Ans to the Q No 1

1.

Let C_1 = Apartment

C_2 = House

C_3 = Condo

Here, the input attributes are ~~conditions~~ continuous valued. So, the conditional probabilities can be modeled with the normal distribution

$$\hat{P}(X_j | C = c_i) = \frac{1}{\sqrt{2\pi}\,\sigma_{ji}} \exp\left(-\frac{(X_j - \mu_{ji})^2}{2\sigma_{ji}^2}\right)$$

Here μ = mean/average

σ = standard deviation

$$= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

For local price feature :

$$\mu_{C_1} = \frac{4.9176 + 4.5573 + 5.0597 + 14.4598 + 5.05 + 8.2464 + 9.0384}{7}$$

$$= 7.333$$

$$\sigma_{C_1} = \sqrt{\frac{(4.9176 - 7.333)^2 + (5.5573 - 7.333)^2 + (5.0597 - 7.333)^2 + (14.4598 - 7.333)^2 + (5.05 - 7.333)^2 + (8.2464 - 7.333)^2 + (9.0384 - 7.333)^2}{7-1}}$$

$$= 3.616$$

Similarly

$$\mu_{C_2} = 5.76 = \frac{5.0208+5.6035+5.8282+5.3003+6.2712+5.6039+6.6969}{7}$$

$$\sigma_{C_2} = 0.57 = \sqrt{\frac{(5.0208-5.76)^2+(5.6035-5.76)^2+(5.8282-5.76)^2+(5.3003-5.76)^2+(6.2712-5.76)^2+(5.6039-5.76)^2+(6.6969-5.76)^2}{7-1}}$$

and

$$\mu_{C_3} = 7.416 = \frac{4.5429+3.891+5.898+16.4202+5.9592+7.784}{6}$$

$$\sigma_{C_3} = 4.611 = \sqrt{\frac{(4.5429-7.416)^2+(3.891-7.416)^2+(5.898-7.416)^2+(16.4202-7.416)^2+(5.9592-7.416)^2+(7.784-7.416)^2}{6-1}}$$

For Bathrooms feature :

$\mu_{C_1} = 1.286$ $\mu_{C_2} = 1.071$ $\mu_{C_3} = 1.333$

$\sigma_{C_1} = 0.567$ $\sigma_{C_2} = 0.189$ $\sigma_{C_3} = 0.606$

For Land Area feature :

$\mu_{C_1} = 6.104$ $\mu_{C_2} = 6.631$ $\mu_{C_3} = 6.025$

$\sigma_{C_1} = 3.259$ $\sigma_{C_2} = 2.249$ $\sigma_{C_3} = 2.545$

For Living area feature :

$\mu_{C_1} = 1.505$ $\mu_{C_2} = 1.392$ $\mu_{C_3} = 1.553$

$\sigma_{C_1} = 0.704$ $\sigma_{C_2} = 0.213$ $\sigma_{C_3} = 0.923$

For # Garages feature :

$\mu_{C_1} = 1.214$ $\mu_{C_2} = 1.07$ $\mu_{C_3} = 1.333$

$\sigma_{C_1} = 0.699$ $\sigma_{C_2} = 0.838$ $\mu_{C_4} = 0.516$

For # Rooms feature

$\mu_{C_1} = 6.857$ $\mu_{C_2} = 6.143$ $\mu_{C_3} = 6.833$

$\sigma_{C_1} = 1.345$ $\sigma_{C_2} = 0.69$ $\sigma_{C_3} = 1.602$

For # Bedrooms feature

$\mu_{C_1} = 3.429$ $\mu_{C_2} = 3$ $\mu_{C_3} = 3.333$

$\sigma_{C_1} = 0.976$ $\sigma_{C_2} = 0.577$ $\sigma_{C_3} = 0.816$

For Age of home feature

$\mu_{C_1} = 38.714$ $\mu_{C_2} = 34.286$ $\mu_{C_3} = 39.667$

$\sigma_{C_1} = 14.68$ $\sigma_{C_2} = 12.724$ $\sigma_{C_3} = 13.952$

For Row 1 of test data :

$$P(\text{local price} = 6.0931 \mid C = C_1) = \frac{1}{\sqrt{2\pi}\,3.616} e^{-\frac{(6.0931 - \mu_{C_1})^2}{2\sigma_{C_1}^2}}$$

