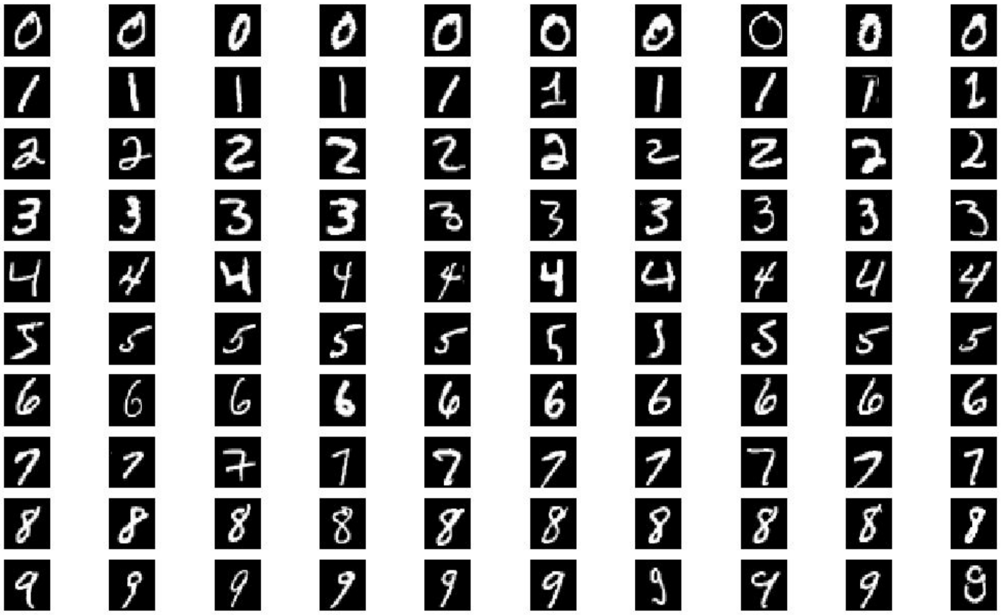
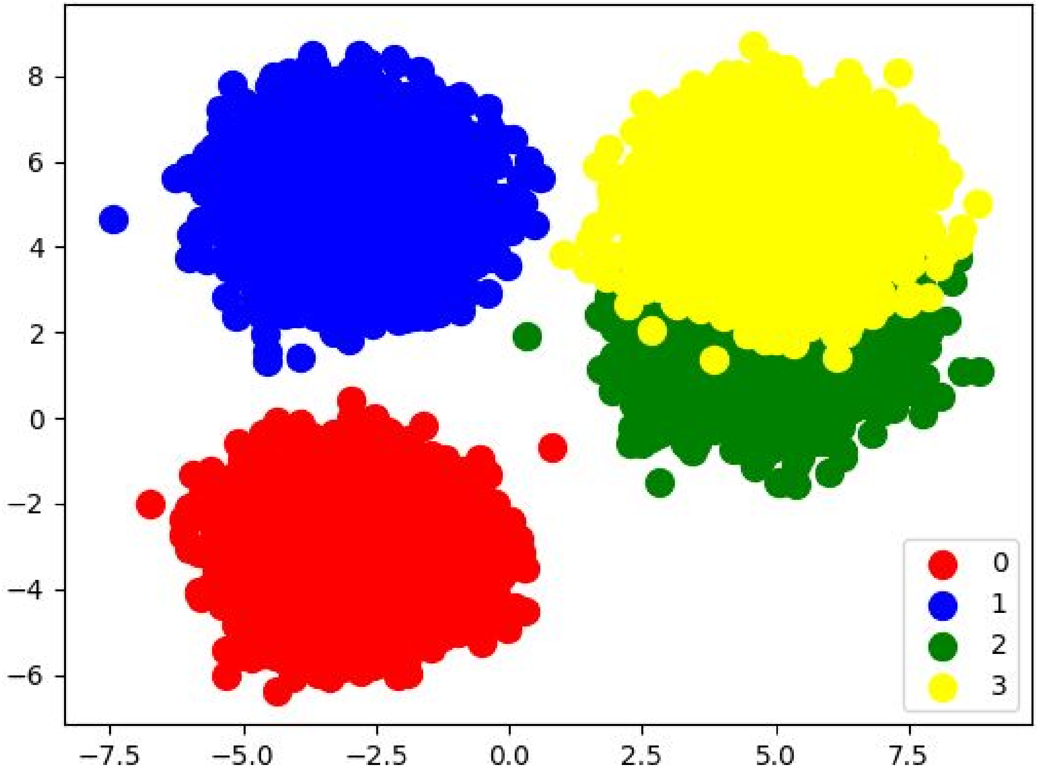
Ques. 1. a)

The given dataset ‘dataset_1’ is MNIST handwritten digits dataset consisting of 28*28 (number of features) sized images of handwritten digits 0 to 9. The dataset consists of 50000 images.



