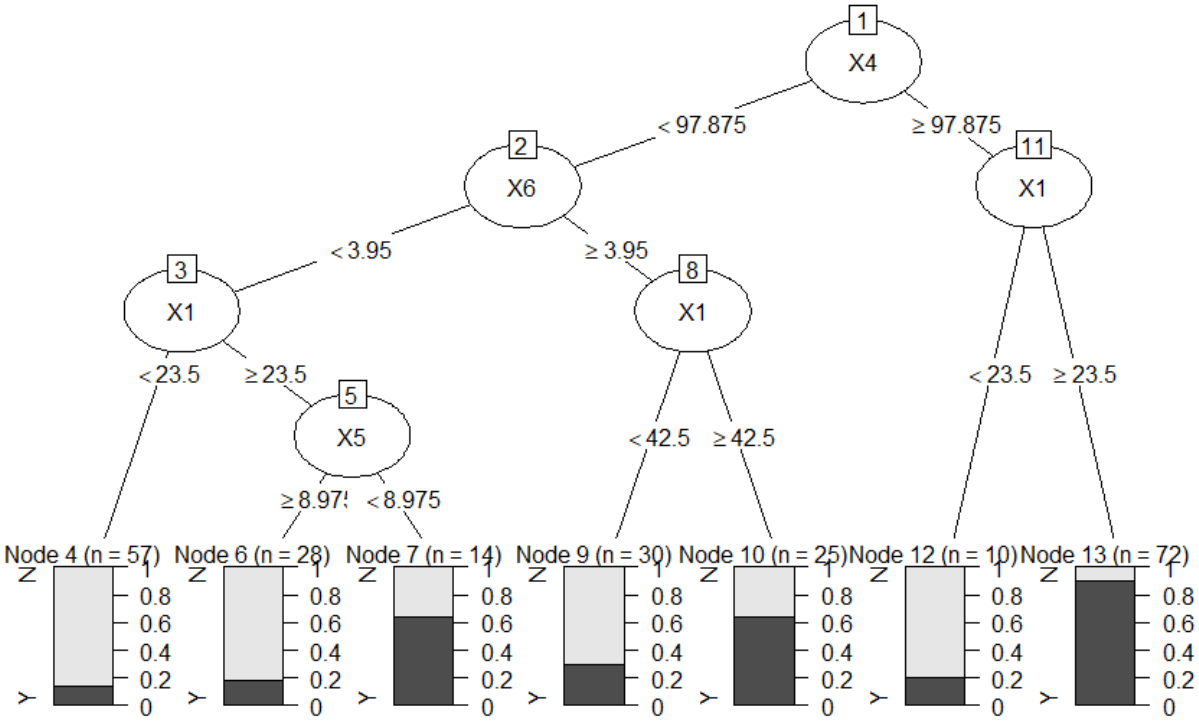
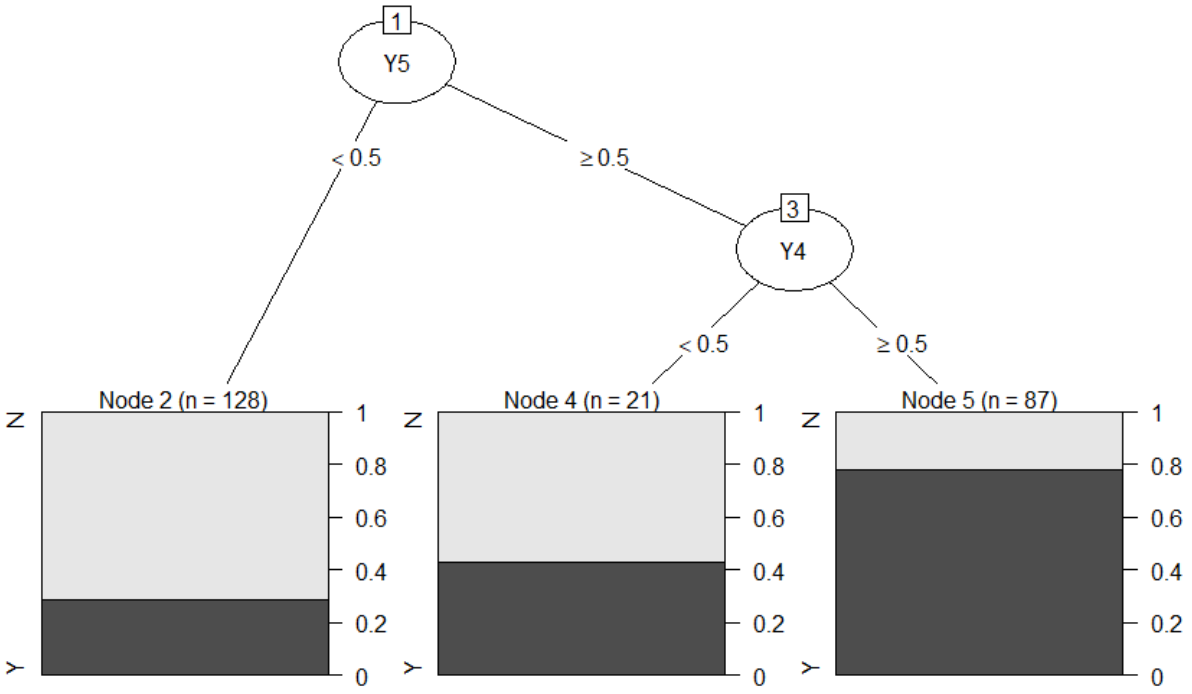
# Project Report for CA 1

