COL774 Assignment 3

Brian Sajeev Kattikat
2021CS50609

November 2023

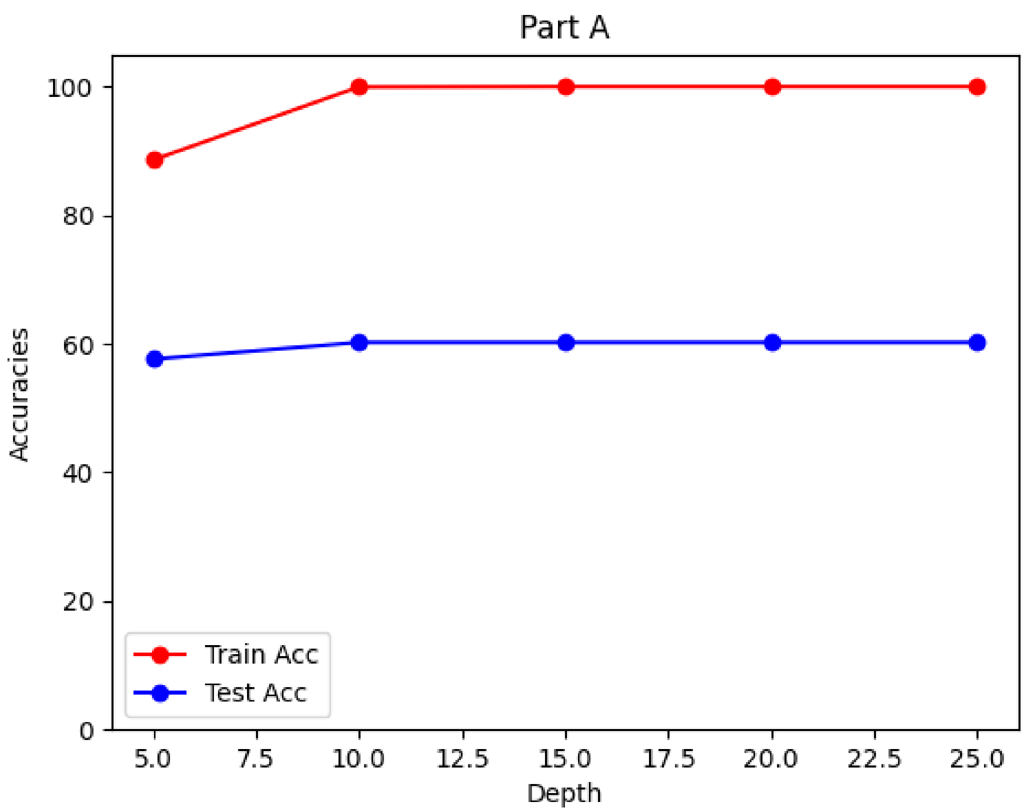
1 Decision Tree

(a) Here we constructed the Decision Tree for varying depths where features to split are determined using highest mutual information metric

i.
- Only Win:
  - Accuracy for in prediction type of only win on training set is 50.3380
  - Accuracy for in prediction type of only win on test set is 49.6380
- Only Loss
  - Accuracy for in prediction type of only loss on training set is 49.6614
  - Accuracy for in prediction type of only loss on test set is 50.3619
- DT with varying depths on training set:
  - Accuracy for depth 5 on training set is 88.55
  - Accuracy for depth 10 on training set is 99.94
  - Accuracy for depth 15 on training set is 100
  - Accuracy for depth 20 on training set is 100
  - Accuracy for depth 25 on training set is 100
  - Accuracy for depth 30 on training set is 100
- DT with varying depths on test set:
  - Accuracy for depth 5 on test set is 57.60
  - Accuracy for depth 10 on test set is 60.19
  - Accuracy for depth 15 on test set is 60.19
  - Accuracy for depth 20 on test set is 60.19
  - Accuracy for depth 25 on test set is 60.19