Ques. 1. b)

The given dataset ‘dataset_2’ consists of 20000 samples having 2 features and 4 classes. The scatter plot shown below shows the distribution of the samples. It can be inferred that out of the total 4 classes present in the dataset, 2 classes are well separated while 2 classes are having some overlapped regions.



Ques. 1. c)

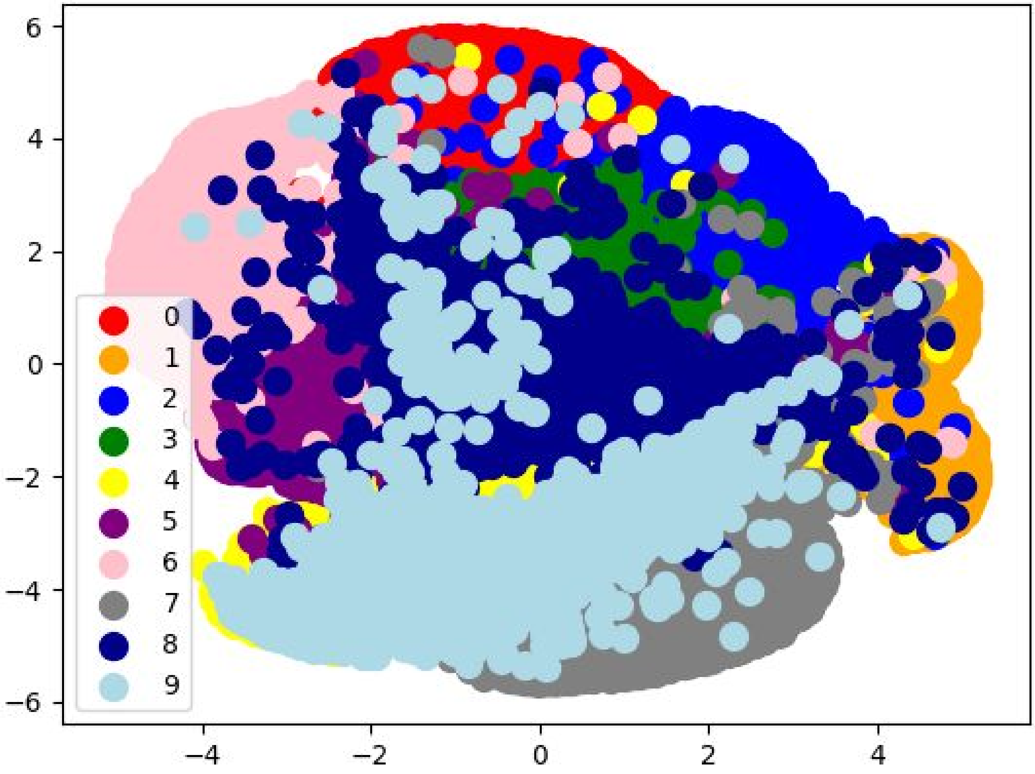
The given dataset ‘dataset_1’ consists of 50000 images of handwritten digits 0 to 9 (28*28 = 784 pixels).

Using the tSNE plot, the dimensions of the given dataset have been reduced to 2 dimensions and plotted.

The scatter plot shown below shows the distribution of the samples belonging to 10 different classes. It can be inferred that the classes are not well separated from each other and they have overlapping regions.

For example: Class 9 (Light blue) and Class 4 (Yellow) are very overlapped which can be verified by the structure of digit 9 and 4 which is quite similar.

