**Statistical Learning Lab**

**Assignment - 2**

**Logistic Regression Assignment**

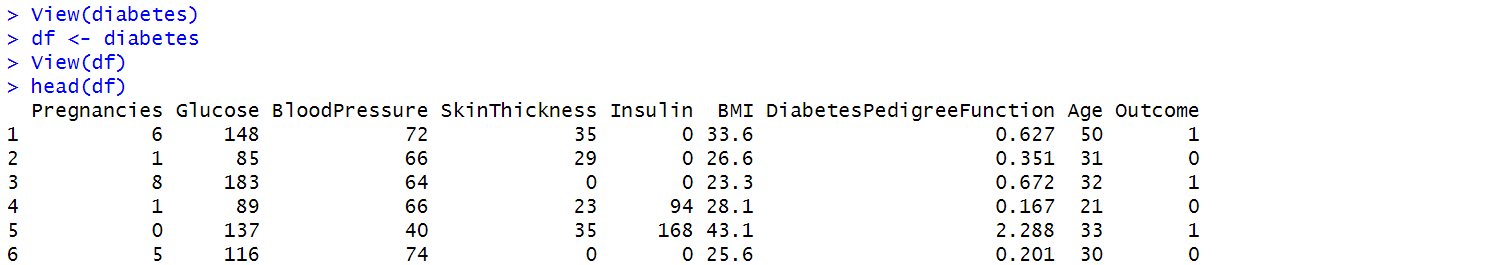
**Name : Sunny Kumar**

**Roll No : 22IM10040**

**Show the code snippets and the corresponding output for the following:**

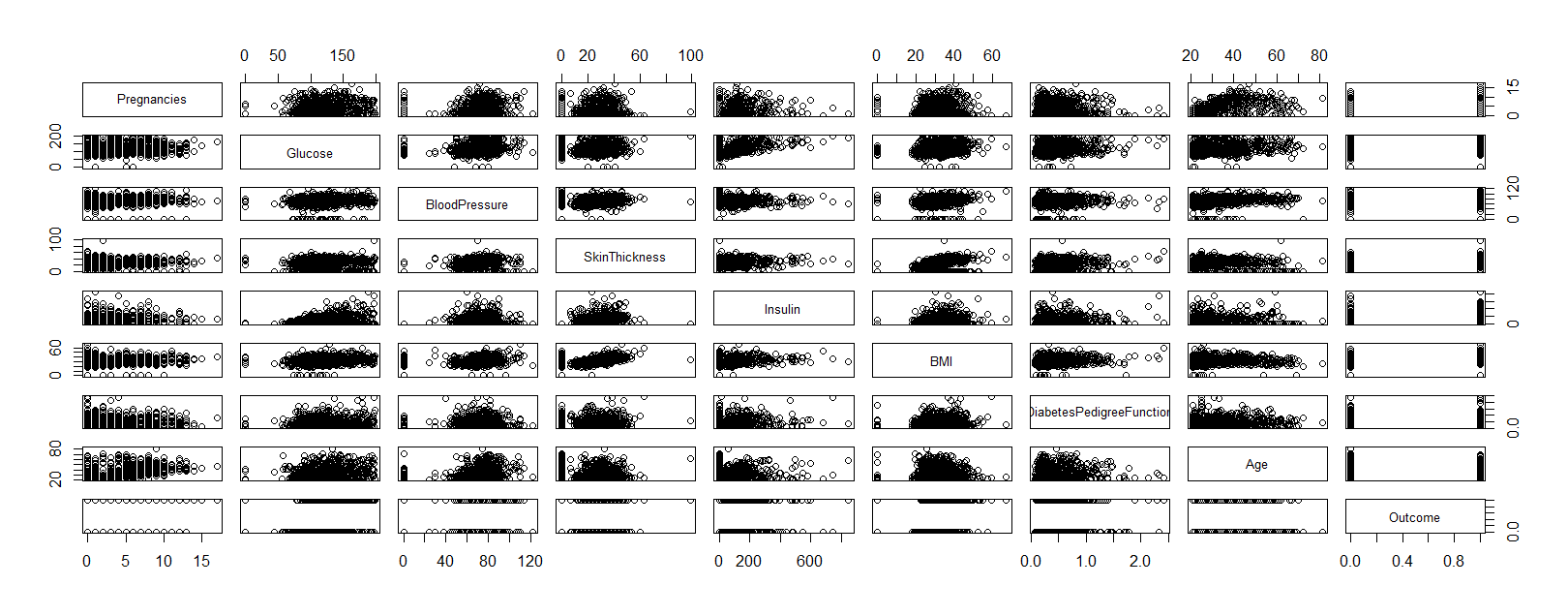
1. **Load the dataset “diabetes.csv”. Display first few rows of the dataset**.

