1a.

Please refer down for question 1a.

1b.

The formula for entropy is as follows

$$H(S) \equiv -p_{\oplus} \log_2 p_{\oplus} - p_{\ominus} \log_2 p_{\ominus}$$

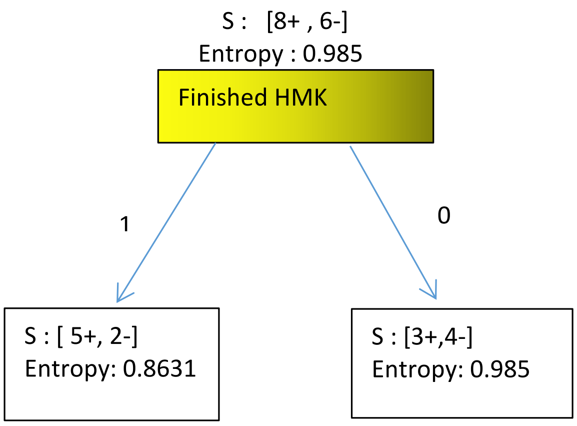
The Formula for Gain is as follows

$$\text{Gain(S,A)} = \text{Entropy(s)} - \sum_{v \in values(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

H(D) = 0.985

H(D / Early) = 0.964 , Gain(D, Early) = 0.021
H(D / FinishedHMK) = 0.924 , Gain(D, FinishedHMK) = **0.061**
H(D / Senior) = 0.973, Gain(D, Senior) = 0.012
H(D / LikesCoffe) = 0.946, Gain(D, LikesCoffee) = 0.039
H(D / LikedLastJedi) = 0.983, Gain(D, LikedLastJedi) = 0.002

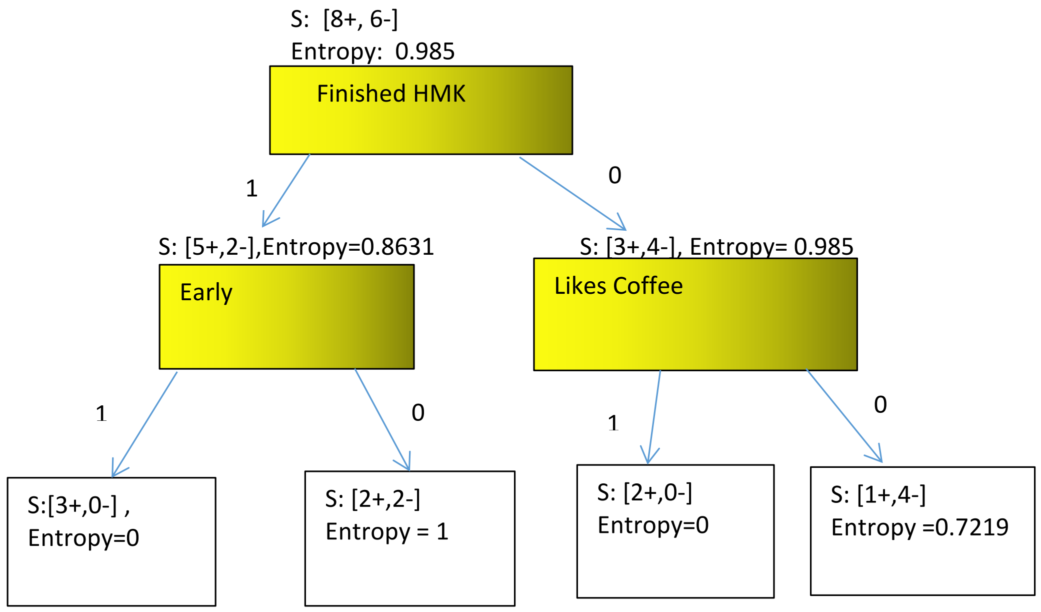
For the Depth 1 we will have graph as follows



