COL 703: Assignment 3

Machine Learning

Part 1: Decision Trees

Command to run is: ./bash run.sh 1 <trainfile> <testfile> <validationfile> <a/b/c/d>

a)

Decision trees were implemented. Best attribute was chosen using the mutual information criteria.

The accuracies obtained using multi class split are:

Accuracy on train set: 0.9996682149966821

Accuracy on test set: 0.8710462287104623

Accuracy on validation set: 0.8728438743918621

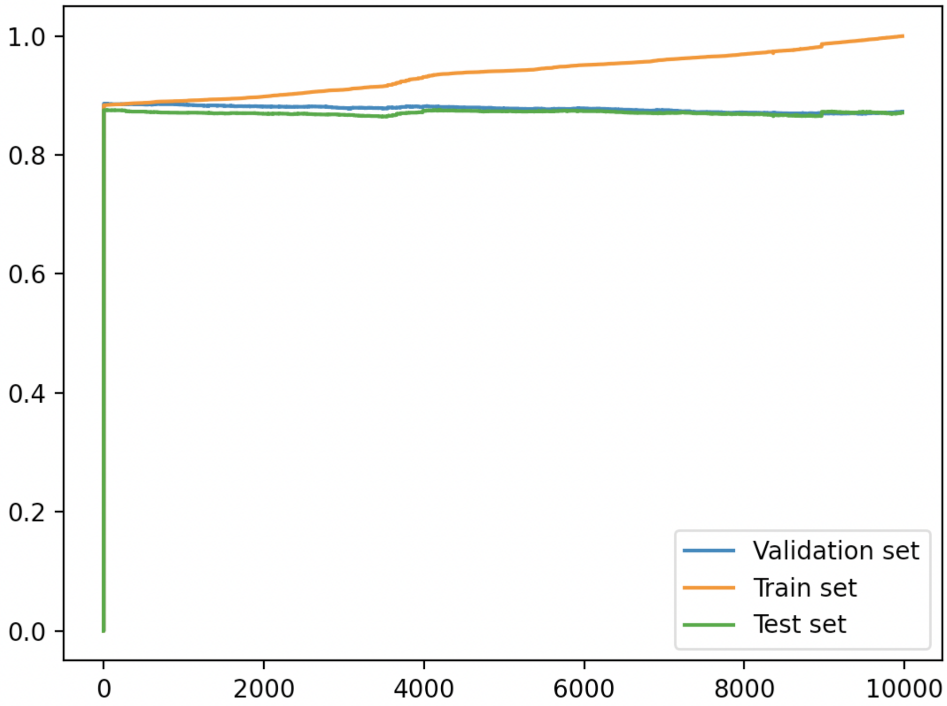
The accuracies obtained using one hot encoding are:

Accuracy on train set: 0.9983963724839637

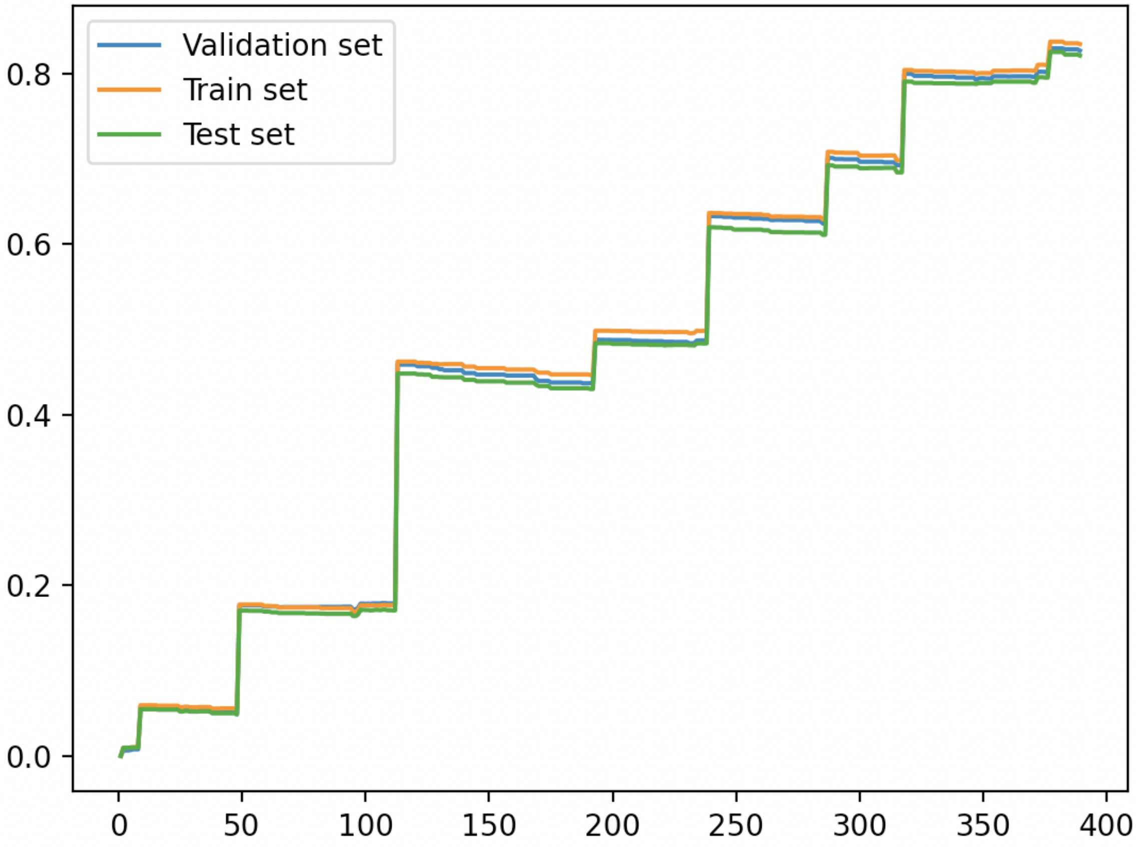
Accuracy on test set: 0.836098208360982

Accuracy on validation set: 0.8505086245024326

Number of nodes and training time is more in case of one hot encoding as features are increased a lot. Multi class split had ~10000 nodes whereas one hot encoding generated ~12000 nodes.



It looks like the model is learning very fast, the train set accuracy is maximum and test set and validation accuracy is ~90%. If the model is trained with a subset of training data, then we get the following graph on training data, test data and validation data. This shows that the model is learning step by step. Accuracies are increasing suddenly because even one split in the right direction makes a lot of difference.



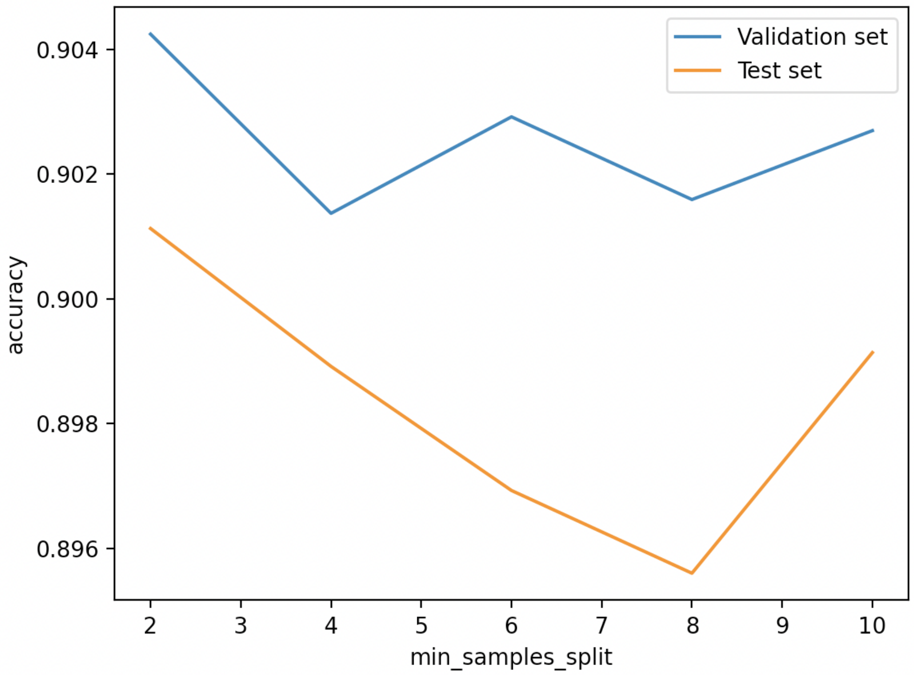
b)
Pruning was done iteratively on each node.
The accuracies obtained after pruning are:
Accuracy on train set after pruning: 0.9146759566467596
Accuracy on test set after pruning: 0.8949347489493474
Accuracy on validation set after pruning: 0.9279080053073862