$$= \frac{1}{\sqrt{2\pi}\,\sigma_{C_1}} e^{-\frac{(6.0931 - 7.333)^2}{2(3.616)^2}}$$

$$= \frac{1}{\sqrt{2\pi} \times 3.616} e$$

$$= 0.104$$

$$P(\text{Bathrooms} = 1.5 \mid C = C_1)$$

$$= \frac{1}{\sqrt{2\pi}(0.567)} e^{-\frac{(1.5 - 1.286)^2}{2(0.567)^2}}$$

$$= 0.655$$

$$P(\text{Local price} = 6.0931 \mid C = C_2)$$

$$= \frac{1}{\sqrt{2\pi}(0.57)} e^{-\frac{(6.0931 - 5.76)^2}{2(0.57)^2}}$$

$$= 0.59$$

$$P(\text{Bathrooms} = 1.5 \mid C = C_2)$$

$$= \frac{1}{\sqrt{2\pi}(0.189)} e^{-\frac{(1.5 - 1.071)^2}{2(0.189)^2}}$$

$$= 0.161$$

$$P(\text{Local price} = 6.0931 \mid C = C_3)$$

$$= \frac{1}{\sqrt{2\pi}(0.466)} e^{-\frac{6.0931 - 7.416)^2}{2(4.611)^2}}$$

$$= \frac{1}{\sqrt{2\pi}(4.611)} e$$

$$= 0.083$$

$$P(\text{Bathrooms} = 1.5 \mid C = C_3)$$

$$= \frac{1}{\sqrt{2\pi}(0.606)} e^{-\frac{(1.5-1.333)^2}{2(0.606)^2}}$$

$$= 0.634$$

Detail calculation has been shown for two features. By following the same procedure, all other conditional probabilities can be determined.

$$P(\text{Land Area} = 6.7265 \mid C = C_1) = 0.120$$

$$P(\text{Land Area} = 6.7265 \mid C = C_2) = 0.177$$

$$P(\text{Land Area} = 6.7265 \mid C = C_3) = 0.151$$

$$P(\text{Living Area} = 1.652 \mid C = C_1) = 0.554$$

$$P(\text{Living Area} = 1.652 \mid C = C_2) = 0.889$$

$$P(\text{Living Area} = 1.652 \mid C = C_3) = 0.43$$

Conditional probability tables are in the next ~~25~~ pages

For Apartment class :

House ID \ Features	Local Price	Bathrooms	Land Area	Living Area	#Garages	#Rooms	#Bedrooms	Age of home
24	0.104	0.655	0.12	0.554	0.545	0.242	0.371	0.025
25	.106	.655	0.079	0.526	0.303	0.207	0.344	0.022
26	.108	.62	.103	0.567	0.303	0.295	0.371	0.001
27	.097	.655	0.114	.509	.525	0.207	.344	.024
28	.048	.655	.116	.516	.303	.242	.371	.023

For House Class

House ID \ Features	Local Price	Bathrooms	Land Area	Living Area	#Garages	#Rooms	#Bed-rooms	Age of home
24	0.59	0.161	0.177	0.889	0.474	0.566	0.691	0.023
25	2.11e-5	0.161	.095	.366	0.258	.015	0.154	0.018
26	1.15e-4	1.967	0.147	1.631	0.258	0.267	0.691	0.002
27	1.59e-8	.161	.169	.224	.418	.015	.154	.03
28	6.62e-27	.161	.136	1.248	.258	.566	.691	.03

For Condo Class

House ID \ Features	Local Price	Bathrooms	Land Area	Living Area	#Garages	#Rooms	#Bed rooms	Age of home
24	.083	.634	0.151	0.43	0.628	0.218	0.45	0.027
25	.085	.634	0.074	0.42	0.335	0.191	0.35	0.024
26	.085	0.586	0.116	0.432	0.335	0.248	0.45	0.001
27	.081	.634	.138	.413	.734	.191	0.35	.024
28	.053	.634	.145	.402	.335	.218	0.45	.022

Results for house ID 24:
Probability of Apartment: 1.9389181719507603e-06
Probability of House: 2.2305995775315273e-05
Probability of Condo: 1.7050080080766167e-06
Class: HOUSE

