Predictive Modeling of Diabetes Onset in Pima Indian Women

Bin Han

2025-06-26

Project Introduction

In this project, we aim to identify key predictors for the onset of diabetes in Pima Indian women using logistic regression. The analysis is based on a medical dataset provided through the R course, which includes various health-related measurements. Through exploratory data analysis, statistical testing, and model selection techniques such as AIC, we determine a set of significant variables. We handle missing data using mean imputation and evaluate the predictive performance of the final model. Achieved a model accuracy of 77.4%, exceeding the 76.5% performance benchmark. Gained experience in data preprocessing, model evaluation, and the importance of separating training and testing data to avoid overfitting.

Running Code

1. Load the data and familiarize yourself with the dataset by performing EDA and reading about the variables on Canvas. What predictors do you think will be relevant for the analysis? How is diabetes diagnosed and when is it measured with respect to the other variables in the dataset? **Glucose, BloodPressure, Pregnancies, might be relevent for the analysis.**

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4
```

```
-- Conflicts --------------------------------------- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
getwd()
```

```
[1] "/home/binhan"
```

```
diabetes_data<-read_csv("a2_diabetes.csv")
```

```
Rows: 500 Columns: 9
-- Column specification --------------------------------------------------------
Delimiter: ","
dbl (9): Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, D...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(diabetes_data)
```

```
# A tibble: 6 x 9
  Pregnancies Glucose BloodPressure SkinThickness Insulin   BMI
        <dbl>   <dbl>         <dbl>         <dbl>   <dbl> <dbl>
1           0     138            60            35     167  34.6
2           8      74            70            40      49  35.3
3           5     143            78            NA      NA  45
4           3      87            60            18      NA  21.8
5           8      85            55            20      NA  24.4
6           5      78            48            NA      NA  33.7
# i 3 more variables: DiabetesPedigreeFunction <dbl>, Age <dbl>, Outcome <dbl>
```

2. Recode the variable Outcome as a factor (but do not recode any other variables). Explore the data to find a set of "good" predictors for diabetes using descriptive statistics, plots, significance, and AIC. This should be done before imputation (next task).

From the boxplot of diabetes outcome VS other predictors, we see the Glucose, Insulin, BMI has better distinction between the diabetes and non-diabetes

In the logistic regression model, Glucose is the strongest independent predictor of diabetes

The stepwise-selected logistic regression model identifies Glucose, Pregnancies, and SkinThickness as significant predictors of diabetes. BMI contributes to the model but isn't statistically significant on its own.

The correlations you observed suggest some moderate multicollinearity, especially between Pregnancies & Age, BMI & SkinThickness and Glucose & Insulin.

```
library(corrplot)
```

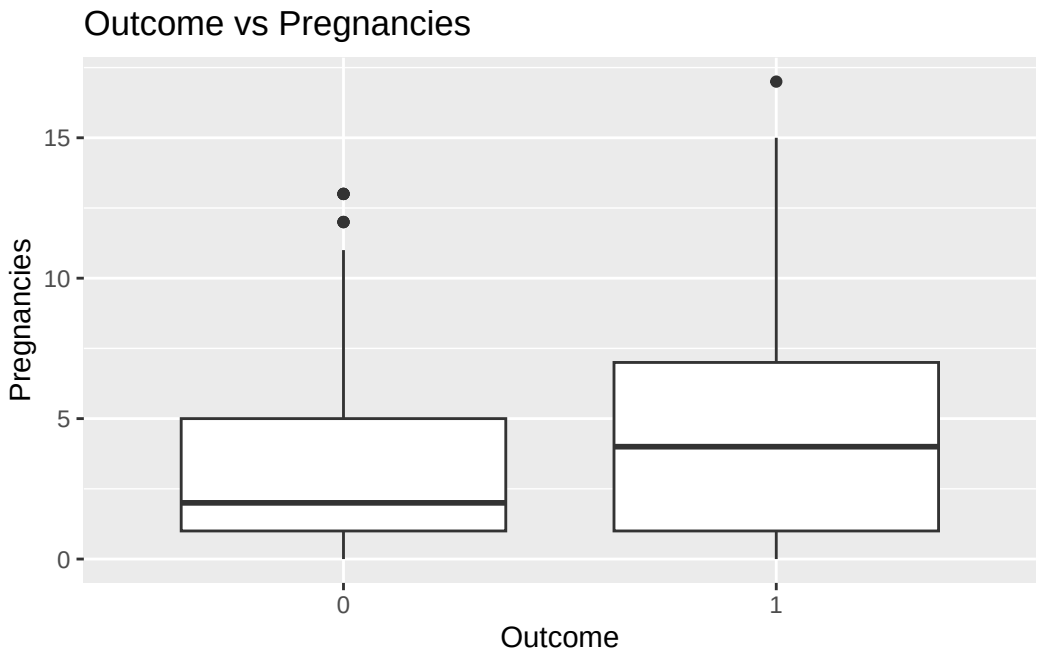
```
corrplot 0.95 loaded
```