By Arambakam Mukesh - 19301497



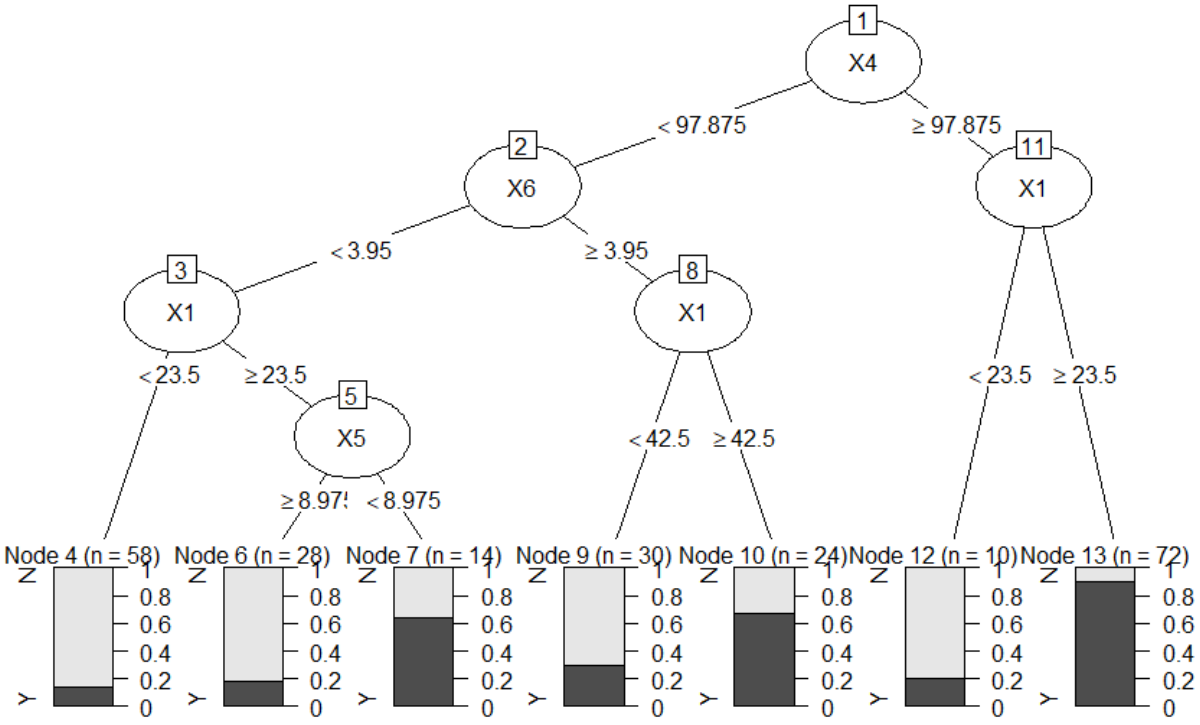

*Plot 1 - DT over only X*

Plot 1 represents the Decision Tree over the entire data set but with the Predictors though X1-X7. This DT predicts with an accuracy of 80% (0.8).