From the data obtained we find that single type prediction(only win, only loss) performs worse compared to Decision Tree classification (DT is almost 2x better in training prediction). As we expect the training accuracy is much better than test accuracy. We also find that the accuracy is almost the same after depth 10-15 (This can be attributed to the aggressive terminating conditions applied on grow_tree / fit function)

ii. The following Accuracy vs depth graph was obtained

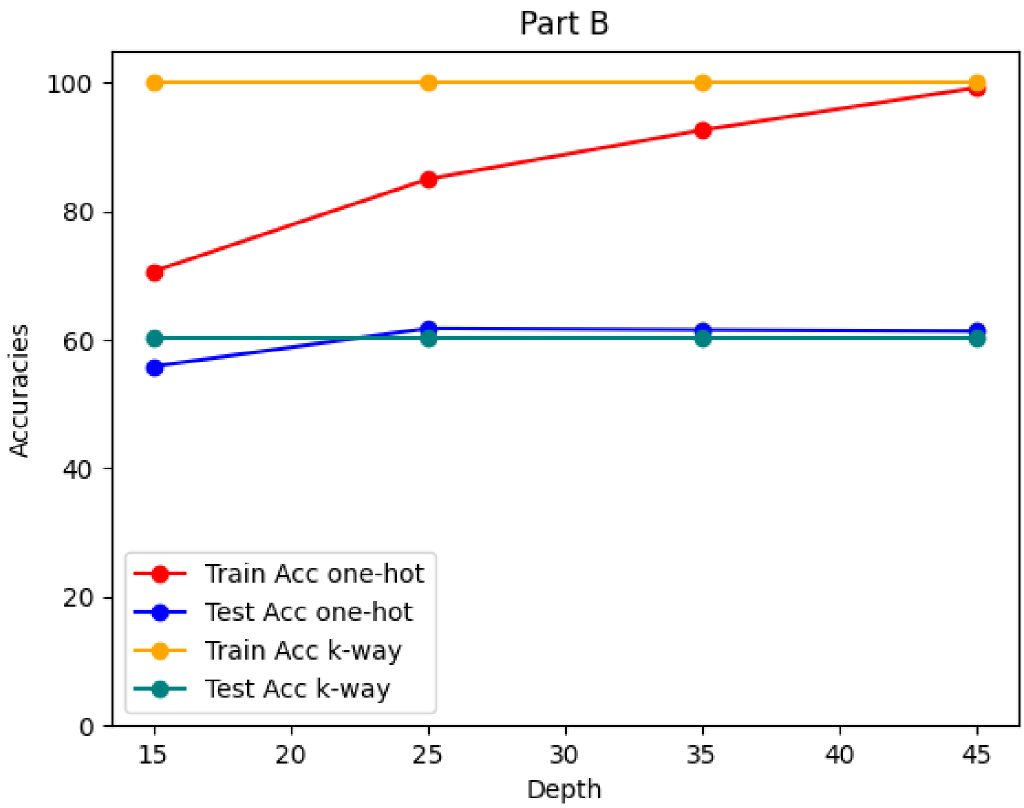


(b) Using one-hot encoding we obtain the following results

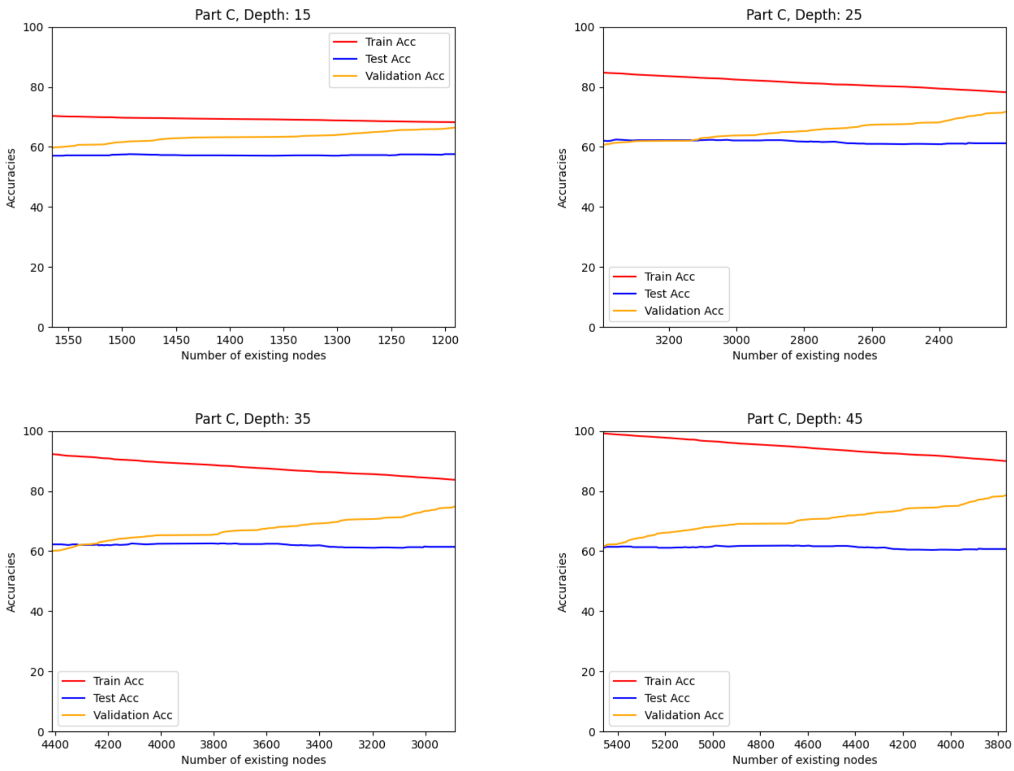
i.
- Only Win:
  - Accuracy for in prediction type of only win on training set is 50.3386
  - Accuracy for in prediction type of only win on test set is 49.6381
- Only Loss
  - Accuracy for in prediction type of only loss on training set is 49.6614
  - Accuracy for in prediction type of only loss on test set is 50.3619
- DT with varying depths on training set:
  - Accuracy for depth 15 on training set is 70.59
  - Accuracy for depth 25 on training set is 84.96
  - Accuracy for depth 35 on training set is 92.60
  - Accuracy for depth 45 on training set is 99.18
- DT with varying depths on test set:
  - Accuracy for depth 15 on test set is 55.84
  - Accuracy for depth 25 on test set is 61.74
  - Accuracy for depth 35 on test set is 61.53
