Question 1

Process Followed and Plots:

1. Import required libraries like pandas, numpy, sklearn and their componants.
2. Get Training and Testing Data from the provided arff files i.e 'training_subsetD.arff' and 'testingD.arff'
3. Pre-processing the data
   3.1 Handling NULL and NaN values
   Since the data set contain many Null and None values, their handling is necessary in getting correct picture of the classification. Technique used for this process is
   1. If Null or NaN value count of any column is more tha 50% of total data points then the column is dropped.
   2. If Null or NaN value count is less than 50%, then the data points are estimated by putting the mode of data points of that specific column.
   Till this point the shape of training data frame is 40000*139 and testing data frame is 25000*139.

   3.2 Labelling

   Since the data available is in the categorical form, it is necessary to convert this into numerical labels. This is done with the help of LabelEncoder of sklearn's preprocessing.

4. Division of Training and Test data into two categories where all columns except last column is in one dataframe as 'X' and the last column as 'Y'.

   ```
   X_df=df.values[:,0:138]
   Y_df=df.values[:,138]
   X_dftest=df_test.values[:,0:138]
   Y_dftest=df_test.values[:,138]
   ```

5. <u>Decision tree classification on training data and prediction on testing data.</u>

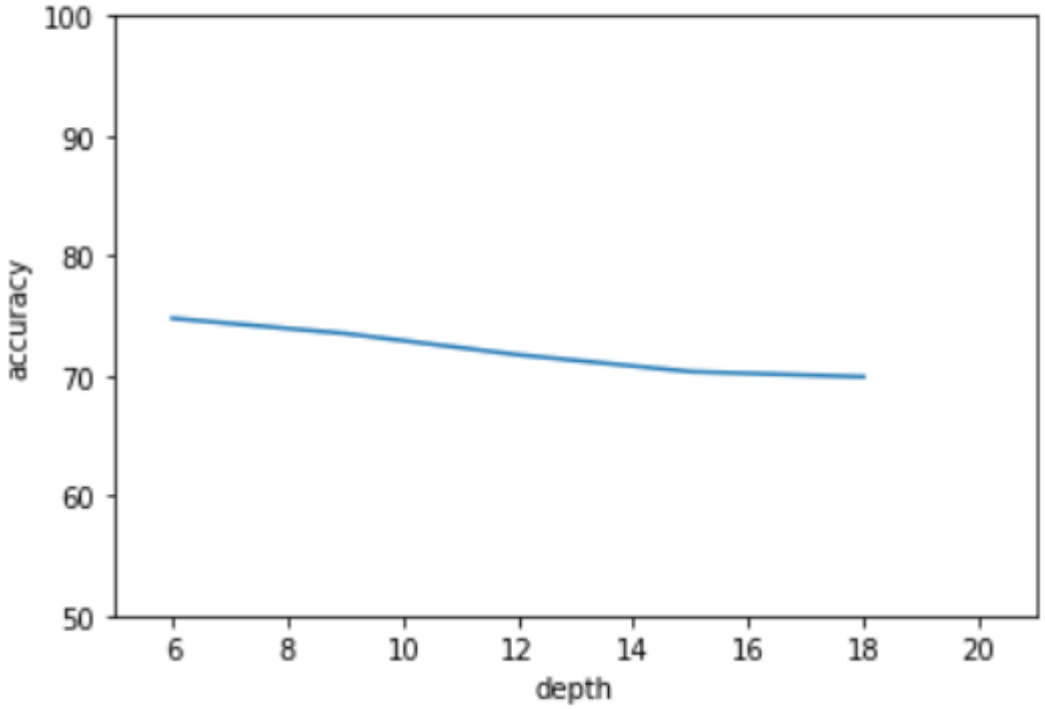
   Decision tree classifier is used which has used 'entropy' as its measure to learn model from training data.

   Then the model is used to predict the last column values of the test data.

The predicted value and original values of last column of the test data are then used to calculate the accuracy.

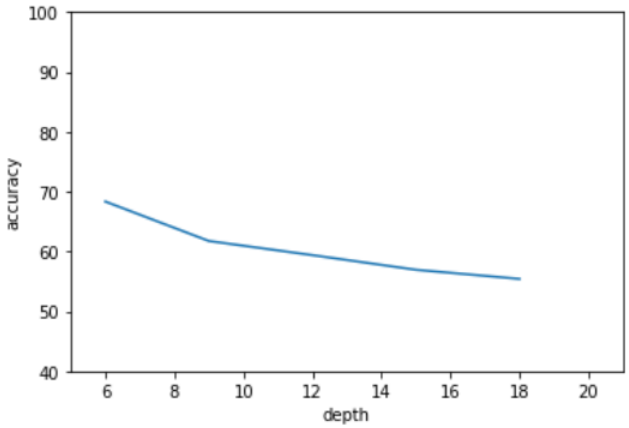
Accuracy for different depth decision tree are stored in a list and the graph between accuracy and depth of the decision tree.

```
[74.78399999999999, 73.52, 71.748, 70.34, 69.908]
[6, 9, 12, 15, 18]
```



6. <u>Decision tree classification using cross validation.</u>
   Firstly the folds are decided K=5.
   Data is split in to 5 divisions and with the loop iteration, one of the 5 was selected as test data and rest as training data.
   Decision tree classifier is applied and accuracy is calculated everytime.
   Mean of accuracy is taken for each decision tree with certain depth during cross validation and stored in a list.
   Graph is plot between depth and mean accuracies.

```
[6, 9, 12, 15, 18]
[68.33809987887656, 61.7450860269154, 59.382700992932826, 56.93786971059171, 55.41285250766176
```



7. Best Decision tree with accuracy is found out which comes as:
   Depth is 6
   Accuracy : 74.78%

Observations and Inferences:

1. When Decision tree classifier is applied on the test data, it was observed that the accuracy was approx. 75% with the depth =6. But when the depth increased iteratively, the accuracy went down with every change increment in the depth of decision tree though the decrement was proportional to the depth.
2. When Decision tree classifier is applied on the training data with 5 fold cross validation, same behaviour of depth and accuracy was observed. But the notable thing is that here the accuracy decrement is form depth=6 to 9 is not proportional.
3. So, It is very clear that the model is headed towards the over fitting due to many reasons:
    a. Noisy data
    b. Unbalanced class problem.
    c. High dimensionality
    d. Bias due to estimation of NaN and Null values