Appendix - Tree Details

The following information is provided to help follow through the Tree Analysis

Tree description

In the tree details we are presented with the node number. Each node has a split indicating what predictor and criteria was used to split the observations to its children nodes.

Following that is the number of observations in that node, n. Our initial n is 430, so 430 observations were used in total and that is the n shown in the root node.

Loss shows how many observations from the n in that node was miss classified. Taking a look at the second node, we can see that the total observations was $n = 240$ and there was a loss of 71.

Next it shows whether the classification was non diabetic, 0, and diabetic, 1. This is referred to as `yval`. On the second node we see 0, indicating that the classification was non diabetic. With this we can assume that the majority will determine the classification. Meaning that since majority led to a classification of non diabetic in node 2, we can assume that the previous loss of 71 observations was a miss classification of 1, diabetic.

The last part of each line shows of the target probability based on what the `yval` was. Looking at node two, the probability of target being 0 is 70% and the miss classification is 30%. This is labeled as `(yprob)`.

Tree Predictions and Variable Importance of Original Tree

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 39 15
##          1 14 38
##
##                Accuracy : 0.7264
##                  95% CI : (0.6313, 0.8085)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 1.726e-06
##
##                   Kappa : 0.4528
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.7358
##             Specificity : 0.7170
##          Pos Pred Value : 0.7222
##          Neg Pred Value : 0.7308
##              Prevalence : 0.5000
##          Detection Rate : 0.3679
##    Detection Prevalence : 0.5094
##       Balanced Accuracy : 0.7264
##
##        'Positive' Class : 0
##

##     Glucose         Age         BMI     Insulin          BP Pregnancies
##   57.838695   20.715961   19.333357   14.853132   14.265176   11.930137
##         DPF          ST
##    6.576736    5.064233
```

Insulin is an Important Variable

In the Variable Importance results we see that Insulin is listed higher above many variables, despite not being used in the branching of the tree This could be due to the fact that variable importance is calculated using "the sum of the goodness of split measures for each split for which it was the primary variable, plus goodness *(adjusted agreement) for all splits in which it was a surrogate" (reference 1). This is different to how the tree calculates which variable to perform the split. The tree can have a variable occur many times, "either as a primary or surrogate variable" (reference 1). We may also have gotten different results for the tree depending on the seed we use.

Tree Predictions and Variable Importance of Pruned Tree

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 39 12
##          1 14 41
##
##                Accuracy : 0.7547
##                  95% CI : (0.6616, 0.8331)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : 7.135e-08
##
##                   Kappa : 0.5094
##
##  Mcnemar's Test P-Value : 0.8445
##
##             Sensitivity : 0.7358
##             Specificity : 0.7736
##          Pos Pred Value : 0.7647
##          Neg Pred Value : 0.7455
##              Prevalence : 0.5000
##          Detection Rate : 0.3679
##    Detection Prevalence : 0.4811
##       Balanced Accuracy : 0.7547
##
##        'Positive' Class : 0
##
```

```
##     Glucose         Age         BMI     Insulin          BP Pregnancies
##   54.264105   19.211565   18.746292   14.695894   13.520873   11.285396
##         DPF          ST
##    6.361822    4.749757
```

References

1. Therneau, T. M., Atkinson, E. J., & Foundation, M. (2022, October 21). An Introduction to Recursive Partitioning Using the RPART Routines. From https://cran.r-project.org/web//packages/rpart/vignettes/longintro.pdf