MTL782: Data Mining
Assignment 1

Saket Kandoi 2021MT60265
Navneet Raj 2021MT10240
Aditya Thomas 2021MT60944

Question 1

Packages Used: pandas, sklearn, tensorflow_decision_forests, tf_keras, maplotlib, random

Random Forest (Default Model)

Accuracy: 86.67%
Number of trees: 300

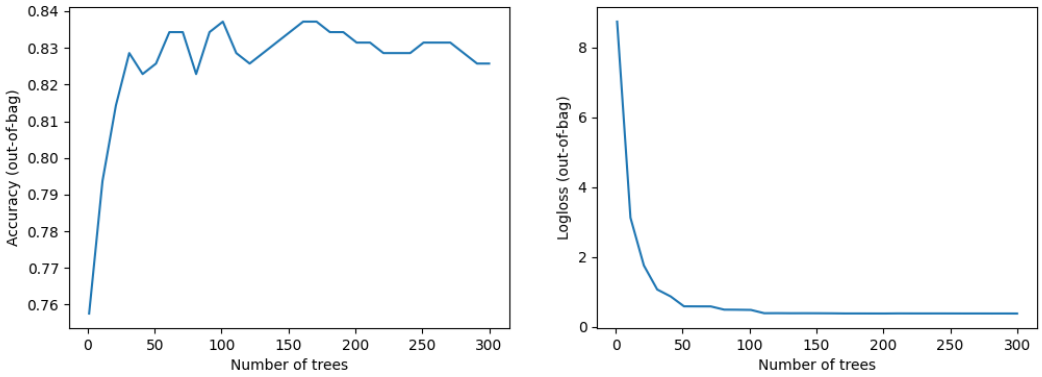


Figure 1: Random Forest (Default Model)

