Statistical Learning Lab

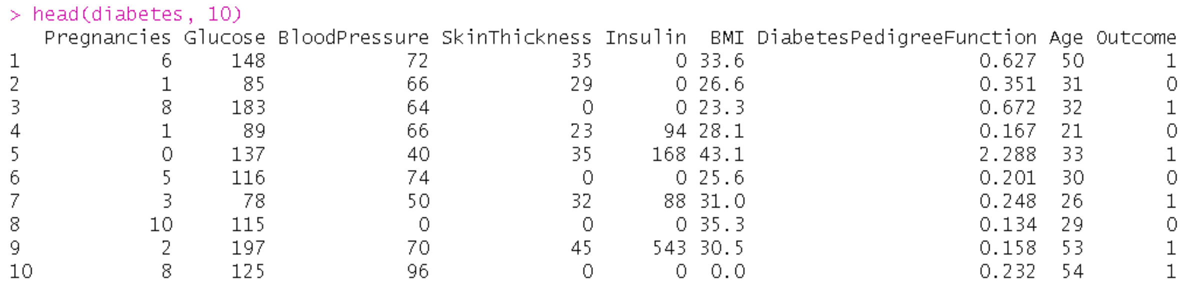
Assignment - 2

Logistic Regression Assignment

Ben Abraham Biju
22IM10048

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

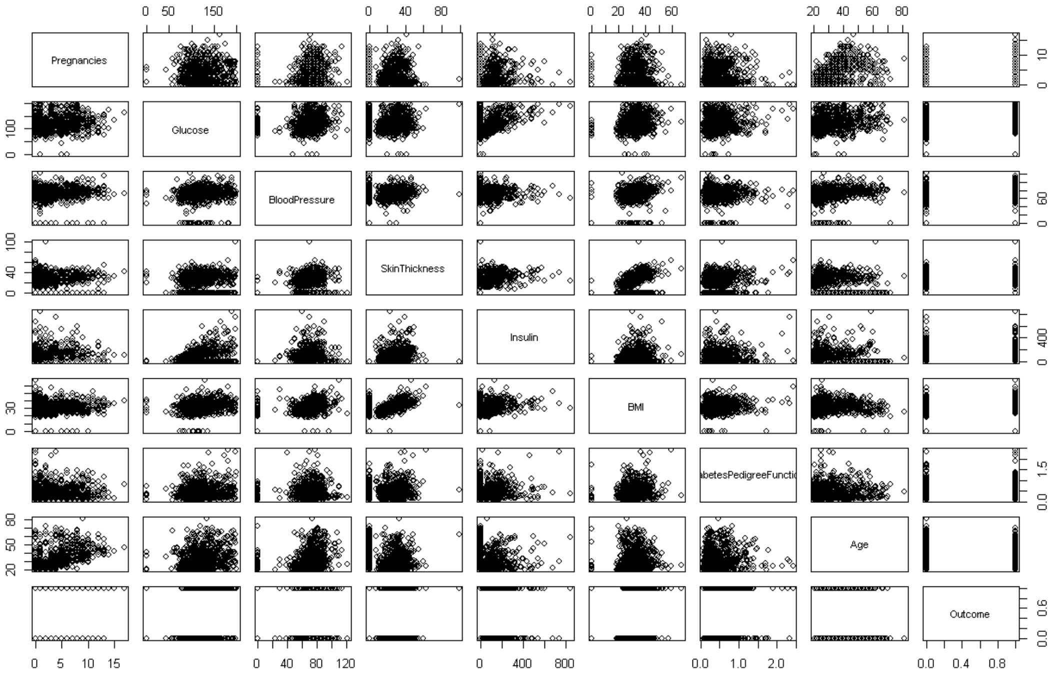
The dataset was loaded using the read.csv() function to the variable *diabetes.*



