DIABETES PREDICTION AND VISUALIZATION

R Programming Miniproject

Project Guide - Prof. Shruti Agrawal

Project Team -

1. Tanaya Desai – 18101B0059
2. Shraddha Babar – 18101B0065
3. Siddhi Suryavanshi – 18101B0071

Introduction

Diabetes occurs when your blood glucose or sugar is too high. One in six people with diabetes in the world is from India. With the development of standards of living, diabetes is gradually increasing in people. Diabetes is one of the major international health problems. So, we decided to predict diabetes using R.

In this project, a diabetes prediction system is implemented for predicting diabetes which comprises of some external features for diabetes alongside regular features like Insulin, Glucose, BMI, Age, etc. The prediction model, which produces highly accurate results, applied and compared various algorithms like Decision Tree, Linear Regression and Naive Bayes to determine Diabetes.

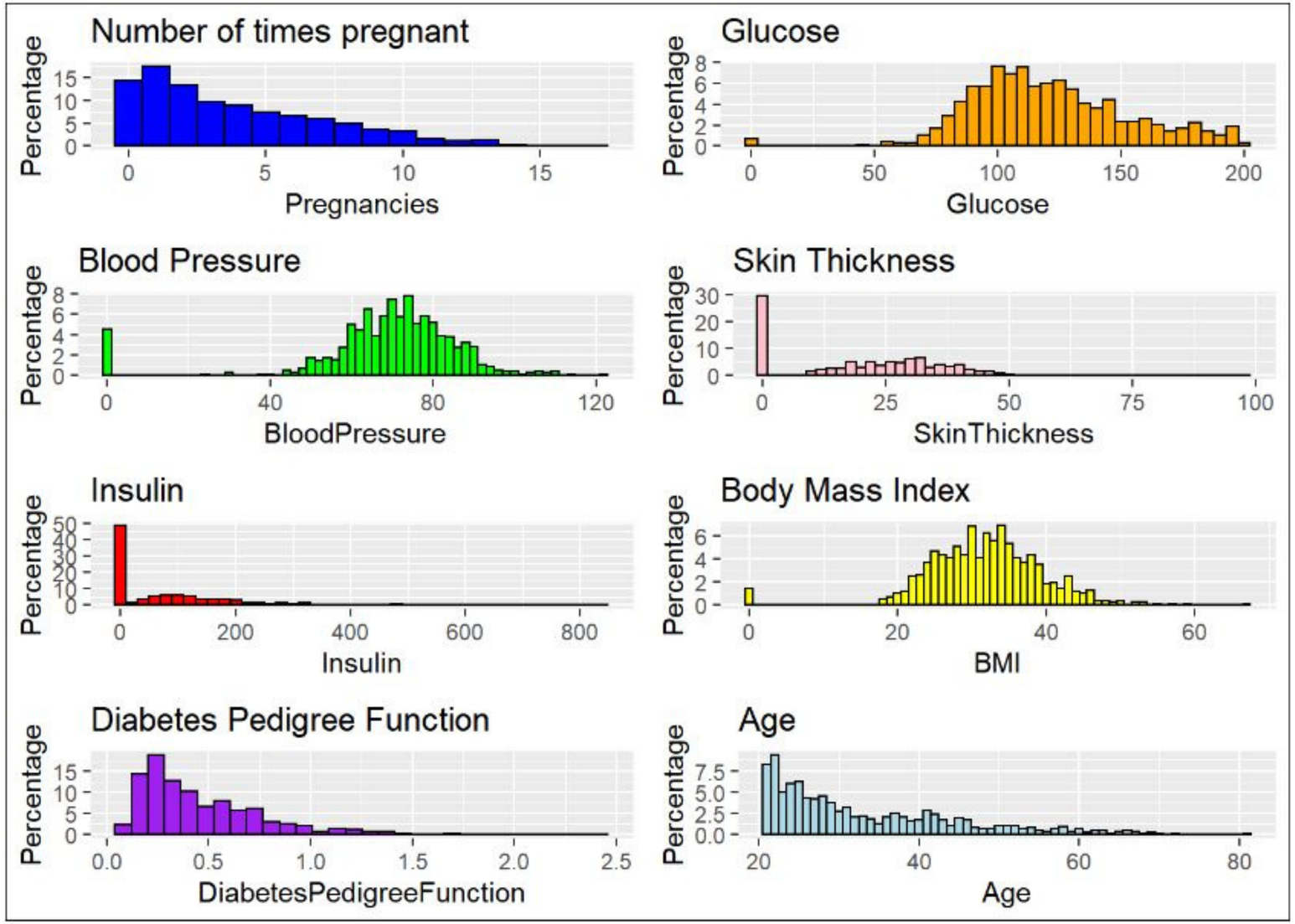
Dataset Used

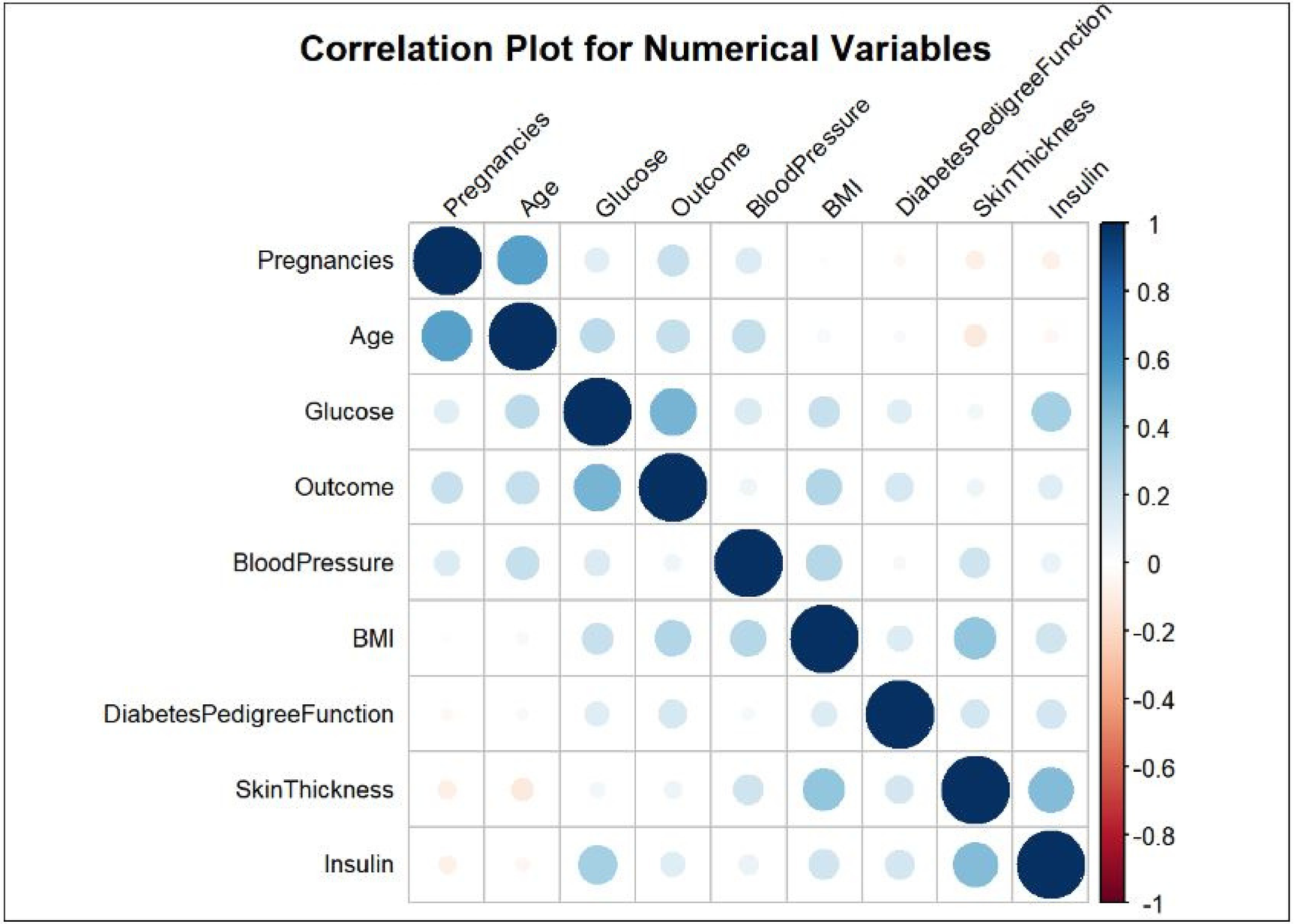
This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes. The data set is taken from UCI machine learning repository.