Results for house ID 25:
Probability of Apartment: 4.793162384239558e-07
Probability of House: 4.434914558885 58e-13
Probability of Condo: 2.7006475427063995e-07
Class: APARTMENT

Results for house ID 26:
Probability of Apartment: 4.538813386026905e-08
Probability of House: 1.8070931962597316e-09
Probability of Condo: 2.704004121945 6005e-08
Class: APARTMENT

Results for house ID 27:
Probability of Apartment: 1.1577186924225047e-06
Probability of House: 9.82506164736096e-16
Probability of Condo: 1.0340305803977789e-06
Class: APARTMENT

Results for house ID 28:
Probability of Apartment: 4.1211461150337135e-07
Probability of House: 1.916641302846243e-31
Probability of Condo: 4.2483171700747804e-07
Class: CONDO

Ans to the Q No 2

1. Using default parameters
Predicted classification on test input: ['Condo' 'Condo' 'Condo' 'Apartment' 'Apartment']
Depth of the tree: 5

(a) What is the accuracy on the training set?
Accuracy score on training set: 1.0
(b) What is the accuracy on the test set?
Accuracy score on test set: 0.4

2. What is the effect of restricting the maximum depth of the tree? Try different depths and find the best value.

Results of varying maximum depth from 1 to 7:

```
Allowed maximum depth: 1
Actual depth of the tree: 1
Predicted classification on test input: ['Apartment' 'Apartment'
'House' 'House' 'House']
Accuracy score on training set: 0.55
Accuracy score on test set: 0.4
Allowed maximum depth: 2
Actual depth of the tree: 2
Predicted classification on test input: ['Apartment' 'Apartment'
'House' 'Apartment' 'Apartment']
Accuracy score on training set: 0.75
Accuracy score on test set: 0.8
Allowed maximum depth: 3
Actual depth of the tree: 3
Predicted classification on test input: ['Condo' 'Condo' 'Condo'
'Apartment' 'Apartment']
Accuracy score on training set: 0.9
Accuracy score on test set: 0.4
Allowed maximum depth: 4
```

```
Actual depth of the tree: 4
Predicted classification on test input: ['Condo' 'Condo' 'Condo'
'Apartment' 'Apartment']
Accuracy score on training set: 0.95
Accuracy score on test set: 0.4
Allowed maximum depth: 5
Actual depth of the tree: 5
Predicted classification on test input: ['Condo' 'Condo' 'Condo'
'Apartment' 'Apartment']
Accuracy score on training set: 1.0
Accuracy score on test set: 0.4
Allowed maximum depth: 6
Actual depth of the tree: 5
Predicted classification on test input: ['Condo' 'Condo' 'Condo'
'Apartment' 'Apartment']
Accuracy score on training set: 1.0
Accuracy score on test set: 0.4
Allowed maximum depth: 7
Actual depth of the tree: 5
Predicted classification on test input: ['Condo' 'Condo' 'Condo'
'Apartment' 'Apartment']
Accuracy score on training set: 1.0
Accuracy score on test set: 0.4
```