  - Accuracy for depth 45 on test set is 61.32

From the data obtained we find that Decision Tree classification performs better compared to single type prediction(only win, only loss). As we expect the training accuracy is much better than test accuracy for a given depth. However contrary to part (a) we find that here the accuracies significantly increase as we increase the depth for the case of training set and for the test set, it increases from 15 to 25, decreases from 25 to 35 and then remains same for the next increment of depth.

ii. The following Accuracy vs depth graph was obtained



(c) The following Accuracy vs nodes graphs were obtained upon performing reduced error pruning for various depths



Some observations to note:

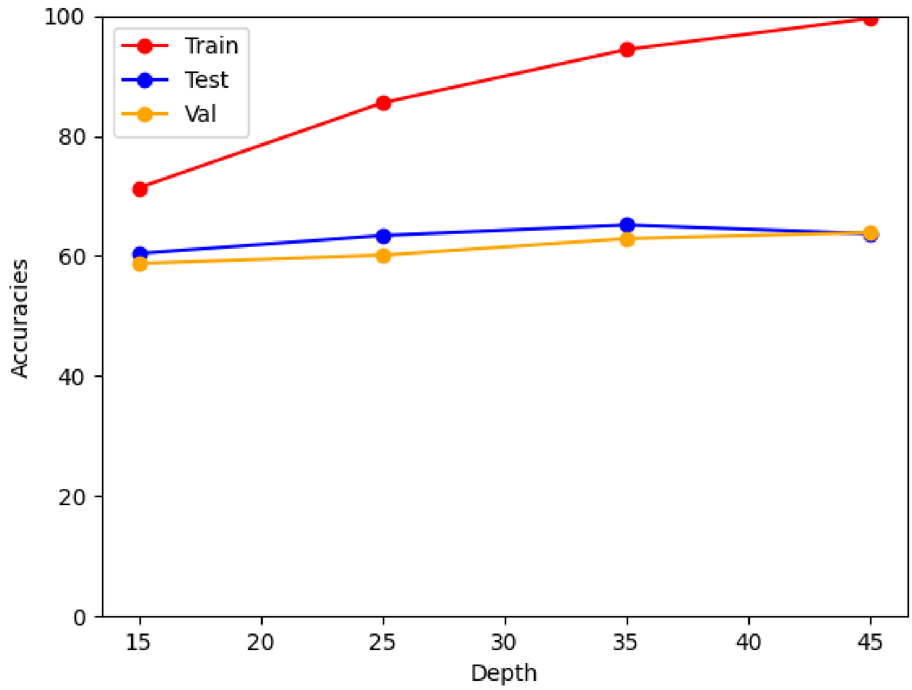
- In all graphs the training accuracy decreases as number of nodes reduce in the graph (which suggests that more nodes could possibly lead to overfitting)
- Both the validation increases as number of nodes reduce in the graph
- we find that the test accuracies remain around the same value throughout the process
- In the initial stage of the DT (before pruning) we find the following order of accuracies:
  - For depth 15: train > val > test accuracy
  - For all other depths: train > test > val accuracy
- As more nodes are pruned we find the following order of accuracies: train > val > test

