**Elements of AI : A4**
Username : sdasara-abkaitha106-jl146

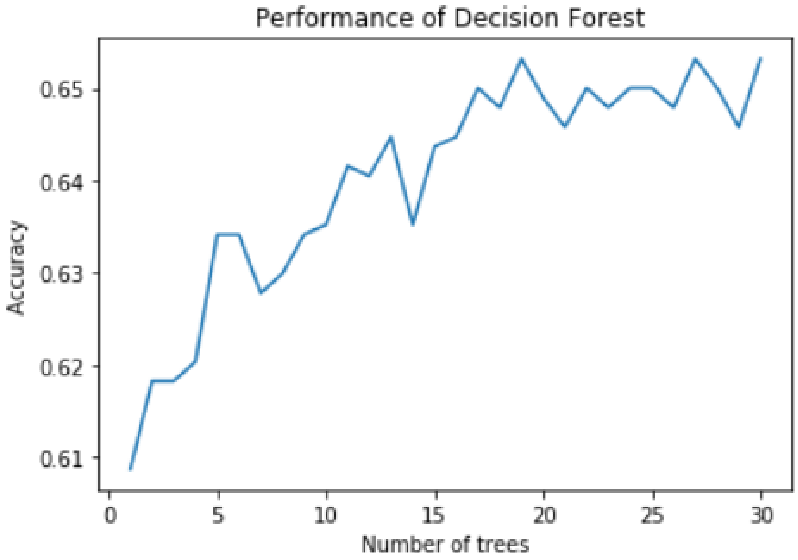
**Decision Forest Report:**

**(1) Description of the problem formulation**

As suggested in the given assignment, I generated random indices for the train data and took this subset of the training data to build the decision trees.

Taking multiple trees and making them vote helps in reducing the variance and does model averaging to get us a more generalised ensemble model to predict.

The parameters to tune in this case are:
Maximum depth of each tree
Number of trees to build
Ratio of train data to be subsetted
Minimum number of samples to be considered for a split



I measured the performance of the decision forest using accuracy metric with respected to number of trees for voting.

As the number of trees increases, rise in the accuracy is noticed.
The final accuracy was 65.32% for 30 trees.

**(2) Description of the working of the program:**

**Train:**
Gini index was used as the splitting criterion.
The decision tree is represented in the form of a nested dictionary containing the keys: index, left and right.
We find the best split feature based on gini index of all features and set the feature with the lowest gini index as the root node and initialse a dictionary.
The rest of the tree is built as nested dicitonary on top of this dictionary.

Index is the node feature index in the attribute list.
Left and right contains the data subsetted based on the being greater than or lesser than the feature threshold value.
The leaf node has the same value for both left and right keys.
The training subset ratio was 0.03 which came out to be 1109 records for each decision tree.

**Test:**
We traverse through the dictionary based on the values for each feature index and we check if we reached the leaf node by using “isinstance” to see if the value of the key is a dictionary. If it is “False” it means we have reached the leaf node and that value is returned as the prediction.
We take the counts for each prediction and return the label with the highest count.

**(3) Discussion of any problems, assumptions or simplifications :**

Thresholds for each feature was taken using the median value for each pixel position.
Only gini index was used for the creation of a decision tree.
If one the two groups after split is empty we make it a leaf node and assign the same label for both the nodes.
The maximum depth was limited to 5 due to computational limits and minimum samples for split was set to 50.
Number of trees was limited to less than 30 due to time constraints.
Training subset ratio was taken to be small for faster computation.