```
diabetes_data<-diabetes_data %>%
  mutate(Outcome=as.factor(Outcome))
head(diabetes_data)
```

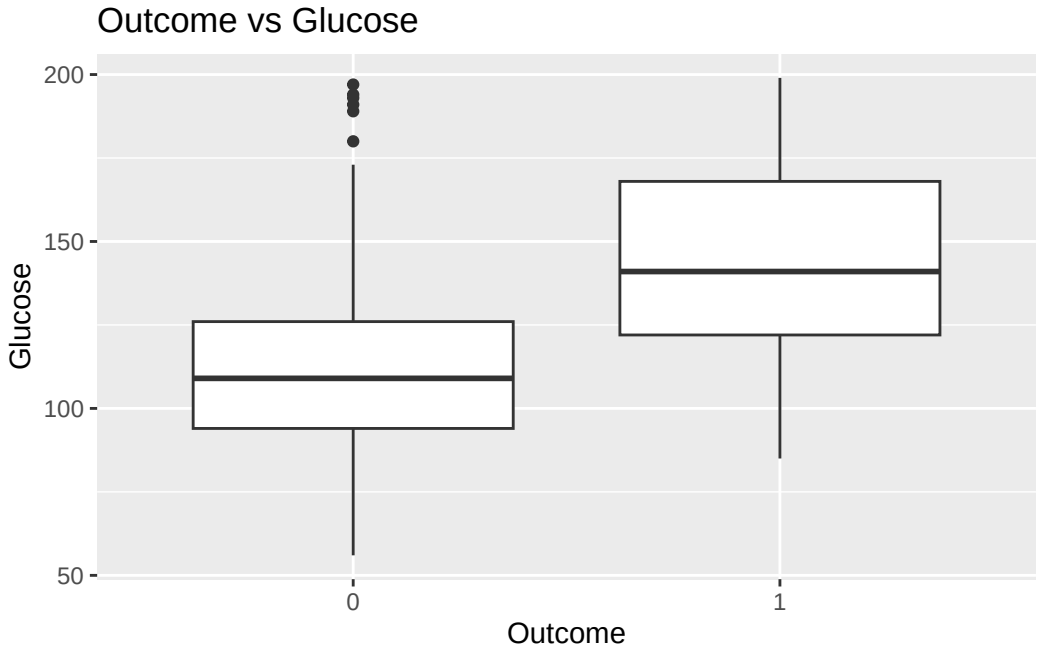
```
# A tibble: 6 x 9
  Pregnancies Glucose BloodPressure SkinThickness Insulin   BMI
        <dbl>   <dbl>         <dbl>         <dbl>   <dbl> <dbl>
1           0     138            60            35     167  34.6
2           8      74            70            40      49  35.3
3           5     143            78            NA      NA  45
4           3      87            60            18      NA  21.8
5           8      85            55            20      NA  24.4
6           5      78            48            NA      NA  33.7
# i 3 more variables: DiabetesPedigreeFunction <dbl>, Age <dbl>, Outcome <fct>
```

```
vars=c("Pregnancies","Glucose", "BloodPressure", "SkinThickness",
           "Insulin", "BMI", "DiabetesPedigreeFunction", "Age")

