PREDICTION OF DIABETES IN FEMALES BASED ON THE DIFFERENT HEALTH PARAMETERS

Name: **RISHIRAJ SUTAR**

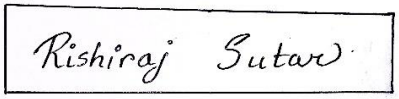
Roll No: **435**

Department: **STATISTICS**

Registration No: **A01-1122-0833-19**

Supervisor: **DR. SURUPA CHAKRABORTY**

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

signature 

CONTENTS:

PAGES

INTRODUCTION

What is diabetes?

Diabetes Mellitus is a physical condition in which the pancreas does not produce enough insulin. Insulin is a hormone which controls the amount of glucose level by converting the excess glucose in the blood into glycogen. If the body lacks insulin, the excess glucose remains in the blood thereby increasing the blood sugar level and resulting in diabetes mellitus. Doctors often use the full name diabetes mellitus instead of only diabetes alone, in order to distinguish this disorder from diabetes insipidus which is a completely different disease.

India is known as the diabetes capital of the world with around 50 million people suffering from it. Diabetes is often called the silent killer since it kills the affected person slowly over the years. It can lead to kidney failure, atherosclerosis, neuropathy in the limbs and blindness. Frequent urination and thirst; loss of weight, blurry vision, dryness of the skin are the major symptoms which helps in suspecting diabetes mellitus.

However, doctors and medical researchers have said that, if the disease can be detected at an early stage followed by proper treatments, then it can go a long way in helping the patients live a normal life.

Diabetes in females

Between the year 1971 and 2000, the Annals of Internal Medicine had conducted a study where it stated that, the Death Rates for men with diabetes had fallen significantly. This decrease reflects the medical advancement in diabetes treatment. But at the same time, the survey also indicated that the Death Rates for women with diabetes did not improve much. The death rates in diabetes are higher among the women for the following reason:

- Diabetes increases the risk of heart disease by almost four times in women than in men. They are also at higher risk of other diabetes-related complications such as blindness, kidney, depression, etc
- Some of the diabetes related complications are difficult to diagnose in women than in men
- Women often develop gestational diabetes during pregnancy which can often become dangerous.
- Hormones and inflammation act differently in women.

Though diabetes is a deadly disease, it often proves to be more fatal in women than in men. There is no permanent cure for this disease and once diagnosed with it, one can only manage its symptoms. In United States, an estimated 7.3 million adults ages 18 years or older have

diabetes but are undiagnosed (21.4 percent of adults with diabetes). Thus, the early detection of diabetes is crucial part in the diagnosis of diabetes.

SOURCE OF THE DATA

The dataset used in this project has been taken from the National Institute of Diabetes and Kidney Diseases.

DATA DESCRIPTION

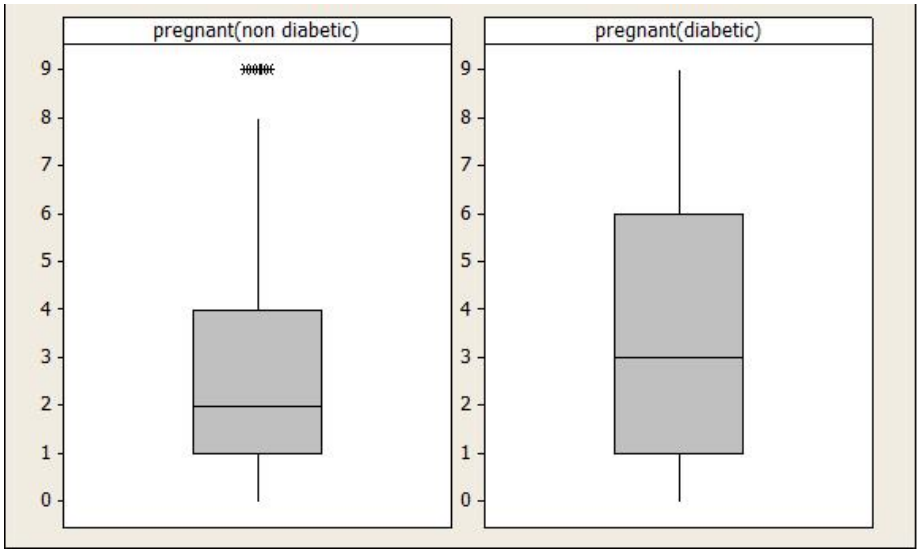
The dataset used in this project contains information of 376 women across 8 variables. The outcome tested was Diabetes, 115 of them tested positive and 261 of them tested negative. Therefore, there is one response (dependent) variable and 7 health parameters that could and should be related to the onset of diabetes and its future complications.

The binary response variable (denoted as Y) is defined as:

$$Y = \begin{cases} 1, \text{ the patient has diabetes} \\ 0, \text{ otherwise} \end{cases}$$

The 7 covariates are defined below:

1. Pregnancies: It displays the number of times the woman got pregnant. It is a discrete variable and is labelled as x_1 in our data. Gestational diabetes is a type of diabetes that can develop in a pregnant woman who don't already have diabetes. Every year, almost 10% of the pregnant women gets affected by the gestational diabetes. So, this covariate has been included in the dataset.