Saket Kandoi 2021MT60265
Navneet Raj 2021MT10240
Aditya Thomas 2021MT60944

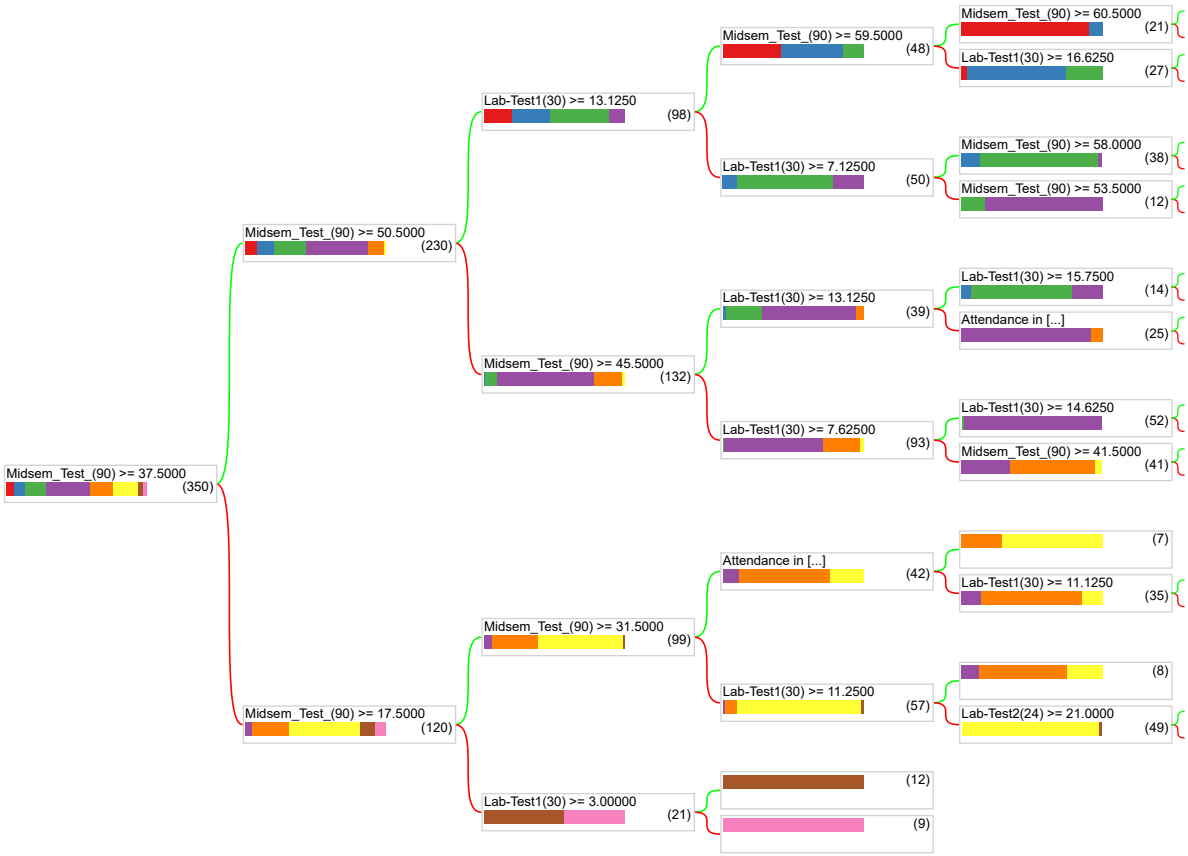


Figure 2: Random Forest (Default Model)

Random Forest (Tuned Model)

Accuracy: 87.33%
Number of trees: 300
The drop in accuracy could be due to overfitting.

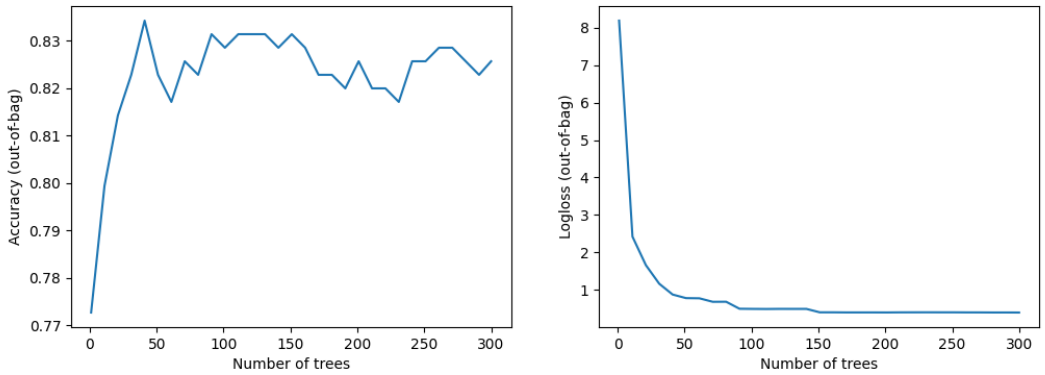


Figure 3: Random Forest (Improved Model)