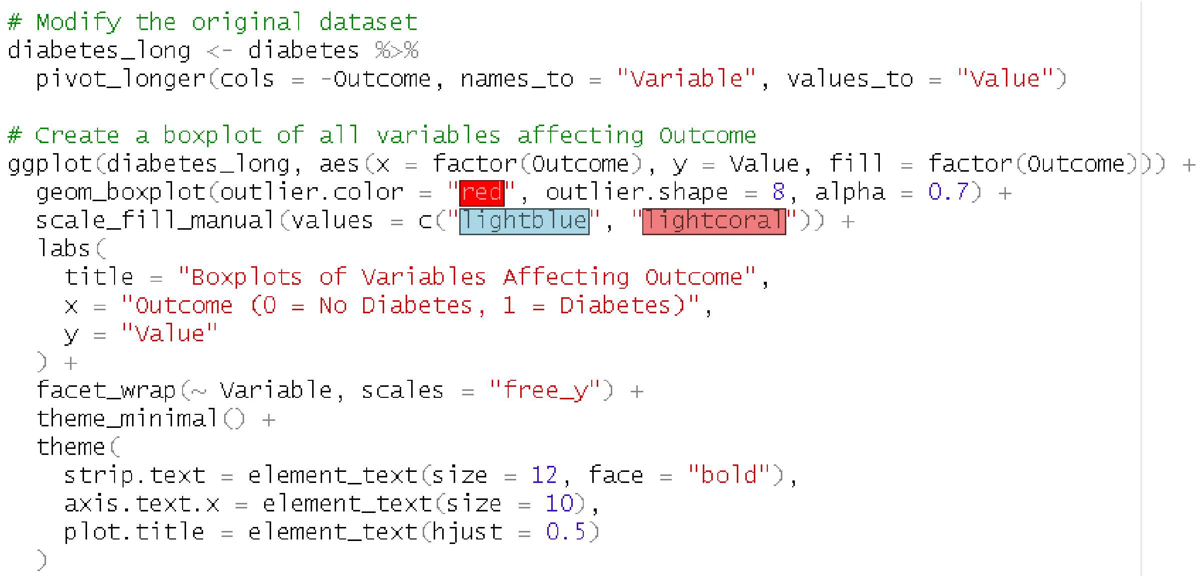
```
> head(diabetes, 10)
   Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI DiabetesPedigreeFunction Age Outcome
1            6     148            72            35       0 33.6                    0.627  50       1
2            1      85            66            29       0 26.6                    0.351  31       0
3            8     183            64             0       0 23.3                    0.672  32       1
4            1      89            66            23      94 28.1                    0.167  21       0
5            0     137            40            35     168 43.1                    2.288  33       1
6            5     116            74             0       0 25.6                    0.201  30       0
7            3      78            50            32      88 31.0                    0.248  26       1
8           10     115             0             0       0 35.3                    0.134  29       0
9            2     197            70            45     543 30.5                    0.158  53       1
10           8     125            96             0       0  0.0                    0.232  54       1
```

2. 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.