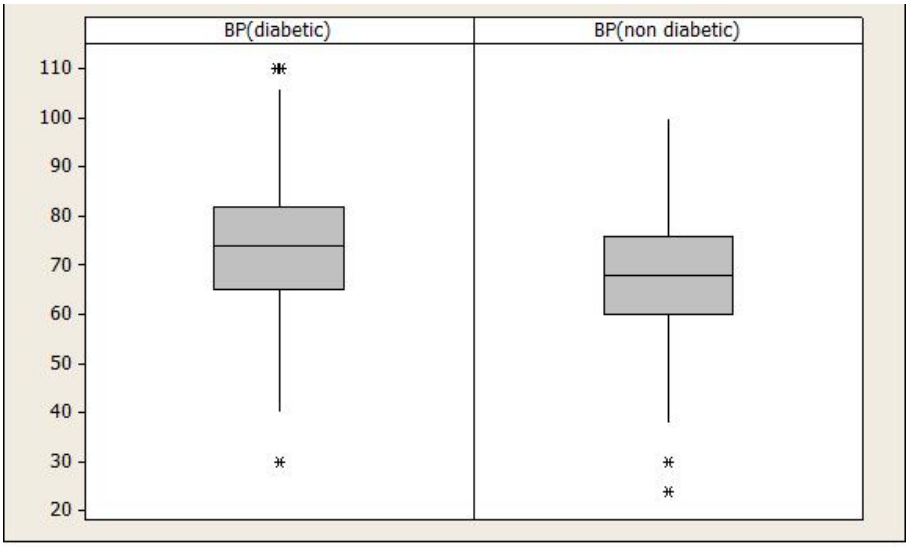


From the above graph, we can observe that, on an average the number of pregnancies in a diabetic female is slightly greater than that of a non-diabetic female. Thus, our data supports the fact that the chances of getting diabetes increase with the increase in the number of pregnancies for a female.

2. Glucose concentration: It displays the concentration of glucose in the blood (in mg/dL) and it is measured by 2 hours of oral glucose tolerance test (OGTT). It is a discrete variable. The normal fasting blood glucose level is lower than 110 mg/dL. A person having an OGTT greater than 145mg/dL is likely to have diabetes.

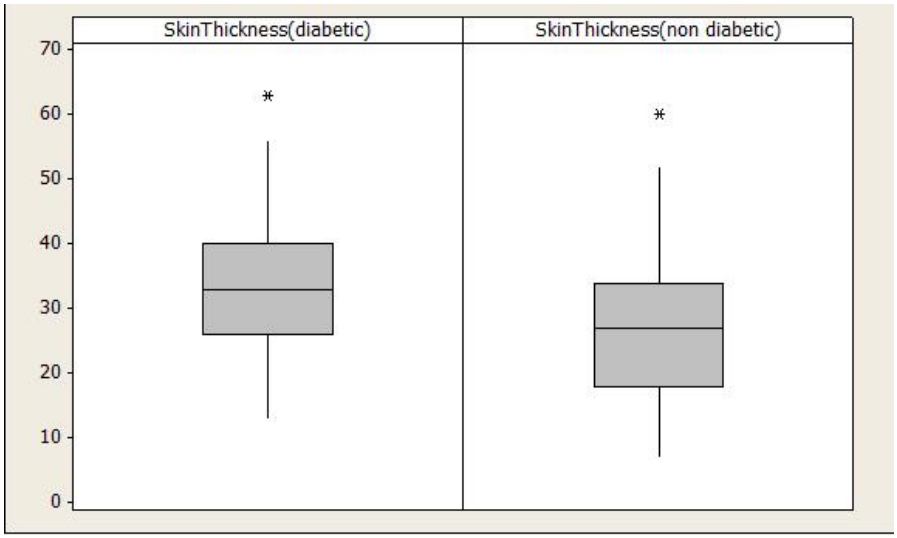
From the above graph, we can conclude that, on an average the diabetic person has a higher glucose concentration in their blood compared to that of non-diabetic person in our dataset.

3. Blood Pressure: It displays the diastolic blood pressure (in mm Hg) of the person. It is a discrete variable. The normal diastolic blood pressure is generally lesser than 80 mm Hg. Diabetes results in narrowing of the arteries and this condition is known as atherosclerosis. This causes High Blood Pressure, and it can often lead to heart attacks and renal failure. Thus, high blood pressure can be an indication of diabetes.



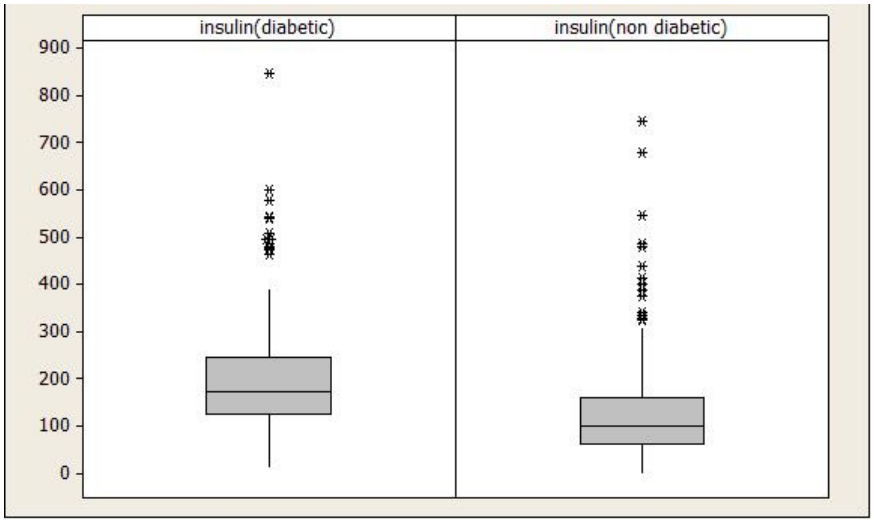
From the graph, we can say that the average Blood Pressure of diabetic females in the dataset is slightly more than that of non-diabetic females. Though there are out few outliers present in both the cases.

