# Contents

Mohan Arora

# Project - Early Detection of Diabetes

## 1 Introduction

High blood sugar levels are a hallmark of diabetes, a long-term condition caused by the body's inability to produce or efficiently use insulin. Diabetes has been steadily increasing worldwide, with an estimated 422 million adults in 2014 suffering from the disease, according to the World Health Organization (WHO). By 2040, this number is expected to reach 642 million, making diabetes a major issue for public health.

Diabetes has increased as a result of a number of factors, including population growth, aging, urbanization, unhealthy diets, and inactivity. Obesity and metabolic disorders, which are major risk factors for type 2 diabetes, have increased as a result of these factors.

With more than 1.3 billion people, India has the second highest total population in the world. In 2017, it was estimated by the International Diabetes Federation that 72.9 million adults in India had diabetes.

Age-standardized incidence and mortality rates of diabetes in India increased during the NCBI study period. The incidence rate increased from 199.14 to 317.02 per 100,000 people, and mortality increased from 22.30 to 27.35 per 100,000 people. Therefore, diabetes is unquestionably a worrying condition in today's world. and it is also regarded as the **"Greatest epidemic of Human Era"**.

The results of diabetes can be serious and perilous, including cardiovascular sickness, kidney disappointment, visual impairment, and lower appendage removals. Additionally, diabetes has a significant financial impact on both individuals and healthcare systems worldwide. In order to improve global health outcomes, this dissertation examines the factors that contribute to the rise in diabetes prevalence worldwide and potential diabetes prevention and management strategies.

Data Description

## 1.1 Context

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes for women.

## 1.2 Content

- Pregnancies: Number of times pregnant
- Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
- BloodPressure : Diastolic blood pressure (mm Hg)
- SkinThickness: Triceps skin fold thickness (mm)
- Insulin: 2-Hour serum insulin (mu U/ml)
- BMI: Body mass index (weight in $kg$/(height in m)$^2$)
- DiabetesPedigreeFunction: Diabetes pedigree function
- Age: Age(years)
- Outcome : 0 or 1

## 1.3 Nature of Covariates:-

Here,we have 8 covariates namely Pregnancies,Age,BMI,Glucose, SkinThickness,Insulin,Blood Pressure,Diabetes Pedigree Function. The following table shows the actual name ,the symbolic name and the nature of the covariates.

Table 1:

| Name of The covariate | Symbolic name of the covariate | Nature of the Covariate |
|---|---|---|
| Pregnancies | X1 | Discrete |
| Age | X2 | Continuous |
| BMI | X3 | Continuous |
| Glucose | X4 | Continuous |
| SkinThickness | X5 | Continuous |
| Insulin | X6 | Continuous |
| Blood Pressure | X7 | Continuous |
| Diabetes Pedigree Function | X8 | Continuous |

The detailed descrption of the covariates is given below:-

- The discrete variable Pregnancies(X1) denotes the no. of time a women has been pregnant
- The continous variable Age(X2) denotes the age of the women
- The continous variable BMI(X3) is Body Mass Index which is person's weight in kilograms (or pounds) divided by the square of height in meters
- The continuous variable Glucose (X4) is the main type of sugar in the blood measured in mg(milligram)
- Thickness of the skin is SkinThickness(X5) which is a continuous variable
- The continous variable Insulin (X6) ,it is a hormone which controls the blood sugar level
- The continous variable Blood Pressure (X7),is Diastolic blood pressure measured in mm Hg.

- The continuous variable Diabetes Pedigree Function(X8) indicates the function which scores likelihood of diabetes based on family history.

Nature of the Response Variable

5mm In our dataset presence of diabetes is the response variable .We denote the response variable as 'Y'.Y is defined as follows:-

$$Y = \begin{cases} 0, & \text{if individual doesnt suffer from diabetes} \\ 1, & \text{if individual suffers from diabetes} \end{cases} \tag{1}$$

Objectives:-

The objectives of this project is as follows:-

- To study the role of several biological factors (i.e the covariates) in causing the presence or absence of diabetes
- To find which covariates play significant role in diabetes
- To analyse the Goodness of Fit

Methodology:-

- First we will fit a binary logistic regression to our data
- In the next step we will find out if the effects of different covariates on diabetes are significant and do further analysis of the significant covariates
- In the next step we will see how well the model is fitted to the data and compare with other models.

# 2 Fitting binary regression model to the data

In regression theory,our main problem is to predict the value of one variable known as the dependent variable on the basis of given set of values of one or more variables generally known as the predictor .

In binary regression model, the response variable takes two values or it has two outcomes which is represented numerically as 0 or 1 .

We are representing our response variable as Y.

$$Y = \begin{cases} 0, & \text{if individual doesnt suffer from diabetes} \\ 1, & \text{if individual suffers from diabetes} \end{cases} \tag{2}$$

Here Y is binary in nature .So we fit a binary regression model to our data.

$\pi = P(Y = 1 \mid X1, X2, ..., Xp) = \frac{e^{\eta}}{1+e^{\eta}} - - - (1)$

here $\eta$ ,is a function of all the predictor variables.

Where $\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$

$\beta_i$ is the regression coefficient corresponding to Xi ,i= 1(1)p and $\beta_0$ is a constant.

p=8 which is the no. of independent variables under consideration

We define $\beta_i$ as the change in $\eta$ due to one unit change in Xi,i=1(1)p keeping other covariates fixed.

## 2.1 The fitted regression model using logistic regression

The regression model is given by:- P(Y=1$\mid X1, X2, ..., Xp) = \frac{e^{\eta}}{1+e^{\eta}}$

$\eta = -9.0984 + 0.1249X_1 + 0.0130X_2 + 0.1003X_3 + 0.0385X_4 - 0.0029X_5 - 0.0015X_6 - 0.0132X_7 + 0.9451X_8$

The regression table is given by:-

Table 2:

| Covariates | Estimates | Standard Error | Z values | P values |
|---|---|---|---|---|
| Intercept | -9.0984 | 0.812 | -11.201 | 0.000 |
| X1 | 0.1249 | 0.0320776 | 3.840 | 0.000123 |
| X2 | 0.0130 | 0.0093348 | 2.370 | 0.0171192 |
| X3 | 0.1003 | 0.0180876 | 5.686 | 0.000 |
| X4 | 0.0385 | 0.0037087 | 9.841 | 0.000 |
| X5 | -0.0029 | 0.011 | -2.252 | 0.00801 |
| X6 | -0.0015 | 0.001 | -1.392 | 0.164 |
| X7 | -0.0132 | 0.0052336 | -2.540 | 0.011072 |
| X8 | 0.9451 | 0.2991475 | 3.160 | 0.001580 |

- Interpretation of regression coefficients

We know that $\beta_i$ is the change in $\eta$ due to one unit change in Xi ,i=1(1)p,keeping the other covariates fixed.We can interpret the beta coefficients with respect to log odds and odds.