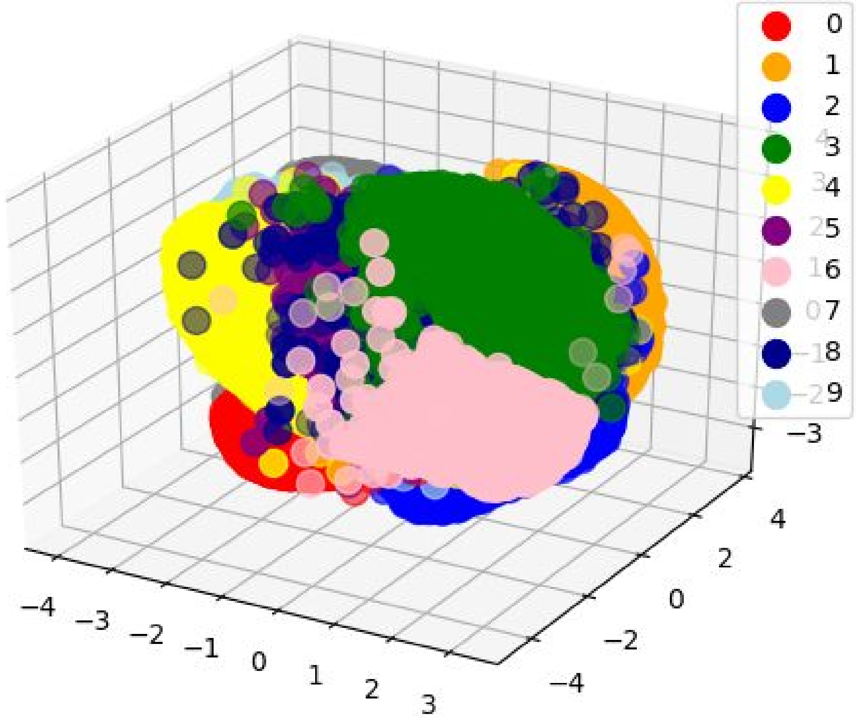
Class 6 (Light pink on extreme left) and 1 (orange on extreme right) are well separated classes.



Ques. 1. d)

The given dataset ‘dataset_1’ consists of 50000 images of handwritten digits 0 to 9 (28*28 = 784 pixels).
Using the tSNE plot, the dimensions of the given dataset have been reduced to 3 dimensions and plotted.
The scatter plot shown below shows the distribution of the samples belonging to 10 different classes. It can be inferred that the classes are not well separated from each other and they have some overlapping region.
For example: Class 6 (Light pink) and Class 8 (blue) are very overlapped which can be verified by the structure of digit 6 and 8 which is quite similar.

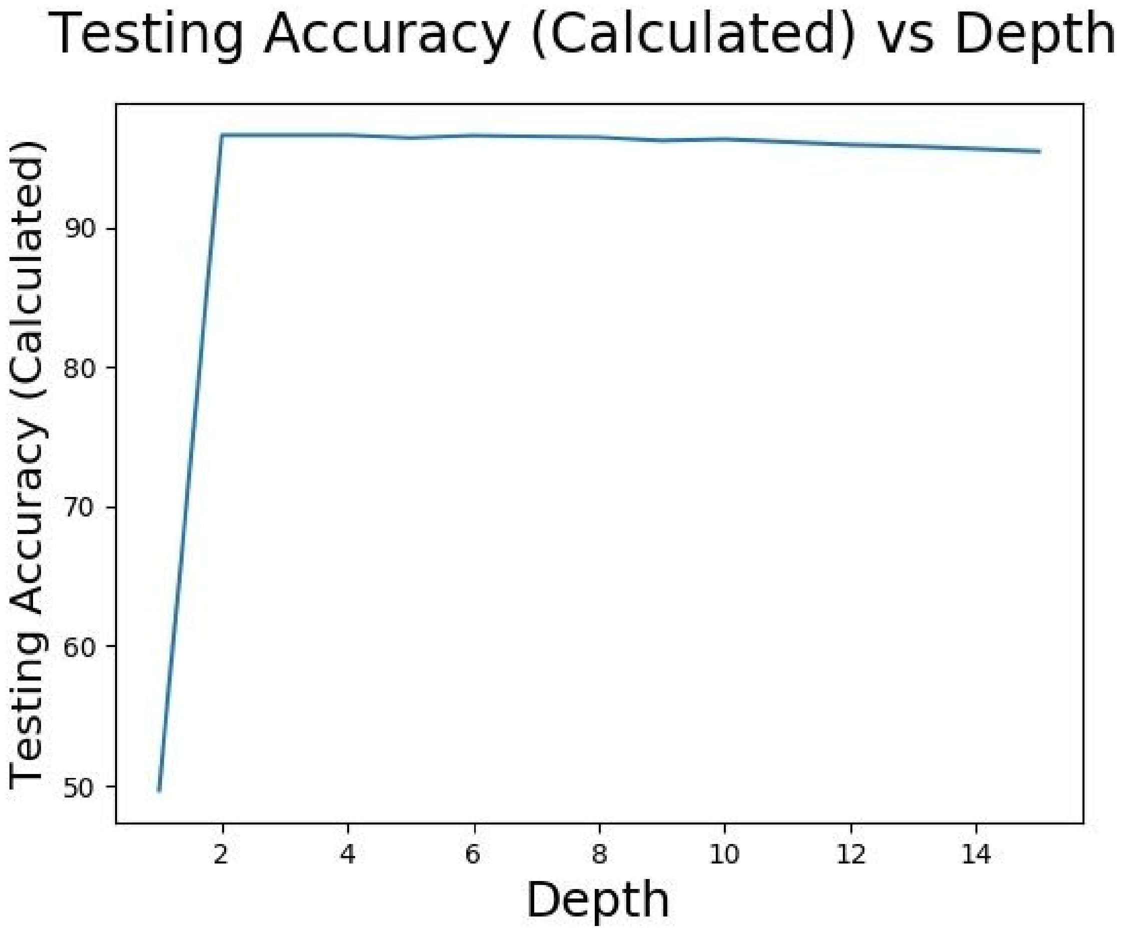
The part c) scatter plot gives more clarity on the extent of overlapping between different classes while part d) plot gives compact view and even some classes are not visible in this view (e.g. class 7 and 9).