4. Skin thickness: It displays the triceps skin fold thickness (in mm). To be more precise, it is a measure of the obesity. It is a discrete variable. Almost 90% of the diabetic patients are obese. Overweight causes increased levels of fatty acids and inflammation which in turn leads to insulin resistance, thereby leading to diabetes mellitus.



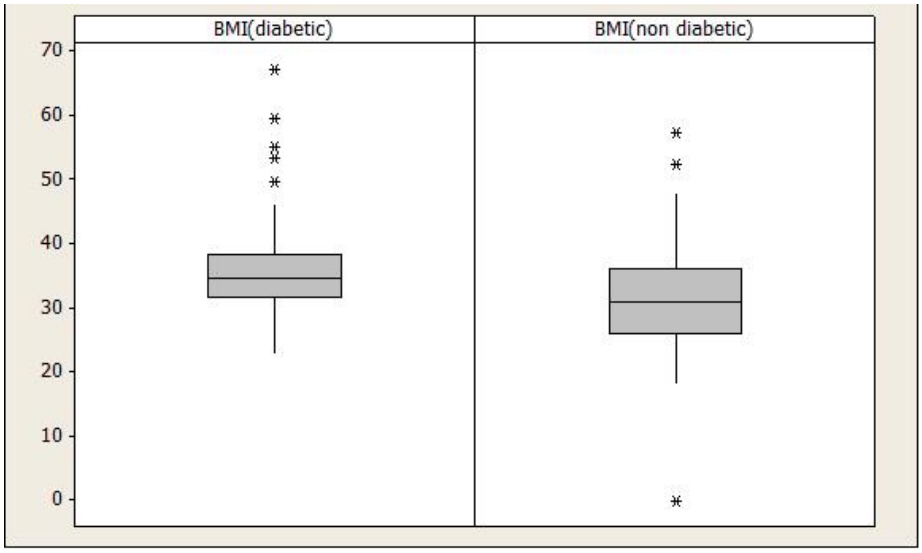
From the boxplot, we observe that the average triceps skinfold thickness is slightly more in diabetic females than the non-diabetic females in the dataset. Thus, our dataset supports the fact that obesity can be a reason for the occurrence of diabetes in females.

5. Insulin: It displays the amount of insulin concentration in the blood (in mu U/ml). The two-hour glucose tolerance test with Insulin levels is used to assess how an individual processes glucose and how the insulin in the body responds to those glucose levels.

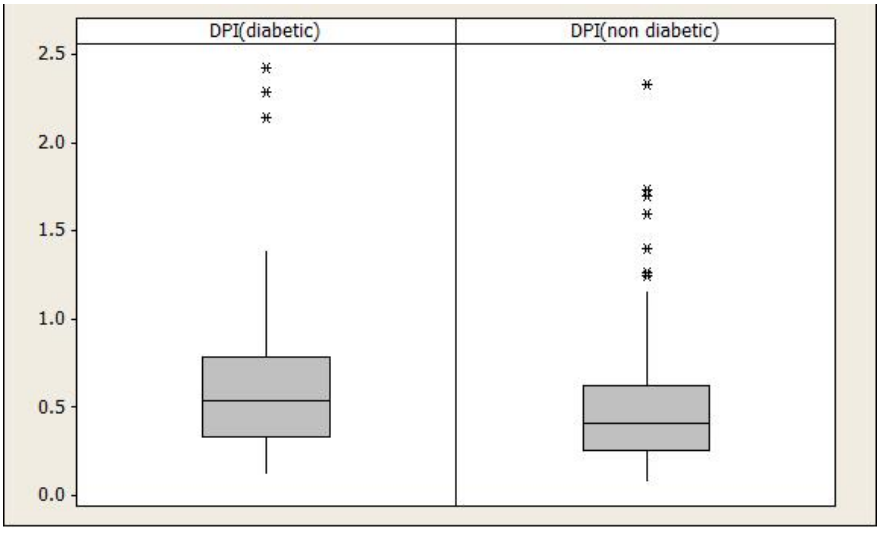


From the above boxplot, we can observe that the average insulin concentration in the blood for diabetic females is more than that of non-diabetic females in the dataset. Information obtained from the dataset is quite unexpected because diabetes mellitus is the condition in which the body fails to produce much insulin. As a result, a diabetic patient lacks insulin in their body. However, we can also see several outliers present the dataset.

6. Body Mass Index (BMI): It displays the body mass index of the person (in Kg/m^2). BMI is calculated as the weight of the person (in kg) divided by their height in (metre sq.). Therefore, a person who is overweight has a high BMI and a person who is underweight has low BMI.



7. Diabetes pedigree function: it is a function which scores the likelihood of diabetes based on the family history. It provided some data on diabetes mellitus history in relatives and the genetic relationship of those relatives to the person. This measure of genetic influence gave us an idea of the hereditary risk one might have with the onset of diabetes mellitus.