for (var in vars) {
  print(ggplot(diabetes_data, aes_string(x="Outcome", y=var))+
          geom_boxplot()+
          labs(title=paste("Outcome vs", var))
        )
}
```

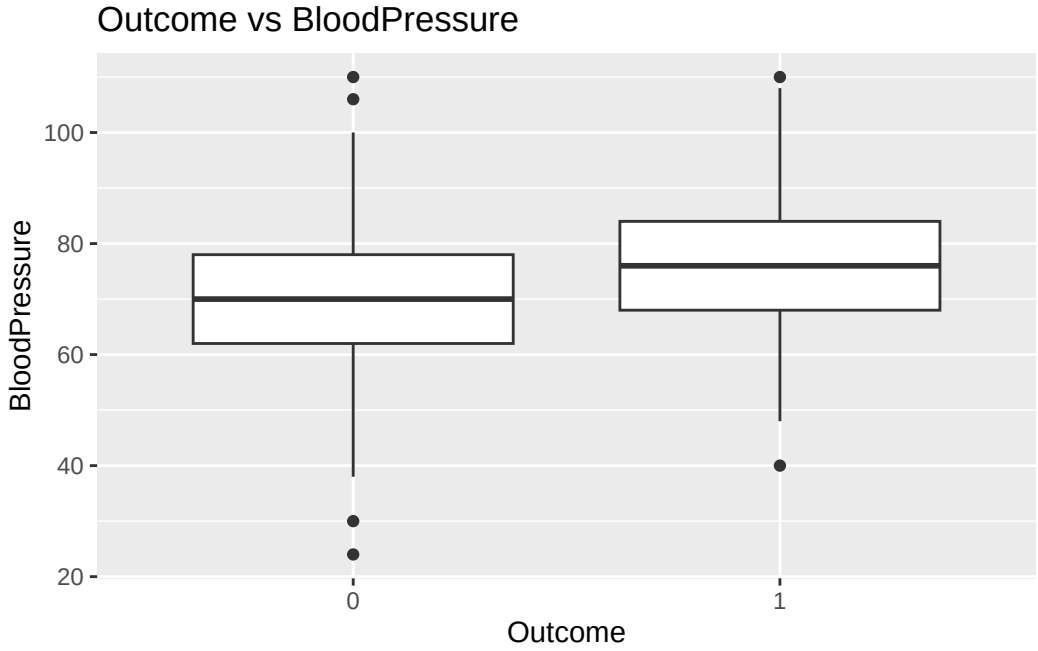
```
Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
i Please use tidy evaluation idioms with `aes()`.
i See also `vignette("ggplot2-in-packages")` for more information.
```



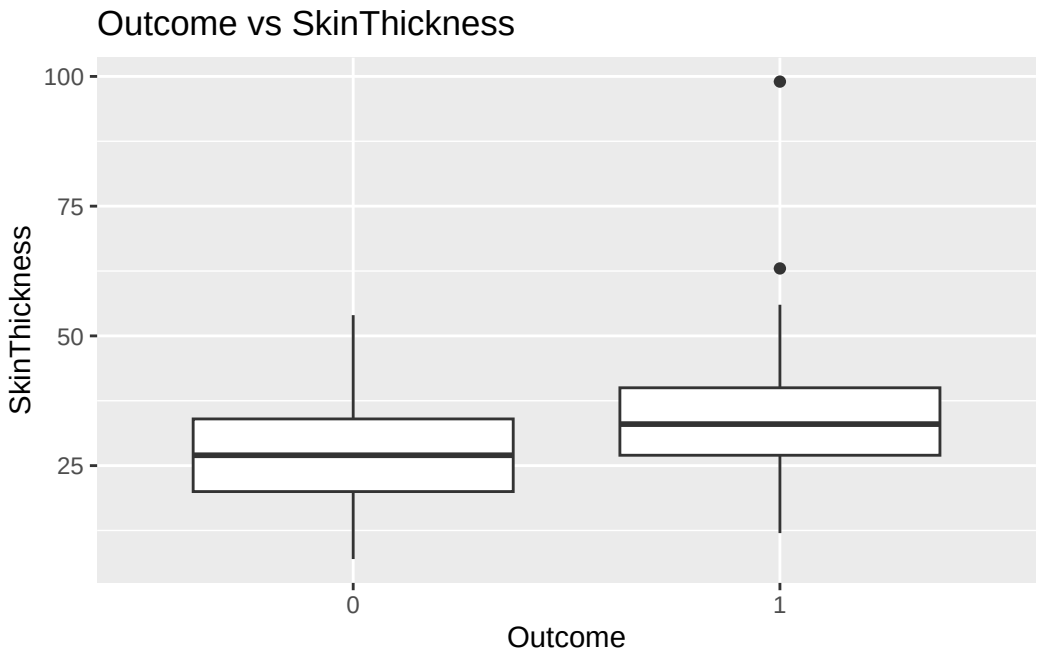

```
Warning: Removed 4 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



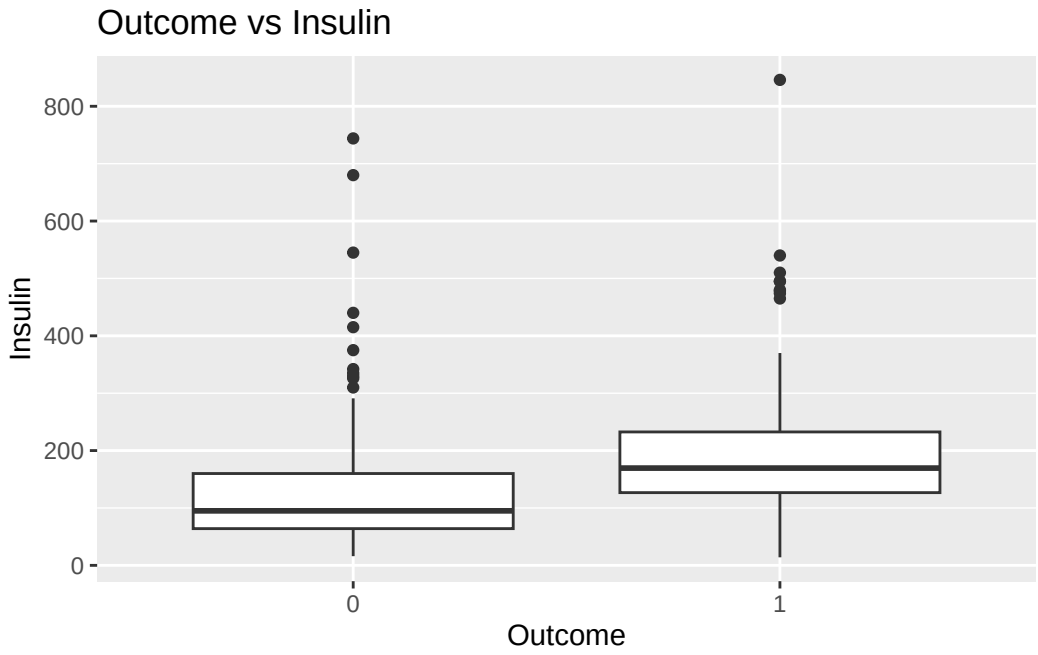
```
Warning: Removed 24 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



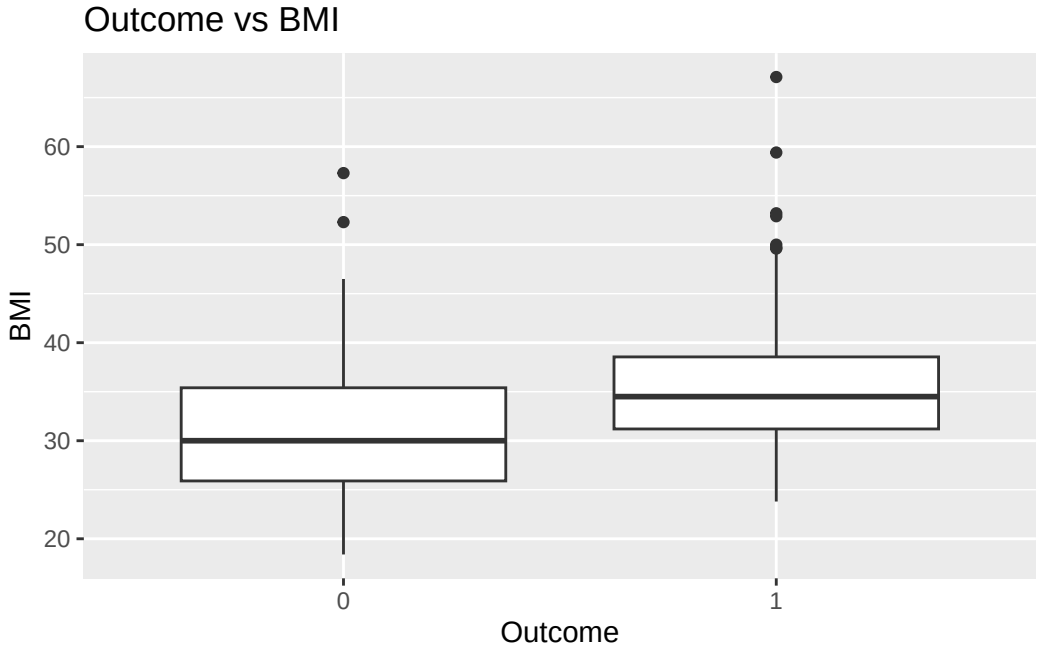

```
Warning: Removed 148 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```

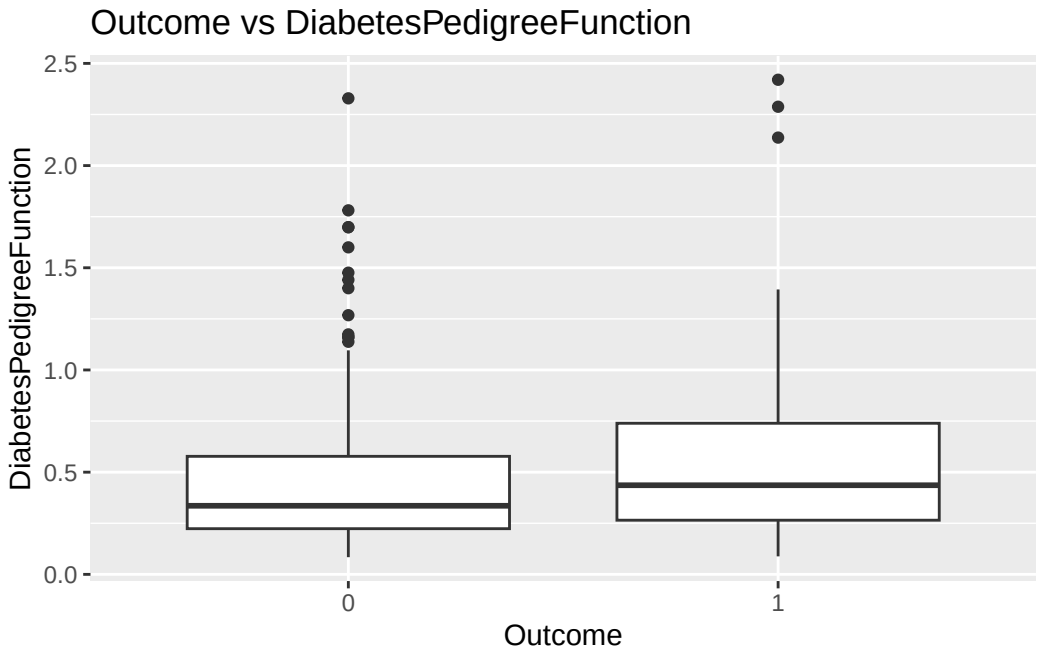


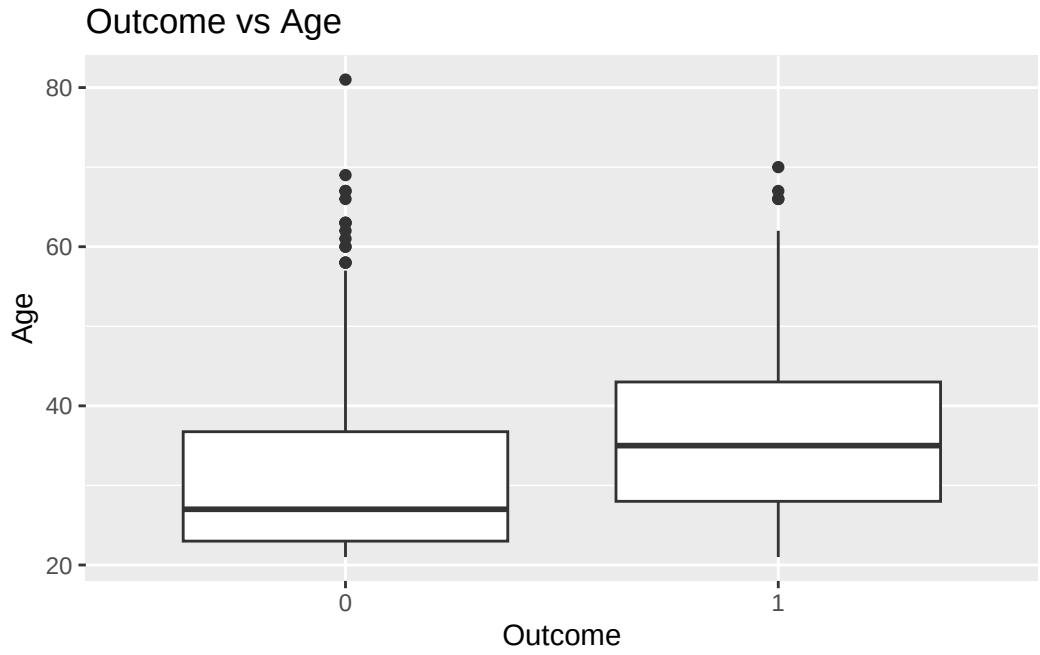

```
Warning: Removed 239 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```



```
Warning: Removed 7 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```








```
#from the boxplot, we see the different median in pregnancies,
#Glucose, BMI, DiabetesPedigreeFunction, Age, Pregnancies have more difference.
#Insulin, SkinThickness, BloodPressure have less difference
summary(glm(Outcome~ ., data=diabetes_data, family="binomial"))
```

```
Call:
glm(formula = Outcome ~ ., family = "binomial", data = diabetes_data)

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.012e+01  1.474e+00  -6.865 6.67e-12 ***
Pregnancies               9.778e-02  6.753e-02   1.448    0.148
Glucose                   4.308e-02  7.345e-03   5.866 4.46e-09 ***
BloodPressure             2.724e-03  1.482e-02   0.184    0.854
SkinThickness             3.683e-02  2.138e-02   1.722    0.085 .
Insulin                  -9.287e-04  1.549e-03  -0.600    0.549
BMI                       4.694e-02  3.282e-02   1.430    0.153
DiabetesPedigreeFunction  5.361e-01  4.820e-01   1.112    0.266
Age                       1.300e-02  2.023e-02   0.642    0.521
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 333.63  on 260  degrees of freedom
Residual deviance: 225.82  on 252  degrees of freedom
  (239 observations deleted due to missingness)
AIC: 243.82

Number of Fisher Scoring iterations: 5
```

```r
# only Glucose has p value< 0.05
clean_data<-na.omit(diabetes_data)
full_model<-glm(Outcome~ ., data=clean_data, family="binomial")
step_model<-step(full_model, direction="both")
```

```
Start:  AIC=243.82
Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness +
    Insulin + BMI + DiabetesPedigreeFunction + Age

                          Df Deviance    AIC
- BloodPressure            1   225.85 241.85
- Insulin                  1   226.18 242.18
- Age                      1   226.24 242.24
- DiabetesPedigreeFunction 1   227.09 243.09
<none>                         225.82 243.82
- BMI                      1   227.92 243.92
- Pregnancies              1   227.95 243.95
- SkinThickness            1   228.82 244.82
- Glucose                  1   269.34 285.34

Step:  AIC=241.86
Outcome ~ Pregnancies + Glucose + SkinThickness + Insulin + BMI +
    DiabetesPedigreeFunction + Age

                          Df Deviance    AIC
- Insulin                  1   226.22 240.22
- Age                      1   226.30 240.30
- DiabetesPedigreeFunction 1   227.10 241.10
<none>                         225.85 241.85
- Pregnancies              1   228.07 242.07
- BMI                      1   228.44 242.44
- SkinThickness            1   228.85 242.85
```

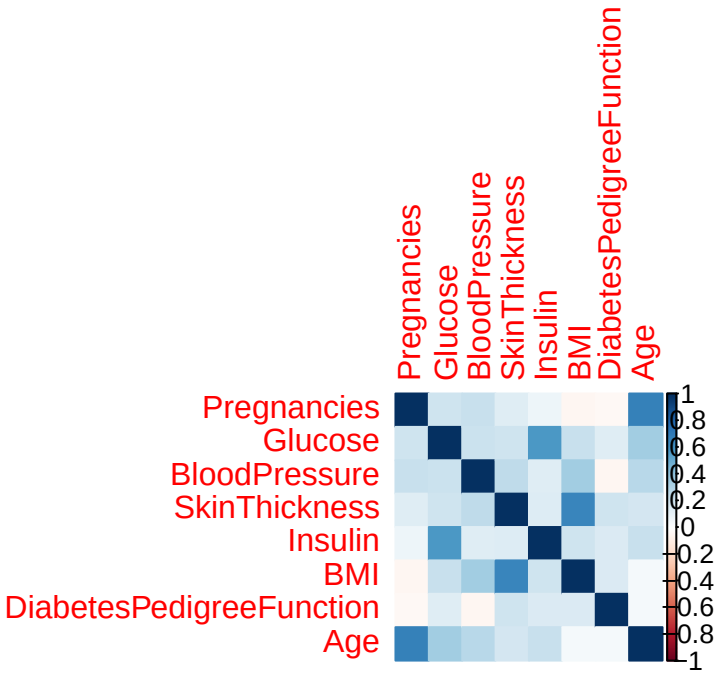
```
+ BloodPressure             1   225.82 243.82
- Glucose                   1   269.76 283.76

Step:  AIC=240.22
Outcome ~ Pregnancies + Glucose + SkinThickness + BMI + DiabetesPedigreeFunction +
    Age

                           Df Deviance    AIC
- Age                       1   226.61 238.61
- DiabetesPedigreeFunction  1   227.38 239.38
<none>                          226.22 240.22
- BMI                       1   228.55 240.55
- Pregnancies               1   228.60 240.60
- SkinThickness             1   229.38 241.38
+ Insulin                   1   225.85 241.85
+ BloodPressure             1   226.18 242.18
- Glucose                   1   280.44 292.44

Step:  AIC=238.61
Outcome ~ Pregnancies + Glucose + SkinThickness + BMI + DiabetesPedigreeFunction

                           Df Deviance    AIC
- DiabetesPedigreeFunction  1   227.91 237.91
<none>                          226.61 238.61
- BMI                       1   228.85 238.85
- SkinThickness             1   230.10 240.10
+ Age                       1   226.22 240.22
+ Insulin                   1   226.30 240.30
+ BloodPressure             1   226.53 240.53
- Pregnancies               1   233.41 243.41
- Glucose                   1   287.27 297.27

Step:  AIC=237.91
Outcome ~ Pregnancies + Glucose + SkinThickness + BMI

                           Df Deviance    AIC
<none>                          227.91 237.91
- BMI                       1   230.12 238.12
+ DiabetesPedigreeFunction  1   226.61 238.61
+ Age                       1   227.38 239.38
+ Insulin                   1   227.68 239.68
+ BloodPressure             1   227.89 239.89
- SkinThickness             1   231.97 239.97
```

```
- Pregnancies                1   234.24 242.24
- Glucose                    1   290.12 298.12
```

```
summary(step_model)
```

```

Call:
glm(formula = Outcome ~ Pregnancies + Glucose + SkinThickness +
    BMI, family = "binomial", data = clean_data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -9.412623   1.243684  -7.568 3.78e-14 ***
Pregnancies    0.123504   0.049891   2.475   0.0133 *
Glucose        0.042212   0.006265   6.737 1.61e-11 ***
SkinThickness  0.041834   0.020954   1.997   0.0459 *
BMI            0.044278   0.030155   1.468   0.1420
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 333.63  on 260  degrees of freedom
Residual deviance: 227.91  on 256  degrees of freedom
AIC: 237.91

Number of Fisher Scoring iterations: 5
```

```
# final step model, Glucose, Pregnancies, SkinThickness has stronger effect.
#BMI has less effect.
#library(corrplot)
multicollinearity_plot<-diabetes_data %>%
  select(-Outcome) %>%
  cor(, use="complete.obs") %>%
  corrplot(, method="color")
```



3. Impute missing values in all variables using mean imputation and store the imputed data in a variable called imputed_data. Make sure you do your data exploration before imputation, otherwise you will get biased results.

Each NA in a numeric column is replaced by the mean of that column.

```
imputed_data <-diabetes_data %>%
  mutate(across(where(is.numeric), ~ifelse(is.na(.), mean(., na.rm=TRUE), .)))
```

4. Find your final logistic regression model using imputed_data where all predictors have statistically significant relationships with the outcome at a 5% significance level. Save the final model in a variable called logistic_model.

Used stepwise selection (AIC-based) to find the most parsimonious model.

