Assignment -2: Decision Tree

Deevyansh Khadria

October 2024

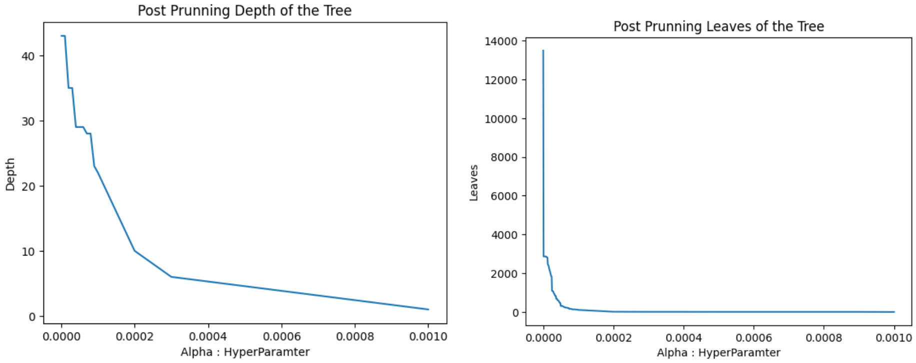
1 Approach

- Handling the categorical (encoding) and numerical features (scaling).
- Writing the function to calculate both the Entropy and Gini Impurity.
- Implementing the recursion for the Decision Tree Node Creation

2 Prunning Approach

- Max Depth is defined to introduce the pre-prunning strategy.
- The hyperparameter (alpha) is introduced that calculates the reward $(R(t) - R(T))$ and penalty $(\alpha(|T| - 1))$ per node.
- The post-running is implemented afterwards for a specific alpha.$\alpha_e f f = (R(t) - R(T))/(|T| - 1)$

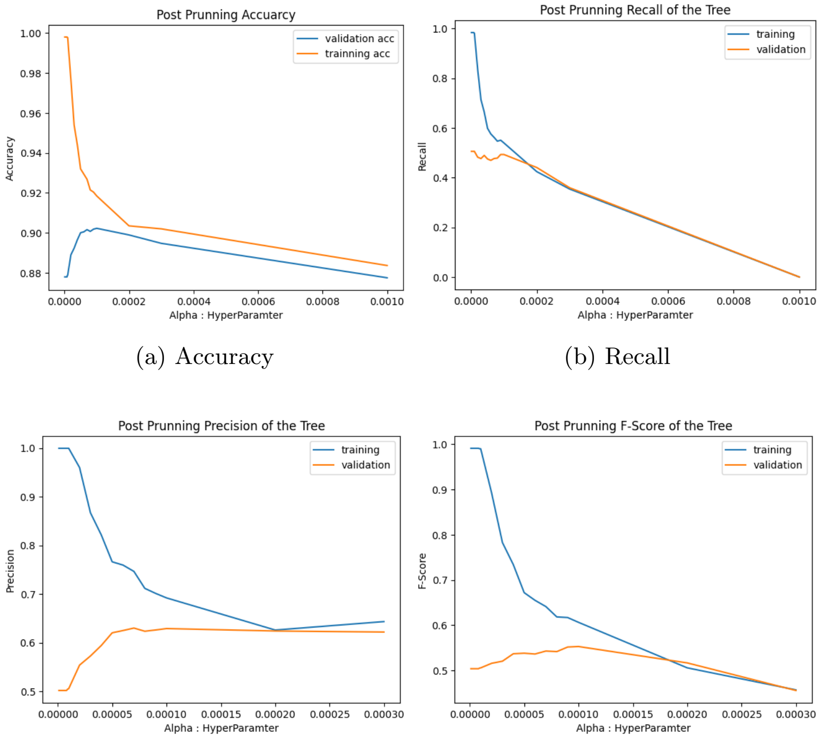
3 How Prunning Affects the Size of the Tree



(a) Depth of the Tree

(b) Leaves of the Tree

4 The differences in model performance before and after pruning.



(a) Accuracy

(b) Recall

(c) Precision

(d) F-Score

Figure 2: Evaluation metrics for the model with pruning (ENTROPY)

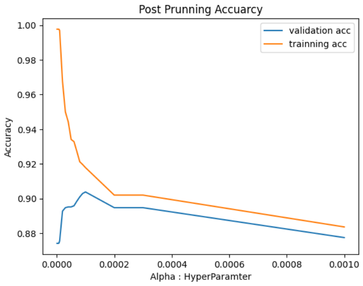


Figure 3: Accuracy Graph for gini impurity

5 Cross Validation for finding the Optimal Depth

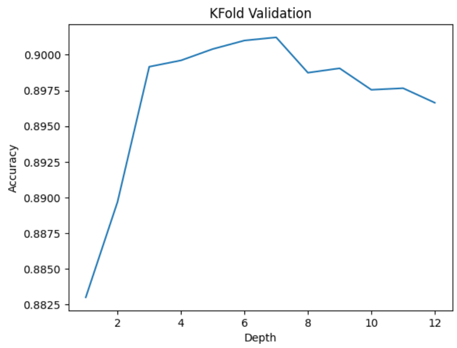


Figure 4: 7 comes out to be optimal depth for the decision tree(gini Impurity)

6 Whether pruning helped in mitigating overfitting and improving generalization

- Before Prunning the max accuracy on the Validation set is 0.873. After Prunning, it increases to 0.910 (for gini impurity)
- Before Prunning the max F-Score obtained was 0.469 on Validation set. After Prunning it increases to 0.546.
- Prunning reduces the accuracy and F-Score on the trainning data , reducing the variance.
- Cross Validation for the optimal depth increases the accuracy from the 0.883 to 0.901 (for the gini impurity).