Saket Kandoi 2021MT60265
Navneet Raj 2021MT10240
Aditya Thomas 2021MT60944

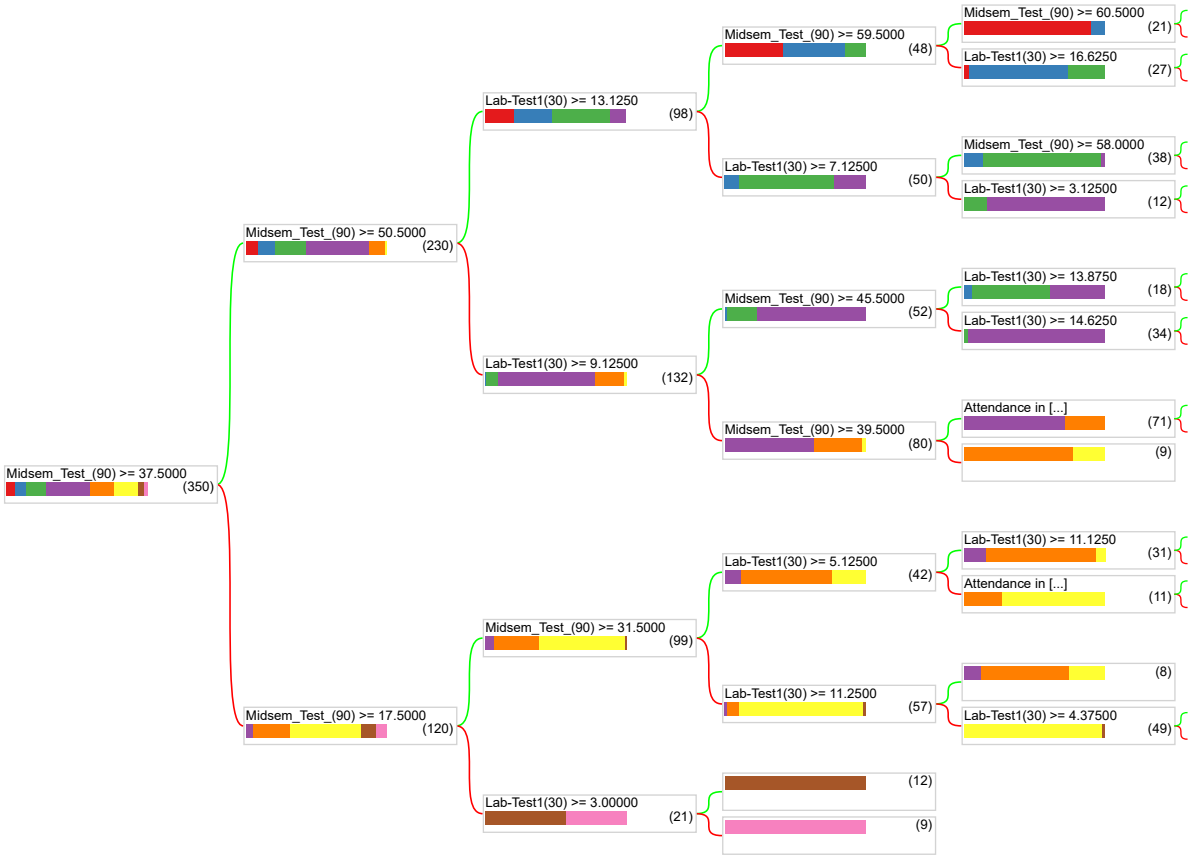


Figure 4: Random Forest (Improved Model)

Random Forest (Reduced Model)

Accuracy: 88.00%
Number of trees: 30
Restricting the number of trees to 30 gives a much better accuracy, since on 300 trees model could be overfitting due to limited sample size of data.

