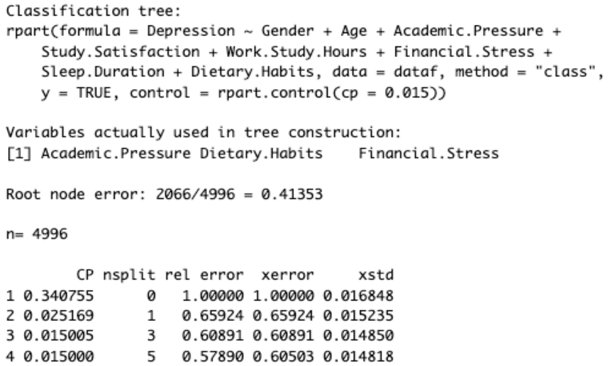
Decision Trees

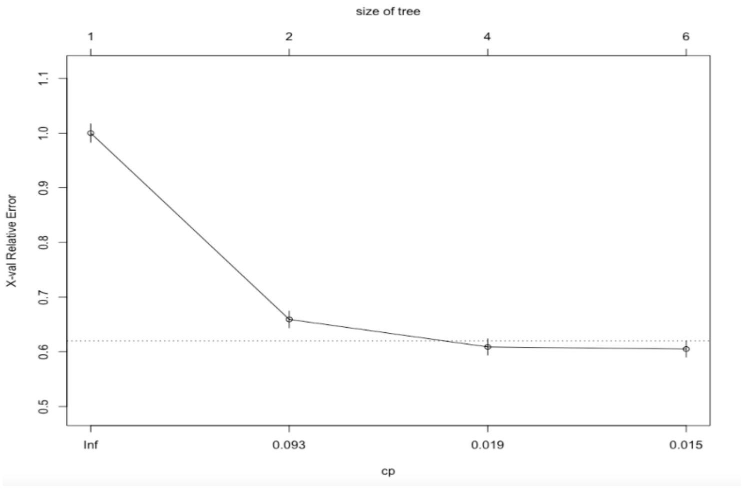
Decision trees are a simple yet powerful machine learning algorithm that split data into smaller subsets based on feature values, forming a tree-like structure. They are easy to interpret and effective for both classification and regression tasks, making them ideal for identifying patterns in complex datasets.



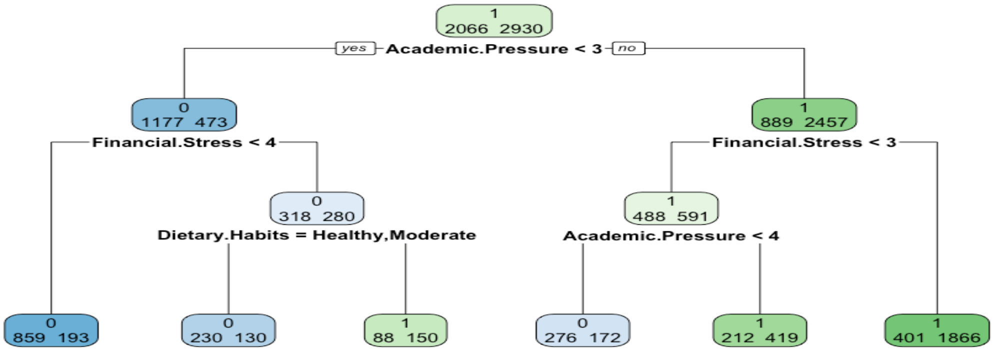
```
Classification tree:
rpart(formula = Depression ~ Gender + Age + Academic.Pressure +
    Study.Satisfaction + Work.Study.Hours + Financial.Stress +
    Sleep.Duration + Dietary.Habits, data = dataf, method = "class",
    y = TRUE, control = rpart.control(cp = 0.015))

Variables actually used in tree construction:
[1] Academic.Pressure Dietary.Habits    Financial.Stress

Root node error: 2066/4996 = 0.41353

n= 4996

        CP nsplit rel error  xerror     xstd
1 0.340755      0   1.00000 1.00000 0.016848
2 0.025169      1   0.65924 0.65924 0.015235
3 0.015005      3   0.60891 0.60891 0.014850
4 0.015000      5   0.57890 0.60503 0.014818
```

The decision tree was constructed using three variables: academic pressure, dietary habits, and financial stress, which were identified as the most impactful factors influencing the predictions. The root node error represents the proportion of misclassifications if the tree did not make any splits, 41% of the observations would have been misclassified at the root node, highlighting the importance of further splitting the data.