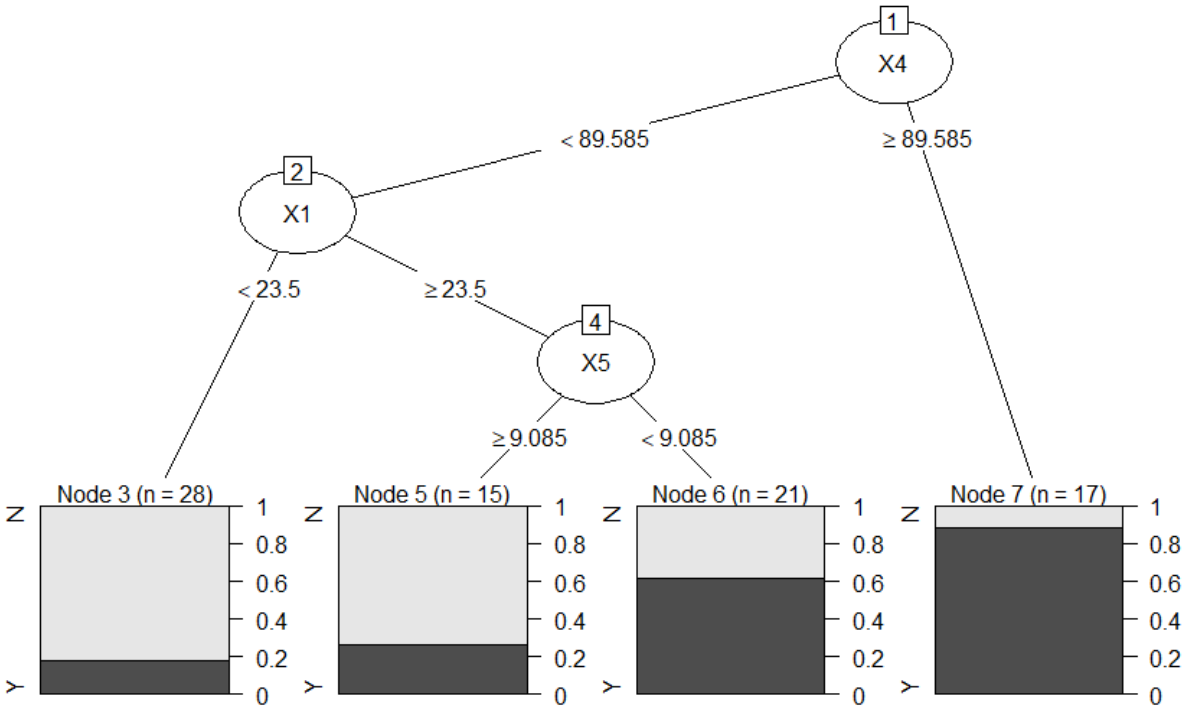

*Plot 2 - DT over only Y*

Plot 2 represents the Decision Tree over the entire data set but with the Predictors though Y1-Y7. This DT predicts with an accuracy of 80% (0.8).