## 2.2 Interpretation using log odds

From (1) we observe that, $\frac{\pi}{1-\pi}$= e$^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}$

$\frac{\pi}{1-\pi}$ is called the odds of positive response log-odds=ln $\left(\frac{\pi}{1-\pi}\right)$=$\beta_0+\beta_1 X_1+\cdots+\beta_p X_p$

Thus $\beta_i$ is the amount by which ln $\left(\frac{\pi}{1-\pi}\right)$ changes due to unit change in $X_i$keeping other covariates fixed,i=1(1)p

## 2.3 Interpretation using odds

From (1) we observe that, $\frac{\pi}{1-\pi}$= e$^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}$

Due to one unit increment in $X_i$ keeping the other covariates fixed , the odds of positive response increases by a factor $e_i^{\beta}$

# 3 Testing of hypothesis to find the significant predictors

Here our main focus is to find out which of the covariates have a significant effect on the presence of the response variable . So we will apply some methodology regarding this :-

- From the binary regression fit we will obtain the p values corresponding to regression coefficients.We test the hypothesis on the basis of those p values and then we find which covariates have significant effect on the response variable
- We can draw different scattered bar plot to observe which covariates have significant effect on the presence of diabetes in a women along with comparison for non diabetic patient
- We can also check the significance of covariates by calculating the odds ratio as well as we can do Pearson Chi Sqauare's Test to find out the dependency of the response variable on the covariates

## 3.1 Procedure:-

Here we proceed to test whether a particular covariate has a significant effect on the response variable or not.
Our test procedure follows:

$$H_{oi}{:}\beta_i = 0 \text{ vs } H_{1i}{:}\beta_i \neq 0, i = 1(1)8$$

Here $\beta_i$ is the regression coefficient corresponding to ith explanatory variable .

If $H_{oi}$ is rejected then we can claim that the regression coefficient $\beta_i$ is not 0 and the corresponding covariate has a significant role in the presence of diabetes in women.

## 3.2 P values testing Method

The probability of obtaining results from a statistical hypothesis test that are at least as extreme as those that were observed, assuming the null hypothesis is true, is represented by the p values.
The smallest level of significance at which the null hypothesis would be rejected is given by the p value as an alternative to rejection points.

- Testing Rule:

We reject $H_{oi}$ in the favour of $H_{1i}$ iff the P value corresponding to i th predicting variable is less than $\alpha$ where $\alpha$ being the desired level of significance . We take $\alpha$=0.05

The following table shows the predictors, Their corresponding P values and also,the decision regarding predictors:-

| Predictors | P values | Decision of Hypothesis | Significant or Non significant |
|---|---|---|---|
| Glucose | 0.000 | Reject | Significant |
| Blood Pressure | 0.011072 | Reject | Significant |
| Insulin | 0.164 | Accept | Not Significant |
| BMI | 0.000 | Reject | Significant |
| Age | 0.0171 | Reject | Significant |
| Skin Thickness | 0.0801 | Reject | significant |
| Diabetes Pedigree Function | 0.0015 | Reject | Significant |
| Pregnancies | 0.000123 | Reject | Significant |

Table 3: An 8x4 table

Therefore in the light of given data we can observe that Glucose,Blood Pressure,BMI,Diabetes Pedigree Function,Pregnancies,Age increases the chances of diabetes in a woman and has significant effect.Later we will again fit a binary logistic regression check the accuracy of the model here both insulin and skin thickness are inidcator for diabetes in real life so we will not exclude them from the model

## 3.3 Test Using Pearsonian Chi Square

Classifying the effects of covariates is an additional method for analyzing them. Using the Pearson's chi square test, we categorize the covariates at various levels to determine whether they are dependent or independent on the significant covariates.

Formula for chi square is denoted by:-

$\chi^2$=$\sum_i (O_i - E_i)^2/E_i$

where $O_i$ is the observed value and $E_i$ is the expected value.

- Methodology:-
- First we will observe the odds that the covariate has an effect on the response variable
- After calculating the odds we will confirm it with a more conclusive evidence ,Pearson's Chi square Test
- Effect of no. of pregnancies(X1) in women on the presence of diabetes or not

From this data we observe that no. of pregnancies in women varies from 0 to 17 , so for observing the effect we classify the data into 2 groups :-

- Grp1:-Women with no. of pregnancies$\geq$ 7
- Grp2:-Women with no. of pregnancies$<$ 7 and form the following contingency table and check whether they have diabetes or not.

Table 4:

| DiabetesOutcome (Y)\|Pregnancy (X1) | <7 times (0) | $\geq$ 7 times (1) | Total |
|---|---|---|---|
| Absent(0) | 426 | 74 | 500 |
| Present(1) | 173 | 95 | 268 |
| Total | 599 | 169 | 768 |

So from the given table we will first compute the odds of diabetes in women with multiple pregnancies in consideration

- The odds of presence of diabetes when the no. of pregnancy in women is less than 7 :-

$\frac{P(Y=1|X1=0)}{P(Y=0|X1=0)}$ = $\frac{P(Y=1,X1=0)}{P(Y=0,X1=0)}$=$\frac{173}{426}$= 0.406

- The odds of presence of diabetes when the no. of pregnancy in women is greater than equals 7 :-

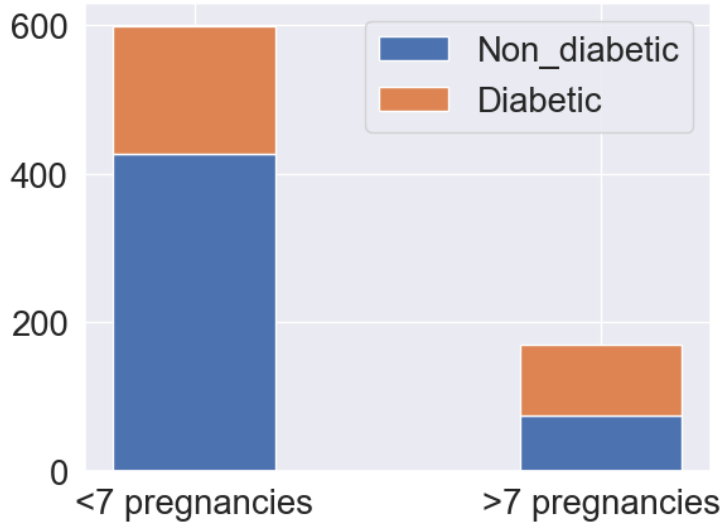


Figure 1: Stacked bar diagram of response variable against Pregnancies

$\frac{P(Y=1|X1=1}{P(Y=0|X1=1)}$ = $\frac{P(Y=1,X1=1)}{P(Y=0,X1=1)}$=$\frac{95}{74}$= 1.283

- To test: $H_0$ : The independence of the covariate and the response variable,presence of Diabetes against $H_1$ : not $H_0$

  Here the test statistic follows chi square distribution with degrees of freedom = 1

  The observed value of the pearsonian chi square is given by :——

  $\chi^2_{observed}$ = 42.146 , $\chi^2_{0.05,1}$=3.841 and Pvalue=0.00

- Descision: $\chi^2_{observed}$> $\chi^2_{0.05,1}$,we reject $H_0$

- Conclusion: In the light of the given data Diabetes status depends on the no. of times a woman has been pregnant.