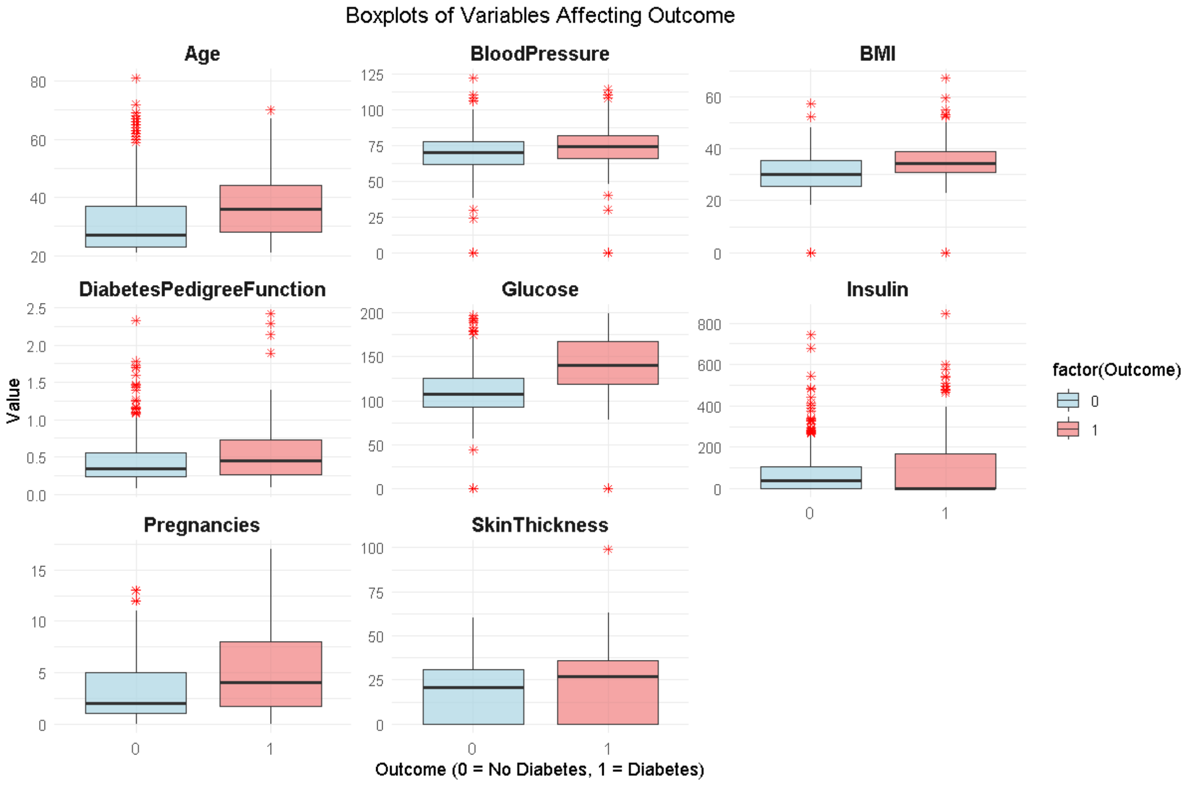
> plot(diabetes)



The relations of different variables with *Outcome* can be seen through boxplots.

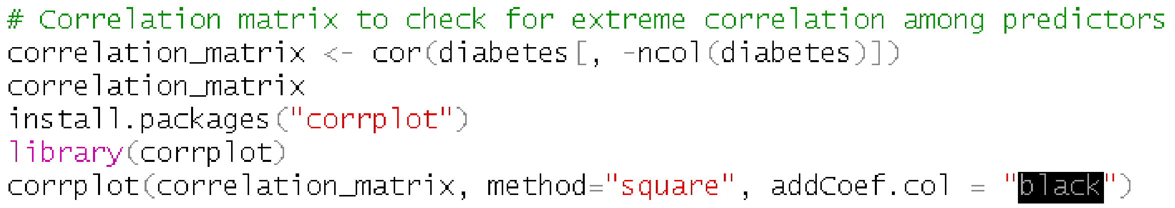


```
# Modify the original dataset
diabetes_long <- diabetes %>%
  pivot_longer(cols = -Outcome, names_to = "Variable", values_to = "Value")

# Create a boxplot of all variables affecting Outcome
ggplot(diabetes_long, aes(x = factor(Outcome), y = Value, fill = factor(Outcome))) +
  geom_boxplot(outlier.color = "red", outlier.shape = 8, alpha = 0.7) +
  scale_fill_manual(values = c("lightblue", "lightcoral")) +
  labs(
    title = "Boxplots of Variables Affecting Outcome",
    x = "Outcome (0 = No Diabetes, 1 = Diabetes)",
    y = "Value"
  ) +
  facet_wrap(~ Variable, scales = "free_y") +
  theme_minimal() +
  theme(
    strip.text = element_text(size = 12, face = "bold"),
    axis.text.x = element_text(size = 10),
    plot.title = element_text(hjust = 0.5)
  )
```