The data set consists of 9 attributes: number of times pregnant, plasma glucose concentration, diastolic blood pressure, triceps skin folds thickness, serum insulin, body mass index, pedigree type, age and class. Here, the class label is binary classification. It has two values

- Tested positive (1) which means diabetic
- Tested negative (0) which says nondiabetic

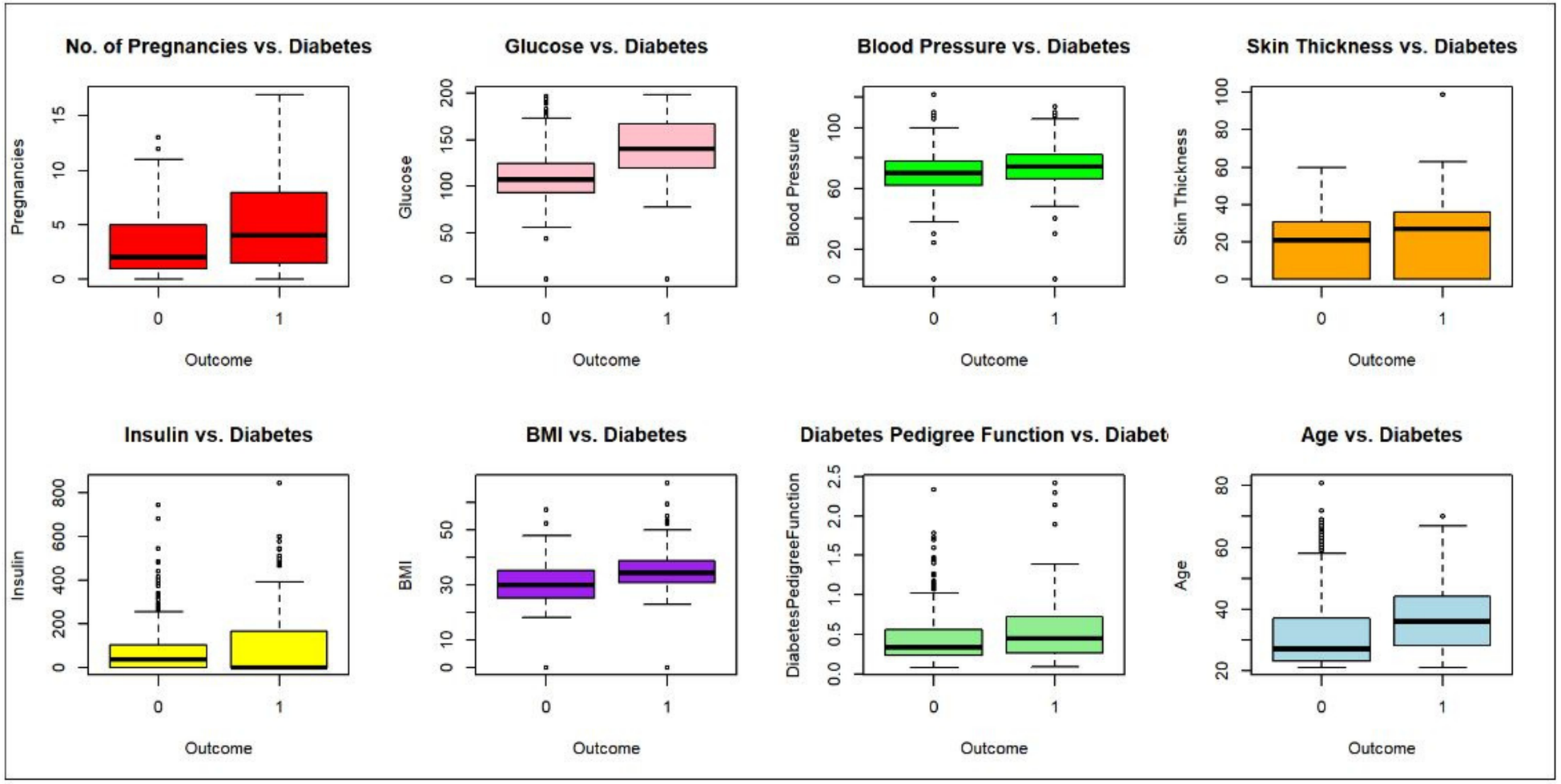
Histograms for Numeric Values





Correlation between Numeric Values

Correlation between Numeric Values and Outcomes



Conclusion

Linear Regression

```
confusionMatrix(conf_matrix_logi)
```

```
## Confusion Matrix and Statistics
##
##
## fitted.results   0   1
##              0 136  34
##              1  14  44
##
##                Accuracy : 0.7895
##                  95% CI : (0.7307, 0.8405)
##     No Information Rate : 0.6579
##     P-Value [Acc > NIR] : 9.506e-06
##
##                   Kappa : 0.5016
##
##  Mcnemar's Test P-Value : 0.006099
##
##             Sensitivity : 0.9067
##             Specificity : 0.5641
##          Pos Pred Value : 0.8000
##          Neg Pred Value : 0.7586
##              Prevalence : 0.6579
##          Detection Rate : 0.5965
##    Detection Prevalence : 0.7456
##       Balanced Accuracy : 0.7354
##
##        'Positive' Class : 0
##
```

Decision Tree

```
confusionMatrix(conf_matrix_dtree)
```

```