  The odds incase of diabetes is 3 times more if the no. of pregnancy in a women is greater than equal to 7 than that of less than 7.

- Effect of no. of Age(X2) in women on the presence of diabetes or not

From this data we observe that age of women varies from 21 to 81 , so for observing the effect we classify the data into 2 groups :-

- Grp1:-Women with age $\geq 35$
- Grp2:-Women with age $< 35$ and form the following contingency table and check whether they have diabetes or not.

Table 5:

| DiabetesOutcome (Y)\|Age (X1) | <35 years (0) | $\geq 35$ years (1) | Total |
|---|---|---|---|
| Absent(0) | 362 | 138 | 500 |
| Present(1) | 126 | 142 | 268 |
| Total | 488 | 280 | 768 |

So from the given table we will first compute the odds of diabetes in women with age $\geq 35$

- The odds of presence of diabetes when the age of women is less than 35 :-

$\frac{P(Y=1|X2=0)}{P(Y=0|X2=0)} = \frac{P(Y=1,X2=0)}{P(Y=0,X2=0)} = \frac{126}{362} = 0.348$

- The odds of presence of diabetes when the age of women is greater than equals 35 :-

$\frac{P(Y=1|X1=1}{P(Y=0|X1=1)} = \frac{P(Y=1,X1=1)}{P(Y=0,X1=1)} = \frac{142}{138} = 1.02$

- To test: $H_0$ : The independence of the covariate and the response variable,presence of Diabetes against $H_1$ : not $H_0$

  Here the test statistic follows chi square distribution with degrees of freedom = 1

  The observed value of the pearsonian chi square is given by :——

  $\chi^2_{observed}$ = 47.44 , $\chi^2_{0.05,1}$=3.841 and Pvalue=0.00

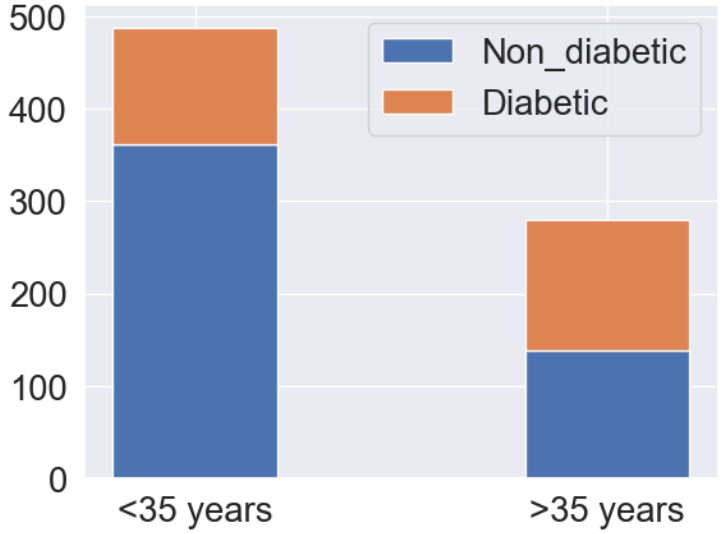


Figure 2: Stacked Bar plot of response variable against age

- Descision: $\chi^2_{observed} > \chi^2_{0.05,1}$,we reject $H_0$
- Conclusion: In the light of the given data Diabetes status depends on the age of the woman.

  The odds incase of diabetes is very less if the age of woman is less than 35 than that of less than that of greater than 35 ,though from the result we cant be sure that with age diabetic rate increase but we can surely tell that diabetic rate is lower incase of younger women.

- Effect of no. of BMI(X3) in women on the presence of diabetes or not

From this data we observe that BMI of women varies from 18 $kg/m^2$ to 67 $kg/m^2$ , so for observing the effect we classify the data into 2 groups :-

- Grp1:-Women with BMI $\geq$ 30 $kg/m^2$
- Grp2:-Women with BMI $<$ 30 $kg/m^2$ and form the following contingency table and check whether they have diabetes or not.

Table 6:

| DiabetesOutcome (Y)\|BMI (X3) | <30 $Kg/m^2(0)$ | $\geq$ 30Kg/m$^2$ (1) | Total |
|---|---|---|---|
| Absent(0) | 238 | 262 | 500 |
| Present(1) | 47 | 221 | 268 |
| Total | 285 | 483 | 768 |

So from the given table we will first compute the odds of diabetes in women with BMI $\geq$ 30 $Kg/m^2$

- The odds of presence of diabetes when the BMI of women is less than 30 $Kg/m^2$ :-

$\frac{P(Y=1|X3=0)}{P(Y=0|X3=0)}$ = $\frac{P(Y=1,X3=0)}{P(Y=0,X3=0)}$=$\frac{47}{238}$= 0.197

- The odds of presence of diabetes when the BMI Of women is greater than equals 30 Kg :-

$\frac{P(Y=1|X3=1}{P(Y=0|X3=1)}$ = $\frac{P(Y=1,X3=1)}{P(Y=0,X3=1)}$=$\frac{221}{262}$= 0.84

- To test: $H_0$ : The independence of the covariate and the response variable,presence of Diabetes against $H_1$ : not $H_0$

  Here the test statistic follows chi square distribution with degrees of freedom = 1

  The observed value of the pearsonian chi square is given by :——

  $\chi^2_{observed}$ = 66.284 , $\chi^2_{0.05,1}$=3.841 and Pvalue=0

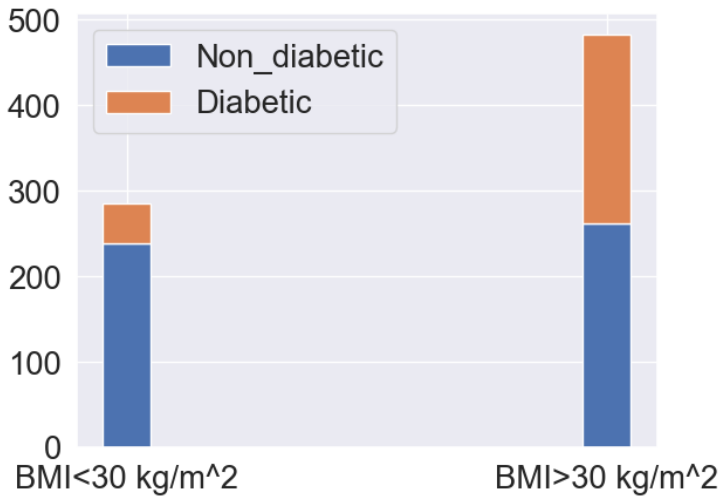


Figure 3: Stacked bar diagram of response variable against BMI

- Descision: $\chi^2_{observed} > \chi^2_{0.05,1}$,we reject $H_0$
- Conclusion: In the light of the given data Diabetes status depends on the BMI of the woman.

  When we compute the odds ratio ,we observe that incase of diabetic patients , The BMI is around 4.5 times than that of non diabetic patients.