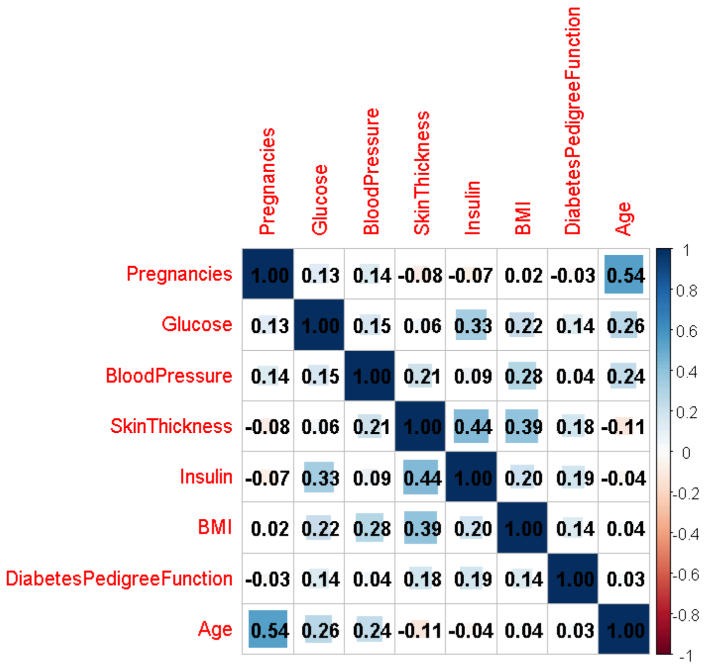


This boxplot visualizes the distribution of key variables based on diabetes outcome (0 = No Diabetes, 1 = Diabetes). Individuals with diabetes (red) tend to have higher Glucose, BMI, Age, DiabetesPedigreeFunction, and Pregnancies compared to those without diabetes (blue). BloodPressure and SkinThickness show slight increases, while Insulin has a wider spread among diabetic individuals. Outliers are present in multiple variables, showing some extreme values in the dataset.

To check for extreme correlations among predictors, the correlation matrix can be plotted.



```
# Correlation matrix to check for extreme correlation among predictors
correlation_matrix <- cor(diabetes[, -ncol(diabetes)])
correlation_matrix
install.packages("corrplot")
library(corrplot)
corrplot(correlation_matrix, method="square", addCoef.col = "black")
```

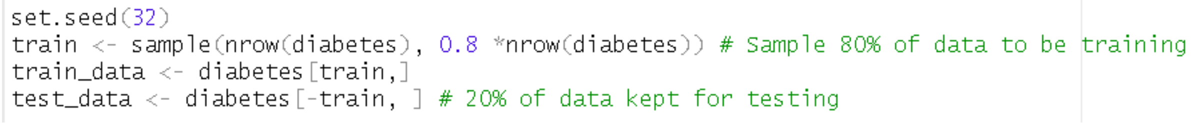


	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
Pregnancies	1.00	0.13	0.14	-0.08	-0.07	0.02	-0.03	0.54
Glucose	0.13	1.00	0.15	0.06	0.33	0.22	0.14	0.26
BloodPressure	0.14	0.15	1.00	0.21	0.09	0.28	0.04	0.24
SkinThickness	-0.08	0.06	0.21	1.00	0.44	0.39	0.18	-0.11
Insulin	-0.07	0.33	0.09	0.44	1.00	0.20	0.19	-0.04
BMI	0.02	0.22	0.28	0.39	0.20	1.00	0.14	0.04
DiabetesPedigreeFunction	-0.03	0.14	0.04	0.18	0.19	0.14	1.00	0.03
Age	0.54	0.26	0.24	-0.11	-0.04	0.04	0.03	1.00

No severe correlation (>0.7) can be found among the predictors.

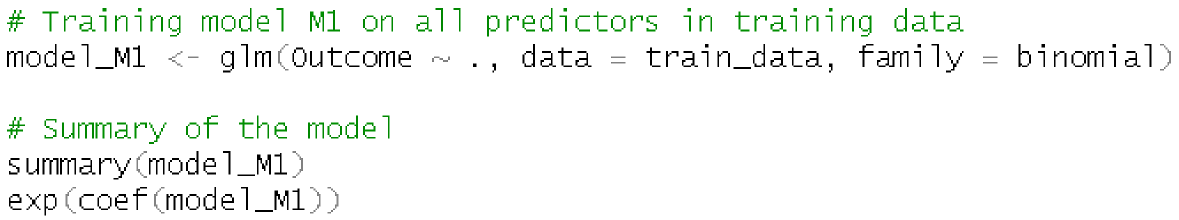
3. **Randomly sample 80% of the data as training data and rest as test data. Fit a Logistic Regression model with all the predictors on training data. From the summary which factors seem to be significant? Explain how the predictors impact the log-odds ratio of diagnosed with diabetes (Outcome)**