(d) Decision Tree using sci-kit learn

i. Varying Max-Depth

- Training Set Accuracies:
  - Training Accuracy for max_depth = 15 is 71.32
  - Training Accuracy for max_depth = 25 is 85.49
  - Training Accuracy for max_depth = 35 is 94.35
  - Training Accuracy for max_depth = 45 is 99.51
- Test Set Accuracies:
  - Test Accuracy for max_depth = 15 is 60.29
  - Test Accuracy for max_depth = 25 is 63.81
  - Test Accuracy for max_depth = 35 is 65.15
  - Test Accuracy for max_depth = 45 is 64.12
- Validation Set Accuracies:
  - Validation Accuracy for max_depth = 15 is 58.16
  - Validation Accuracy for max_depth = 25 is 61.61
  - Validation Accuracy for max_depth = 35 is 62.53
  - Validation Accuracy for max_depth = 45 is 61.95

The obtained graph is as follows:

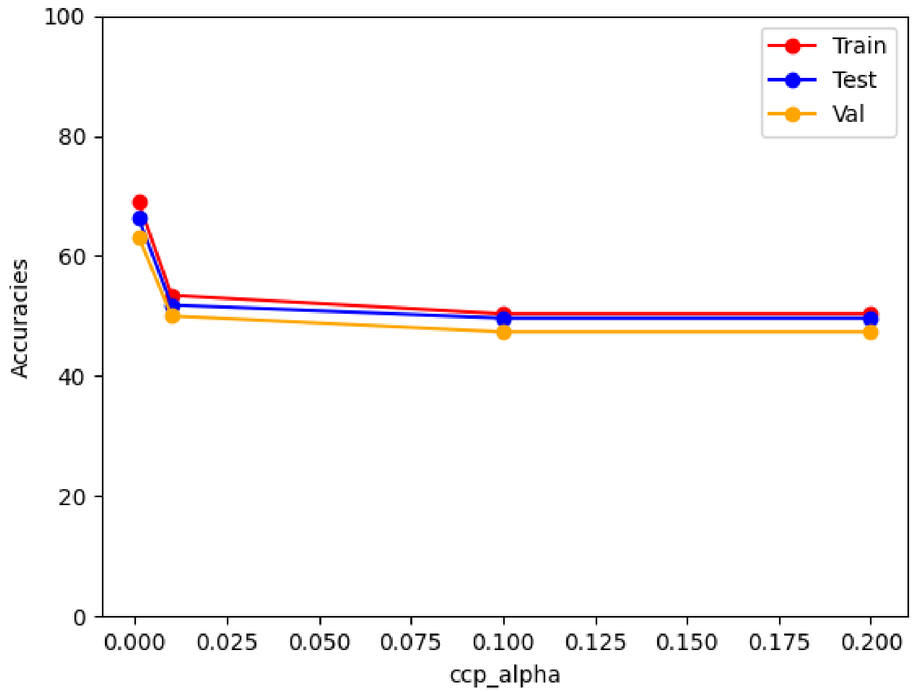


We find that the best max_depth obtained using the validation set is 35

ii. Varying ccp_alpha

- Training Set Accuracies:
  - Training Accuracy for ccp_alpha = 0.001 is 68.94
  - Training Accuracy for ccp_alpha = 0.01 is 53.44
  - Training Accuracy for ccp_alpha = 0.1 is 50.33
  - Training Accuracy for ccp_alpha = 0.2 is 50.33
- Test Set Accuracies:
  - Test Accuracy for ccp_alpha = 0.001 is 66.29
  - Test Accuracy for ccp_alpha = 0.01 is 51.81
  - Test Accuracy for ccp_alpha = 0.1 is 49.64
  - Test Accuracy for ccp_alpha = 0.2 is 49.64
- Validation Set Accuracies:
  - Validation Accuracy for max_depth = 15 is 63.22
  - Validation Accuracy for max_depth = 25 is 50.00
  - Validation Accuracy for max_depth = 35 is 47.36
  - Validation Accuracy for max_depth = 45 is 47.36