- Effect of Glucose(X4) in women on the presence of diabetes or not

From this data we observe that Glucose concentration in the body of a women varies from 44 mg to 200 mg , so for observing the effect we classify the data into 2 groups :-

- Grp1:-Women with Glucose concentration $\geq 100$ mg
- Grp2:-Women with Glucose Concentration $< 100mg$ and form the following contingency table and check whether they have diabetes or not.

Table 7:

| DiabetesOutcome (Y)\|Glucose (X4) | <100 mg (0) | $\geq 100$ mg (1) | Total |
|---|---|---|---|
| Absent(0) | 178 | 322 | 500 |
| Present(1) | 14 | 254 | 268 |
| Total | 192 | 576 | 768 |

So from the given table we will first compute the odds of diabetes in women with Glucose concentration $\geq 100$mg

- The odds of presence of diabetes when the Glucose conncentration in women is less than 100mg :-

$\frac{P(Y=1|X4=0)}{P(Y=0|X4=0)}$ = $\frac{P(Y=1,X4=0)}{P(Y=0,X4=0)}$=$\frac{14}{178}$= 0.07

- The odds of presence of diabetes when the concentration of gluocose in women is greater than equals to 100 mg :-

$\frac{P(Y=1|X4=1}{P(Y=0|X4=1)}$ = $\frac{P(Y=1,X1=1)}{P(Y=0,X1=1)}$=$\frac{254}{322}$= 0.788

- To test: $H_0$ : The independence of the covariate and the response variable,presence of Diabetes against $H_1$ : not $H_0$

  Here the test statistic follows chi square distribution with degrees of freedom = 1

  The observed value of the pearsonian chi square is given by :——

  $\chi^2_{observed}$ = 84.25 , $\chi^2_{0.05,1}$=3.841 and Pvalue=0.00

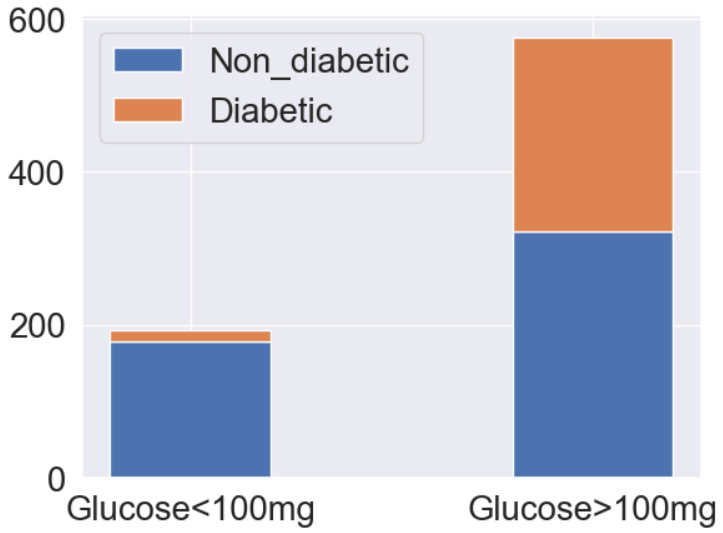


Figure 4: Stacked bar diagram of response variable against Glucose conc.

- Descision: $\chi^2_{observed} > \chi^2_{0.05,1}$,we reject $H_0$
- Conclusion: In the light of the given data Diabetes status depends on the Glucose Concentration.

  When we compute the odds ratio ,we observe that incase of diabetic patients , The Glucose concentration in their body is around 11 times than that of non diabetic patients.

- Effect of Skin Thickness(X5) in women on the presence of diabetes or not

From this data we observe that Skin Thickness in the body of a woman varies from 7 mm to 99 mm , so for observing the effect we classify the data into 2 groups :-

- Grp1:-Women with Skin Thicness $\geq 24$ mm
- Grp2:-Women with Skin Thickness $< 24$ mm and form the following contingency table and check whether they have diabetes or not.

Table 8:

| DiabetesOutcome (Y)\|Skin Thickness (X5) | <24 mm (0) | $\geq 24$ mm (1) | Total |
|---|---|---|---|
| Absent(0) | 284 | 216 | 500 |
| Present(1) | 115 | 153 | 268 |
| Total | 399 | 369 | 768 |

So from the given table we will first compute the odds of diabetes in women with Skin Thickness $\geq 24$mm

- The odds of presence of diabetes when the Skin Thickness in women is less than 24mg :-

$\frac{P(Y=1|X5=0)}{P(Y=0|X5=0)}$ = $\frac{P(Y=1,X5=0)}{P(Y=0,X5=0)}$=$\frac{115}{284}$= 0.404

- The odds of presence of diabetes when the Skin Thickness in women is greater than equals to 24 mm :-

$\frac{P(Y=1|X5=1}{P(Y=0|X5=1)}$ = $\frac{P(Y=1,X5=1)}{P(Y=0,X5=1)}$=$\frac{153}{216}$= 0.708

- To test: $H_0$ : The independence of the covariate and the response variable,presence of Diabetes against $H_1$ : not $H_0$

  Here the test statistic follows chi square distribution with degrees of freedom = 1

  The observed value of the pearsonian chi square is given by :——

  $\chi^2_{observed}$ = 12.93 , $\chi^2_{0.05,1}$=3.841 and Pvalue=0.00032

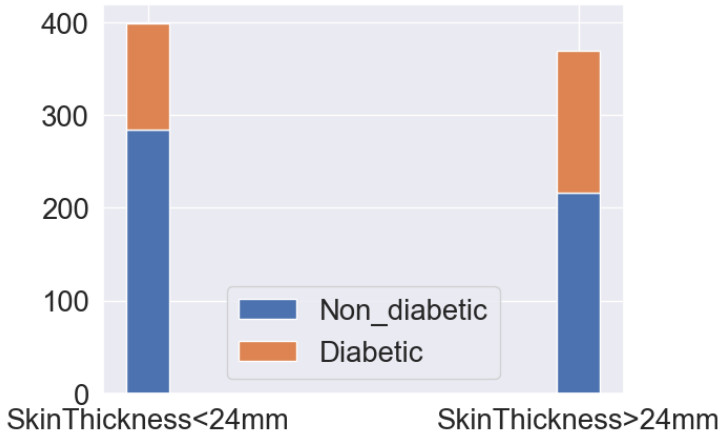


Figure 5: Stacked Bar diagram of response variable against Skin Thickness

- Descision: $\chi^2_{observed} > \chi^2_{0.05,1}$,we reject $H_0$
- Conclusion: In the light of the given data Diabetes status depends on the Skin Thickness.

  When we compute the odds ratio ,we observe that incase of diabetic patients , The Skin Thickness in their body is around 2 times than that of non diabetic patients.

-

- Effect of Blood Pressure (X7) in women on the presence of diabetes or not

From this data we observe that Blood Pressure in the body of a woman varies from 24 mm Hg to 124 mm Hg , so for observing the effect we classify the data into 2 groups :-

- Grp1:-Women with Blood Pressure $\geq 75$ mm Hg
- Grp2:-Women with Blood Pressure $< 75$ mm Hg and form the following contingency table and check whether they have diabetes or not.