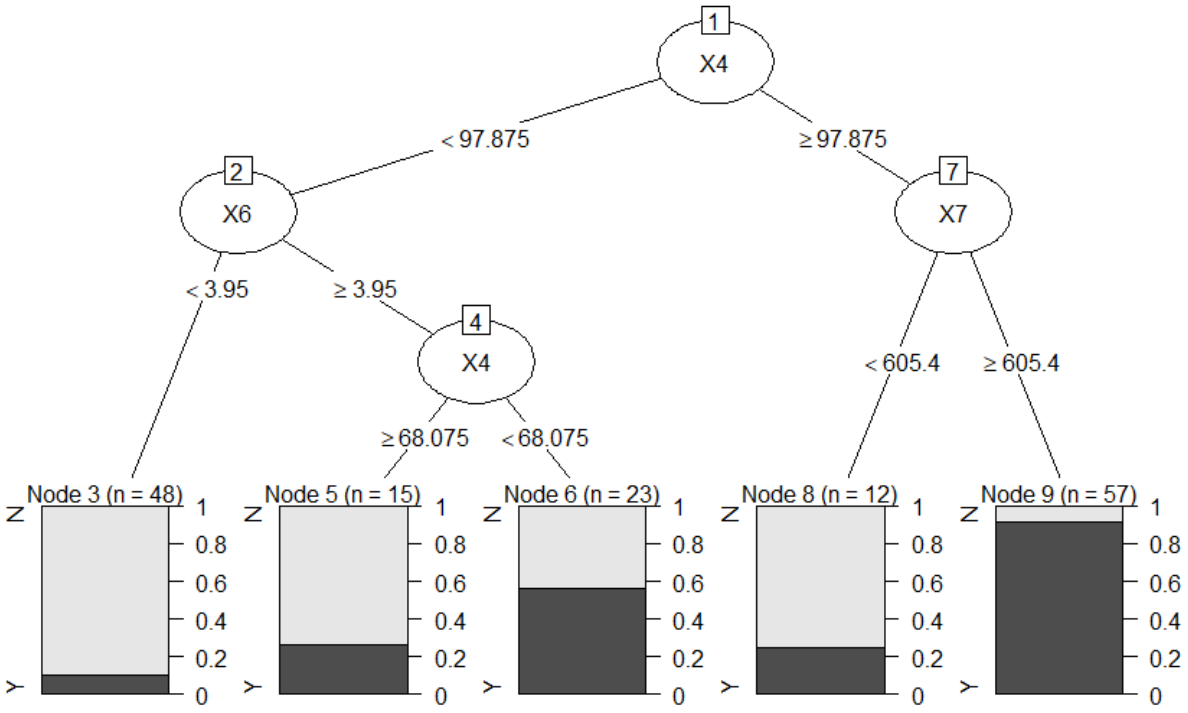

*Plot 3 - DT over X and Y*

Plot 3 represents the Decision Tree over the entire data set but with the Predictors though X1-X7 and Y1-Y7. This DT predicts with an accuracy of 80% (0.8).