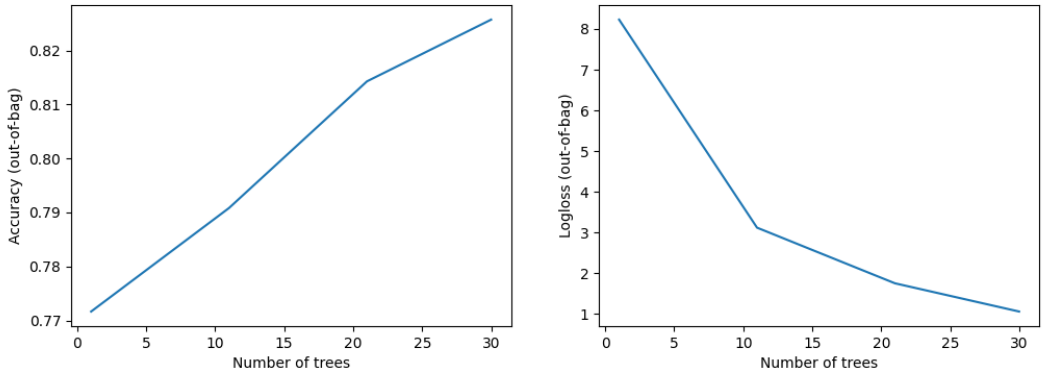
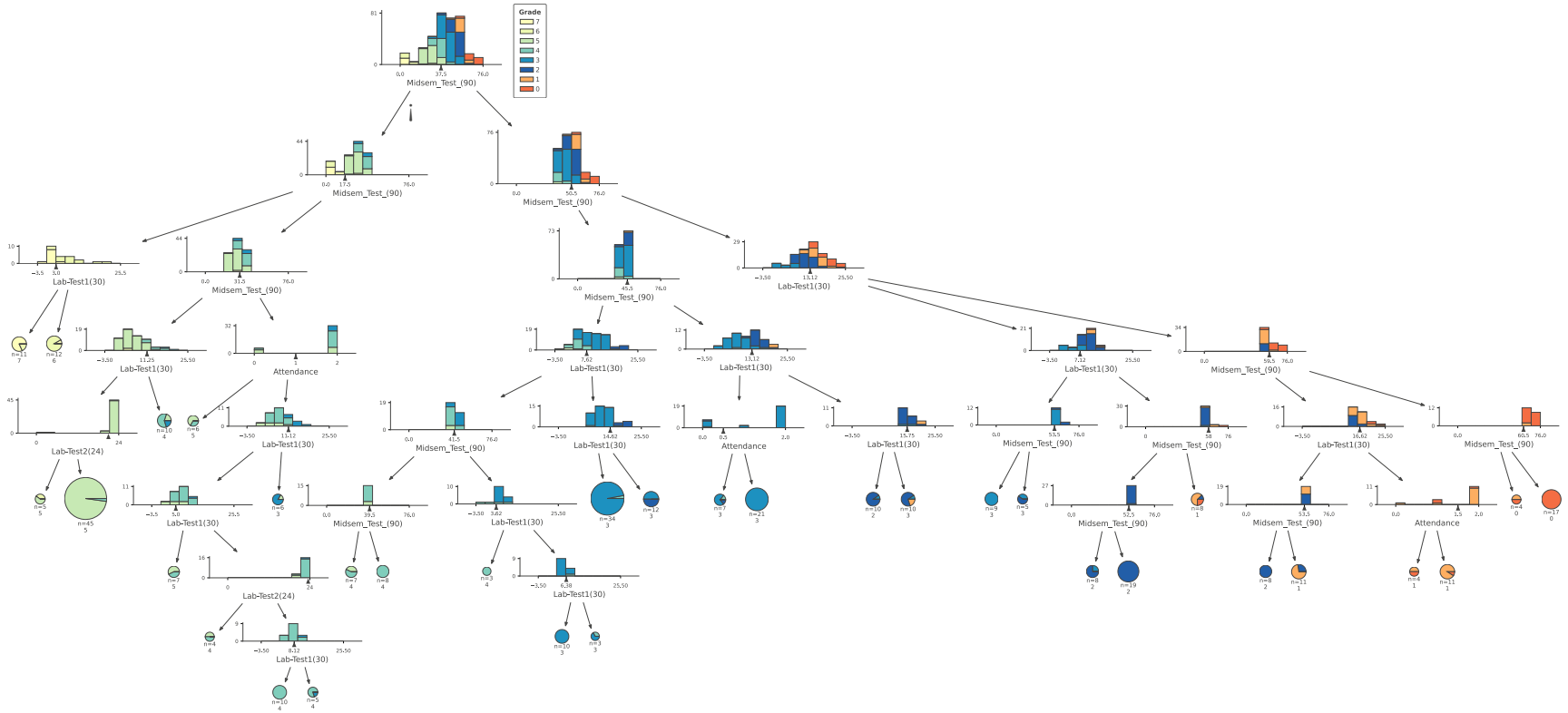


Figure 5: Random Forest (Reduced Model)