Ques. 2. a)

The grid search on the hyperparameter “depth” of the decision tree has been performed for 18 values (1 to 15, 20, 25, 30) of depth.
The optimal depth value comes out to be **2** achieving the best testing accuracy of **96.63.**

As can be inferred from the plot between testing accuracy and depth shown below, the depth > 2 does not have any significant impact on the performance of the decision tree classifier. Infact, after increasing the depth above 4, the test accuracy starts decreasing slightly. The reason can be inferred as the given dataset has only 4 features.
Yes, the best performance is consistent with the distribution of samples of dataset_2 shown in Q1b) as the four classes could not be classified with a linear decision boundary (depth=1). We can see the decision tree gives 50% accuracy with depth 1 which is consistent with the distribution of classes. The best accuracy achieved is 96.69 as two classes have overlapping region in the distribution.



Ques. 2. b)

Depth of Decision Tree	Training Accuracy	Testing Accuracy
1	50.14	49.67
2	96.69	96.63
3	96.69	96.63
4	96.7	96.62
5	96.79	96.43
6	96.94	96.58
7	97.04	96.53
8	97.27	96.48
9	97.46	96.23
10	97.7	96.3
11	97.97	96.22
12	98.17	96.07
13	98.41	95.82
14	98.63	95.6
15	98.81	95.57
20	99.6	95.22
25	98.95	95.27
30	100	95.1

From the table, it can be seen that with increasing depth of the decision tree, training accuracy increases and achieves 100% accuracy at depth = 30. On the other hand, testing accuracy at depth 2 is 96.63 while at depth 30, testing accuracy is 95.1, this means that at depth 30, decision tree is overfitting.

At depth 1, decision tree is underfitting as the training and testing accuracy both are around 50%.

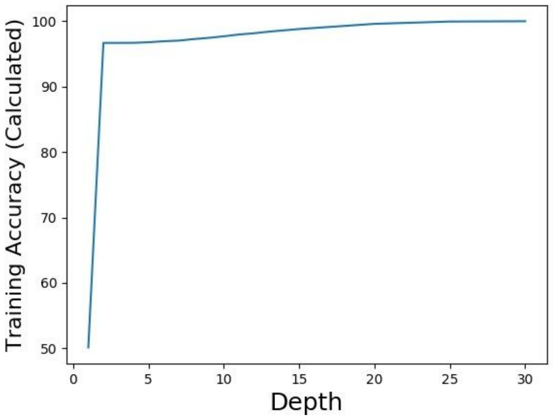
With increasing depth >=3, decision trees are overfitting as with increase in depth, the training accuracy increases while testing accuracy decreases.

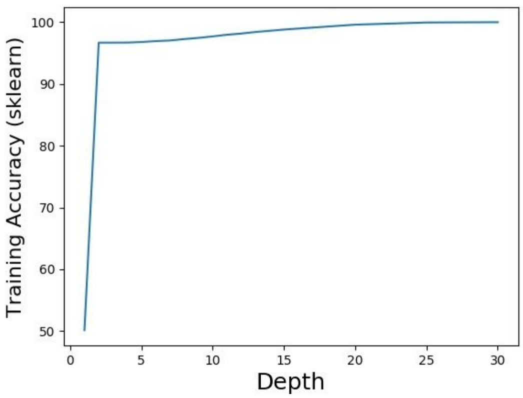
Ques.2. c)

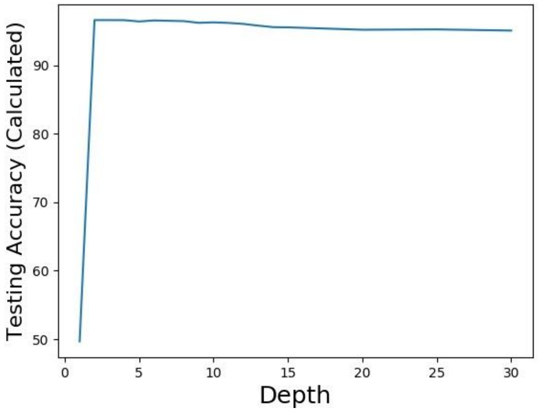
The part b is replicated using sklearn “accuracy” function. The table is shown below. There was no deviation between the results of accuracy implemented function and sklearn accuracy function.