From the above plot, we can observe that Diabetes Pedigree Function (DPI) is slightly more for the diabetic females than that of non-diabetic females. The median DPI for the diabetic females is 0.553 whereas, the median DPI for the non-diabetic females is 0.405. Thus, our dataset supports the claim that a person with family history of diabetes is more likely to develop diabetes.

The first 12 rows of the dataset are given below:

	A	B	C	D	E	F	G	H
1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Outcome
2	1	89	66	23	94	28.1	0.167	0
3	0	137	40	35	168	43.1	2.288	1
4	3	78	50	32	88	31	0.248	1
5	2	197	70	45	543	30.5	0.158	1
6	1	189	60	23	846	30.1	0.398	1
7	5	166	72	19	175	25.8	0.587	1
8	0	118	84	47	230	45.8	0.551	1
9	1	103	30	38	83	43.3	0.183	0
10	1	115	70	30	96	34.6	0.529	1
11	3	126	88	41	235	39.3	0.704	0
12	1	97	66	15	140	23.2	0.487	0

OBJECTIVE

Diabetes is one of the most common diseases around the world and if detected earlier, it may prevent the progression of the disease and other complications. Thus, the main aim of this project is to help in the early detection of diabetes. Our primary objective is to identify which health parameters provided in the dataset plays the most significant role in the occurrence of diabetes in females. After identifying the significant covariates, now we can use suitable statistical tools to predict whether a woman has a potential chance of developing diabetes or not based on the information provided on the dataset.

MULTICOLLINEARITY

Before we fit a suitable regression model, we need to ensure whether we require all the covariates for our study or we can eliminate a few of them. That is, we need to find out if a subset of the given set of covariates is sufficient enough to predict whether a female has a high chance of getting diabetes or not. In this context, we will study the Multicollinearity within the covariates.

Multicollinearity is a phenomenon in which there are high correlations between two or more predictor variables, i.e., one covariate can be used to predict the other. As a result, the estimates of the coefficients of the predictor variables in the regression equation may get changed and it affects our interpretations regarding the individual effects of the predictors on the response

variable. If the correlation coefficient(r) between the pairs of predictor variables is exactly +1 or -1, then it is called perfect multicollinearity. In practice, we will rarely find any pair of predictor variables which are perfectly correlated. Thus, if 'r' is close to +1 or -1, one of the variables should be removed from the regression model if at all possible.

Detection of multicollinearity

One way of detecting multicollinearity is by calculating the Variance Inflation Factor (VIF) for each of the predictor variables. It is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically, the VIF for a regression model is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.

The Variance Inflation Factor is given by:

$$\text{VIF} = \frac{1}{1-R^2}$$

Where the value of R^2 is obtained by regressing one predictor variable on the remaining predictor variables.

The rule of thumb for interpreting the Variance Inflation Factor:

- 1 => not correlated
- Between 1 and 5 => moderately correlated
- Greater than 5 => highly correlated

PREDICTOR VARIABLE	VALUE OF VIF
PREGNENCIES	1.091834
GLUCOSE	1.424761
BMI	1.800348
INSULIN	1.417086
BLOOD PRESSURE	1.140995
SKIN THICKNESS	1.598570
DIABETES PEDIGREE FUNCTION	1.030273

From the above table, we can observe that the VIF lies between 1 and 1.6 for all the predictor variables. Thus, we can conclude that no multicollinearity is absent within the predictor variables. As a result, we need to work with all the predictor variables provided in the dataset and cannot drop any of them while fitting a suitable regression model.

ANALYZING THE SIGNIFICANCE OF THE PREDICTOR VARIABLES USING GLM

After we have checked for the Multicollinearity, we will now proceed to fit an appropriate regression model. Mathematically, a binary logistic model has a dependent variable with two possible outcomes such as present/absent which is represented by an indicator variable, where the two values are labelled as '1' and '0'. Since, for our data, the response variable is the presence or absence of diabetes denoted by '1' or '0', we proceed to fit a binary logistic regression to our data. The fitted model will give us an idea about which of the covariates have most significant impact on the occurrence of diabetes in females.

Let us define the random variables:

- Y: random variable denoting the diabetes status of a randomly selected female.
- X_1: random variable denoting the number of pregnancies of a randomly selected female.
- X_2: random variable denoting the glucose concentration in the blood of a randomly selected female
- X_3: random variable denoting the Body Mass Index of a randomly selected female.
- X_4: random variable denoting the insulin concentration in the blood of a randomly selected female.
- X_5: random variable denoting the diastolic blood pressure of a randomly selected female.
- X_6: random variable denoting the triceps skin fold thickness of a randomly selected female.
- X_7: random variable denoting the Diabetes Pedigree Function of a randomly selected female.

We need to fit a multiple logistic regression model. Hence, the model is given by -

$$P\,(Y = 1 \mid x_1, x_2, x_3, x_4, x_5, x_6, x_7) = \frac{e^{\eta}}{1+e^{\eta}} + \varepsilon \quad \ldots\ldots\ldots\ldots\ldots \quad (1)$$

where, $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$

Here, $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ are unknown parameters which are to be estimated.