So from the given table we will first compute the odds of diabetes in women with Blood Pressure $\geq 75$mm Hg

Table 9:

| DiabetesOutcome (Y)\|Blood Pressure (X7) | <75 mm Hg (0) | $\geq$ 75 mm Hg (1) | Total |
|---|---|---|---|
| Absent(0) | 329 | 171 | 500 |
| Present(1) | 142 | 126 | 268 |
| Total | 471 | 297 | 768 |

- The odds of presence of diabetes when the Blood Pressure in women is less than 75mm Hg :-

$\frac{P(Y=1|X7=0)}{P(Y=0|X7=0)}$ = $\frac{P(Y=1,X7=0)}{P(Y=0,X7=0)}$=$\frac{142}{329}$= 0.43

- The odds of presence of diabetes when the Blood Pressure in women is greater than equals to 75 mm Hg :-

$\frac{P(Y=1|X7=1}{P(Y=0|X7=1)}$ = $\frac{P(Y=1,X7=1)}{P(Y=0,X7=1)}$=$\frac{126}{171}$= 0.73

- To test: $H_0$ : The independence of the covariate and the response variable,presence of Diabetes against $H_1$ : not $H_0$

  Here the test statistic follows chi square distribution with degrees of freedom = 1

  The observed value of the pearsonian chi square is given by :——

  $\chi^2_{observed}$ = 11.54 , $\chi^2_{0.05,1}$=3.841 and Pvalue=0.00067

- Descision: $\chi^2_{observed}$> $\chi^2_{0.05,1}$,we reject $H_0$

- Conclusion: In the light of the given data Diabetes status depends on the Blood Pressure.

  When we compute the odds ratio ,we observe that incase of diabetic patients , The Blood Pressure(>75 mmHg) in their body is around 1.69 times than that of non diabetic patients.

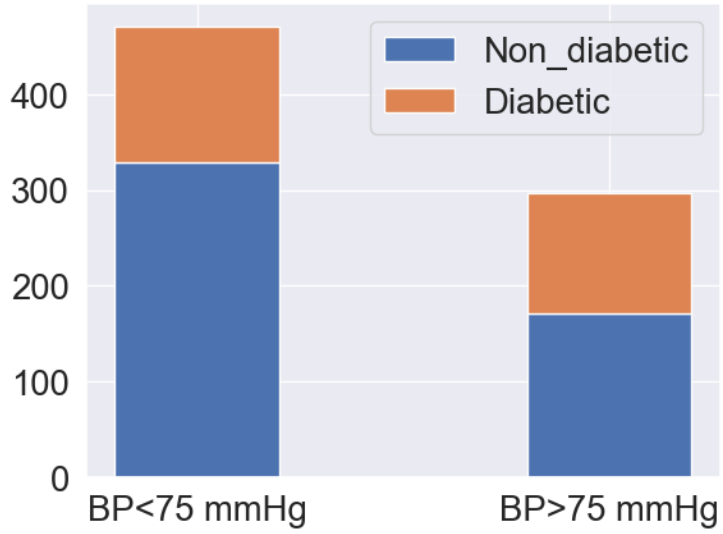


Figure 6: Stacked bar diagram of response variable against Blood Pressure

- Effect of Diabetes Pedigree Function (X8) in women on the presence of diabetes or not

From this data we observe that Diabetes Pedigree Function(DPF) of a woman varies from 0.07 to 2.4 , so for observing the effect we classify the data into 2 groups :-

- Grp1:-Women with DPF $\geq 0.5$
- Grp2:-Women with DPF $< 0.5$ and form the following contingency table and check whether they have diabetes or not.

Table 10:

| DiabetesOutcome (Y)\|DPF (X8) | <0.5 (0) | $\geq 0.5$ (1) | Total |
|---|---|---|---|
| Absent(0) | 349 | 151 | 500 |
| Present(1) | 142 | 126 | 268 |
| Total | 491 | 277 | 768 |

So from the given table we will first compute the odds of diabetes in women with DPF score $\geq 0.5$

- The odds of presence of diabetes when the DPF score in women is less than 0.5 :-

$\frac{P(Y=1|X8=0)}{P(Y=0|X8=0)}$ = $\frac{P(Y=1,X8=0)}{P(Y=0,X8=0)}$=$\frac{142}{349}$= 0.40

- The odds of presence of diabetes when the DPF score in women is greater than equals to 0.5 :-

$\frac{P(Y=1|X8=1}{P(Y=0|X8=1)}$ = $\frac{P(Y=1,X8=1)}{P(Y=0,X8=1)}$=$\frac{126}{151}$= 0.83

- To test: $H_0$ : The independence of the covariate and the response variable,presence of Diabetes against $H_1$ : not $H_0$

  Here the test statistic follows chi square distribution with degrees of freedom = 1

  The observed value of the pearsonian chi square is given by :——

  $\chi^2_{observed}$ = 20.67 , $\chi^2_{0.05,1}$=3.841 and Pvalue=0.000

- Descision: $\chi^2_{observed}$> $\chi^2_{0.05,1}$,we reject $H_0$

- Conclusion: In the light of the given data Diabetes status depends on the DPF score.

  When we compute the odds ratio ,we observe that incase of diabetic patients , The DPF score - is around 2 times than that of non diabetic patients.

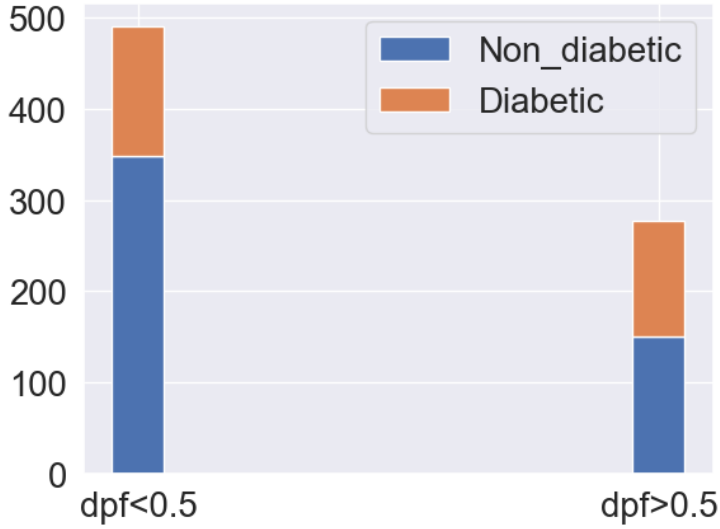


Figure 7: Stacked Bar diagram of response variable against Diabetes Pedigree Function

# 4 Comparison of Logit Model,Probit Model and KNN Model

## 4.1 Logit Model

Here in this section we will see how well the logistic regression model fits the data.By a good fitted model we mean a model where the absolute difference between the actual value and the predicted value of the response variable is low on average.