For Depth 2 The tree will look like

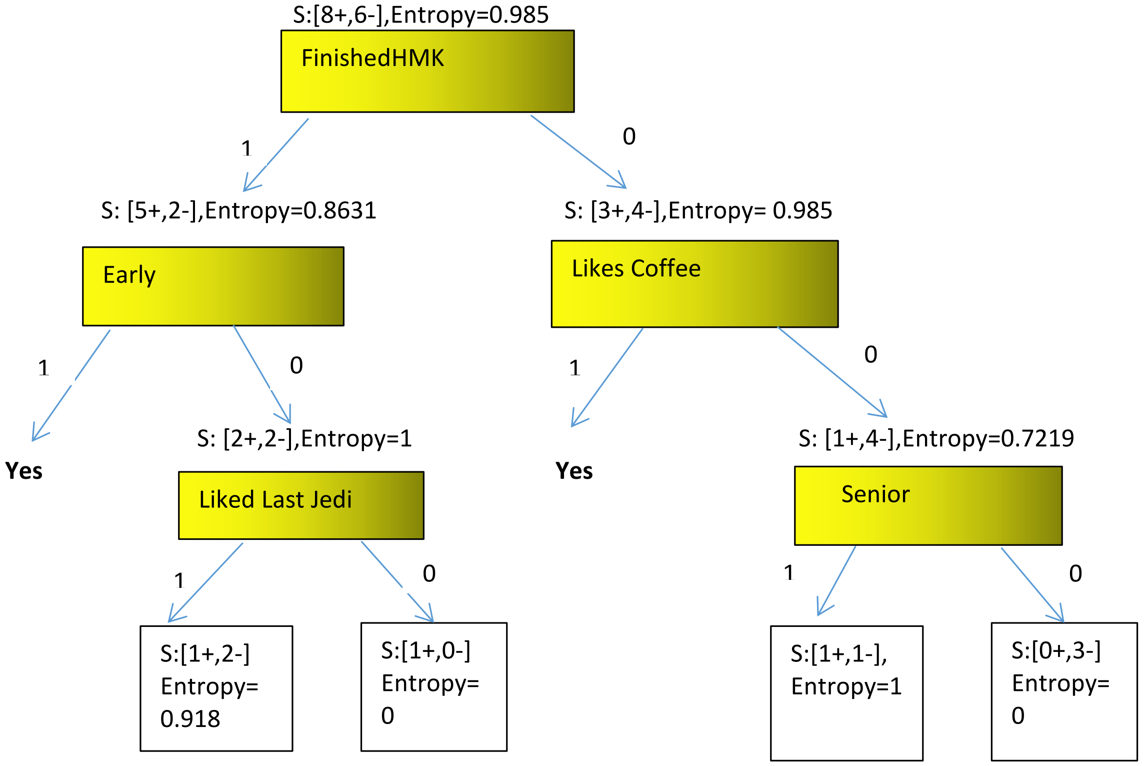
H(FinishedHMK =1) = 0.8631
H(FinishedHMK =1/ Early) = 0.571 , Gain(FinishedHMK =1, Early) = **0.292**
H(FinishedHMK =1/ senior) = 0.6935 , Gain(FinishedHMK =1, senior) = 0.170
H(FinishedHMK =1/ LikesCoffee) = 0.8031 , Gain(FinishedHMK =1, LikesCoffee) = 0.062
H(FinishedHMK =1/ LikedLastJedi) = 0.6935 , Gain(FinishedHMK =1, LikedLastJedi) = 0.170

H(FinishedHMK =0) = 0.985
H(FinishedHMK =0/ Early) = 0.571 , Gain(FinishedHMK =0, Early) = 0.414
H(FinishedHMK =0/ senior) = 0.857 , Gain(FinishedHMK =0, senior) = 0.128
H(FinishedHMK =0/ LikesCoffee) = 0.515 , Gain(FinishedHMK =0, LikesCoffee) = **0.470**
H(FinishedHMK =0/ LikedLastJedi) = 0.964 , Gain(FinishedHMK =0, LikedLastJedi) = 0.021



1c.