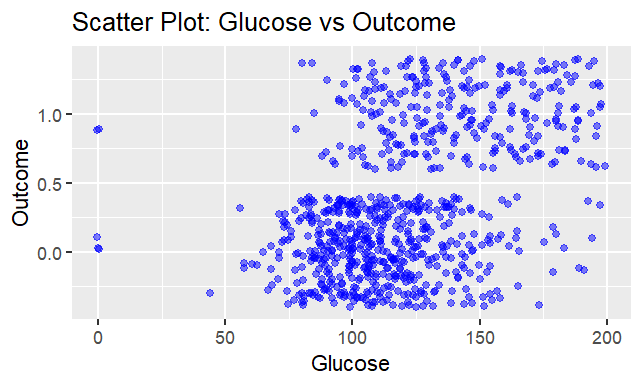
Ans : Loaded the dataset using environment -> import dataset and imported dataset.

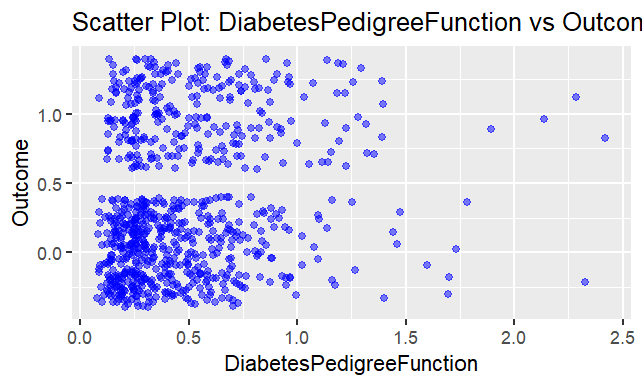


1. **Perform preliminary analysis to show how the variables are related to each other. Use scatter plot, box plot etc. to visualize how different variables impact the “Outcome” variable.**

Ans : Scatter plot of among different variables is given below:







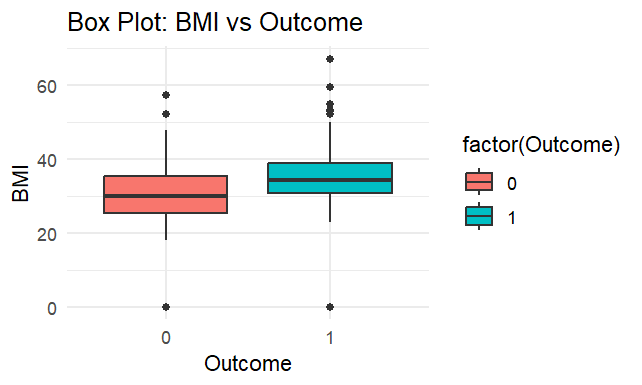
From the above two scatter plots we, cannot say anything that some particular column is impacting the outcome variable.

For lower boxplot : *ggplot(diabetes, aes(x = factor(Outcome), y = BMI, fill = factor(Outcome))) +*

