# COL341: Assignment 4 Report

Nihar Patel

2017MT60780

Splitting for Decision Tree:

- I one hot encoded all the features (except continuous one). So each feature split at node has two choices- 0 or 1, giving us a binary decision tree
- For continuous features, I tried three types of splitting:
    - Threshold-based splitting: Here I sorted a column of features, then for each case where y label changes from 0 to 1 or 1 to 0, I computed gain for the split there and then took the split with maximum gain
    - Mean splitting: computed mean of the feature values, and then split by comparing it with feature values. Left for less than mean and right for greater than mean
    - Median splitting: Same as mean splitting but instead of mean, I computed median of the feature column
- Here I used Information Gain (entropy) to split the node by maximising gain
- Pruning:
    - I started by post traversing the decision tree. At each node, I computed the error on validation data at that node, and compared it with the error for if the node was replaced by a leaf.
    - If the error by replacing the node with a leaf is less, I replaced the node with leaf
    - After fully traversing the tree for pruning, I computed error on whole validation data.

- I kept pruning (by post traversing) the decision tree till the validation error kept decreasing.

Results:

| Splitting method | Pruning | Test accuracy | Validation accuracy |
|---|---|---|---|
| Threshold-based splitting | No | 0.81 | 0.813 |
| Threshold-based splitting | Yes | 0.78933 | 0.83173 |
| Mean splitting | No | 0.8103 | 0.802 |
| Mean splitting | Yes | 0.76933 | 0.90567 |
| Median splitting | No | 0.80733 | 0.79933 |
| Median splitting | Yes | 0.77533 | 090667 |