For Depth 3 The Tree will Look like



H(Early=0) =1
H(Early=0 / Senior) = 0.688 , Gain(Early=0,Senior) = 0.312
H(Early=0 / LikesCoffee) = 1, Gain(Early=0,LikesCoffee) = 0
H(Early=0 / LikedLastJedi) = 0.688 ,Gain(Early=0,LikedLastJedi) = **0.312**

H(LikesCoffee=0) = 0.7219
H(LikesCoffee=0 / Early)=0.5509 , Gain(LikesCoffee=0 / Early) = 0.1710
H(LikesCoffee=0 / Senior)=0.4, Gain(LikesCoffee=0 / Senior) = **0.3219**
H(LikesCoffee=0 / LikedLastJedi)=0.5509 , Gain(LikesCoffee=0 / LikedLastJedi) = 0.1710

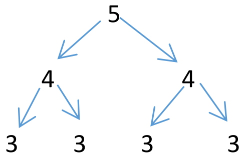
I would prefer the tree with Depth-2 since it is kind of stable and will be able to predict the results with less error. If we have case like Early 1 or LikesCoffee 1 we can classify the case as 1 with 100 percent accuracy. When we go to Dept-3, I believe there is a kind of instability at Depth-3, for Finished HMK=1 and Early =0 we can have child node as either senior or LikedLastJedi, this make me feel like there is some kind of instability at Depth-3 and there might be a chance of overfitting. Furthur analysis is needed to check if there is overfitting or not. If there is no overfitting at Depth-3, Depth-3 would be my choice because it covers more data and the results would be more accurate.

1d.

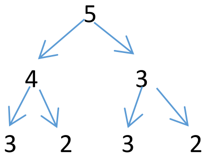
A Dataset will not be realizable if there are conflicting examples in the dataset. Let us consider an example where we have only one attribute as input and we want to predit one variable from it. In this example we have for x=0 or x=1 if y=1 then this is not a realizable dataset because y is like independent of x. In the dataset given above for Depth-3 we can see that we have two choices between senior and liked last jedi when FinishedHMK=1 and Early=0. If we have conflicting examples for this scenario then we are in a state that we cannot classify the data correctly. So to avoid this we should not have conflicting examples in our dataset and if the attributes are very large then we should either have the dataset very large such that we have a clear distinction of which attribute should be placed at what depth. We can use Occam Razors rule to find out the optimal depth of tree.

1a.

I have two cases for this question which can be visualized as follows



Fig(1)



fig(2)

Where the numbers 5,4,3…,2 are the no of possible values at that particular depth. In figure 1 I am considering the possibility of same attribute happening at different time so that they are different. In figure 2, I am assuming that same attribute cannot be at the same level.

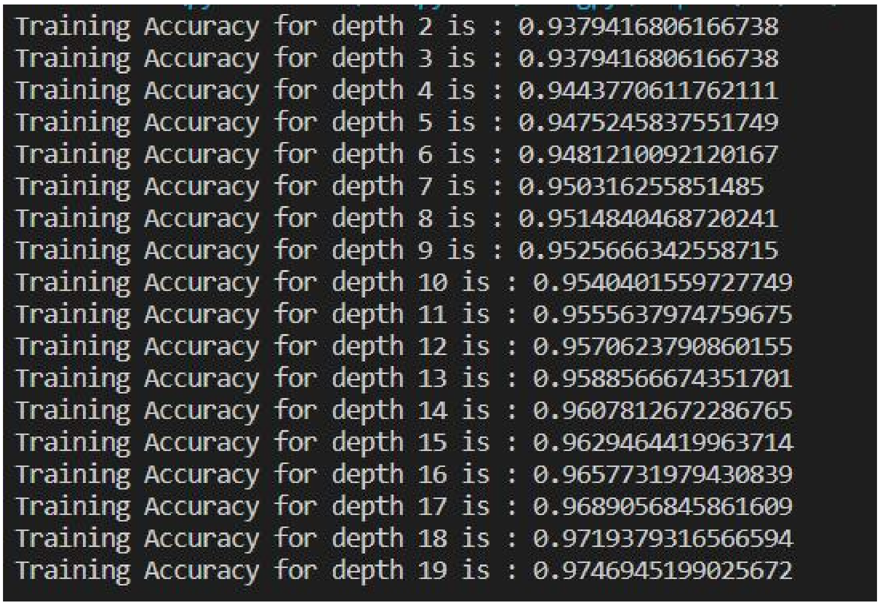
The total no of trees possible for figure 1 = 5* (4*4) *(3*3*3*3) = 6480
The total no of trees possible for figure 2 = 5* (4*3) *(3*2*3*2) = 2160

The general formula for fig 1 would be no of trees = $\prod_{i=1}^{D}(A-i-1)^{i}$ where D = depth
I am not able to get a general formula for fig 2

2a.