OUTPUT TABLE

PARAMETER	ESTIMATED VALUE OF PARAMETER	STANDARD ERROR OF ESTIMATE (S.E)	Z VALUE
β_0	-9.9287	1.22657	-8.095
β_1	0.1167	0.05564	2.099
β_2	0.0409	0.00601	6.808
β_3	0.0596	0.02740	2.176
β_4	-0.0009	0.00131	-0.688
β_5	0.00545	0.01209	0.451
β_6	0.02149	0.01743	1.233
β_7	1.2332	0.43520	2.834

Testing of hypothesis

To Test:

H_{0xi}: The covariate x_i is not significant i.e., $\beta_i = 0$, $i = 1(1)7$

ag

H_{0xi}: The covariate x_i is significant i.e., $\beta_i \neq 0$, $i = 1(1)7$

Test Statistic:

Here, $Z = \frac{Estimate}{Standard\ Error}$ denotes the test statistic for the above testing problem

where $Z \sim N(0,1)$

Critical Region:

The critical point is $\tau_{0.025} = 1.96$ i.e., we reject H_0 if $Z_{obs} > 1.96$ at 5% level of significance.

Interpretation

1. Here, $|Z_{\text{pregnancies}}| = 2.099 > 1.96$ and hence we reject H_{0x_1}, at 5% level. So, in the light of the given data, we can say that the number of pregnancies plays a significant rule in the occurrence of diabetes in females. Further, we can say that the chances of diabetes increases with the increase in the number of pregnancies.

2. Here, $|Z_{\text{glucose}}| = 6.808 > 1.96$ and hence we reject H_{0x_2}, at 5% level. So, in the light of the given data, we can say that the glucose concentration in the blood plays a significant rule in the occurrence of diabetes in females. Further, we can say that the chances of diabetes increases with the increase in the amount of blood glucose concentration.

3. Here, $|Z_{\text{BMI}}| = 2.176 > 1.96$ and hence we reject H_{0x_3}, at 5% level. So, in the light of the given data, we can say that the body mass index(BMI) plays a significant rule in the occurrence of diabetes in females. Further, we can say that the chances of diabetes increases with the increase in the body mass index.

4. Here, $|Z_{\text{insulin}}| = 0.688 < 1.96$ and hence we accept H_{0x_4}, at 5% level. So, in the light of the given data, we can say that the insulin concentration in the blood does not play a significant rule in the occurrence of diabetes in females. Further, we can say that the chances of diabetes increases with the decrease in the amount of blood insulin concentration.

5. Here, $|Z_{\text{pressure}}| = 0.451 < 1.96$ and hence we accept H_{0x_5}, at 5% level. So, in the light of the given data, we can say that the blood pressure does not play a significant rule in the occurrence of diabetes in females.

6. Here, $|Z_{\text{skin\_thickness}}| = 1.233 < 1.96$ and hence we accept H_{0x_6}, at 5% level. So, in the light of the given data, we can say that the skin thickness does not play a significant rule in the occurrence of diabetes in females.

7. Here, $|Z_{DPI}|$ = 2.834 > 1.96 and hence we reject H_{0x_7}, at 5% level. So, in the light of the given data, we can say that the family history of diabetes plays a significant rule in the occurrence of diabetes in females. Further, we can say that the chances of diabetes increases if many of the family members or closer relatives have been diagnosed with diabetes.

MODEL FITTING USING SIGNIFICANT COVARIATES

Thus, from the above model, we conclude that the most significant covariates are:

- Pregnancies
- Glucose
- BMI
- Diabetes Pedigree Function

Since, we have obtained the significant covariates, our next task is to fit a fresh binary logistic regression model using only those covariates.

Let us define the random variables:

- Y: random variable denoting the diabetes status of a randomly selected female.
- X_1: random variable denoting the number of pregnancies of a randomly selected female.
- X_2: random variable denoting the glucose concentration in the blood of a randomly selected female.
- X_3: random variable denoting the Body Mass Index of a randomly selected female.
- X_4: random variable denoting the Diabetes Pedigree Function of a randomly selected female.

We need to fit a multiple logistic regression model. Hence, the model is given by -

$$P(Y = 1 \mid x_1, x_2, x_3, x_4) = \frac{e^{\eta}}{1+e^{\eta}} + \varepsilon \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

where, $\eta = \beta^*_0 + \beta^*_1 x_1 + \beta^*_2 x_2 + \beta^*_3 x_3 + \beta^*_4 x_4$

Here, β^*_0, β^*_1, β^*_2, β^*_3 and β^*_4 are the unknown parameters which are to be estimated.

The Output Table is given by:

PARAMETER	ESTIMATED VALUE OF PARAMETER	STANDARD ERROR OF ESTIMATE (S.E)	Z VALUE
β_0	-9.5502	1.0609	-9.001
β_1	0.1291	0.05455	2.367
β_2	0.03901	0.005075	7.687
β_3	0.08027	0.02098	3.826
β_4	1.26705	0.43457	2.916

Thus, from the above output table, we observe that all the covariates, namely, Pregnancies, Glucose Conc., BMI and DPI have significant impact on the response variable 'y'.

Therefore, we would be using the logistic model (2) for further analysis.

BINARY CLASSIFICATION

We have fitted a binary logistic regression model using the significant covariates given in the dataset and have estimated the coefficients. Now, our next and the most crucial task is to predict whether a female has diabetes or not using our regression model.

