Data Science
Lab Exercise (Decision Tree)

1. UCI ML Repository contains many datasets for classification. You need to find 5 datasets with at least 10 attributes
   https://archive.ics.uci.edu/ml/datasets.php

   Complete the following tables and calculate accuracy using (1) Use 10 x 10 Fold CV (ii) 70% Holdout approach repeated 100 times

   Show all the standard deviations in the table. Briefly discuss advantages / disadvantages of hold out and cross validation approach. Analysis the result. Which approach is good and why? Why some approaches unable to perform well in some data sets.

| | Dataset1 | Dataset2 | Dataset3 | Dataset4 | Dataset5 |
|---|---|---|---|---|---|
| DT using gini (without pruning) | | | | | |
| DT using gini (with pruning) | | | | | |
| DT using entropy (without pruning) | | | | | |
| DT using entropy (with pruning) | | | | | |

   Hint: Check ccp_alpha parameter for pruning. Use ccp_alpha = 0.015 for pruning