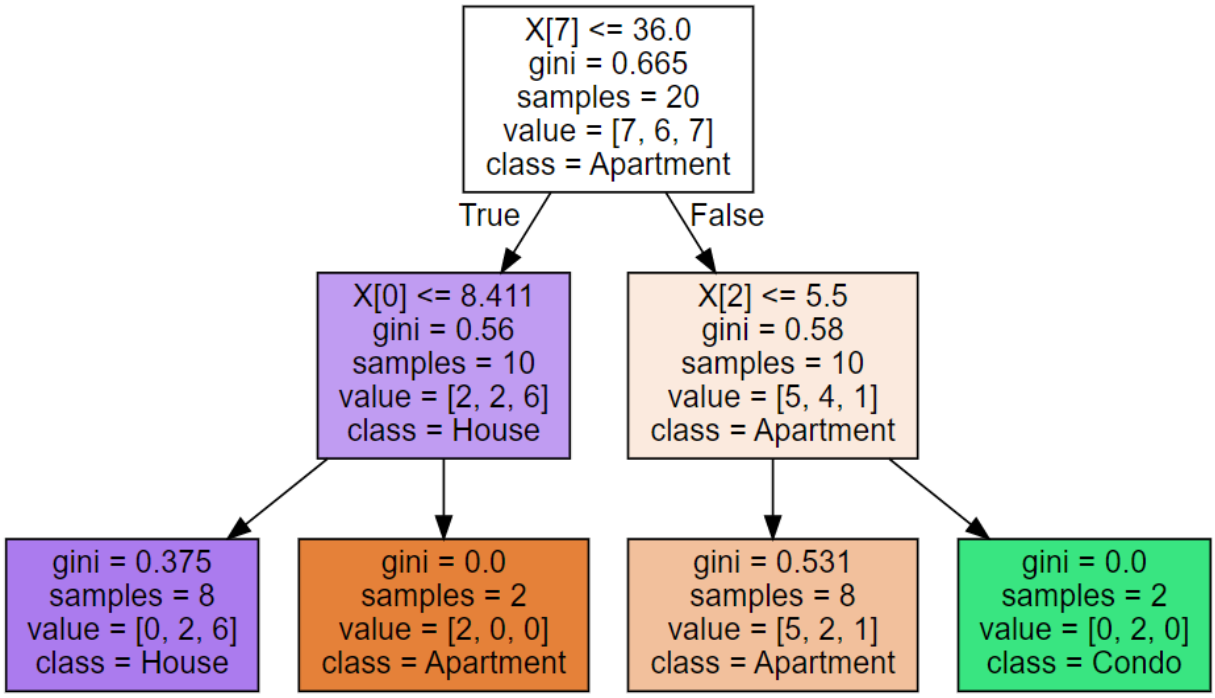
We see that with varying maximum depth the performance of the classifier changes. We see that the best test performance occurs at maximum depth of 2 where we obtain a test accuracy of 0.8. After that the accuracy reduces with increasing depth size and the maximum depth saturates at 5.

3. Why does restricting the depth have such a strong effect on the classifier performance?
With high depth of the tree the classifier overfits on the training data and thus performs poorly due to lack of generalization. By restricting the depth size, we can reduce overfitting and thus test performance can be improved.

4. Visualize the resulting tree. Perform the inference on this tree manually (i.e. show/trace the path taken towards classification) and provide a classification for the following example:

Local Price	9.0384
Bathrooms	1
Land Area	7.8
Living area	1.5
# Garages	1.5
# Rooms	7
# Bedrooms	3
Age of home	23

Visualization of the tree for depth=2 which yield the best result:

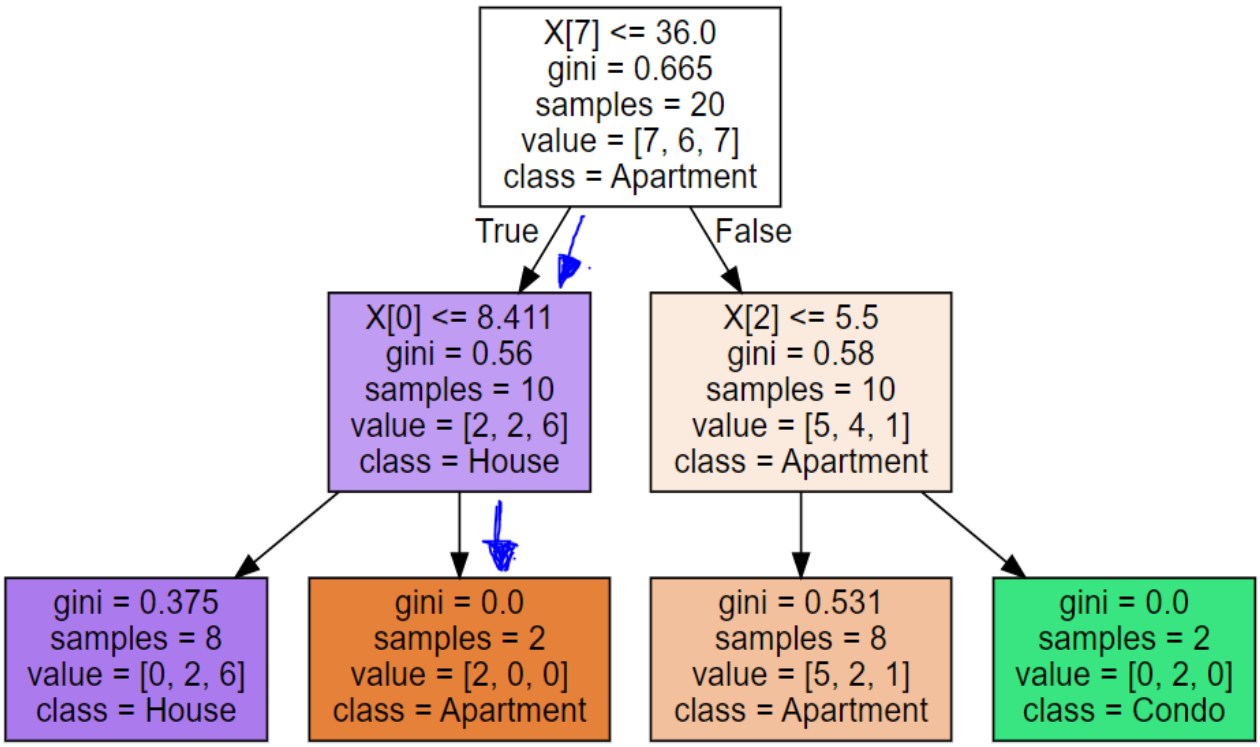


Here,
x[0] = Local Price
x[1] = Bathrooms
x[2]= Land Area
x[3] = Living area
x[4] = # Garages
x[5] = # Rooms
x[6] = # Bedrooms
x[7] = Age of home

For the following inputs:

Local Price 9.0384 , Bathrooms 1, Land Area 7.8 , Living area 1.5 , # Garages 1.5 , # Rooms 7 , #Bedrooms 3 , Age of home 23
The classification is: Apartment.

Below is shown the path for the classification in the tree.



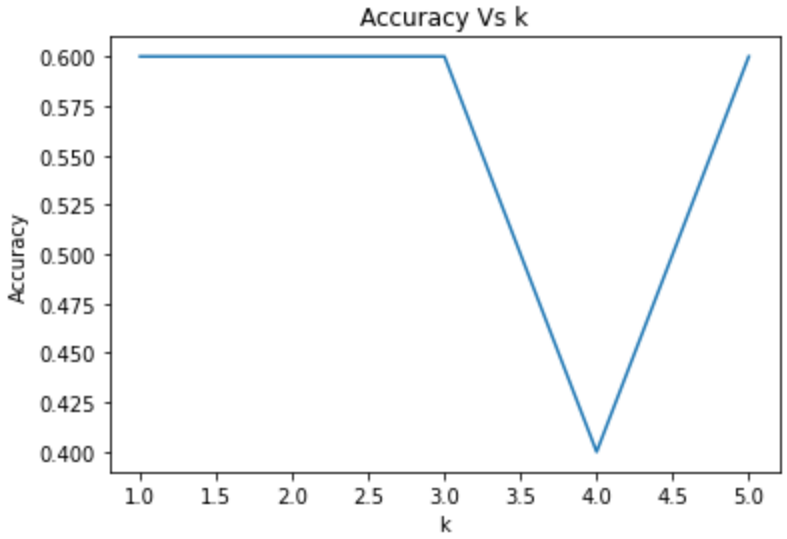
Ans to the Q No 3

Results of classification using KNN is shown below for varying number of K.

```
K = 1
Test set row 1 , Votes for: Apartment 4 House 1 and Condo 3
Test set row 2 , Votes for: Apartment 6 House 2 and Condo 0
Test set row 3 , Votes for: Apartment 7 House 1 and Condo 0
Test set row 4 , Votes for: Apartment 5 House 2 and Condo 1
Test set row 5 , Votes for: Apartment 4 House 2 and Condo 2
Accuracy: 0.6
['Apartment' 'Apartment' 'Apartment' 'Apartment' 'Apartment']


K = 2
```