Binary logistic regression does not return directly, the class of observations. But it allows us to estimate the probabilities of class membership, i.e., it will give us an estimate of the probability of occurrence diabetes in a particular female based on the independent predictor variables. Here, we need to follow a specific algorithm to decide whether the woman is diabetic or not.

Binary Classification is the method of classifying the elements of the given set into two groups on the basis of a classification rule. It is a dichotomization applied to a practical situation. Typical Binary Classification problems include:

- Medical Testing to determine if a patient has a certain disease or not
- Quality control in industry, deciding whether the specification has been met or not.

Classification Threshold

For binary classification, we need to define a classification threshold (also known as the decision threshold) denoted by "Th". A value of the fitted probability which is less than the threshold indicates the absence of diabetes and is labelled as '0' while, a value of the fitted probability which is greater than the threshold indicates the presence of diabetes and is labelled as '1', i.e.,

$$Y_{predict} = \begin{cases} 1 & \text{, fitted probability} > \text{threshold} \\ 0 & \text{, otherwise} \end{cases}$$

Confusion Matrix

Based on this threshold, we now construct a classification matrix (also known as the confusion matrix). In the field of machine learning and especially for the problem of statistical classification, a confusion matrix is a table layout that allows visualization of the performance of the fitted model. It consists of two rows and two columns where each row of the matrix represents the instances in an actual class while each column represents the instances of the predicted class.

Some terminologies related to confusion matrix:

- True Positive (TP): A test result that correctly indicates the presence of a condition or characteristic.
- False Positive (FP): A test result that wrongly indicates the presence of a condition or characteristic.
- True Negative (TN): A test result that correctly indicates the absence of a condition or characteristic.
- False Negative (FN): A test result that wrongly indicates the absence of a condition or characteristic.

- True Positive Rate (TPR) or Sensitivity: It measures how appropriate the model is in detecting events in the positive class. Thus, it quantifies how many diabetic females are correctly predicted as diabetic.

$$\text{Sensitivity/TPR} = \text{Prob}(Y_{predict}=1 | Y_{actual}=1) = \frac{\text{Prob}(Ypredict=1, Yactual=1)}{\text{Prob}(Yactual=1)} = \frac{TP}{TP+FN}$$

- Specificity: It measures how appropriate the model is in detecting events in the positive class, i.e., it shows the percentage chance that the test will correctly identify a person who is disease-free

$$\text{Specificity} = \text{Prob}(Y_{predict}=0 \mid Y_{actual}=0) = \frac{\text{Prob}(Ypredict=0, Yactual=0)}{\text{Prob}(Yactual=0)} = \frac{TN}{TN+FP}$$

- False Positive Rate (FPR): False Positive rate (also known as the false alarm ratio) is the probability of falsely predicting the presence of a condition or characteristic.

$$FPR = \text{Prob}(Y_{predict}=1 | Y_{actual}=0) = \frac{\text{Prob}(Ypredict=1, Yactual=0)}{\text{Prob}(Yactual=0)} = \frac{FP}{FP+TN}$$

- Accuracy: it denotes the accuracy of the model, i.e., how often the classifier makes correct prediction.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Therefore, 2X2 confusion matrix is given as,

PREDICTED / ACTUAL	$Y_{predict}=1$	$Y_{predict}=0$
$Y_{actual}=1$	TP	FN
$Y_{actual}=0$	FP	TN

Now, as we have defined classification threshold and confusion matrix, we will perform the Binary Classification by selecting a specific threshold, i.e., we will predict the presence or absence of Diabetes on the basis of the estimated probabilities we have obtained by fitting a logistic regression model.

For every classification problem, the default choice of the Threshold is 0.5.

Thus, fixing the threshold at 0.5, let us draw the confusion matrix.

CONFUSION MATRIX for THRESHOLD=0.5

PREDICTED / ACTUAL	$Y_{predict}=1$	$Y_{predict}=0$	TOTAL
$Y_{actual}=1$	63 (TP)	52 (FN)	115
$Y_{actual}=0$	23 (FP)	238 (TN)	261
TOTAL	86	290	376

(All computations are done using the R Studio)

- Sensitivity / TPR $= \frac{\mathbf{TP}}{\mathbf{TP+FN}} = \frac{63}{63+52} = 0.5478$
- FPR $= \frac{\mathbf{FP}}{\mathbf{FP+TN}} = \frac{23}{23+238} = 0.088$
- Specificity $= 1 - FPR = 0.912$
- Accuracy $= \frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FP+FN}} = \frac{63+238}{23+52+63+238} = 0.8$

As we observe that the accuracy of the model is 0.8 which is pretty good. It implies that the classification is giving accurate results 80% of the times. Moreover, the Specificity of the model is 0.912 which is also quite high.

Though we have got desirable values of accuracy and specificity, we would still reject this threshold. This is because, our main aim is to predict the presence of diabetes in any female with more precision. The value of sensitivity/TPR is 0.5478. It implies that our model can correctly predict the presence of diabetes in any female 54.78% of the times which is quite low. In order to increase the sensitivity of our model, we would proceed to our next section of Finding an Optimum Threshold.

Finding An Optimum Threshold

The main objective of our project is to correctly detect the presence of diabetes. Previously, we saw that, for the threshold 0.5, the accuracy of the model was high, but the sensitivity came out to be very low. As we go on increasing the threshold from 0 to 1, the sensitivity of the model goes on decreasing while the specificity increases. i.e., sensitivity and specificity are inversely related to each other.