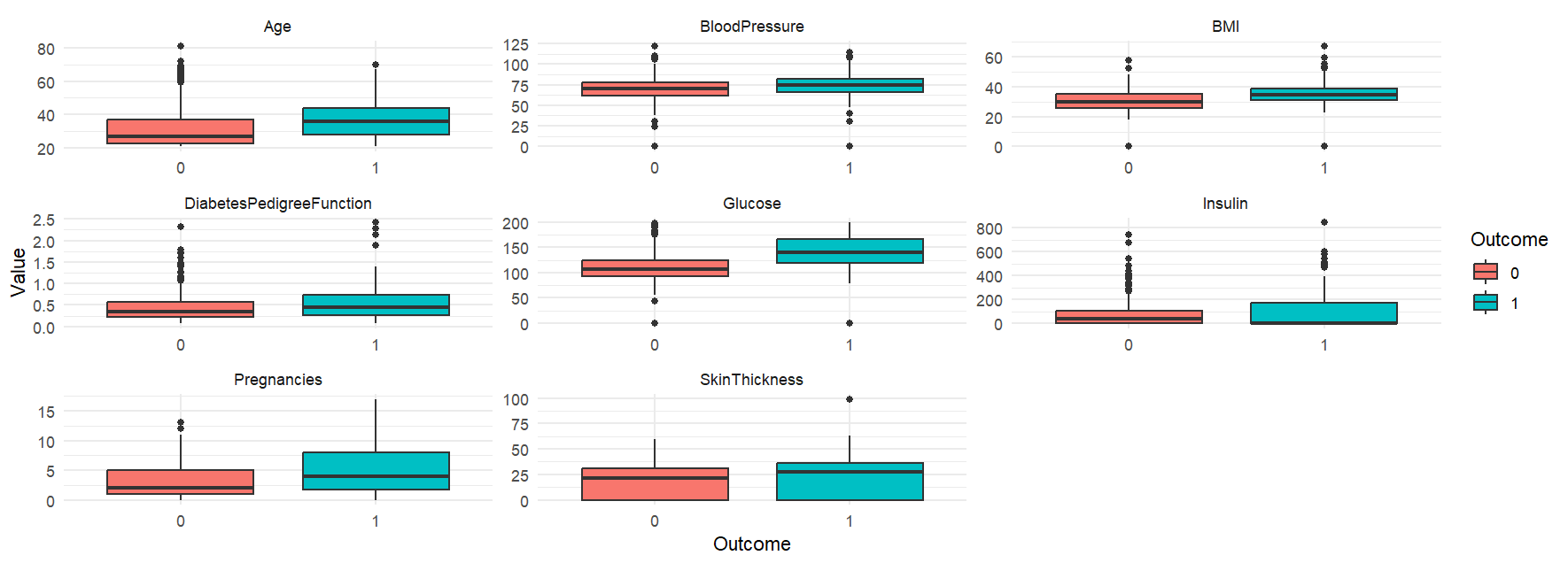
*geom\_boxplot() +*

*labs(title = "Box Plot: BMI vs Outcome", x = "Outcome", y = "BMI") +*

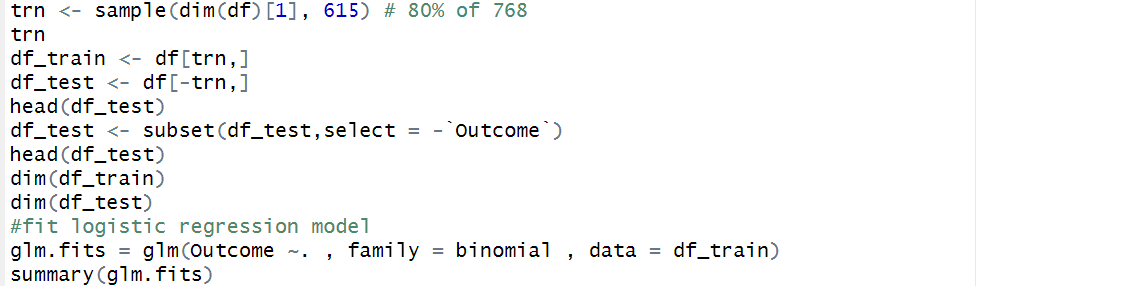
*theme\_minimal()*

**

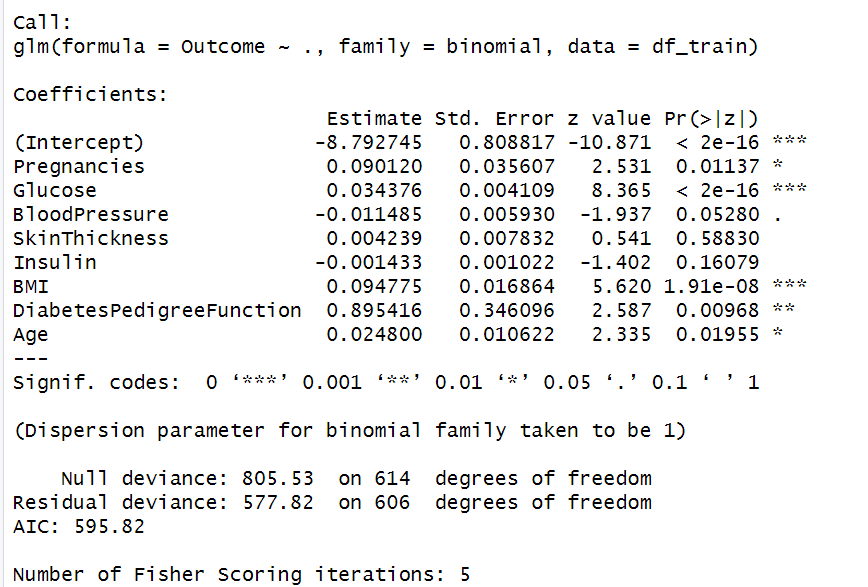
My Inference : The median BMI for individuals with Outcome = 1 (diabetic) is higher than for individuals with Outcome = 0 (non-diabetic). This suggests that diabetics tend to have a higher BMI on average. BMI appears to have a relationship with the Outcome variable. Higher BMI values are more associated with diabetes (Outcome = 1). Similarly, All the plots are shown below.