Code to split the dataset into training and testing datasets:



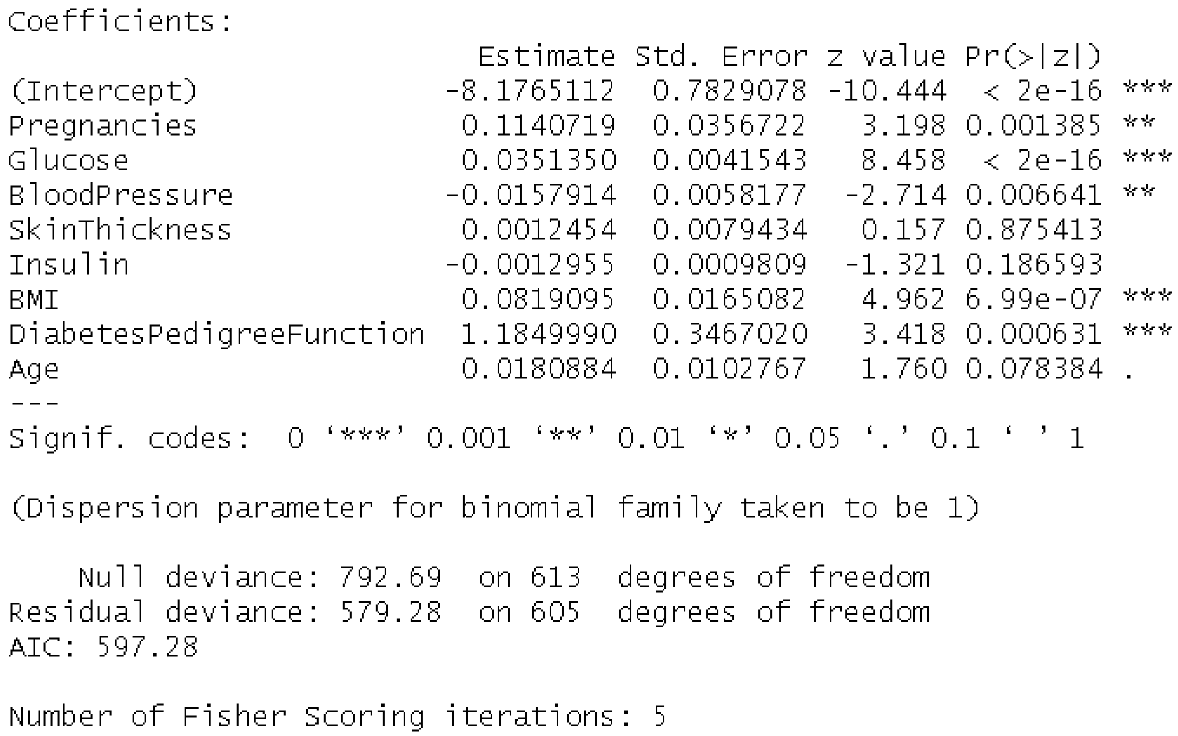
```
set.seed(32)
train <- sample(nrow(diabetes), 0.8 *nrow(diabetes)) # Sample 80% of data to be training
train_data <- diabetes[train,]
test_data <- diabetes[-train, ] # 20% of data kept for testing
```

A logistic regression model, M1 was trained on all the predictor variables.
Code:



```
# Training model M1 on all predictors in training data
model_M1 <- glm(Outcome ~ ., data = train_data, family = binomial)

# Summary of the model
summary(model_M1)
exp(coef(model_M1))
```