Usually we use the coefficient of determination $R^2$ or the adjusted $R^2$ to find out how well the model fits to the given data.$R^2$ measures the total variation in the response variable that is explained by the fitted regression model.

Note:- In this data insulin is an insignificant covariate according to the data but we will still keep insulin in our model as it still holds a lot of information on diabetes because in real life insulin plays a major role in determining diabetes.

Here we split the data into Train Data and Test Data where the training set contains 70% of the information of the data and the test data contains 30% of the information.

We will define the Training Set and the Test set below as follows:-

- Training Dataset:-

Algorithms are used in machine learning to learn from datasets of data. They discover patterns, acquire comprehension, make choices, and evaluate those choices.

In AI, datasets are parted into two subsets.

The main subset is known as the preparation information - it's a piece of our real dataset that is taken care of into the AI model to find and learn designs. Our model is trained in this manner.

The testing data is the other subset. We'll cover more on this underneath.

Testing data tend to be smaller than training data. This is due to our desire to provide the model with as much data as possible in order to identify and comprehend meaningful patterns. When information from our datasets are taken care of to an AI calculation, it gains designs from the information and decides.

Machines can solve problems based on previous observations thanks to algorithms. Similar to learning from people's examples. The only difference is that machines require significantly more examples to learn and recognize patterns.

Over time, machine learning models get better as they are exposed to more relevant training data.

- Test Dataset:-

  You need unseen data to test your machine learning model after it has been constructed using your training data. This information is called trying information, and you can utilize it to assess the exhibition and progress of your calculations' preparation and change or enhance it for further developed results.

  Two main criteria for testing data exist. It ought to:

  Represent the actual dataset Be large enough to make accurate predictions As previously stated, this dataset must be brand-new, "unseen" data. This is on the grounds that your model definitely "knows" the preparation information. You will be able to determine whether it is functioning accurately or whether it needs additional training data to meet your requirements based on how it performs on new test data.

Test data is a final, real-world check of a dataset that has never been seen to ensure that the machine learning algorithm was properly trained.

In data science, it's common practice to divide your data into training and testing groups of 70 percent. and 30 percent

Now we will consider the logit model:-
P(Y=1$|\ X1, X2, ..., Xp) = \frac{e^{\eta}}{1+e^{\eta}}$

$\eta = -9.0984 + 0.1249X_1 + 0.0130X_2 + 0.1003X_3 + 0.0385X_4 - 0.0029X_5 - 0.0015X_6 - 0.0132X_7 + 0.9451X_8$

since in this case,our response variable is binary in nature,finding the goodness of fit using the $R^2$ values may not give satisfactory results.

There are other measures to find the goodness of fit in case of logistic regression model.Here we will the confusion matrix to see how the model fits to the data

Confusion Matrix Layout

| Y \| $\hat{Y}$ | Absent (0) | Present (1) |
|---|---|---|
| Absent (0) | TN | FP |
| Present (1) | FN | TP |

Y and $\hat{Y}$ are the actual and the predicted value of the response variable .

- True Positive (TP):- It is the case when we predict that diabetes is present and actual present in the original data.

  TPR(True Positive Rate)=P($\hat{Y} = 1 \mid Y = 1$)

- True Negative (TN):- It is the case when we predict that diabetes is absent and actual absent in the original data.

  TNR(True Negative Rate)=P($\hat{Y} = 0 \mid Y = 0$)

- False Positive (FP):- It is the case when we predict that diabetes is present and actual absent in the original data.

FPR(False Positive Rate)=P($\hat{Y} = 1 \mid Y = 0$)

- False Negative (FN):- It is the case when we predict that diabetes is Absent and actual Present in the original data.

FNR(False Negative Rate)=P($\hat{Y} = 0 \mid Y = 1$)

Now as we have defined the whole layout of confusion matrix ,thus we will present the confusion matrix in case of logit regression model and check the model accuracy.Here we will get the confusion matrix on the basis of the test data which is 231 as we have taken 30% of the data at random:-

| Y \| $\hat{Y}$ | Absent(0) | Present(1) | Total |
|---|---|---|---|
| Absent(0) | 125 | 27 | 152 |
| Present(1) | 37 | 42 | 79 |
| Total | 162 | 69 | 231 |

Now we will calculate the model accuracy :-

$$\text{Accuracy=} \frac{\text{No. of observations predicted correctly}}{231} * 100 = 72.29\%$$

so here we can observe that the accuracy of the model is 72.29% which is a fairly good prediction if not the best as we have included insulin in the model which is insignificant with respect to data but it still holds a lot of valuable information to the real life.

- ROC CURVE:-

ROC curve depicts a binary classifier system's diagnostic capabilities as its discrimination threshold is varied. TPR vs. FPR are plotted on a ROC graph at various classification levels. More objects are classified as positive when the classification criterion is lowered, which raises the number of both False Positives and True Positives.

Area under the curve (AUC):-

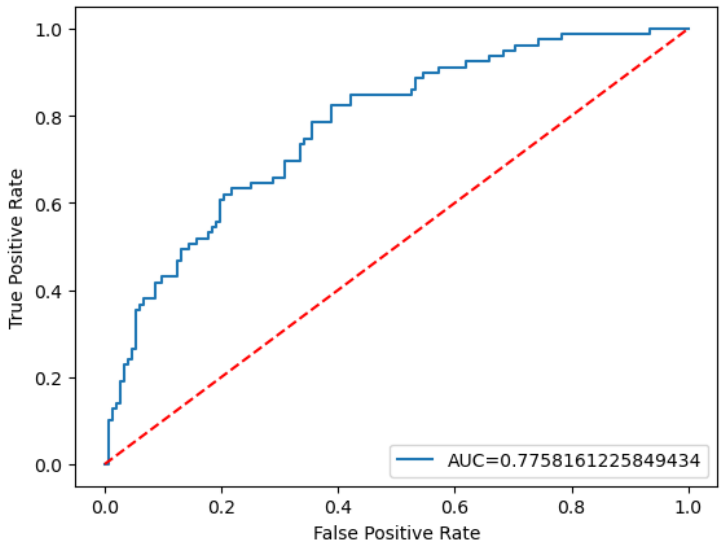


Figure 8: ROC curve under logit model

The abbreviation "Area under the ROC Curve" is AUC. In other words, AUC gauges the complete two-dimensional region beneath the entire ROC curve from (0,0) to (1,1) (contemplate integral calculus). (1,1).

The success across all potential classification levels is measured overall by AUC. AUC can be understood as the likelihood that the model values a randomly chosen positive example higher than a randomly chosen negative example. It is also used as a measure of overall performance of a classifier .

- Interpretation:-

In the above Logit model the **AUC** is 0.775 which means 77.5% time the model will assign a higer absolute risk to a randomly selected women with diabetes rather than randomly selected women without diabetes and the model has a strong predictive power as area is near 0.8.