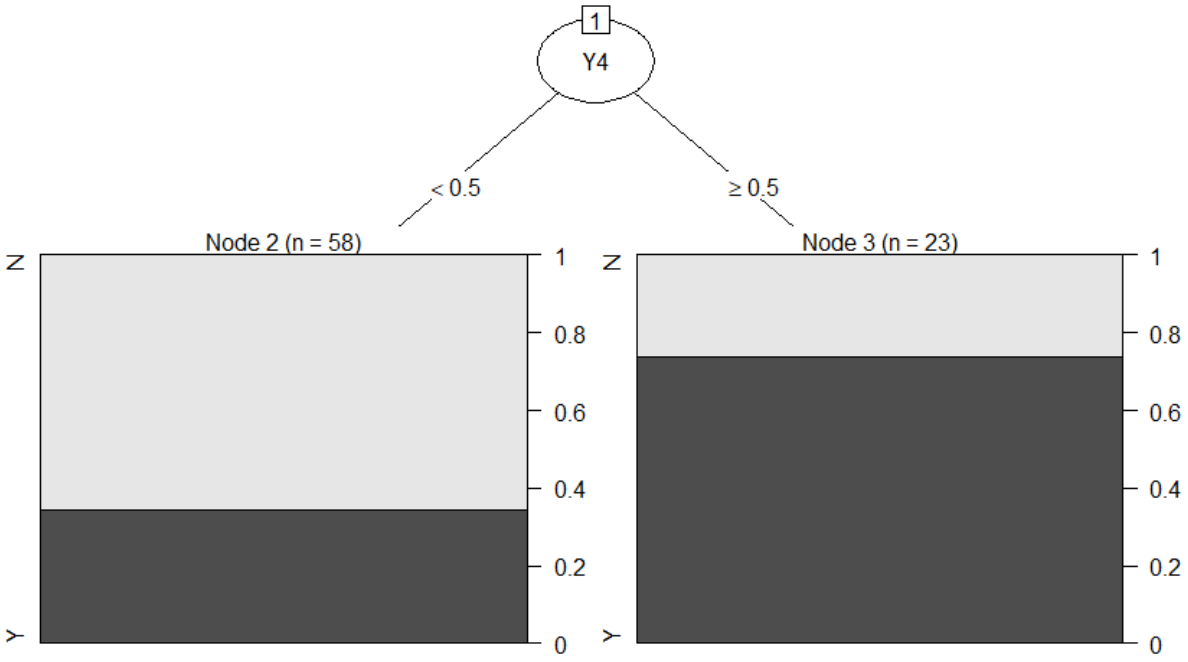

*Plot 4 - DT over X with Group 0*

Plot 4 represents the Decision Tree over Group 0 set but with the Predictors though X1-X7. This DT predicts with an accuracy of 77% (0.77).



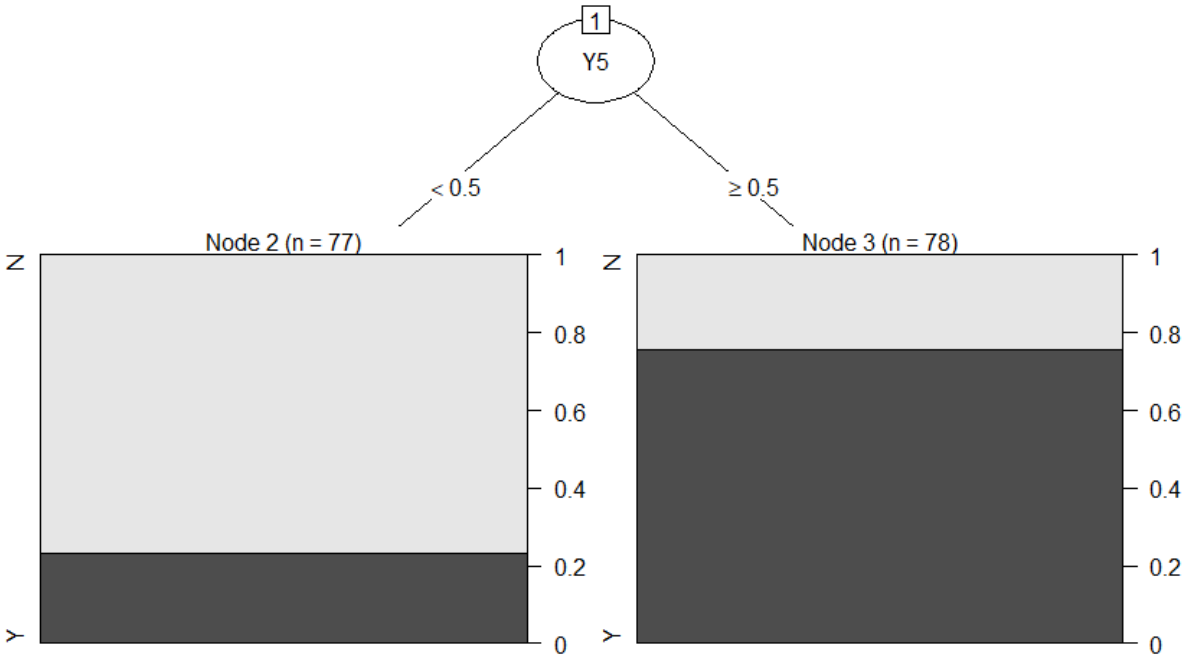

*Plot 5 - DT over X with Group 1*

Plot 5 represents the Decision Tree over Group 1 set but with the Predictors though X1-X7. This DT predicts with an accuracy of 78% (0.78).