Saket Kandoi 2021MT60265
Navneet Raj 2021MT10240
Aditya Thomas 2021MT60944



First tree in the trained Random Forest

Saket Kandoi 2021MT60265
Navneet Raj 2021MT10240
Aditya Thomas 2021MT60944

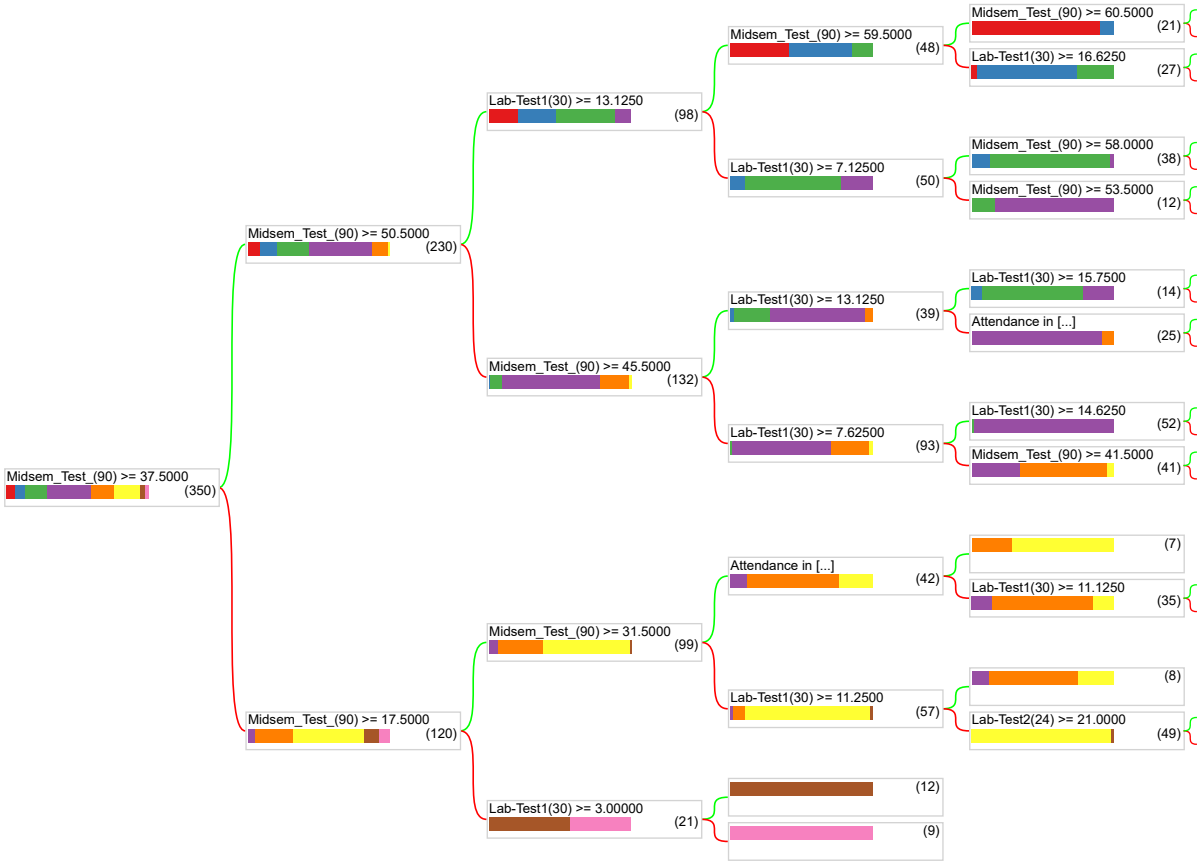


Figure 6: Random Forest (Reduced Model)

Random Forest (Without encoding)

Accuracy: 87.33%
Though we expect a decrease in accuracy, due to randomness and limited data, there is no significant change.

