Data Exploration:

The dataset has 9 variables as shown in the image below:

Dataset statistics

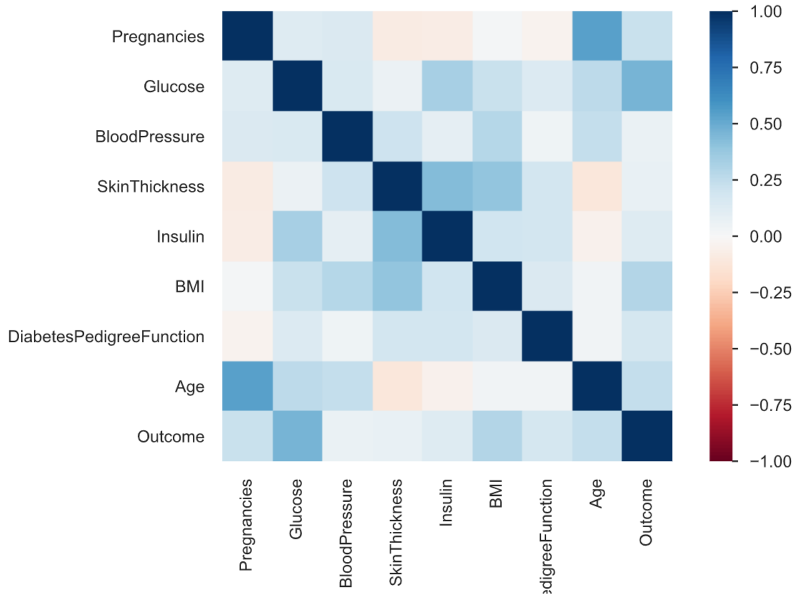
Number of variables	9
Number of observations	768
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	54.1 KiB
Average record size in memory	72.2 B

The 9 variables are:

1. Pregnancies: Number of times pregnant
2. Glucose: Plasma glucose concentration 2 hours in an oral glucose tolerance test
3. Blood Pressure: Diastolic blood pressure (mm Hg)
4. Skin Thickness: Triceps skin fold thickness (mm)
5. Insulin: 2-Hour serum insulin (mu U/ml)
6. BMI: Body mass index (weight in kg / (height in m) ^2)
7. Diabetes Pedigree Function: Diabetes pedigree function
8. Age: Age (years)
9. Outcome: 1 as diabetes detected & 0 as not detected

Correlation:

The image below shows correlation between variables:



It is noticeable that the variable 'Glucose' has the highest correlation with 'Outcome' which is our target variable.

Modeling:

To generate the decision tree, the 'tree' package was imported from the scikit learn library. We first need to split the dataset into training set and test set using the train_test_split function from train_test_split package.

The next step is to apply the DecisionTree function and store into a variable. We now fit training set and testing set into the dataframe using the fit function.

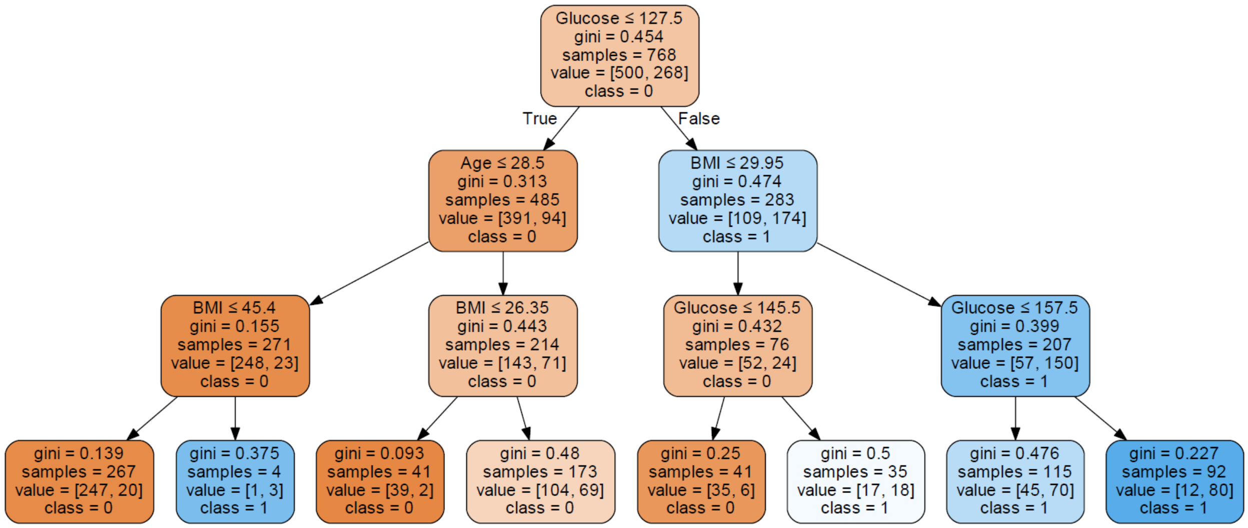
Use the graphviz library to plot the decision tree.

Initially the tree obtained had a depth of 13, which looked a bit more complex and the accuracy of the decision tree was 68%.

To increase the accuracy and make the model purer, we use max_depth in DecisionTree function. In this case the max_depth was set to 3.

Evaluation:

The image below shows the decision tree obtained for the diabetes dataset.



The first question considering the parent node; is the Glucose <= 127.5, if true we check whether the Age is <= 28.5 otherwise, if false we check if the BMI is <= 29.95.

Coming to the second phase we check if the Age is <= 28.5 if yes, we then check if the BMI is <=45.4, if no, we check if the BMI is <=26.35. On the other hand, we check if BMI is <=29.95, if true we check if Glucose is <=145.5, if BMI is >29.95, we check if Glucose is <=157.5.

We now check the third and final phase. There are four conditions: BMI <=45.4, BMI <=26.35, Glucose <=145.5, Glucose <= 157.5. If BMI<= 45.4 then no diabetes is detected else diabetes is detected. For BMI <= 26.35 diabetes will be not be detected, similarly for Glucose <=157.5, diabetes will be detected whether true or false.

The model can be used to determine the ranges in different variables to avoid diabetes. For example: For someone to stay in the non-diabetic range, the person should have Glucose <= 127.5, the age of the person should be less than or equal to 28.5 and the BMI should be less than equal to 45.4.

For someone who has BMI <= 29.95 but has Glucose <= 145.5, he/she can be detected of diabetes.

Accuracy and Error Matrix:

The accuracy of this model is:

Accuracy: 0.7204724409448819

	precision	recall	f1-score	support
0	0.77	0.80	0.78	162
1	0.62	0.59	0.60	92

The error matrix in this case is as shown: