Machine Learning

Decision Trees

Dataset-F : Adult Census Dataset

Group 0076: Aayush Prasad(18CS30002) and Rajdeep Das(18CS30034):

Part 1

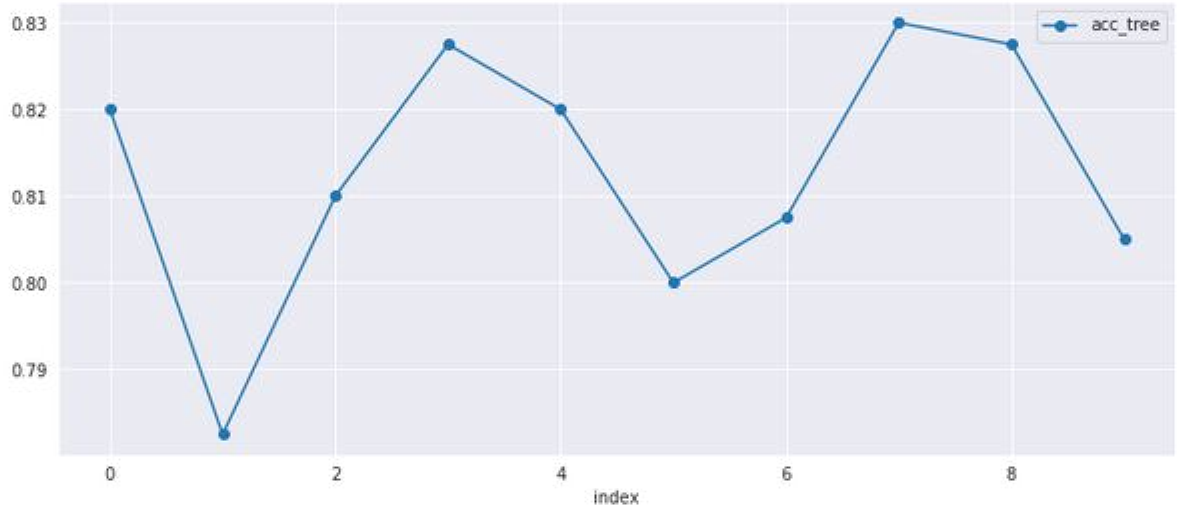
Build a decision tree by taking as input a maximum depth and by randomly splitting the dataset as 80/20 split i.e., 80% for training and 20% for testing. Provide the accuracy by averaging over 10 random 80/20 splits. Consider that particular tree which provides the best test accuracy as the desired one.

Procedure

- The data was stored in a dataframe and header names were added.
- The missing values were handled.
    - All the attributes having missing values were categorical so mode was used to replace the missing values
    - Following attributes had missing values:
        - Workclass
        - Occupation
        - Native country
- The maximum depth limit was taken as input.
    - If the input was -1, the entire tree will be computed (to it's maximum depth)
- The data was split as 80/20 split i.e. 80% for training and 20% for testing.
- The accuracy was calculated by averaging over 10 random 80/20 splits.

Results

- On inputting max_depth = 6, we got the the following results:
  - Average accuracy over 10 random splits: 0.813
  - Best accuracy = 0.83
- The tree giving the best accuracy was printed



Part 2

What is the best possible depth limit to be used for your dataset. Provide a plot explaining the same.

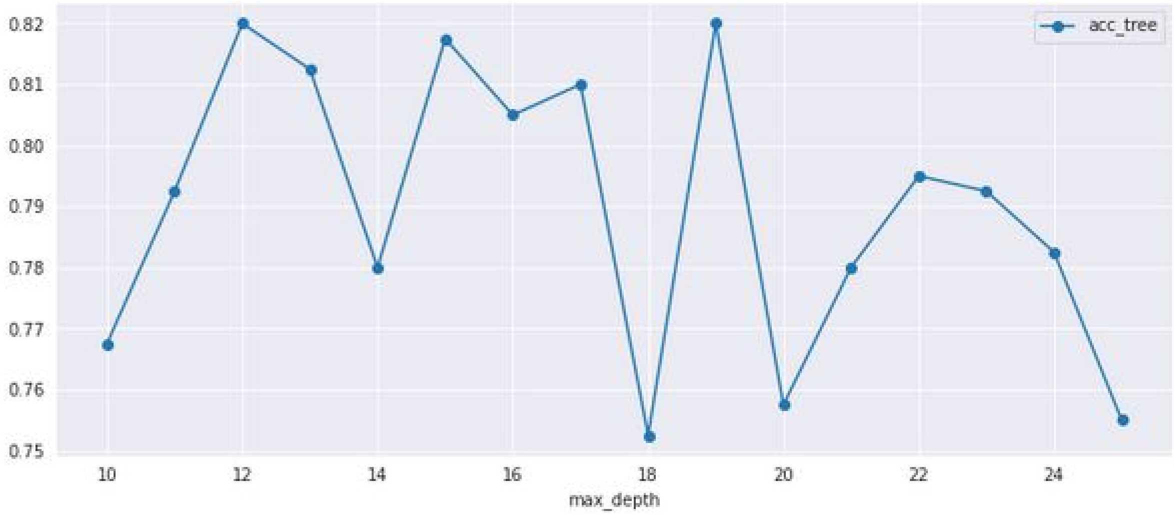
Procedure

- The accuracy of the Decision Tree Model formed considering each max-depth limit from 10 to 25 was computed
- A plot showing the average accuracy against the max-depth limit was shown.