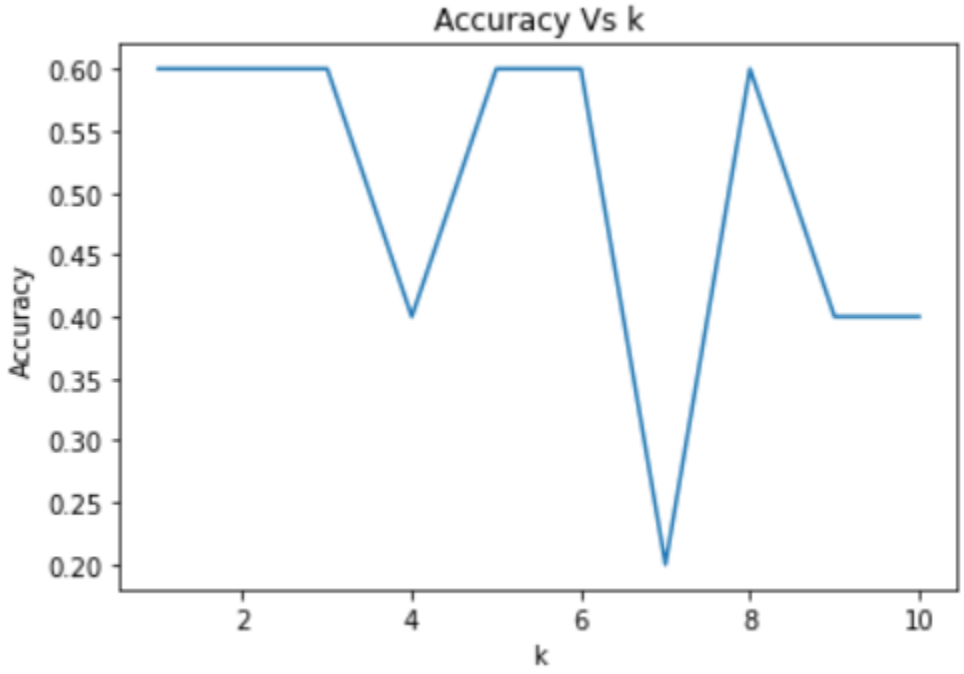
```
Test set row 1 , Votes for: Apartment 7 House 5 and Condo 4
Test set row 2 , Votes for: Apartment 9 House 6 and Condo 1
Test set row 3 , Votes for: Apartment 7 House 6 and Condo 3
Test set row 4 , Votes for: Apartment 8 House 6 and Condo 2
Test set row 5 , Votes for: Apartment 7 House 7 and Condo 2
Accuracy: 0.6
['Apartment' 'Apartment' 'Apartment' 'Apartment' 'Apartment']


K = 3
Test set row 1 , Votes for: Apartment 9 House 7 and Condo 8
Test set row 2 , Votes for: Apartment 12 House 8 and Condo 4
Test set row 3 , Votes for: Apartment 11 House 8 and Condo 5
Test set row 4 , Votes for: Apartment 11 House 9 and Condo 4
Test set row 5 , Votes for: Apartment 12 House 8 and Condo 4
Accuracy: 0.6
['Apartment' 'Apartment' 'Apartment' 'Apartment' 'Apartment']


K = 4
Test set row 1 , Votes for: Apartment 13 House 9 and Condo 10
Test set row 2 , Votes for: Apartment 15 House 11 and Condo 6
Test set row 3 , Votes for: Apartment 14 House 10 and Condo 8
Test set row 4 , Votes for: Apartment 12 House 13 and Condo 7
Test set row 5 , Votes for: Apartment 14 House 9 and Condo 9
Accuracy: 0.4
['Apartment' 'Apartment' 'Apartment' 'House' 'Apartment']


K = 5
Test set row 1 , Votes for: Apartment 15 House 12 and Condo 13
Test set row 2 , Votes for: Apartment 18 House 15 and Condo 7
Test set row 3 , Votes for: Apartment 15 House 17 and Condo 8
Test set row 4 , Votes for: Apartment 14 House 19 and Condo 7
Test set row 5 , Votes for: Apartment 17 House 13 and Condo 10
Accuracy: 0.6
['Apartment' 'Apartment' 'House' 'House' 'Apartment']
```

Accuracy vs k is plotted below:



There is no clear trend in accuracy vs k plot. Accuracy remained 0.6 from k=1 to 3. Then for k=4 it decreased to 0.4 then again increased for k=5.

Plotting Accuracy vs k for k values upto 10



We see that accuracy value fluctuates with increase in k value.