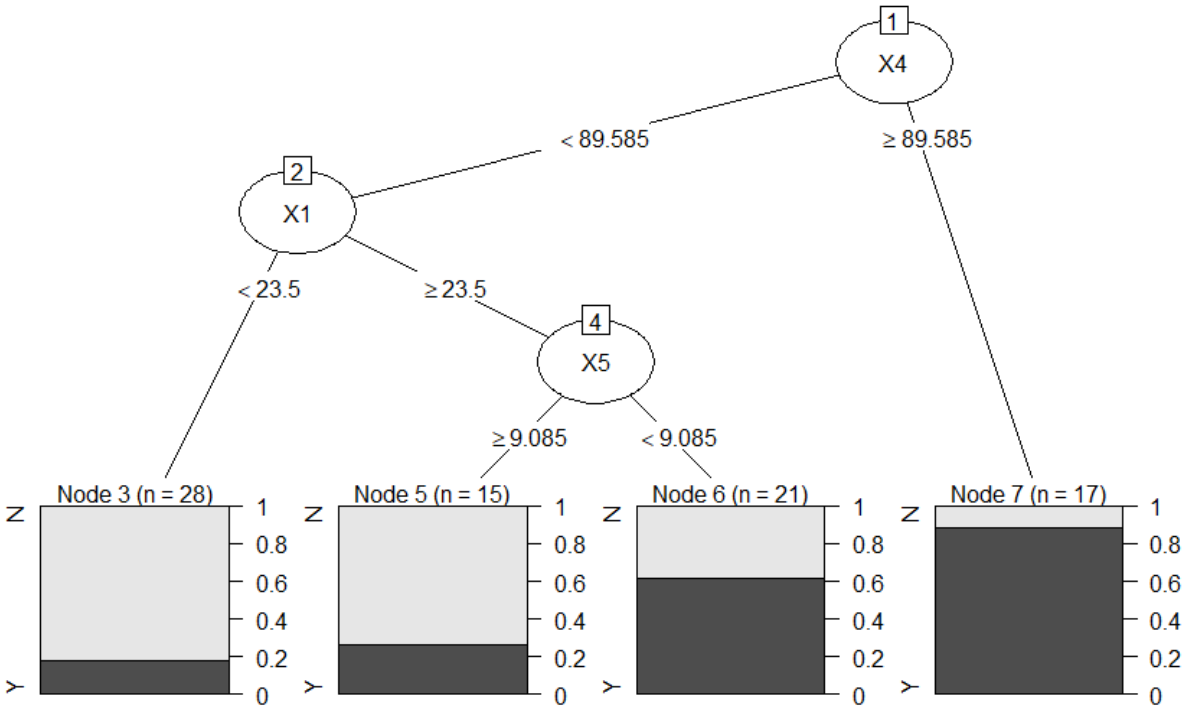

*Plot 6 - DT over Y with Group 0*

Plot 6 represents the Decision Tree over Group 0 set but with the Predictors though Y1-Y7. This DT predicts with an accuracy of 72% (0.72).