## 4.2 Probit Model:-

Here in this section we will see how well the Probistic regression model fits the data.By a good fitted model we mean a model where the absolute difference between the actual value and the predicted value of the response variable is low on average. Here we split the data into Train Data and Test Data where the training set contains 70% of the information of the data and the test data contains 30% of the information.
Now we will consider the Probit model:-
P(Y=1| $X1, X2, ..., Xp$) =$\Phi(\eta)$ where $\eta = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$
so the $\eta = 0.0837X_1 - 0.0042X_2 + 0.0180X_3 + 0.0128X_4 - 0.0140X_5 + 0.0006X_6 - 0.0336X_7 + 0.0990X_8$
Now to check the accuracy of the model we will get the confusion matrix of the predicted and the observed values of the response variable.

- Confusion Matrix for Probit Model:-

| Y \| $\hat{Y}$ | Absent(0) | Present(1) | Total |
|---|---|---|---|
| Absent(0) | 127 | 23 | 150 |
| Present(1) | 41 | 40 | 81 |
| Total | 168 | 63 | 231 |

Now we will calculate the model accuracy :-

Accuracy= $\frac{\text{No. of observations predicted correctly}}{231} * 100 = 72.29\%$
so here we can observe that the accuracy of the model is 72.29% also for the probit model as well which is a fairly good prediction if not the best as we have included insulin in the model which is insignificant with respect to data but it still holds a lot of valuable information to the real life.

- ROC CURVE:-

- Interpretation:-

In the below Probit model the **AUC** is 0.67 which means 67% time the model will assign a higer absolute risk to a randomly selected women with diabetes rather than randomly selected women without diabetes

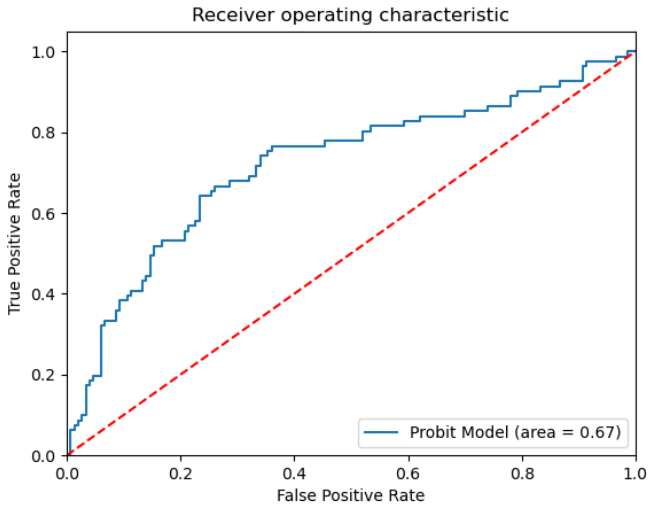


Figure 9: ROC curve under Probit model

### 4.3 KNN Classification:-

A supervised learning classifier, the k-nearest neighbors algorithm, also known as KNN or k-NN, uses proximity to classify or predict the grouping of a single data point. It can be used for either classification or regression problems, but most of the time it is used as a classification algorithm because it assumes that similar points can be found close to each other.Here also we are splitting the data into training dataset and test dataset in 70 : 30 ratio.

The idea behind regression problems and classification problems is similar, but in this case, the average of the k closest neighbors is used to predict a classification. Classification is used for discrete values, whereas regression is used for continuous ones, which is the main difference here. However, the distance must be defined before a classification can be made. The most common method is the Euclidean distance, which we'll discuss in greater detail below.

It's likewise important that the KNN calculation is likewise essential for a group of "lazy learning" models, implying that it just stores a preparation dataset as opposed to

going through a preparation stage. This also indicates that when a classification or prediction is made, the entire computation takes place. It is also known as an instance-based or memory-based learning method because all of its training data is stored in memory. However, as a dataset grows, KNN becomes increasingly inefficient, compromising overall model performance. It is commonly used for simple recommendation systems, pattern recognition, data mining, financial market predictions, intrusion detection, and more.

- KNN distance metrics :-

  The k-nearest neighbor algorithm's objective is to determine the query point's closest neighbors so that a class label can be assigned to that point. KNN needs to meet a few things in order to accomplish this:

- Euclidean Distance(p=2):- This is the distance measure that is used the most, but it only works with real-valued vectors. It measures a straight line between the query point and the other point being measured using the formula below. The power parameter that is utilized in the Minkowski distance metric to determine the distance that separates two points in a space with multiple dimensions is referred to as "p." The Minkowski distance, which is an extension of the Euclidean distance, can be calculated in the following manner: The general expression for Euclidean Distance :-

  d(x,y)=$\left|\left|\sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}\right|\right|$

  where $y_i$ and $x_i$ are the two points between which the distance is measured.

- Minkowski Distance:- The generalized version of the Euclidean and Manhattan distance metrics is this distance measure. Other distance metrics can be created thanks to the parameter p in the formula below.

  d(x,y)=$\left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{\frac{1}{p}}$

We are here using the minkowski distance formula with p=2.
Now we have to check the accuracy of the model .Thus, we create a confusion matrix of the predicted values and the observed values of the response variable.

| Y \| $\hat{Y}$ | Absent(0) | Present(1) | Total |
|---|---|---|---|
| Absent(0) | 122 | 32 | 154 |
| Present(1) | 31 | 46 | 77 |
| Total | 153 | 78 | 231 |

Now we will calculate the model accuracy :-

Accuracy= $\frac{\text{No. of observations predicted correctly}}{231} * 100 = 72.27\%$

so here we can observe that the accuracy of the model is 72.27% also for the KNN Classification as well which is a fairly good prediction if not the best as we have included insulin in the model which is insignificant with respect to data but it still holds a lot of valuable information to the real life.

- ROC CURVE:-

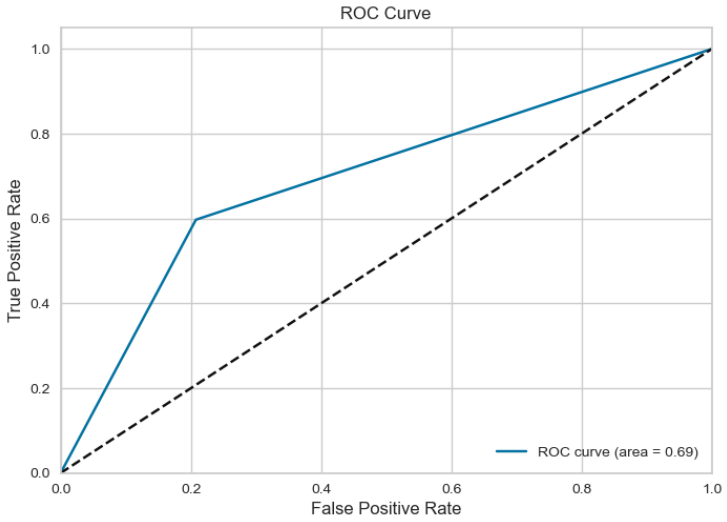


Figure 10: ROC curve under KNN Classification model

- Interpretation:-

In the above KNN Classification model the **AUC** is 0.69 which means 69% time the

model will assign a higer absolute risk to a randomly selected women with diabetes rather than randomly selected women without diabetes