Let us consider a varying range of threshold and for each value of threshold, we would calculate the sensitivity and the specificity. Now, we would choose that value of threshold for which

T = (sensitivity*specificity) is maximum.

The threshold table is given below:

THRESHOLD	SENSITIVITY	SPECIFICITY	(SENSITIVITY*SPECIFICITY)	ACCURACY
0.01	1	0.0268	0.0268	0.3244
0.11	0.9565	0.4329	04141	0.5930
021	0.8521	0.6704	0.5713	0.7260
0.31	0.7913	0.7969	0.6306	0.7951
0.41	0.6521	0.8620	0.5622	0.7978
0.51	0.5478	0.9157	0.5016	0.8031
0.61	0.4521	0.9386	0.4244	0.7898
0.71	0.3478	0.9578	0.3331	07712
0.81	0.2347	0.9846	0.2311	0.7553
0.91	0.0521	0.9923	0.0517	07047

(All computations are done in R Studio)

Thus, from the threshold table, we can observe that the product of sensitivity and specificity is maximum for Threshold = 0.31.

Therefore, Threshold=0.31 is chosen as an optimum threshold for binary classification.

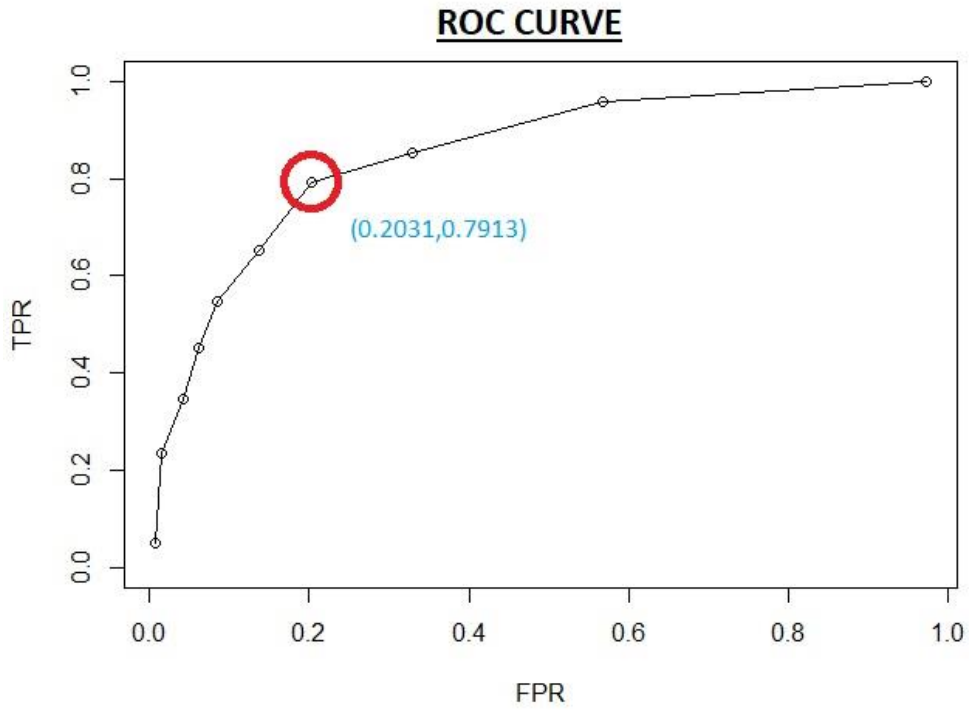
Receiver Operating Characteristic Curve (ROC Curve)

ROC curve in logistic regression is used to determine the optimum threshold for binary classification. It is a plot of the False Positive Rate on the X-axis versus the True Positive Rate on the Y-axis for a number of different classification threshold values ranging from 0 to 1. A model with perfect skill is represented at a point (0,1), i.e., the point at the top-left corner of

the ROC curve while a model with the worst skill is represented at a point (1,0), i.e., the point at the bottom-right corner of the ROC curve.

In reality, we cannot construct a model which is perfectly skilled, i.e., it is not possible to reach the top-left corner of the ROC curve. In order to get an optimum threshold for classification, we need to select that point on the ROC curve which is closest to the point (0,1), i.e., top-left corner of the ROC curve.

Now, let us plot the ROC curve based on the threshold table given above.



From the above ROC curve, we observe that the point circled in red, i.e., FPR=0.2031 and TPR=0.7913 is closest to the top-left corner of the graph. This point corresponds to the threshold 0.31

Thus, from the ROC curve, we verify that 0.31 is the value of the optimum threshold for classification

Confusion Matrix for The Optimum Threshold.

Let us now construct a confusion matrix based on the optimum threshold to visualize the improved performance of the fitted model.

CONFUSION MATRIX for OPTIMUM THRESHOLD

PREDICTED / ACTUAL	$Y_{predict}=1$	$Y_{predict}=0$	TOTAL
$Y_{actual}=1$	91 (TP)	24 (FN)	115
$Y_{actual}=0$	53 (FP)	208 (TN)	261
TOTAL	86	290	376

(All computations are done using the R Studio)

- Sensitivity / TPR $= \frac{\mathbf{TP}}{\mathbf{TP+FN}} = \frac{91}{91+24} = 0.7913$
- FPR $= \frac{\mathbf{FP}}{\mathbf{FP+TN}} = \frac{53}{53+208} = 0.203$
- Specificity $= 1 - FPR = 0.796$
- Accuracy $= \frac{\mathbf{TP+TN}}{\mathbf{TP+TN+FP+FN}} = \frac{91+208}{91+24+53+208} = 0.7952$

Thus, we observe that the sensitivity (or true positive rate) of the model for the optimum threshold has improved. Our model can correctly predict the presence of diabetes in any female almost 80% of the times which is quite good. Moreover, the accuracy of the model is almost 0.8. Therefore, our model will give correct results 80% of times.