The obtained graph is as follows:



We find that the best ccp_alpha obtained using the validation set is 0.001

iii. Observations to note:

- We find that the training data prediction accuracy is lesser in the sci-kit learn model compared to the model developed in both part b and c
- On the other hand we find that the test data prediction accuracy is higher in the sci-kit model as compared to the model developed in both b and c

(e) Random Forests: Using out of box accuracies and grid search over the parameter space we observe the following result:

- Best Parameters:
  - max_features: 0.7
  - min_samples_split: 8
  - n_estimators: 150
- Out of Box Accuracy: 71.8922
- Training Accuracy: 98.7990
- Test Accuracy: 71.7684
- Validation Accuracy: 69.5402

Compared to the previous parts, we find that the training, test and validation accuracy obtained here is higher than what was obtained in part d. However validation accuracy is lesser than what was obtained in part c (Training and Test accuracy are higher)

2 Neural Networks

(a) A general Neural network architecture was created with configurable parameters such as:

- Mini-batch size: M
- Number of features: n
- Hidden layer architecture: [h1, h2, h3, ...]
- Target labels: r (number of target labels [1, 2, 3, ... r])
- Activation function: sigmoid, relu
- Learning rate, η: constant, adaptive

The terminating condition was fixed as 200 epochs.
Incomplete: I was able to make the structure in a way that it runs, but despite lot of efforts trying to debug, my accuracies weren't increasing and were changing around randomly between 10% to 30% for every epoch, and only predicting one class for all entries in a set of inputs for an epoch.

(b) Incomplete

(c) Incomplete

(d) Incomplete

(e) Incomplete

(f) MLPClassifier The following tables were obtained for various number of layers:

i. 1 layer

Class	Precision	Recall	F1Score	Support
1.0	0.70	0.91	0.79	1971
2.0	0.52	0.44	0.48	1978
3.0	0.47	0.34	0.40	1952
4.0	0.44	0.33	0.38	2008
5.0	0.58	0.79	0.67	2091

Table 1: train

Class	Precision	Recall	F1Score	Support
1	0.91	0.90	0.74	230
2	0.67	0.70	0.69	190
3	0.55	0.58	0.57	190
4	0.57	0.49	0.53	218
5	0.66	0.72	0.69	172

Table 2: test

ii. 2 layers

Class	Precision	Recall	F1Score	Support
1.0	0.76	0.90	0.83	1971
2.0	0.58	0.53	0.55	1978
3.0	0.46	0.41	0.43	1952
4.0	0.46	0.36	0.40	2008
5.0	0.63	0.78	0.70	2091

Table 3: train

Class	Precision	Recall	F1Score	Support
1	0.93	0.90	0.71	236
2	0.71	0.68	0.70	206
3	0.52	0.58	0.55	178
4	0.56	0.49	0.52	211
5	0.65	0.72	0.68	169

Table 4: test

iii. 3 layers

Class	Precision	Recall	F1Score	Support
1.0	0.80	0.89	0.84	1971
2.0	0.61	0.58	0.59	1978
3.0	0.50	0.44	0.47	1952
4.0	0.46	0.42	0.44	2008
5.0	0.65	0.74	0.69	2091

Table 5: train

Class	Precision	Recall	F1Score	Support
1	0.90	0.92	0.91	223
2	0.70	0.71	0.70	197
3	0.54	0.61	0.57	175
4	0.47	0.50	0.48	176
5	0.80	0.66	0.72	229

Table 6: test

iv. 4 layers

Class	Precision	Recall	F1Score	Support
1.0	0.82	0.88	0.85	1971
2.0	0.62	0.62	0.62	1978
3.0	0.50	0.45	0.47	1952
4.0	0.46	0.43	0.45	2008
5.0	0.66	0.73	0.69	2091

Table 7: train

Class	Precision	Recall	F1Score	Support
1	0.86	0.93	0.90	212
2	0.66	0.70	0.68	187
3	0.51	0.59	0.55	174
4	0.51	0.49	0.50	195
5	0.81	0.66	0.73	232

Table 8: test

We obtain the following graph: