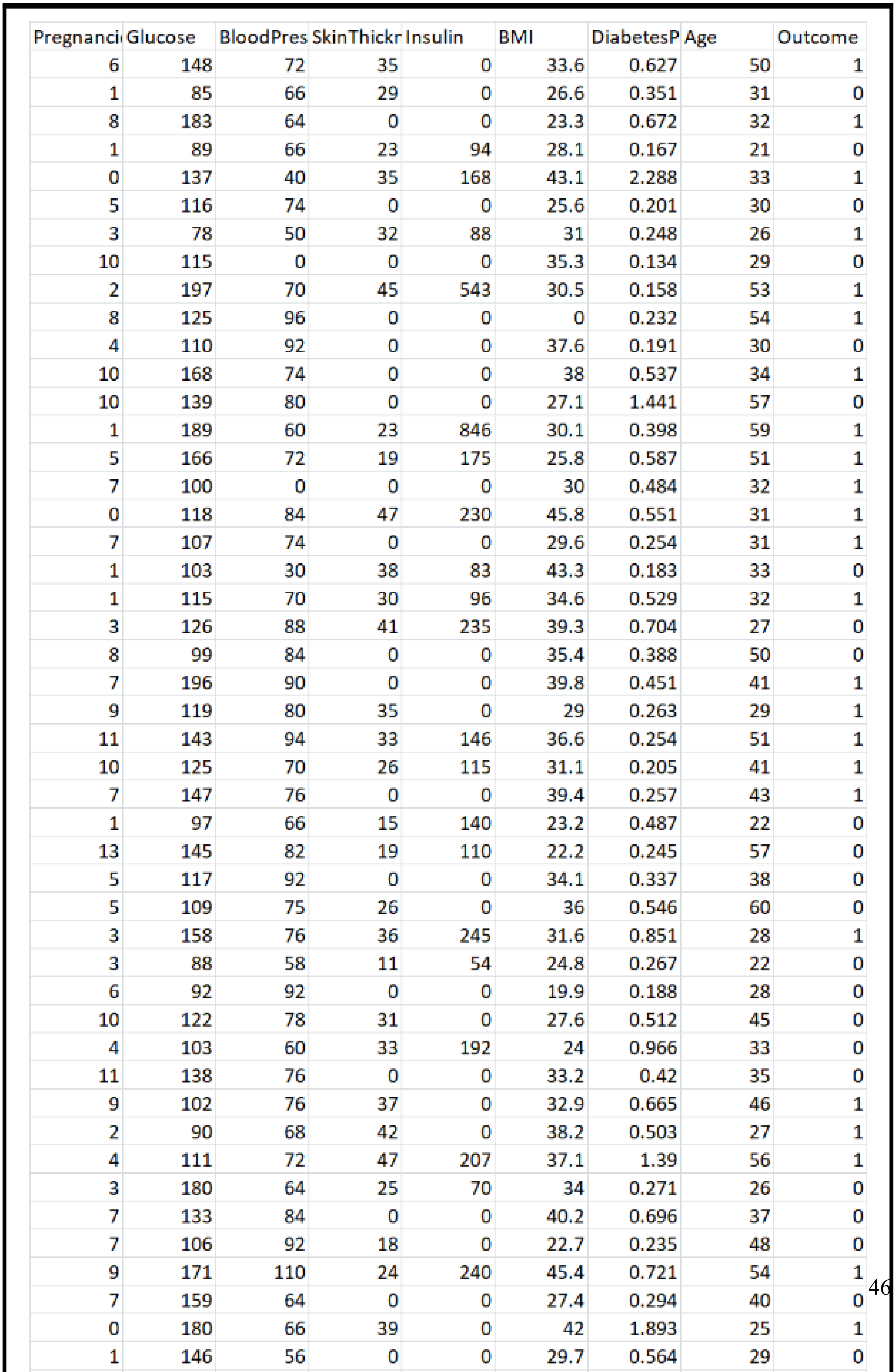
I am also grateful to our wonderful principal, Rev. Dr. Dominic Savio, S.J, who has created a nurturing and creative environment in our college.

I would also like to take a moment to acknowledge my other professors in the Department of Statistics at St. Xavier's College, Kolkata, Prof. Dr. Durba Bhattacharya, Prof. Dr. Ayan Chandra, Prof. Debjit Sengupta, Prof. Dr. Surupa Chakraborty, Prof. Pallabi Ghosh, Prof. Madhura Dasgupta, Prof. Dr. Sancharee Basak & Prof. Rahul Roy, who all have helped me to develop the mind-set prone to research, which has made it possible for me to complete the project.

It would be unfair to not mention the support and help I had received from my friends group, as they presented their ideas, which was duly put to use in my research work. Lastly, big thanks go to my parents for their supporting nature and constant guidance throughout my undergraduate course

APPENDIX

A snapshot of the 1st part of our data set is given below. Since the size of our data set is 768 which is quite big so I have given a snapshot of the 1st part otherwise the project will become huge.



Pregnanci	Glucose	BloodPres	SkinThickr	Insulin	BMI	DiabetesP	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1
9	119	80	35	0	29	0.263	29	1
11	143	94	33	146	36.6	0.254	51	1
10	125	70	26	115	31.1	0.205	41	1
7	147	76	0	0	39.4	0.257	43	1
1	97	66	15	140	23.2	0.487	22	0
13	145	82	19	110	22.2	0.245	57	0
5	117	92	0	0	34.1	0.337	38	0
5	109	75	26	0	36	0.546	60	0
3	158	76	36	245	31.6	0.851	28	1
3	88	58	11	54	24.8	0.267	22	0
6	92	92	0	0	19.9	0.188	28	0
10	122	78	31	0	27.6	0.512	45	0
4	103	60	33	192	24	0.966	33	0
11	138	76	0	0	33.2	0.42	35	0
9	102	76	37	0	32.9	0.665	46	1
2	90	68	42	0	38.2	0.503	27	1
4	111	72	47	207	37.1	1.39	56	1
3	180	64	25	70	34	0.271	26	0
7	133	84	0	0	40.2	0.696	37	0
7	106	92	18	0	22.7	0.235	48	0
9	171	110	24	240	45.4	0.721	54	1
7	159	64	0	0	27.4	0.294	40	0
0	180	66	39	0	42	1.893	25	1
1	146	56	0	0	29.7	0.564	29	0

R codes:

```
rm(list=ls())
library(corrplot)
library(car)
library(dplyr)
library(InformationValue)
library(caret)
library(ISLR)
library(cutpointr)




setwd("C:/Users/USE/Downloads")
getwd()
data=read.csv(file="diabetes.csv")
View(data)
attach(data)
data.500=read.csv(file="diabetes.subset.csv")
data.500
View(data.500)
data.267=read.csv(file="diabetes.data.267.csv")
data.267
View(data.267)
Preg=0
for(i in 1:length(Pregnancies))
{
 if(Pregnancies[i]<=3)
 Preg[i]=0
 else
 Preg[i]=1
}

Preg
plot(data)
hist(Age,xlab = "Age",col = "pink",main = "Histogram of Age",breaks =20,bg=2)




# The max no. of patients suffering from Diabetes belongs to ages between
21-30



boxplot(BMI~as.factor(Age),col=3:8,xlab = "Age(Grouped)",main="Boxplots of
BMI for Different Age Groups")
```

```
boxplot(BMI~Preg,main="Boxplot of BMI for different Pregnancy
group",col=2:3,boxwex=0.5)


boxplot(DiabetesPedigreeFunction~Outcome,main="Boxplot of Diabetes Pedigree
Function for different Outcome",xlab =
"Outcome(0-Non-Diabetic,1-Diabetic)",col=4:5)


## OBTAINING RESULTS USING TEST DATA

#fitting of model, multicollinearity, test of significance ##


## LOGIT MODEL ##


model=glm(Outcome~.,family = binomial(link = "logit"),data=data.500)
summary(model)
vif(model)


## PROBIT MODEL ##


model2=glm(Outcome~.,family = binomial(link = "probit"),data=data.500)
summary(model2)
vif(model2)


#Prediction and evaluation of quality of model (logit model) ##
## LOGIT OPTIMAL ##

mean(predicted)

median(predicted)

predicted=predict(model,data.500,type = "response")

optimal=optimalCutoff(data.500$Outcome, predicted)

p0=optimal


Y.hat.logit=ifelse(predicted>p0,1,0)

Y.hat.logit
t.logit=table(data.500$Outcome,Y.hat.logit)
t.logit

TPR.logit=t.logit[2,2]/(t.logit[1,2]+t.logit[2,2])
FPR.logit=t.logit[2,1]/(t.logit[2,1]+t.logit[1,1])
TPR.logit
FPR.logit
```

```
m.error.logit=1-TPR.logit-FPR.logit
m.error.logit



## TRYING OUT VARIOUS P* VALUES ##


## LOGIT MODEL ##


P_star.logit=c(0.24,0.36,0.5,0.61)
TPR.l=FPR.l=m.logit=array(0)
k=1
for(i in p.logit)
{
 Y.hat.l=ifelse(predicted>i,1,0)
 t.logit=table(Y.hat.l,data.500$Outcome)
 TPR.l[k]=t.logit[2,2]/(t.logit[1,2]+t.logit[2,2])
 FPR.l[k]=t.logit[2,1]/(t.logit[2,1]+t.logit[1,1])
 m.logit[k]=1-TPR.l[k]-FPR.l[k]
 k=k+1
 }
Y.hat.l
TPR.l
FPR.l
m.logit=c(0.2685,0.1869,0.1103,0.0291)
plot(p_star.logit,m.logit,type="l",ylab="Misclassification Error for Logit
model",xlab="pi hat values")



## FOR THE VALUE p0=0.61 WE HAVE LEAST MISCLASSIFICATION ERROR IN CASE OF
LOGIT MODEL ##






## APPLIED DATA SET ##

## LOGIT MODEL ##

model.267=glm(Outcome~.,family = binomial(link = "logit"),data=data.267)
summary(model.267)
vif(model.267)

## PROBIT MODEL ##

model2=glm(Outcome~.,family = binomial(link = "probit"),data=data.267)
summary(model2)
vif(model2)
```

```
predicted=predict(model.267,data.267,type = "response")
optimal=optimalCutoff(data.267$Outcome, predicted)
p0=optimal
p.l=0.61


Y.hat.logit.267=ifelse(predicted>0.61,1,0)
Y.hat.logit.267

t.logit.267=table(data.267$Outcome,Y.hat.logit.267)
t.logit.267

TPR.logit.267=t.logit.267[2,2]/(t.logit.267[1,2]+t.logit.267[2,2])
FPR.logit.267=t.logit.267[2,1]/(t.logit.267[2,1]+t.logit.267[1,1])


TPR.logit.267
FPR.logit.267


m.error.logit.267=1-TPR.logit.267-FPR.logit.267
m.error.logit.267
```