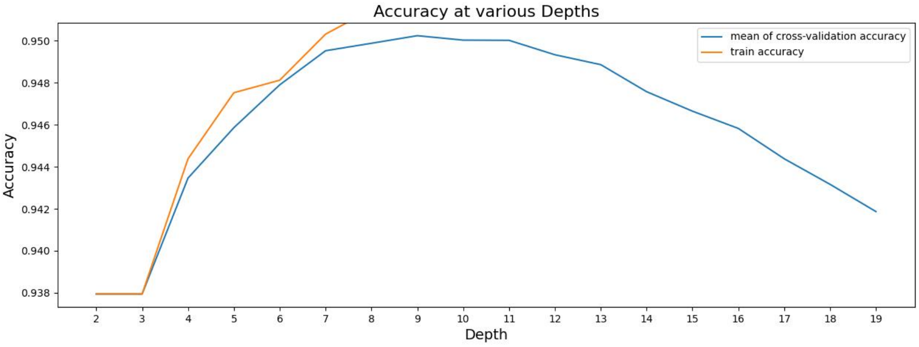
I have used the K-fold cross-validation method to get the optimal depth. For each depth I have taken the mean of accuracies of k-fold and then plotted a graph between depth was mean. The mean is a concave graph and the depth at which we obtain the max mean is the optimal depth.

The training accuracies I got for each depth is as follows.



```
Training Accuracy for depth 2 is : 0.9379416806166738
Training Accuracy for depth 3 is : 0.9379416806166738
Training Accuracy for depth 4 is : 0.9443770611762111
Training Accuracy for depth 5 is : 0.9475245837551749
Training Accuracy for depth 6 is : 0.9481210092120167
Training Accuracy for depth 7 is : 0.950316255851485
Training Accuracy for depth 8 is : 0.9514840468720241
Training Accuracy for depth 9 is : 0.9525666342558715
Training Accuracy for depth 10 is : 0.9540401559727749
Training Accuracy for depth 11 is : 0.9555637974759675
Training Accuracy for depth 12 is : 0.9570623790860155
Training Accuracy for depth 13 is : 0.9588566674351701
Training Accuracy for depth 14 is : 0.9607812672286765
Training Accuracy for depth 15 is : 0.9629464419963714
Training Accuracy for depth 16 is : 0.9657731979430839
Training Accuracy for depth 17 is : 0.9689056845861609
Training Accuracy for depth 18 is : 0.9719379316566594
Training Accuracy for depth 19 is : 0.9746945199025672
```

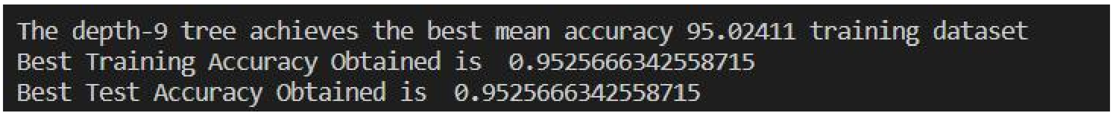
The graph which I obtained for training accuracies, mean and depth is as follows.



From the graph it is clear that the optimal depth is obtained at depth = 9.

2b.

For the optimal depth =9 , I trained the model and tested the model using the test data. The accuracies I got are as follows.



```
The depth-9 tree achieves the best mean accuracy 95.02411 training dataset
Best Training Accuracy Obtained is  0.9525666342558715
Best Test Accuracy Obtained is  0.9525666342558715
```

The Best Test Accuracy Obtained is = **0.9525**

2c.

The over-fitting issue starts from Depth =10 as seen in the graph I attached above. We can see that the mean cross-validation accuracy started to decline from Depth =10 which is caused because of over-fitting of the training data. So Depth =9 is the optimal depth until which over-fitting is not found later on we can see over-fitting of data for higher depths.