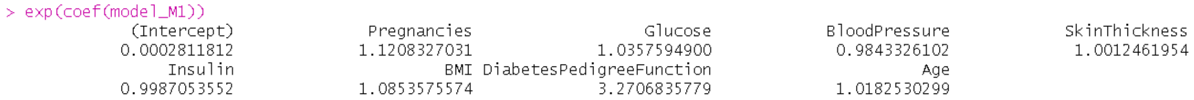
Output:



```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)    
(Intercept)             -8.1765112  0.7829078 -10.444  < 2e-16 ***
Pregnancies              0.1140719  0.0356722   3.198 0.001385 ** 
Glucose                  0.0351350  0.0041543   8.458  < 2e-16 ***
BloodPressure           -0.0157914  0.0058177  -2.714 0.006641 ** 
SkinThickness            0.0012454  0.0079434   0.157 0.875413    
Insulin                 -0.0012955  0.0009809  -1.321 0.186593    
BMI                      0.0819095  0.0165082   4.962 6.99e-07 ***
DiabetesPedigreeFunction 1.1849990  0.3467020   3.418 0.000631 ***
Age                      0.0180884  0.0102767   1.760 0.078384 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 792.69  on 613  degrees of freedom
Residual deviance: 579.28  on 605  degrees of freedom
AIC: 597.28

Number of Fisher Scoring iterations: 5
```

From the summary, we can see that ***Pregnancies, Glucose, BloodPressure, BMI and DiabetesPedigreeFunction*** seem to be significant since their p-value is less than 0.05

The following were the coefficients of the trained model M1.



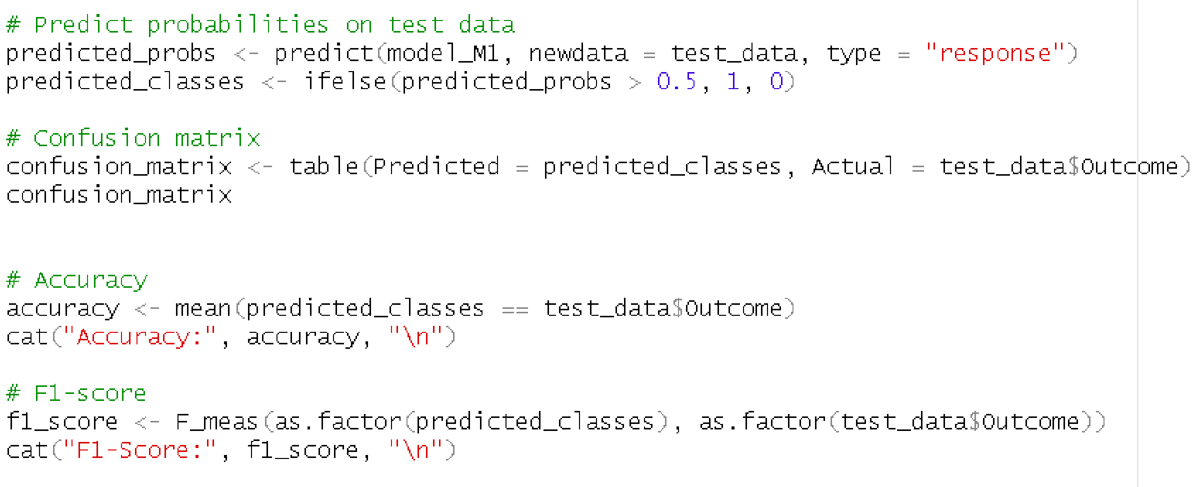
```
> exp(coef(model_M1))
             (Intercept)              Pregnancies                  Glucose            BloodPressure            SkinThickness 
            0.0002811812             1.1208327031             1.0357594900             0.9843326102             1.0012461954 
                 Insulin                      BMI DiabetesPedigreeFunction                      Age 
            0.9987053552             1.0853575574             3.2706835779             1.0182530299 
```

<u>Variable-specific interpretation (Log-odds ratio)</u>

1. Pregnancies (1.1208): Each additional pregnancy increases the odds of the outcome by 1.1208.
2. Glucose (1.0358): A one-unit increase in glucose increases the odds by 1.0358.
3. Blood Pressure (0.9843): A one-unit increase in blood pressure decreases the odds by 0.9843.
4. Skin Thickness (1.0012): A one-unit increase in skin thickness increases the odds by 1.0012.
5. Insulin (0.9987): A one-unit increase in insulin decreases the odds by 0.9987.
6. BMI (1.0854): A one-unit increase in BMI increases the odds by 1.0854.
7. Diabetes Pedigree Function (3.2707): A one-unit increase in this feature increases the odds by 3.2707.
8. Age (1.0183): A one-year increase in age increases the odds by 1.0183.