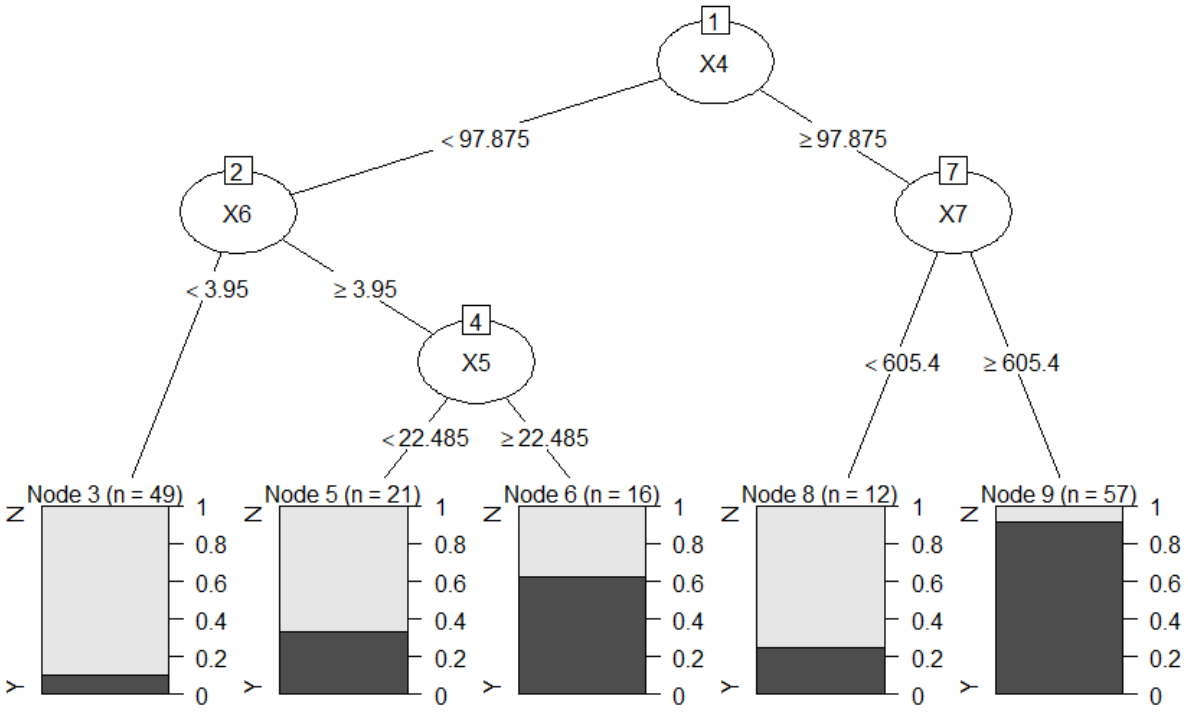

*Plot 7 - DT over Y with Group 1*

Plot 7 represents the Decision Tree over Group 1 set but with the Predictors though Y1-Y7. This DT predicts with an accuracy of 75% (0.75).




*Plot 8 - DT over X and Y with Group 0*

Plot 8 represents the Decision Tree over Group 0 set but with the Predictors though X1-X7 and Y1-Y7. This DT predicts with an accuracy of 77% (0.77).




*Plot 9 - DT over X and Y with Group 1*

Plot 9 represents the Decision Tree over Group 1 set but with the Predictors though X1-X7 and Y1-Y7. This DT predicts with an accuracy of 80% (0.8).

# Conclusion:

The best Decision Tree generated is the DT generated over **Entire Dataset with the Predictors X1-X7** as it has the highest **accuracy of 80%** - see **Plot 1** for the DT. The below is the Decision Tree's summary, indicating the splits:

```
n= 236

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 236 114 N (0.51694915 0.48305085)
   2) X4< 97.875 154  47 N (0.69480519 0.30519481)
     4) X6< 3.95 99  22 N (0.77777778 0.22222222)
       8) X1< 23.5 57   8 N (0.85964912 0.14035088) *
       9) X1>=23.5 42  14 N (0.66666667 0.33333333)
        18) X5>=8.975 28   5 N (0.82142857 0.17857143) *
        19) X5< 8.975 14   5 Y (0.35714286 0.64285714) *
     5) X6>=3.95 55  25 N (0.54545455 0.45454545)
      10) X1< 42.5 30   9 N (0.70000000 0.30000000) *
      11) X1>=42.5 25   9 Y (0.36000000 0.64000000) *
   3) X4>=97.875 82  15 Y (0.18292683 0.81707317)
     6) X1< 23.5 10   2 N (0.80000000 0.20000000) *
     7) X1>=23.5 72   7 Y (0.09722222 0.90277778) *
```

Though the DT's in Plot 3 and Plot 9 also have an accuracy of 80% - Plot 1 is better because it gives a consistent accuracy of 80% when tested with different `rpart` configurations like minsplit, minbucket and maxdepth, more over it has a **Low Variance** between different decision tree when run with different seed values for the data and different rpart configuration. We can also clearly say that X's are the features with **High Information Gain** compared to Y, and also splitting the data into Group's do not show a significant improvement accuracy.

The code for this can be found on my GitHub, please find the link to the repo below.

https://github.com/mukeshmk/r-project