Gradient Boosted Decision Trees

Accuracy: 87.33%
Number of trees: 30

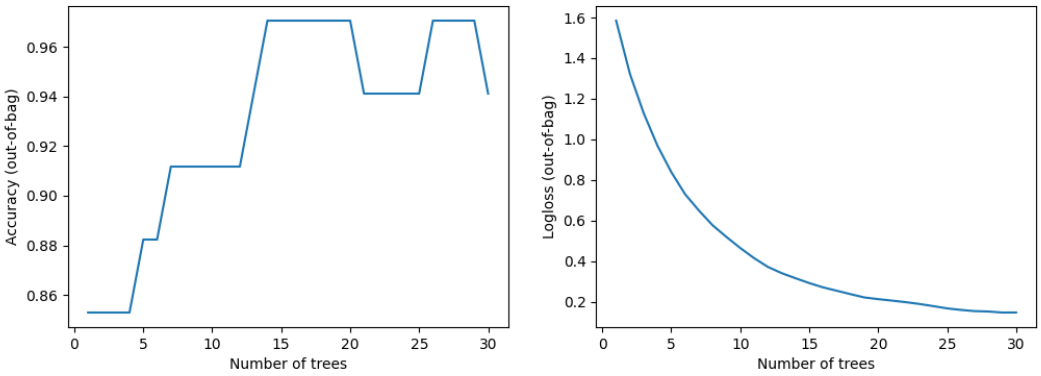


Figure 7: Gradient Boosted Decision Trees

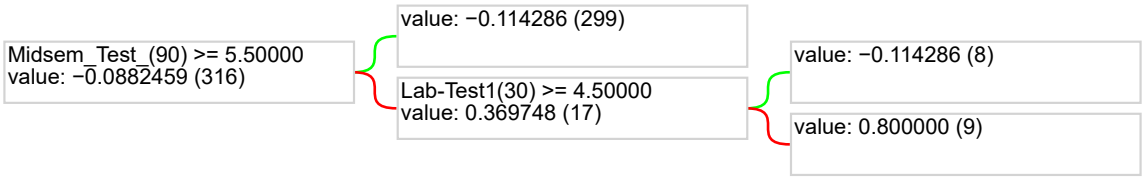


Figure 8: Gradient Boosted Decision Trees

Analysis of Parameters in Random Forests

Train Accuracy = 95.43%
Test Accuracy = 88%
As expected train accuracy is much higher than test accuracy.

Number of Trees

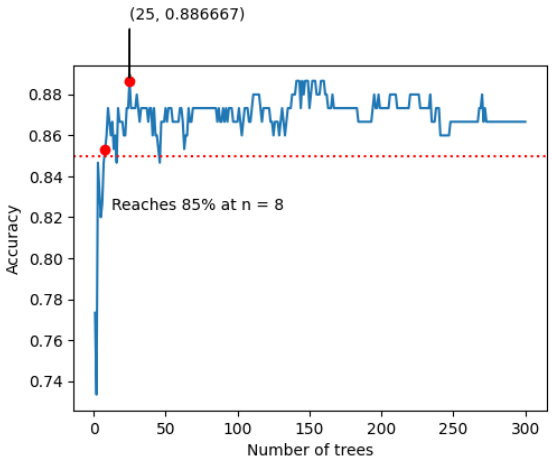


Figure 9: Accuracy with number of trees

Saket Kandoi 2021MT60265
Navneet Raj 2021MT10240
Aditya Thomas 2021MT60944

Maximum Depth

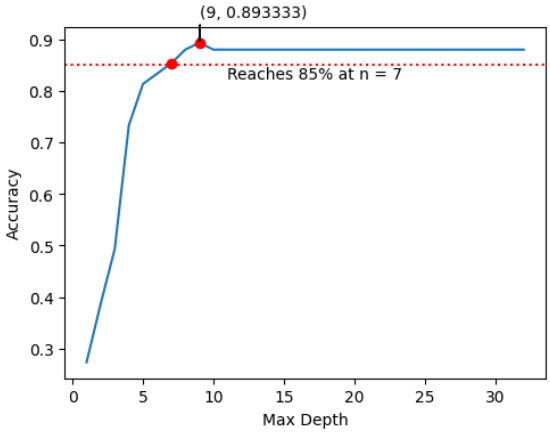


Figure 10: Accuracy with maximum depth, with n = 30

Glossary

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Log Loss} = -\frac{1}{n} \sum (y \times \log(p))$$