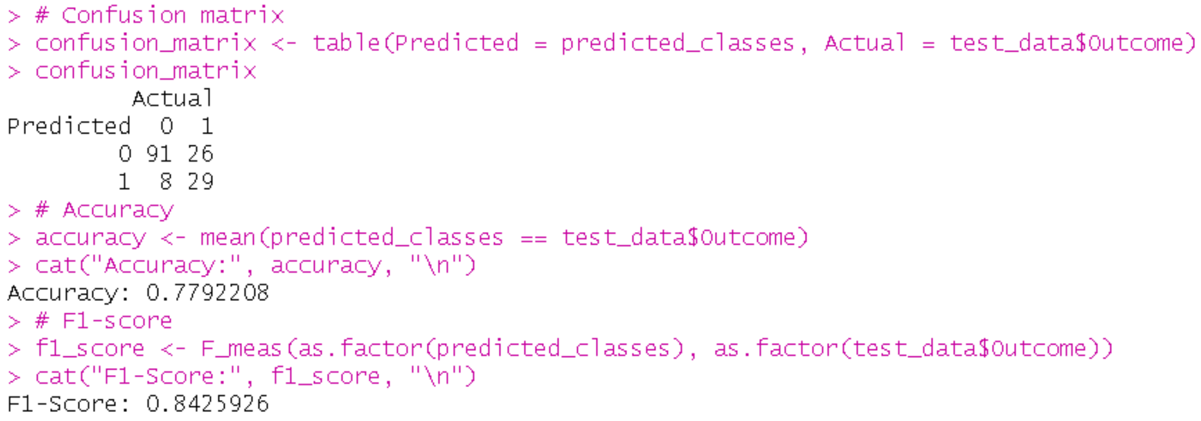
4. From the model fitted in problem 3, derive confusion matrix, accuracy, and F1-score on test data.

Code:



```
# Predict probabilities on test data
predicted_probs <- predict(model_M1, newdata = test_data, type = "response")
predicted_classes <- ifelse(predicted_probs > 0.5, 1, 0)

# Confusion matrix
confusion_matrix <- table(Predicted = predicted_classes, Actual = test_data$Outcome)
confusion_matrix


# Accuracy
accuracy <- mean(predicted_classes == test_data$Outcome)
cat("Accuracy:", accuracy, "\n")

# F1-score
f1_score <- F_meas(as.factor(predicted_classes), as.factor(test_data$Outcome))
cat("F1-Score:", f1_score, "\n")
```

Output:

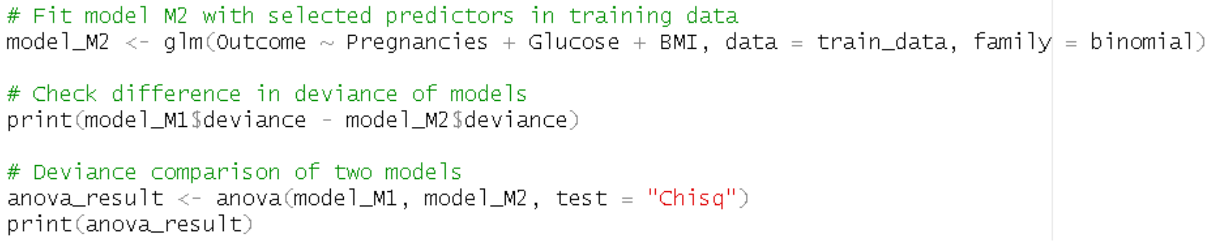


```
> # Confusion matrix
> confusion_matrix <- table(Predicted = predicted_classes, Actual = test_data$Outcome)
> confusion_matrix
         Actual
Predicted  0  1
        0 91 26
        1  8 29
> # Accuracy
> accuracy <- mean(predicted_classes == test_data$Outcome)
> cat("Accuracy:", accuracy, "\n")
Accuracy: 0.7792208
> # F1-score
> f1_score <- F_meas(as.factor(predicted_classes), as.factor(test_data$Outcome))
> cat("F1-Score:", f1_score, "\n")
F1-Score: 0.8425926
```

The F1 score is decently good, showing balance between precision and recall.