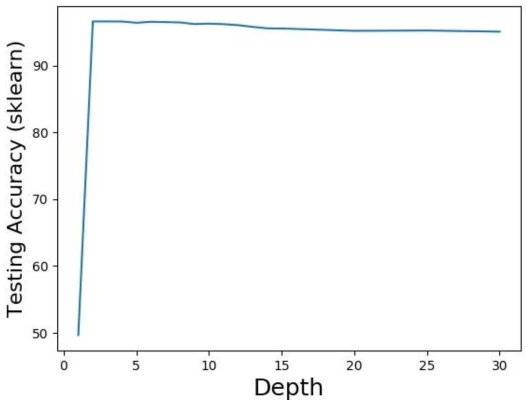
Depth of Decision Tree	Training Accuracy	Testing Accuracy
1	50.14	49.67
2	96.69	96.63
3	96.69	96.63
4	96.7	96.62
5	96.79	96.43
6	96.94	96.58
7	97.04	96.53
8	97.27	96.48
9	97.46	96.23
10	97.7	96.3
11	97.97	96.22
12	98.17	96.07
13	98.41	95.82
14	98.63	95.6
15	98.81	95.57
20	99.6	95.22
25	98.95	95.27
30	100	95.1

Adding the training and testing accuracy plots as well.


Training Accuracy (Calculated) vs Depth


Training Accuracy (sklearn) vs Depth


Testing Accuracy (Calculated) vs Depth


Testing Accuracy (sklearn) vs Depth

Ques. 3. a)

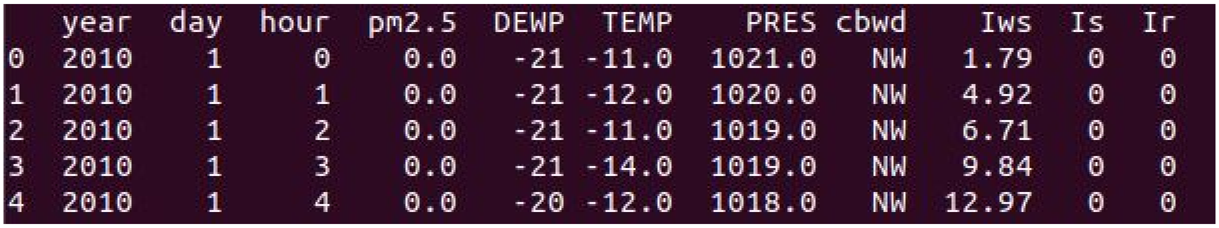
The dataset contains 43824 samples with 11 features.

Preprocessing:

1. “No” column is removed as it is index.
2. “Month” column is used as target class containing 12 different values 1….12 (i.e. 12 classes).
3. Missing values marked as NaN have been filled with 0.
4. “cbwd” column values have been converted to numeric values:

```
'NW':0, 'cv':1, 'NE':2, 'SE':3
```

Dataset after preprocessing:



```
   year  day  hour  pm2.5  DEWP  TEMP    PRES cbwd    Iws  Is  Ir
0  2010    1     0    0.0   -21 -11.0  1021.0   NW   1.79   0   0
1  2010    1     1    0.0   -21 -12.0  1020.0   NW   4.92   0   0
2  2010    1     2    0.0   -21 -11.0  1019.0   NW   6.71   0   0
3  2010    1     3    0.0   -21 -14.0  1019.0   NW   9.84   0   0
4  2010    1     4    0.0   -20 -12.0  1018.0   NW  12.97   0   0
```

Dataset split

The dataset is splitted into 80:20 ratio (training:testing)

('X shape ', (43824, 11))

('Y shape ', (43824,))