1. **Randomly sample 80% of the data as training data and rest as test data. Fit a Logistic Regression model with all the predictors on training data. From the summary which factors seem to be significant? Explain how the predictors impact the log-odds ratio of diagnosed with diabetes (Outcome)**

****

Interpretations: From the summary we can say that Columns : “ Pregnancies , Glucose , BMI , DiabetesPedigreeFunction, Age” are significant , because their P value is less than 0.05.



**Variable-Specific Interpretation(Log Odds Ratio):**

**Pregnancies (0.090120, p = 0.01137)**:

* A one-unit increase in the number of pregnancies increases the log-odds of diabetes diagnosis by 0.090120. This is a statistically significant predictor.

**Glucose (0.034376, p < 2e-16)**\*:

* A one-unit increase in glucose levels significantly increases the log-odds of diabetes diagnosis by 0.034376. This is a highly significant predictor.

**BloodPressure (-0.011485, p = 0.05280)**:

* A one-unit increase in blood pressure slightly decreases the log-odds of diabetes diagnosis by 0.011485. However, this predictor is only marginally significant (. indicates a p-value close to 0.05).

**SkinThickness (0.004239, p = 0.58830)**:

* This predictor has a small positive coefficient but is not statistically significant (p > 0.05), meaning its impact on the Outcome is uncertain.

**Insulin (-0.001433, p = 0.16079)**:

* Insulin has a negligible negative coefficient and is not statistically significant.

**BMI (0.094775, p = 1.91e-08)**\*:

* A one-unit increase in BMI significantly increases the log-odds of diabetes diagnosis by 0.094775. This is a strong and highly significant predictor.

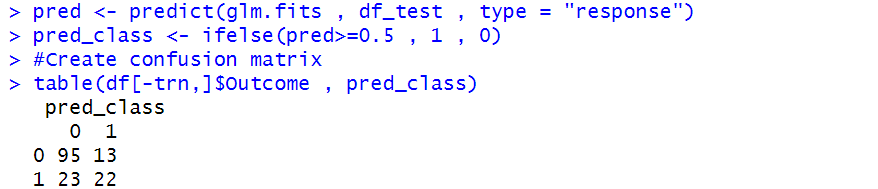
**DiabetesPedigreeFunction (0.895416, p = 0.00968)**:

* A one-unit increase in this metric increases the log-odds of diabetes diagnosis by 0.895416. This variable is statistically significant.

**Age (0.024800, p = 0.01955)**:

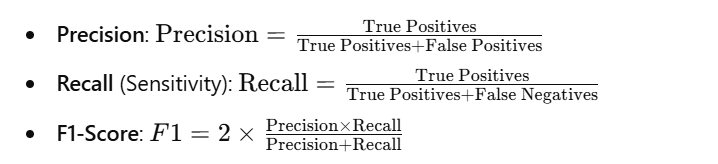
* A one-year increase in age increases the log-odds of diabetes diagnosis by 0.024800. Age is statistically significant.

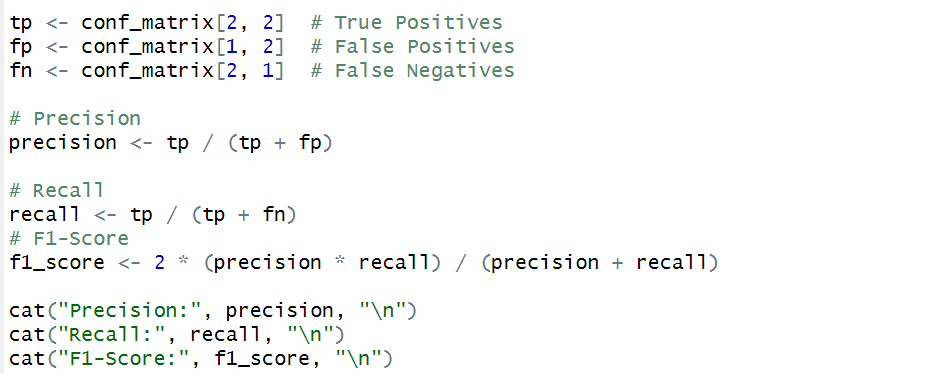
1. **From the model fitted in problem 3, derive confusion matrix, accuracy, and F1-score on test data.**

