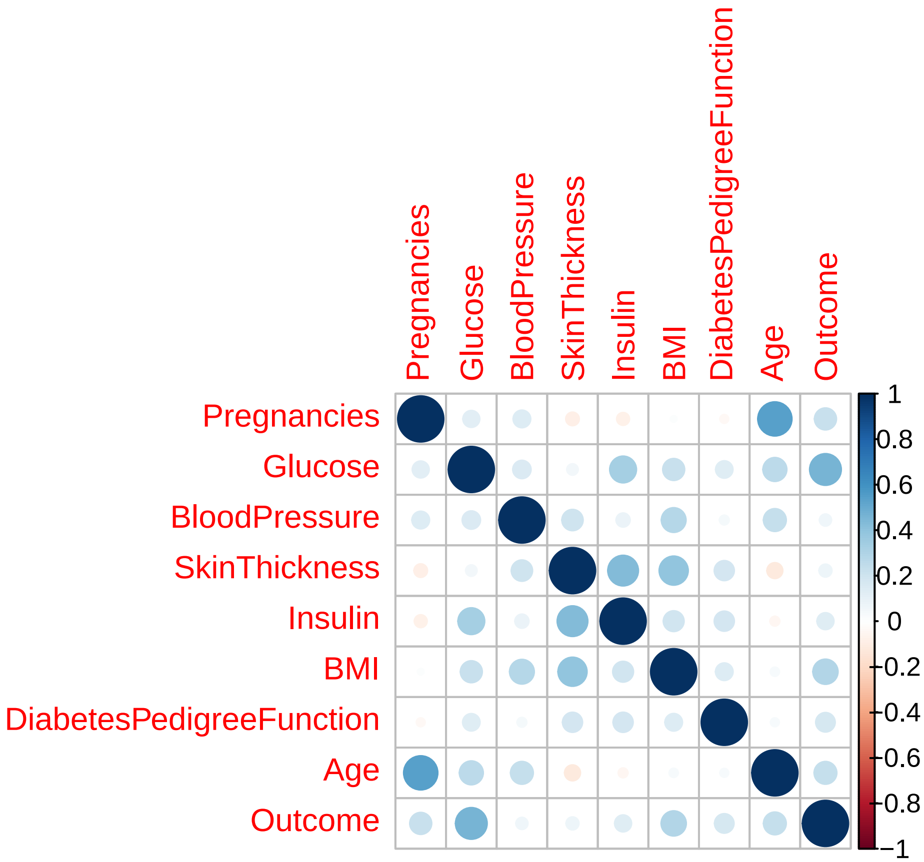


From the correlation matrix, we can see that some of the independent variables are moderately correlated, such as (Age, Pregnancy). This is the result of multicollinearity.

```
## [1] FALSE
```

```
##   Pregnancies        Glucose            BP               ST       
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00  
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00  
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00  
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54  
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00  
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00  
##     Insulin          BMI             DPF              Age        Outcome
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00   0:500  
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00   1:268  
##  Median : 30.5   Median :32.00   Median :0.3725   Median :29.00          
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24          
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00          
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00          
```

Unbalanced distribution, which means about 65% people in this dataset did not have diabetes.

Given the Y(outcome) variable is categorical, we would need to use the logistic regression model.

Using undersampling to reduce bias towards the majority.

```
## Training
```

```
##
##   0   1
## 215 215
```

```
## Testing
```

```
##
##  0  1
## 53 53
```

```
## [1] "train sample size:  430"
```

```
## [1] "test sample size:  106"
```

```
##
##   0   1
## 215 215
```

```
##
##  0  1
## 53 53
```

Generalized Linear Model

Logistic Regression

Using Logit:

```
##
## Call:
## glm(formula = Outcome ~ Pregnancies + Glucose + BP + Insulin +
##     BMI + DPF + Age, family = binomial, data = diabetes.training)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.95529  -0.77910  -0.00446   0.74787   2.71693
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.347946   0.952058  -8.768  < 2e-16 ***
## Pregnancies  0.106247   0.042935   2.475  0.01334 *
## Glucose      0.035361   0.004826   7.327 2.36e-13 ***
## BP          -0.014462   0.006700  -2.159  0.03088 *
## Insulin     -0.001951   0.001074  -1.817  0.06915 .
## BMI          0.090030   0.019238   4.680 2.87e-06 ***
## DPF          1.262500   0.401056   3.148  0.00164 **
## Age          0.031893   0.013066   2.441  0.01465 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 596.11  on 429  degrees of freedom
```

```
## Residual deviance: 424.23  on 422  degrees of freedom
## AIC: 440.23
##
## Number of Fisher Scoring iterations: 5
```

The StepAIC function was used to determine the model goodness of fit between the logit and probit model. This decided the outcome that the logit model is suitable for this specific task as it has a lower AIC compared to the probit model. Also the insignificant variables are the skin thickness and age.

The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable.