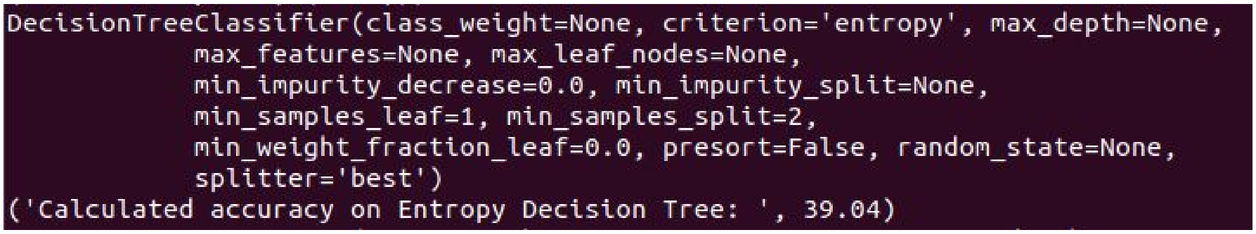
('X train shape: ', (35059, 11))

('X test shape: ', (8765, 11))

('Y train shape: ', (35059,))

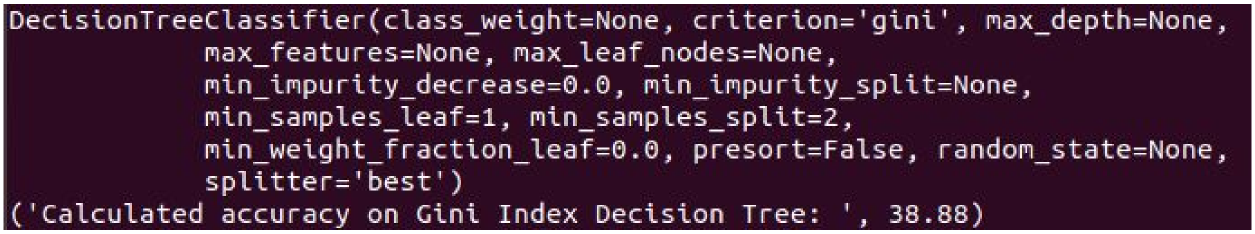
('Y test shape: ', (8765,))

a) **Entropy criteria based decision tree**



```
DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
('Calculated accuracy on Entropy Decision Tree: ', 39.04)
```

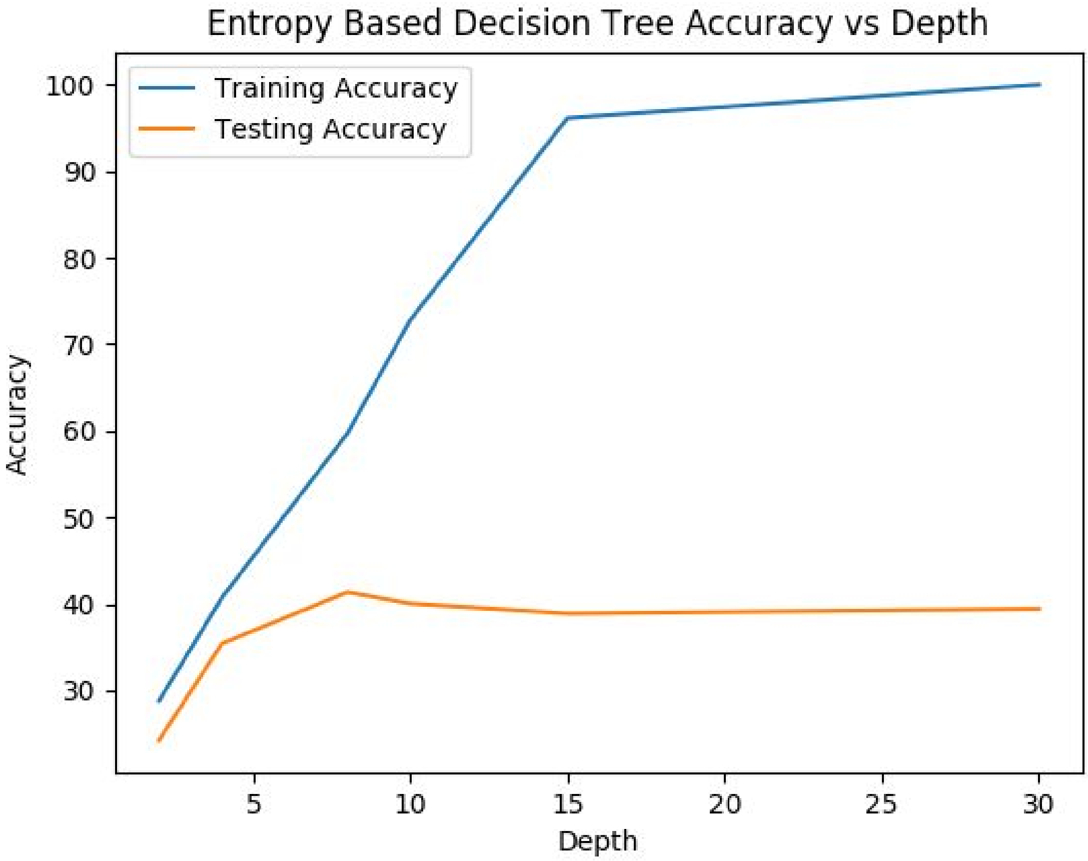
Gini index criteria based decision tree



```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')
('Calculated accuracy on Gini Index Decision Tree: ', 38.88)
```

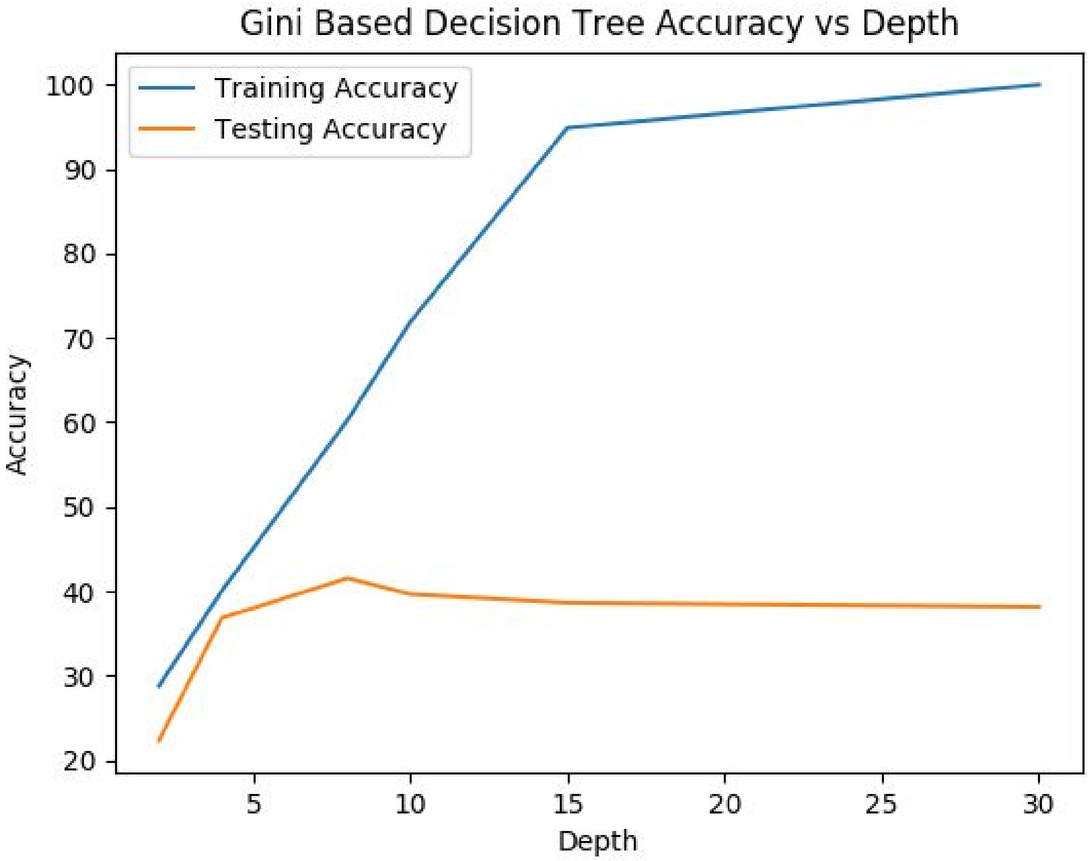
b) Entropy criteria based decision tree

Depth of Decision Tree	Training Accuracy	Testing Accuracy
2	28.8	24.24
4	40.69	35.4
8	59.7	41.35
10	72.8	40.26
15	96.13	39.26
30	100	38.78



Gini index criteria based decision tree

Depth of Decision Tree	Training Accuracy	Testing Accuracy
2	28.8	22.35
4	40.03	36.83
8	60.26	41.52
10	71.89	39.78
15	94.89	38.86
30	100	37.89