Results

- The best accuracy was obtained when the max-depth limit was 19.

- The best Decision Tree considering this depth was formed to be used in Part 3 of the assignment.



Part 3

Perform the pruning operation over the tree obtained in question 2 using a valid statistical test for comparison.

Procedure

- The tree was pruned using the Reduced-Error Pruning statistical method.
- Post-Pruning was done on the decision tree built in part 2 of the assignment
- Reduced Error Pruning
  - Starting at the leaves, each node is replaced with its most popular class. If the prediction accuracy (measured on the validation set) is not affected then the change is kept. While somewhat naive, reduced error pruning has the advantage of simplicity and speed.
- Then the accuracy of the tree before pruning and after pruning was compared

Results

- The accuracy of the tree obtained in part 2 (before pruning) : 0.82
- The accuracy of the tree after pruning : 0.86

Part 4

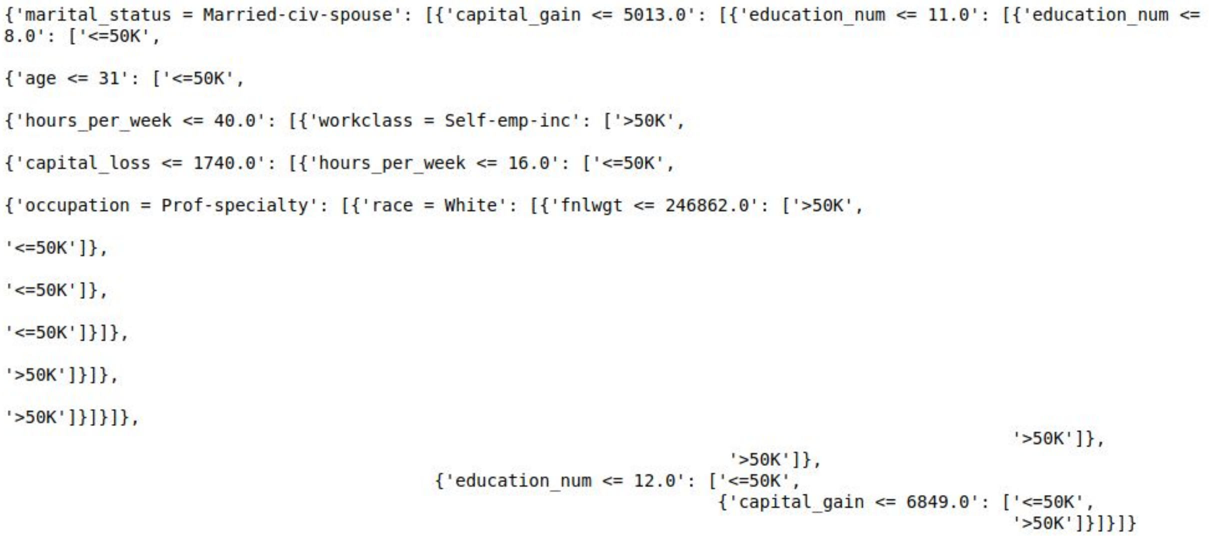
Print the final decision tree obtained from question 3 following the hierarchical levels of data attributes as nodes of the tree.

Procedure

- The pprint function which takes the pruned tree as input was called.
- The tree was represented in the dict-tree data structure
- The pprint function printed the tree following the hierarchical levels of data attributes as nodes of the tree.

Results

- Because of the depth of the tree and the printing area the tree printed was messy on a small screen but it is indeed hierarchical if the spaces and the tabs inserted are analysed.



```
{'marital_status = Married-civ-spouse': [{'capital_gain <= 5013.0': [{'education_num <=
8.0': ['<=50K',

{'age <= 31': ['<=50K',

{'hours_per_week <= 40.0': [{'workclass = Self-emp-inc': ['>50K',

{'capital_loss <= 1740.0': [{'hours_per_week <= 16.0': ['<=50K',

{'occupation = Prof-specialty': [{'race = White': [{'fnlwgt <= 246862.0': ['>50K',

'<=50K']},

'<=50K']},

'<=50K']}]},

'>50K']}]},

'>50K']}]}]},
                                                                                                '>50K']},
                                                                     '>50K']},
                                         {'education_num <= 12.0': ['<=50K',
                                                                    {'capital_gain <= 6849.0': ['<=50K',
                                                                                                '>50K']}]}]}
```