CONCLUSION

From the above study, we can conclude that there is no Multicollinearity present with predictor variables. Moreover, by fitting a logistic regression model, we noted down that 'the no. of pregnancies' , 'Glucose level in blood' , 'Body Mass Index' and the 'Diabetes Pedigree Function' are the most significant predictors in predicting the presence of diabetes in females. By observing the coefficients of the predictor variables, we conclude that increase in the glucose level and BMI; greater number of Pregnancies; low levels of insulin and the family history of diabetes can increase the risk of diabetes in a female.

ACKNOWLEDGEMENT

I would like to show my gratitude to my project guide Dr. Surupa Chakraborty for her guidance throughout the duration of completion of my project. I would also like to thank my parents and my friends who have helped me in various ways during the preparation of my project. Lastly, I would like the thank St. Xavier's College, Kolkata for giving me this opportunity to prepare a dissertation project on this topic.

REFERENCE

- Fundamentals of Statistics, Vol I by Gun, Gupta and Dasgupta
- www.github.com
- https://en.wikipedia.org (for various definitions and facts)
- McCullagh, P & Nelder, J.A.(1995), Generalized Linear Models. Chapman and Hall.
- The Elements of Statistical Learning: Data Mining, Inference, and Prediction by Trevor Hastie, Robert Tibshirani, and Jerome Friedman

R CODE

```
rm(list=ls())
#attaching the data
attach(diabetes_project_)

#displyaing the first 5 rows of the dataset
head(diabetes_project_)

#storing the value of the binary response variable at y
y=Outcome

#fitting of Logistic regression model
m=(glm(y~Pregnancies+Glucose+BMI+BloodPressure+DiabetesPedigreeFunction+SkinThic
kness+Insulin,family=binomial(link = "logit")))
summary(m)

#checking for Multicollinearity
install.packages('car')
library(car)
vif(m)

#fitting the Logistic Regression with only the significant covariates
m1=(glm(y~Pregnancies+Glucose+BMI+DiabetesPedigreeFunction,family=binomial));m
summary(m1)

#storing the fitted probabilities
p_fitted=fitted.values(m1)
p_fitted
```

```
#taking threshold to be 0.5
y_hat=array(0)
threshold=0.5
y_hat=ifelse(p_fitted>threshold,1,0)

#confusion matrix for  threshold=0.5
table(y,y_hat)

#calculating the sensitivity
sensitivity=(table(y,y_hat)[[4]])/((table(y,y_hat)[[4]])+(table(y,y_hat)[[2]]))

#calculating the False Positive Rate
FPR=(table(y,y_hat)[[3]])/((table(y,y_hat)[[3]])+(table(y,y_hat)[[1]]))

#calculating the specificity
specificity= 1- FPR;specificity

#calculating the accuracy of classification
accuracy=(table(y,y_hat)[[1]]+table(y,y_hat)[[4]])/376;accuracy

#calculating TPR and FPR for varying threshold
threshold=seq(0.01,0.99,0.1)
TPR=array(0)
FPR=array(0)
acc=array(0)
F1_SCORE=array(0)
for(i in 1:length(threshold))
  {
   y_hat=array(0)
   for(j in 1:length(p_fitted))
```

```
  {
    if(p_fitted[j]>=threshold[i])
      y_hat[j]=1
    else
      y_hat[j]=0
  }
table(y,y_hat)
TPR[i]=(table(y,y_hat)[[4]])/((table(y,y_hat)[[4]])+(table(y,y_hat)[[2]]))
FPR[i]=(table(y,y_hat)[[3]])/((table(y,y_hat)[[3]])+(table(y,y_hat)[[1]]))
acc[i]=(table(y,y_hat)[[1]]+table(y,y_hat)[[4]])/376
}

specificity=1-FPR
opt=TPR*specificity

#storing the TPR and FPR values for varying thresholds in a data frame
data=data.frame(threshold,TPR,FPR,"ACCURACY"=acc,specificity,opt,j,F1_SCORE)

#Plotting ROC curve for various threshold
plot(FPR,TPR,ylim=c(0,1),type="o")
abline(h=0.7969,v=0.7913)

#misclassification matrix for threshold=0.3
threshold=0.31
y_hat=ifelse(p_fitted>threshold,1,0)
t=table(y,y_hat)

#calculating sensitivity/TPR for optimum threshold
TPR=(table(y,y_hat)[[4]])/((table(y,y_hat)[[4]])+(table(y,y_hat)[[2]]))
```

```
#calculating FPR or (1-specificity) for optimum threshold
FPR=(table(y,y_hat)[[3]])/((table(y,y_hat)[[3]])+(table(y,y_hat)[[1]]))

#calculating accuracy of the classification for optimum threshold
acc=(table(y,y_hat)[[1]]+table(y,y_hat)[[4]])/376
```