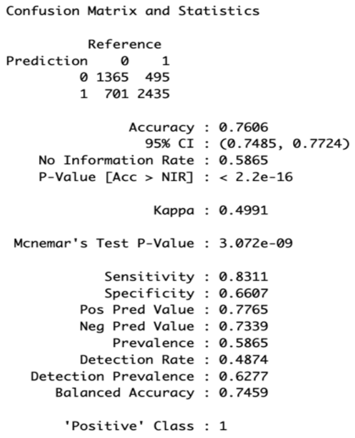
As the tree evolves, smaller complexity parameter values lead to more splits and finer segmentation. The first split significantly reduces errors, showing strong predictive power. Further splits provide smaller improvements, and by the fifth split, the error stabilizes, balancing complexity and accuracy.



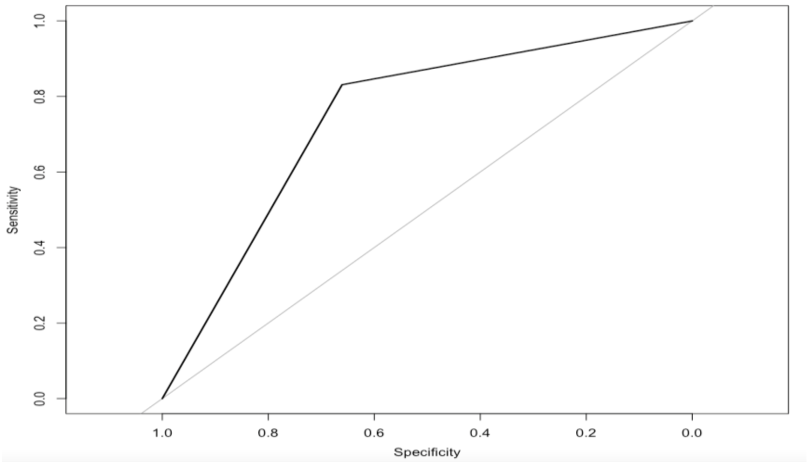
The CP plot shows how the tree's size (number of splits) affects prediction accuracy. Smaller CP values mean more splits and complexity, while larger CP values represent simpler trees. The error drops significantly after the first split (CP = 0.093), but further splits bring only minimal improvement. The best tree size is the simplest one within one standard error of the minimum error, balancing accuracy and simplicity.



The decision tree highlights that Academic Pressure is the most critical predictor of depression, with Financial Stress and Dietary Habits also playing significant roles. Low Academic Pressure is associated with reduced depression rates, while high Financial Stress strongly correlates with increased risk. Additionally, having healthy or moderate dietary habits helps lower depression risk, especially when combined with low Financial Stress. Overall, the tree reveals that high Academic Pressure and Financial Stress are primary drivers of depression, while low Academic Pressure, low Financial Stress, and healthy Dietary Habits act as protective factors.



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 1365  495
         1  701 2435

               Accuracy : 0.7606
                 95% CI : (0.7485, 0.7724)
    No Information Rate : 0.5865
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4991

 Mcnemar's Test P-Value : 3.072e-09

            Sensitivity : 0.8311
            Specificity : 0.6607
         Pos Pred Value : 0.7765
         Neg Pred Value : 0.7339
             Prevalence : 0.5865
         Detection Rate : 0.4874
   Detection Prevalence : 0.6277
      Balanced Accuracy : 0.7459

       'Positive' Class : 1
```



The model's AUC value of approximately **74**% indicates a moderate ability to distinguish between positive (depression) and negative cases. This means that the model has a 74% chance of correctly ranking a randomly chosen positive case higher than a randomly chosen negative case. While the sensitivity (83.11%) shows strong detection of positive cases, the specificity (66.07%) suggests some difficulty in correctly identifying negative cases, which impacts the AUC. Overall, the model demonstrates decent performance, but there is room for improvement, especially in reducing false positives.