5. **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 tests to show whether M1 is significantly more informative than M2.**

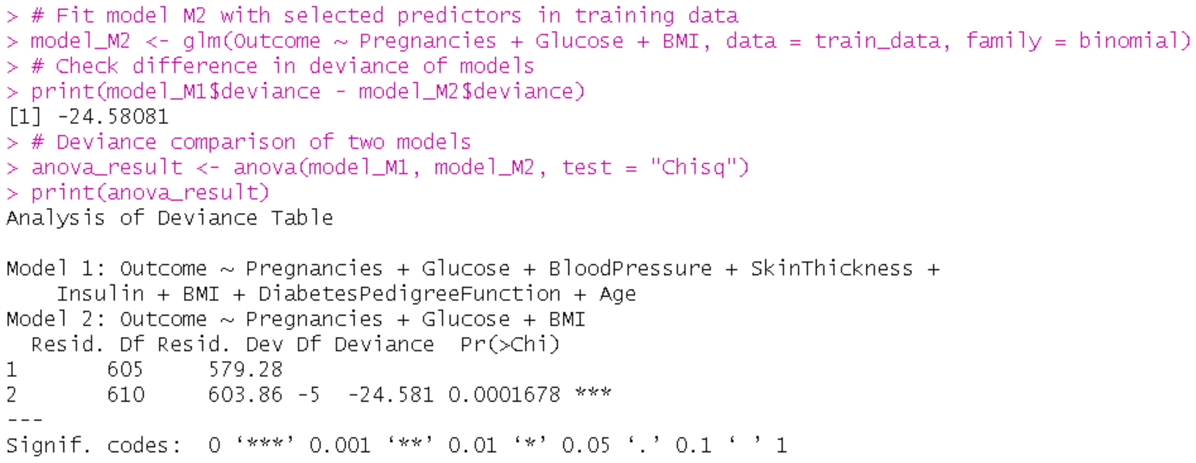
Code:



```
# Fit model M2 with selected predictors in training data
model_M2 <- glm(Outcome ~ Pregnancies + Glucose + BMI, data = train_data, family = binomial)

# Check difference in deviance of models
print(model_M1$deviance - model_M2$deviance)

# Deviance comparison of two models
anova_result <- anova(model_M1, model_M2, test = "Chisq")
print(anova_result)
```

The model M2 was trained on *"Pregnancies", "Glucose" and "BMI"*, and ANOVA test was done on both the models, M1 and M2.

Output:



```
> # Fit model M2 with selected predictors in training data
> model_M2 <- glm(Outcome ~ Pregnancies + Glucose + BMI, data = train_data, family = binomial)
> # Check difference in deviance of models
> print(model_M1$deviance - model_M2$deviance)
[1] -24.58081
> # Deviance comparison of two models
> anova_result <- anova(model_M1, model_M2, test = "Chisq")
> print(anova_result)
Analysis of Deviance Table

Model 1: Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness + 
    Insulin + BMI + DiabetesPedigreeFunction + Age
Model 2: Outcome ~ Pregnancies + Glucose + BMI
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)    
1       605     579.28                          
2       610     603.86 -5  -24.581 0.0001678 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Inference:

M1 has a lower deviance, indicating that it provides a better fit for the data. The p-value of **0.0001678** is highly significant (< 0.001), showing that the improvement in model fit from M2 to M1 is statistically significant.

Conclusion:

The full model (M1) is significantly more informative than model M2. The addition of predictors—BloodPressure, SkinThickness, Insulin, DiabetesPedigreeFunction, and Age—enhances the model’s ability to explain variations in the Outcome variable. Even though some predictors are not individually significant, the collective effect of all predictors leads to better predictive performance.