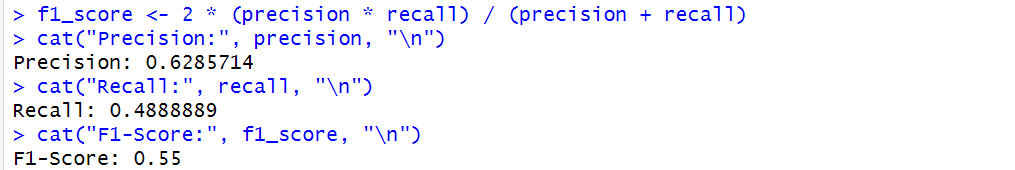


Ans : Accuracy = (95+22)/(95+22+13+23) = 0.76470 . So, Accuracy = 76.47 %.



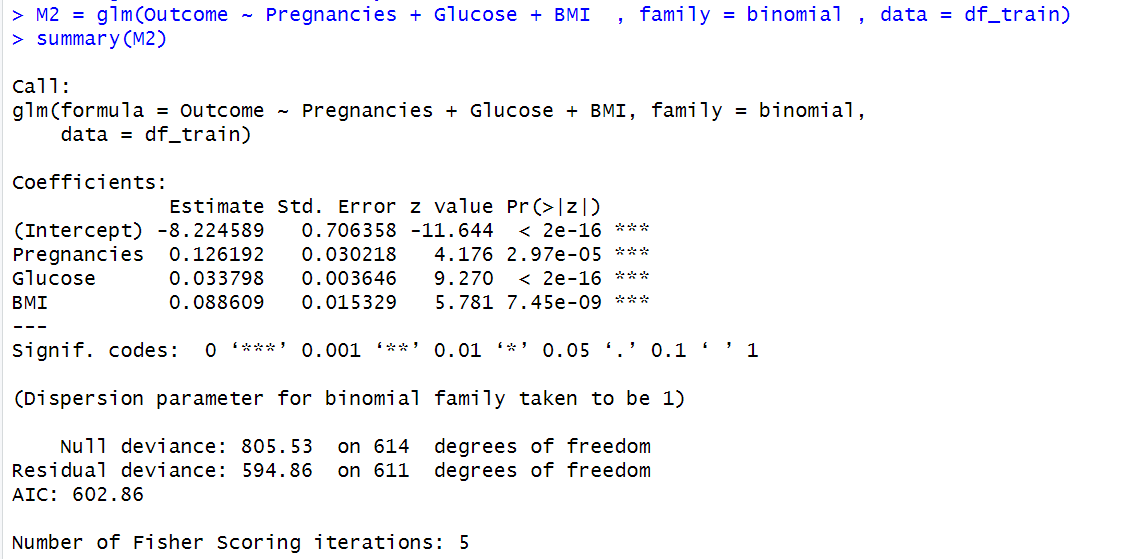


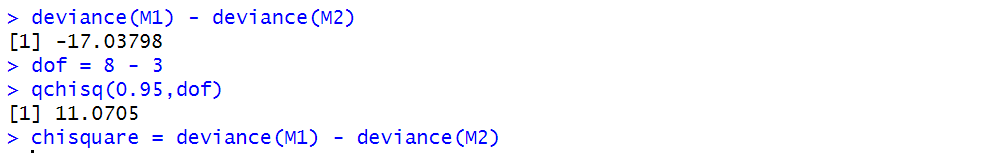
​



So, we get the F1-score as 0.55.

1. **Let’s call the model fitted in problem 3 M1. Now choose predictors “Pregnancies”, “Glucose” and “BMI” and fit a model (M2). Compare the deviances among these two models and perform hypothesis test to show whether M1 is significantly more informative than M2.**





**Interpretation:**

From the above picture we can say that , deviance is less than critical value(-17.03 < 11.07) . So , fail to reject the null hypothesis and hypothesis testing. So, We can say that Our simple model (M2) with three columns **“Pregnancies”, “Glucose” and “BMI”** Is better than Full model (M1) using hypothesis testing. Or say M2 is significantly more informative than M1.

Description of the study:

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988, November). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care* (p. 261). American Medical Informatics Association.