The maximum likelihood estimation can be expressed as:

$$\ln \frac{\pi}{1-\pi} = -8.653 + 0.113X_1 + 0.0438X_2 - 0.0110X_3 - 0.00224X_4 + 0.0938X_5 + 1.174X_6$$

```
##  (Intercept)  Pregnancies      Glucose           BP      Insulin          BMI
## 0.0002368826 1.1120965563 1.0359936362 0.9856419140 0.9980507767 1.0942070495
##          DPF          Age
## 3.5342470033 1.0324074648
```

Interpretation of step model:

- For every one unit increase in pregnancies, there is an increase change in $(1.12 - 1) * 100 = 12\%$ in odds ratio
- For every one unit increase in glucose, there is an increase change in $(1.04 - 1) * 100 = 4\%$ in odds ratio
- For every one unit increase in BP, there is decrease change of 1.1% in odds ratio
- For every one unit increase in Insulin, there is decrease change of 0.2% in odds ratio
- For every one unit increase in BMI, there is an increase change of 9.8% in odds ratio
- For every one unit increase in DPF, there is an increase change in 223% in odds ratio

Statistical Inference:

```
##                     2.5 %        97.5 %
## (Intercept) -10.304333456 -6.5643927482
## Pregnancies   0.023037125  0.1917972781
## Glucose       0.026289814  0.0452457449
## BP           -0.028127001 -0.0017158519
## Insulin      -0.004048190  0.0001865544
## BMI           0.053878350  0.1294103525
## DPF           0.489182126  2.0637607030
## Age           0.006631532  0.0580279643
```

Prediction:

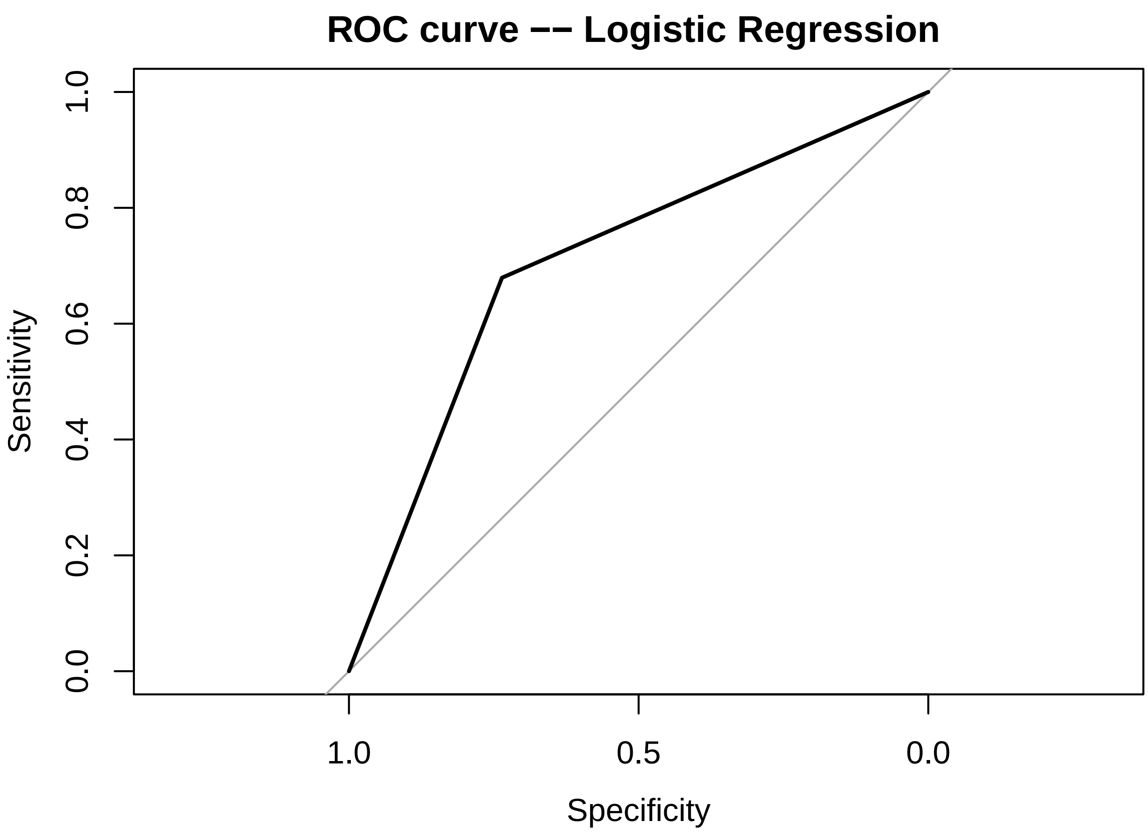
```
##           pred.classes
##            Pred. neg Pred. pos
##   Obs. neg        39        14
##   Obs. pos        17        36
```

Average Predicted Probability:

```
## [1] 0.6168049
```

Accuracy of the step-wise multiple Logistic Regression Model:

```
## [1] 0.7075472
```




```
##
## Call:
## roc.formula(formula = diabetes.testing$Outcome ~ pred.classes)
##
## Data: pred.classes in 53 controls (diabetes.testing$Outcome 0) < 53 cases (diabetes.testing$Outcome 
## Area under the curve: 0.7075
```

Given the plot and AUC, the value 0.7075 indicates that this is a good predictive model.