Assignment 4

Question 1:

1. Assume that the distribution is normal, so that we can use the Gaussian Distribution Equation to calculate the probability distribution table.

   Example for two features: (I've typed out the calculation part for the mean and standard deviation of apartment)

   Apartment:

   Mean is calculated as the sum of all occurrences for a column divided by the number of occurrences.

   To calculate mean for local prices for apartment = (4.9176+4.5573+5.0597+14.4598+5.05+8.2464+9.0384) / 7 = 7.3327428571

   Again

   Standard Deviation is calculated as:

   SD = {[(4.9176-7.33274286)^2 + (4.5573-7.33274286)^2 + (5.0597-7.33274286)^2 + (14.4598-7.33274286)^2 + (5.05-7.33274286)^2 + (8.2464-7.33274286)^2 + (9.0384-7.33274286)^2]/(7-1)}^(1/2)

   = 3.34776292

   For the rest of the normal distribution, I've used excel to calculate the mean and the standard deviation. The mean can be calculated as AVERAGE(range of appropriate column), and the standard deviation as STDEV.P(range of appropriate column)

   The normal distribution received from the calculation is given below:

	Local Price	Bathroom	Land Area	Living Area	#Garages	#Rooms	#Bedrooms	Age of home
Apartment								
SD	3.34776292	0.52489066	3.01679359	0.65187532	0.64681322	1.2453997	0.9035079	13.5932156
Mean	7.33274286	1.28571429	6.10385714	1.505	1.21428571	6.85714286	3.42857143	38.7142857
House								
SD	0.52782973	0.17496355	2.08214461	0.19712919	0.77591289	0.63887656	0.53452248	11.7803018
Mean	5.76074286	1.07142857	6.6309	1.39171429	1.07142857	6.14285714	3	34.2857143
Condo								
SD	4.20947412	0.5527708	2.32305328	0.84298273	0.47140452	1.46249406	0.74535599	12.7366488
Mean	7.4159	1.33333333	6.02466667	1.55333333	1.33333333	6.83333333	3.33333333	39.6666667

2. A sample of conditional probability table is given below:

```
***Condition Probability Table***

Condition Probability for input: [6.0931, 1.5, 6.7265, 1.652, 1, 6, 3, 44, 'Apartment']
Apartment
[0.11127082517956667, 0.6992782413085956, 0.12945370424459396, 0.5966274344740321, 0.5838455060157502, 0.25277963962488087, 0.3945662922932312, 0.02721161725
House
[0.7545594017109934, 2.104839173109561, 0.1915817079993319, 1.988674254041068, 0.5130172863835589, 0.6242746037092262, 0.7463526649356675, 0.0361645152103794
Condo
[0.09327671140519034, 0.7160973198349625, 0.17057055271021157, 0.4722970817944242, 0.820246942054623, 0.2707612547524127, 0.5325677283621869, 0.0311360429247
```

Question 2:

1. a) The accuracy on the training set is 0.7
   b) The accuracy on the test set is 0.8

2. Restricting the depth of the tree effects the accuracy of the model. Here are the different values observed for the depth:

   Depth: 1
   Training Accuracy: 0.55
   Test Accuracy: 0.4

   Depth 2
   Training Accuracy: 0.7
   Test Accuracy: 0.8

   Depth 3
   Training Accuracy: 0.8
   Test Accuracy: 0.8

   Depth 4:
   Training Accuracy: 0.95
   Test Accuracy: 0.6

   Depth 5:
   Training Accuracy: 1.0
   Test Accuracy: 0.4

   As you can see, the accuracy was best at depth 2, or depth 3. But when the depth increased, the test accuracy started going down even though training accuracy went up. This is probably because the model started overfitting.

   The plots when depth of the tree = 2 and depth = 4, is given below:

```
Training Accuracy- 0.7
Test Accuracy- 0.8
```

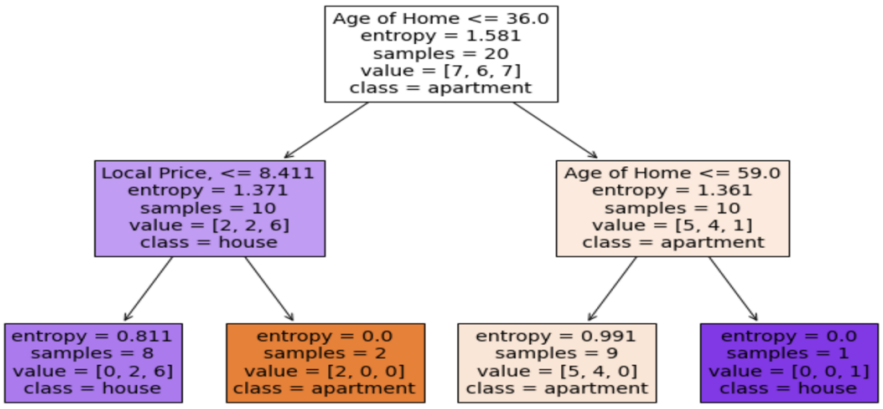


Figure 1 When depth = 2

```
Training Accuracy- 0.95
Test Accuracy- 0.4
```

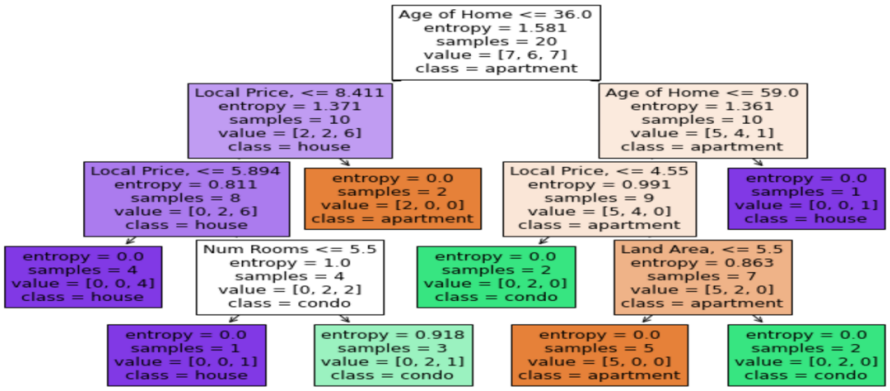
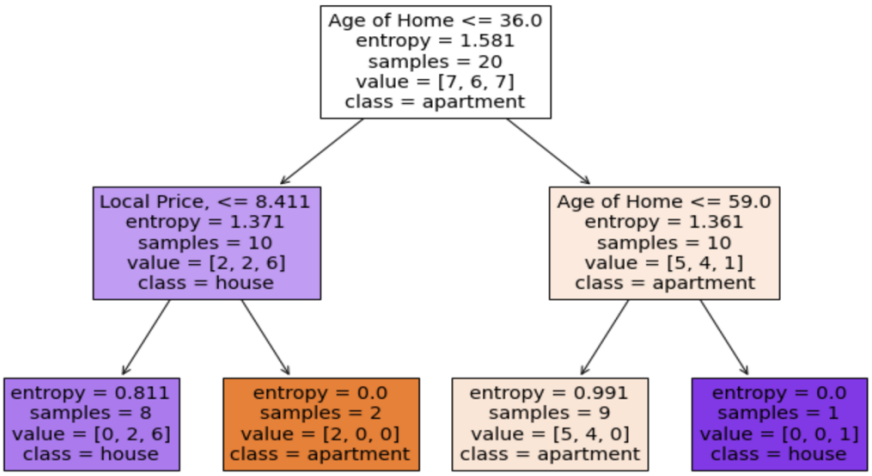


Figure 2 When depth = 4

3. For a decision tree the deeper the tree, the more complex your model becomes. As we observed with our model, when the depth was 2 or 3, the test and training accuracy was optimal. But with the increase in depth, it caused the model to overfit. This will cause it to get the best accuracy for training but decrease accuracy for test set.

4. The resultant tree is observed as below:

```
Training Accuracy- 0.7
Test Accuracy- 0.8
```

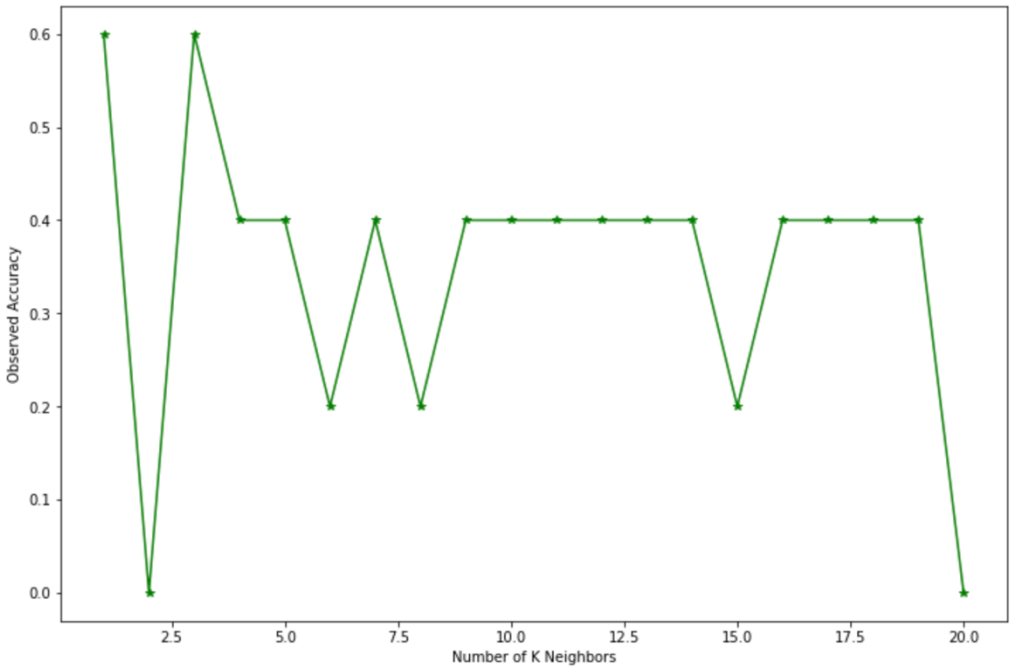


- With respect to the tree, we first look at root node and evaluate the age of the home. Since in the given table, the age of the home is 23, we move towards the left node.

- Then we check the local price. The local price is 9.0384, which is greater than 0.8411, so we move right this time. The predicted class going right would be apartment.

Question 3:

I used the Euclidean distance to calculate the neighboring points. This is the plot that was produced:



This is the plot of accuracy against k number of neighbors. This was done in a range of 20 or in other words, 20 nearest neighbors. The number of ties or uncertain guesses does seem to bring the accuracy down. But for the most part, the accuracy is stable. We could certainly get an even more accurate model, if we increase the number of K since it would decrease the uncertainty of the mode.