Best testing accuracy of entropy based Decision tree 41.35 at optimal depth of 8

Best testing accuracy of Gini based Decision tree 41.52 at optimal depth of 8

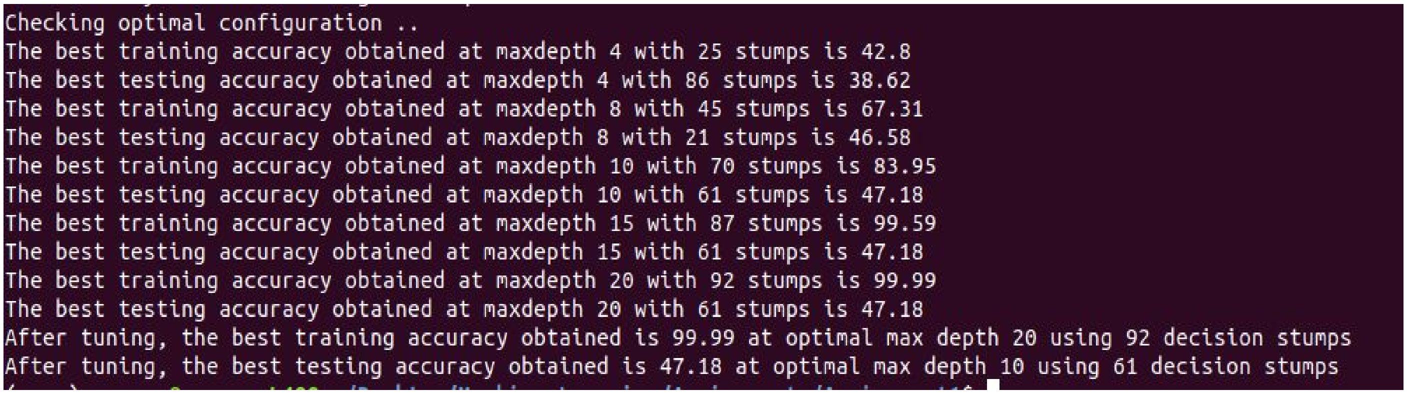
Ques.3. c)

'The training accuracy after ensembling technique: ', **35.94**
'The testing accuracy after ensembling technique: ', **31.83**

The testing accuracy in **part a) is 38-39%** while in **part b), the testing accuracy is 41-42%**. In **part c), after ensembling using 100 decision stumps, the testing accuracy is 31.83%** which is significantly less than the testing accuracy of part a) and b).

Ques. 3. d)

After tuning, the best testing accuracy obtained is **47.18** at optimal **max depth 10 using 61 decision stumps.**



```
Checking optimal configuration ..
The best training accuracy obtained at maxdepth 4 with 25 stumps is 42.8
The best testing accuracy obtained at maxdepth 4 with 86 stumps is 38.62
The best training accuracy obtained at maxdepth 8 with 45 stumps is 67.31
The best testing accuracy obtained at maxdepth 8 with 21 stumps is 46.58
The best training accuracy obtained at maxdepth 10 with 70 stumps is 83.95
The best testing accuracy obtained at maxdepth 10 with 61 stumps is 47.18
The best training accuracy obtained at maxdepth 15 with 87 stumps is 99.59
The best testing accuracy obtained at maxdepth 15 with 61 stumps is 47.18
The best training accuracy obtained at maxdepth 20 with 92 stumps is 99.99
The best testing accuracy obtained at maxdepth 20 with 61 stumps is 47.18
After tuning, the best training accuracy obtained is 99.99 at optimal max depth 20 using 92 decision stumps
After tuning, the best testing accuracy obtained is 47.18 at optimal max depth 10 using 61 decision stumps
```

After tuning the depths and number of decision stumps, the best testing accuracy achieved is **47.18%** which is a significant increase than the testing accuracy of part a), b) and c).
It can be easily inferred that at depth 20, training accuracy reached 99.99% while testing accuracy remains 47.18% i.e. the model is overfititng.

Comparison of Classification Models

1. In part a), we implement Decision tree classifier with default parameters:
   For criteria Entropy, testing accuracy is 39.04
   For criteria Gini index, testing accuracy is 38.88

2. In part b), we tune the depth of decision tree for both the criteria - Entropy and Gini index
   For Entropy based Decision tree, testing accuracy is 41.35 at optimal depth of 8
   For Gini index based Decision tree, testing accuracy is 41.52 at optimal depth of 8

3. In part c), with ensembling technique using 100 decision stumps trained over 50% random training data, testing accuracy is 31.83

4. In part d), after tuning the depth and number of decision stumps, the best testing accuracy is 47.18 using 61 decision stumps of max depth 10.

<u>RankWise Decision Tree Classifier</u>

1. Part d) classifier : 47.18 (testing accuracy)
2. Part b) classifier : 41.38/41.52 (testing accuracy)
3. Part a) classifier : 39.04/38.88 (testing accuracy)
4. Part c) classifier : 31.83 (testing accuracy)