# 5 Conclusion

Now we have entered the final section of the project.Here we will summarise our findings from the previous discussions.We had data on women if they had diabetes or not and we have 8 covariates under study collected on 768 observations.We did testing using p values and chi square test of independence .After that we fit 3 models for classification purpose Logit Model,Probit Model and KNN Classification model on our test data and compare the accuracy of the fitted model and then calculate the AREA UNDER CURVE for comparing the goodness of the fitted model. Now we will finally list down all our finding from the analysis we have conducted

1. It is observed that for higher no. of pregnancies in women increases the diabetes rate significantly.
2. Incase of age factor also there is a considerable increase in diabetic women after the age of 35 .
3. Glucose concentration in blood also has a significant effect on diabetes rate .It is considered that 100 mg is an optimum concentration of glucose in human body.
4. It is also noticeable that diabetic patients have a thicker skin than that of non diabetic patients.
5. Here we also know that diabetic patients take insulin more than non diabetic patients.
6. Incase of blood pressure it is also observed that women with diastolic bp more than 70mmHg have more diabetic rate than with less than 70 mmHg .70 mmHg is also considered as an optimal diastolic blood pressure .
7. Body mass index is a very important measure which is a great indicator of a healthy body . So with Bmi greater than $30kg/m^2$ has shown a greater diabetic

rate in women.

8. Diabetes Pedigree Function has a significant effect on a person with diabetes .It tells us that if our ancestors are also affected with diabetes then there is a high probability that in future we can also have diabetes

9. After we have fitted the 3 classification model it is observed that the rate of accuracy of all he models is nearly 73% but the AUC from the ROC Curve is 77.5% incase of logit model which is higher than those of probit model and KNN classification model. Thus logit model is more accurate for this dataset incase of predictive analysis of diabetes.

# 6 Appendix

Python Code:-

```
import os
os.chdir(r"D:\dissertation 2nd project 2023")
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.gridspec as gridspec # subplots
from matplotlib.ticker import FormatStrFormatter



#Import models from scikit learn module:
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import KFold   #For K-fold cross validation
from sklearn.ensemble import RandomForestClassifier
from sklearn.tree import DecisionTreeClassifier, export_graphviz
from sklearn import metrics
```

```
import statsmodels.api as sm
import scipy.stats as st

import seaborn as sns
from sklearn.metrics import confusion_matrix

diabetes=pd.read_csv("diabetes.csv")
diabetes
diabetes["P_I"]=np.where(diabetes.Pregnancies>=7,1,0)
diabetes
tab=pd.crosstab(diabetes.Outcome, diabetes.P_I, margins=True)

## same code used for other variables for contingency table construction


# Code for stacked bar diagram construction for other variables
and also testing using chi square
tab
tab1=tab.iloc[:-1,:-1]
tab1
st.chi2_contingency(tab1)
Non_diabetic=[426,74]
Diabetic=[173,95]
l=["<7 pregnancies",">7 pregnancies"]
plt.bar(1,Non_diabetic,0.4,label="Non_diabetic")
plt.bar(1,Diabetic,0.4,bottom=Non_diabetic,label="Diabetic")

plt.legend()

#classification
```

```
#Logit model:-

import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
import matplotlib.pyplot as plt

#define the predictor variables and the response variable
X = diabetes.iloc[ :,: 8]
y = diabetes.iloc[:, 8]
#split the dataset into training (70%) and testing (30%) sets
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.3)

#instantiate the model
log_regression = LogisticRegression()

#fit the model using the training data
log_regression.fit(X_train,y_train)
(log_regression.fit(X_train,y_train)).summary()



y_pred_proba = log_regression.predict_proba(X_test)[::,1]
fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)
auc = metrics.roc_auc_score(y_test, y_pred_proba)

#create ROC curve
plt.plot(fpr,tpr,label="AUC="+str(auc))
plt.plot([0, 1], [0, 1],'r--')
```

```
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.legend(loc=4)
plt.show()

y_pred=log_regression.predict(X_test)

#confusion matrix
from sklearn.metrics import confusion_matrix
confusion_matrix=confusion_matrix(y_test,y_pred)
print(confusion_matrix)



#Probit Model:-
probit_model=sm.Probit(y,X)
result=probit_model.fit()
print(result.summary())

from sklearn import metrics
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_
probit=sm.Probit(y_train,X_train)
probit.fit()
print(probit.fit().summary())



result1 = X_test
result1['y_pred'] = result1['Pregnancies'] * 0.0837 + result1['Glucose'] *0.0128
result1
```

```
import scipy.stats as si
def normsdist(z):
    z = si.norm.cdf(z,0.0,1.0)
    return (z)
normsdist(1.96)
result1['y_pred_Probit'] = normsdist(result1['y_pred'])
result1

d = {'y_pred_proba': result1['y_pred_Probit']}
df23 = pd.DataFrame(data=d)
df23 = df23.reset_index()
df23.drop(['index'], axis=1, inplace=True)
df23['y_pred'] = 0.000
for i in range(0,len(df23['y_pred_proba'])):
    if df23['y_pred_proba'][i] > 0.500:
        df23['y_pred'][i] = 1.000
    else:
        df23['y_pred'][i] = 0.000
y_pred = np.array(df23['y_pred'])
y_pred = y_pred.astype('int64')
y_pred

from sklearn.metrics import confusion_matrix
confusion_matrix = confusion_matrix(y_test, y_pred)
print(confusion_matrix)
y_pred_proba = np.array(df23['y_pred_proba'])
y_pred_proba

from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve
```

```
probit_roc_auc = roc_auc_score(y_test, y_pred)
fpr, tpr, thresholds = roc_curve(y_test, y_pred_proba)
plt.figure()
plt.plot(fpr, tpr, label='Probit Model (area = %0.2f)' % probit_roc_auc)
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.savefig('Probit_ROC')
plt.show()


# KNN Classification Model:-
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.30)
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train = sc.fit_transform(X_train)
X_test = sc.transform(X_test)
rom sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 5, metric = 'minkowski', p = 2)
classifier.fit(X_train, y_train)
y_pred = classifier.predict(X_test)
y_pred
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
from sklearn.metrics import accuracy_score
```

```
print ("Accuracy : ", accuracy_score(y_test, y_pred))
cm
roc_auc=auc(fpr,tpr)

# Plot of a ROC curve for a specific class
plt.figure()
plt.plot(fpr, tpr, label='ROC curve (area = %0.2f)' % roc_auc)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
plt.legend(loc="lower right")
plt.show()
```

# 7 Reference:-


1. https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/
2. https://www.obviously.ai/post/the-difference-between-training-data-vs-test-data-in-machine-learning
3. https://www.scribbr.com/statistics/chi-square-test-of-independence/#:~:text=a%20bar%20graph%3A-,Chi%2Dsquare%20test%20of%20independence%20hypotheses,are%20related%20in%20the%20population.