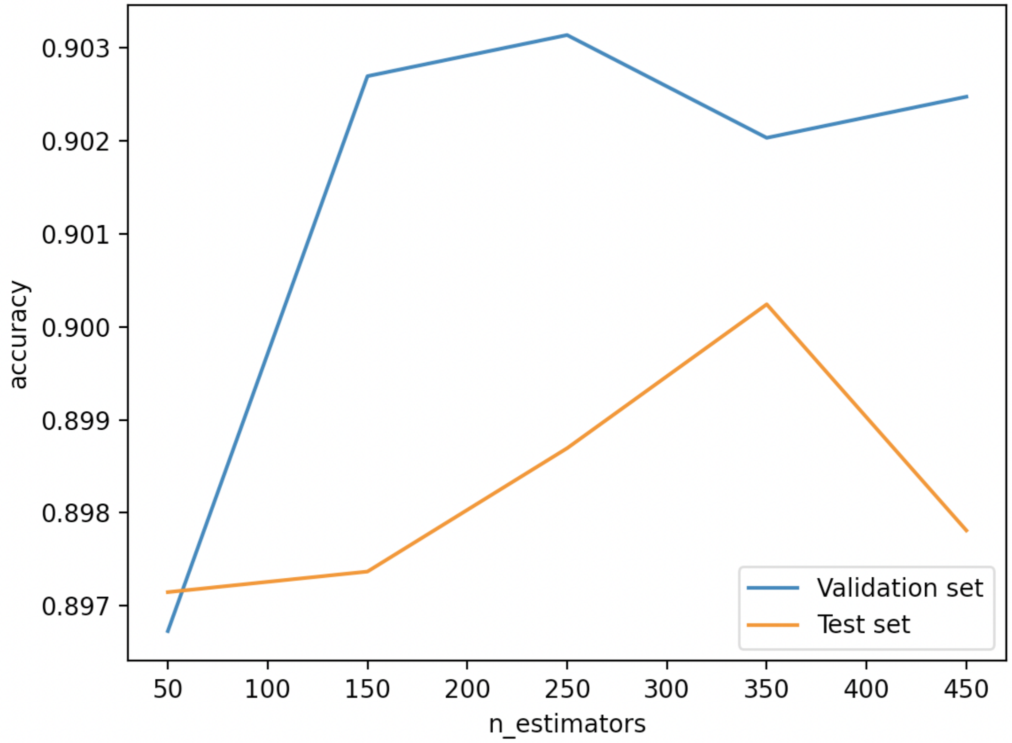
It can be seen that the training set accuracy has decreased which is expected, because now a lot of nodes have been pruned which were initially giving exact labels for training data. Accuracy on validation set and test set have increased which was the goal.

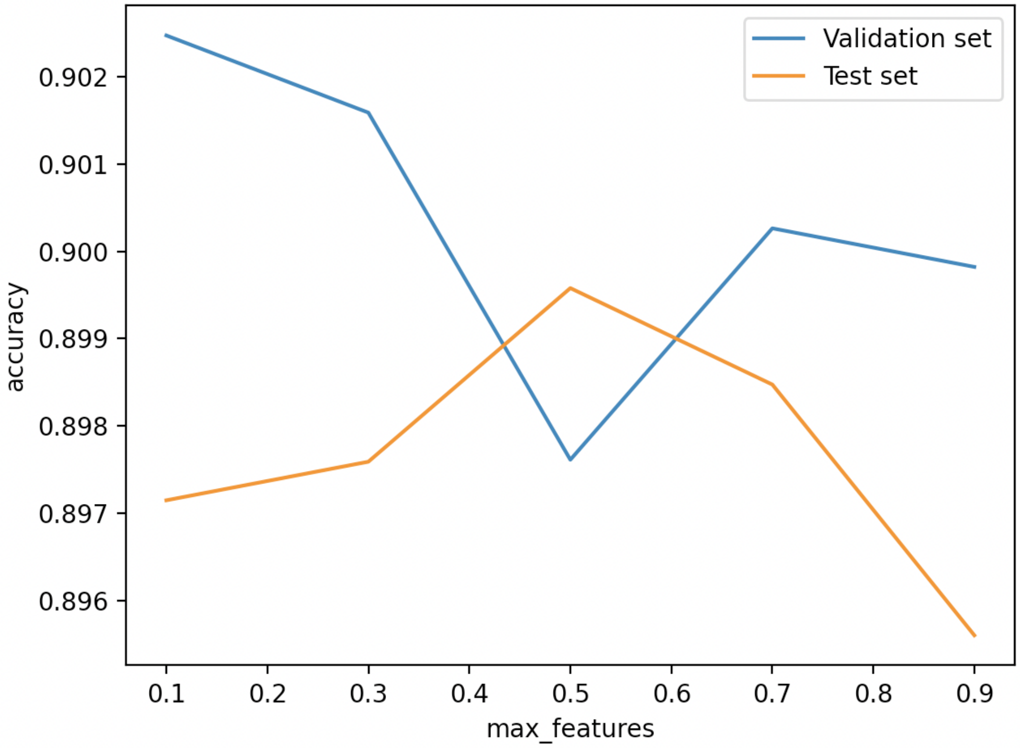
c)
Best parameters obtained are:
n_estimators = 350
max_features = 0.3
min_samples_split = 10

Out-of-bag: 0.9045012165450121
Train: 0.9802864410528644
Validation: 0.9020344980097302
Test: 0.9013492590134926

d)
The graphs for various parameters are attached below. Our model is most sensitive to min_samples_split.




0.903
0.902
0.901
0.900
0.899
0.898
0.897
accuracy
50
100
150
200
250
300
350
400
450
n_estimators
Validation set
Test set


0.902
0.901
0.900
0.899
0.898
0.897
0.896
accuracy
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
max_features
Validation set
Test set

Part 2: Neural Networks

Command to run is: ./bash run.sh 2 <trainfile> <testfile> <a/b/c/d/e/f>
Parts b-e have not been implemented

a)
One hot encoding was implemented from scratch without using the pandas library. The resultant data would be saved in a file named fileencoded.csv, with an input file.

f)
The data observed was:
Training time: 65.46468782424927s
Train accuracy: 0.7175529788084766
Test accuracy: 0.665523