Verified statistical significance ($p < 0.05$) of selected variables.

```
full_model <-glm(Outcome ~ ., data = imputed_data, family = "binomial")
step_model<-step(full_model, direction="both")
```

```
Start:  AIC=481.71
Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness +
    Insulin + BMI + DiabetesPedigreeFunction + Age
```

```
                           Df Deviance    AIC
- BloodPressure             1   463.99 479.99
- SkinThickness             1   464.15 480.15
- Insulin                   1   464.31 480.31
- Age                       1   464.87 480.87
<none>                          463.71 481.71
- DiabetesPedigreeFunction  1   467.26 483.26
- Pregnancies               1   471.01 487.01
- BMI                       1   481.21 497.21
- Glucose                   1   543.78 559.78

Step:  AIC=479.99
Outcome ~ Pregnancies + Glucose + SkinThickness + Insulin + BMI +
    DiabetesPedigreeFunction + Age

                           Df Deviance    AIC
- SkinThickness             1   464.42 478.42
- Insulin                   1   464.55 478.55
- Age                       1   464.97 478.97
<none>                          463.99 479.99
- DiabetesPedigreeFunction  1   467.62 481.62
+ BloodPressure             1   463.71 481.71
- Pregnancies               1   471.13 485.13
- BMI                       1   481.67 495.67
- Glucose                   1   544.04 558.04

Step:  AIC=478.42
Outcome ~ Pregnancies + Glucose + Insulin + BMI + DiabetesPedigreeFunction +
    Age

                           Df Deviance    AIC
- Insulin                   1   465.02 477.02
- Age                       1   465.53 477.53
<none>                          464.42 478.42
+ SkinThickness             1   463.99 479.99
+ BloodPressure             1   464.15 480.15
- DiabetesPedigreeFunction  1   468.15 480.15
- Pregnancies               1   471.68 483.68
- BMI                       1   492.42 504.42
- Glucose                   1   544.66 556.66

Step:  AIC=477.02
Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction +
```

```
    Age

                            Df Deviance    AIC
- Age                        1   466.09 476.09
<none>                           465.02 477.02
+ Insulin                    1   464.42 478.42
+ SkinThickness              1   464.55 478.55
- DiabetesPedigreeFunction   1   468.58 478.58
+ BloodPressure              1   464.78 478.78
- Pregnancies                1   472.36 482.36
- BMI                        1   492.48 502.48
- Glucose                    1   553.99 563.99

Step:  AIC=476.09
Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction

                            Df Deviance    AIC
<none>                           466.09 476.09
+ Age                        1   465.02 477.02
+ SkinThickness              1   465.50 477.50
+ Insulin                    1   465.53 477.53
- DiabetesPedigreeFunction   1   469.98 477.98
+ BloodPressure              1   466.02 478.02
- Pregnancies                1   480.75 488.75
- BMI                        1   492.90 500.90
- Glucose                    1   565.41 573.41
```

```r
summary(step_model)#based on AIC, three predictors p < 0.05
```

```
Call:
glm(formula = Outcome ~ Pregnancies + Glucose + BMI + DiabetesPedigreeFunction,
    family = "binomial", data = imputed_data)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -9.271137   0.866971 -10.694  < 2e-16 ***
Pregnancies               0.126937   0.033744   3.762 0.000169 ***
Glucose                   0.038793   0.004507   8.607  < 2e-16 ***
BMI                       0.088815   0.018198   4.880 1.06e-06 ***
DiabetesPedigreeFunction  0.677158   0.347078   1.951 0.051054 .
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 651.08  on 499  degrees of freedom
Residual deviance: 466.09  on 495  degrees of freedom
AIC: 476.09

Number of Fisher Scoring iterations: 5
```

```r
logistic_model <- glm(Outcome~ Glucose + Pregnancies+BMI, data=imputed_data,
                      family="binomial")
```

5,Compute the accuracy of your model, again using imputed_data. If the accuracy is at least 76.5% when predicting the outcome using a classification threshold of 50%, save it in a numeric variable called accuracy. Otherwise go back to task 10. and find a new set of predictors. Reflect on why it may not always be a good idea to use the same data for fitting the regression model and evaluating its prediction accuracy.

The logistic regression model achieved a prediction accuracy of 77.4%, meaning that approximately 77.4% of the predicted classes matched the actual outcomes in the dataset. This suggests the model has a reasonably good performance in distinguishing between the two outcome classes based on the selected predictors.

```r
predicted_probs <- predict(logistic_model, newdata =imputed_data, type ="response")
predicted_classes<-ifelse(predicted_probs >=0.5, 1, 0)
accuracy <- mean(predicted_classes==imputed_data$Outcome)#0.774
```