## Confusion Matrix and Statistics
##
##
## treePred   0   1
##        0 121  29
##        1  29  49
##
##                Accuracy : 0.7456
##                  95% CI : (0.6839, 0.8008)
##     No Information Rate : 0.6579
##     P-Value [Acc > NIR] : 0.002723
##
##                   Kappa : 0.4349
##
##  Mcnemar's Test P-Value : 1.000000
##
##             Sensitivity : 0.8067
##             Specificity : 0.6282
##          Pos Pred Value : 0.8067
##          Neg Pred Value : 0.6282
##              Prevalence : 0.6579
##          Detection Rate : 0.5307
##    Detection Prevalence : 0.6579
##       Balanced Accuracy : 0.7174
##
##        'Positive' Class : 0
##
```

Naive Bayes

```
confusionMatrix(conf_matrix_naive)
```

```
## Confusion Matrix and Statistics
##
##
## preds_naive   0   1
##           0 129  29
##           1  21  49
##
##                Accuracy : 0.7807
##                  95% CI : (0.7213, 0.8326)
##     No Information Rate : 0.6579
##     P-Value [Acc > NIR] : 3.562e-05
##
##                   Kappa : 0.5005
##
##  Mcnemar's Test P-Value : 0.3222
##
##             Sensitivity : 0.8600
##             Specificity : 0.6282
##          Pos Pred Value : 0.8165
##          Neg Pred Value : 0.7000
##              Prevalence : 0.6579
##          Detection Rate : 0.5658
##    Detection Prevalence : 0.6930
##       Balanced Accuracy : 0.7441
##
##        'Positive' Class : 0
##
```

Thank You!