**Data Mining & Predictive Analytics**

**Midterm Exam**

**Fall 2020**

I hereby certify that I have completed the attached examination materials, using only my own efforts. I have not asked for or received help from any person in completing this exam.

Piyush Garewal October 07, 2020

(Name) (date)

**Q1. Using the concept of overfitting, explain why when a model is perfectly fit to training data, this is typically not good?**

**Answer:** Overfitting a model is a condition where a statistical model begins to describe the random error in the data rather than the relationships between variables. This problem occurs when the model is too complex. Sometimes overfitting can lead to misleading R-Squared values, that it is not an accurate model.

It's not good because when looking at models you want to see the relationship between the data, and if there are zero errors in the data then the information you get is skewed and may not be a true reflection. And can lead to poor predictive performance because it can exaggerate minor fluctuations in the data.

Reference: <https://statisticsbyjim.com/regression/overfitting-regression-models/>

**Q2. Consider the distance between records of the following dataset. Which are the closest records?**

|  |  |  |  |
| --- | --- | --- | --- |
| **Ob. No.** | **Age** | **Income ($)** | **City** |
| 1 | 25 | 49000 | Albany |
| 2 | 56 | 156000 | Albany |
| 3 | 65 | 99000 | NYC |
| 4 | 32 | 192000 | Poughkeepsie |

To calculate the closest records, we will compute the Euclidean distance between each record.

**Answer:** We start by normalizing the data:

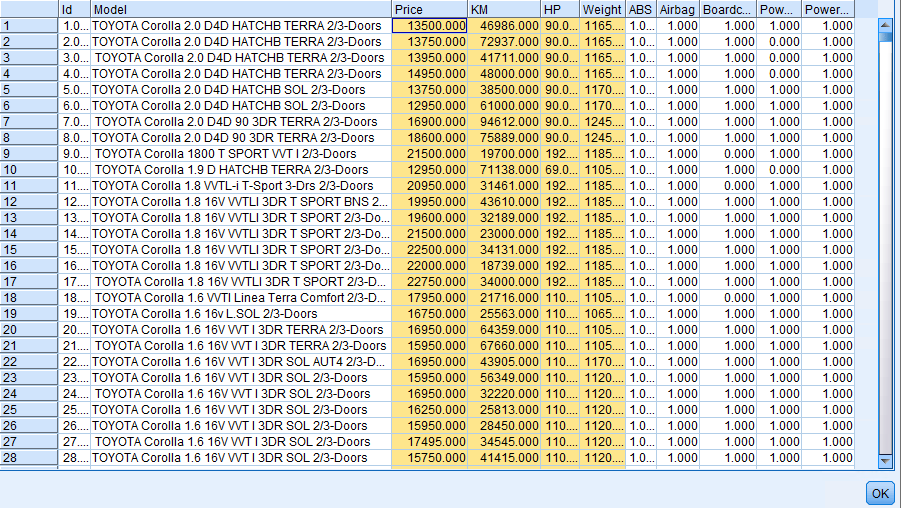
|  |  |  |  |
| --- | --- | --- | --- |
| **Ob. No.** | **Age** | **Income ($)** | **City** |
| 1 | (25 -25)/ (65-25) = 0 | (49000 – 49000) / (192000 - 49000) = 0 | Albany |
| 2 | (56 -25)/ (65-25) = 0.775 | (156000 – 49000) / (192000 - 49000) = 748.251 | Albany |
| 3 | (65 – 25) / (65-25) = 1 | (99000 – 49000) / (192000 - 49000) = 349.65 | NYC |
| 4 | (32 – 25)/ (65 - 25) = 0.175 | (192000 – 49000) / (192000 - 49000) = 100 | Poughkeepsie |

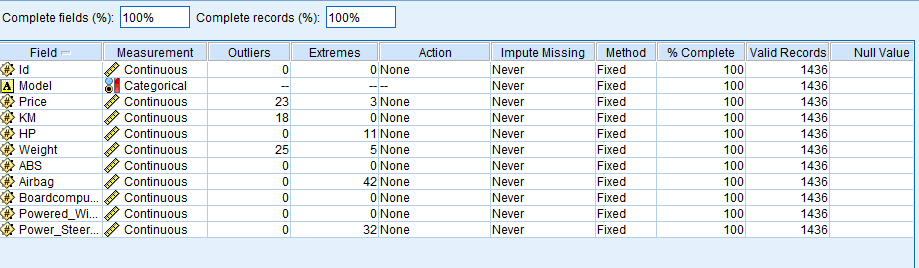
Then we compute the distances:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Distance between Ob. i and j** | ***Age (i) – Age (j)*** | ***Income ($)i – Income ($)j*** | ***City*** | ***Distance*** |
| 1. and 2 | (0 – 0.775) = - 0.775 | (0 - 748.251) = - 748.251 | 0 | D = sqrt (0.6 + 559878.06 + 0) = 748.250 |
| 2 and 3 | 0.775 – 1 = -0.225 | (748.251- 349.65) = 398.601 | 1 | D = sqrt (0.05 + 158,882.757+ 1) = 398.602 |
| 3 and 4 | (1 – 0.175) = 0.825 | (349.65 – 100) = 249.65 | 1 | D = sqrt (0.05 + 158,882.757+ 1) = 249.66 |
| 1 and 4 | (0 –0.175) = -0.175 | (0 – 100) = -100 | 1 | D = sqrt (0.05 + 158,882.757+ 1) = 100.005 |

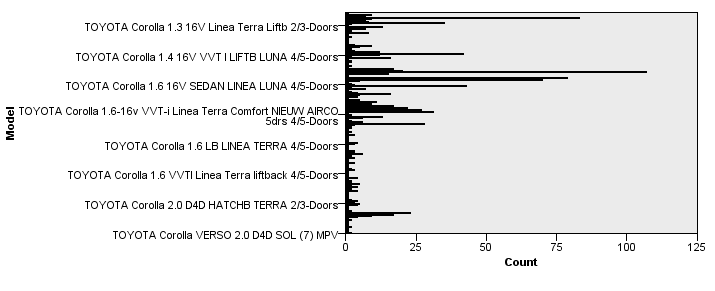
From the above Object Records 1 and 4 are the closest, as the distance between the two is the smallest of them all (100.005)

**Q3. The dataset cars.xlsx contains data on used cars on sale. Explore the numeric data using data visualization capabilities available in SPSS Modeler: which of the pairs of variables seem to be correlated? Verify it numerically.**

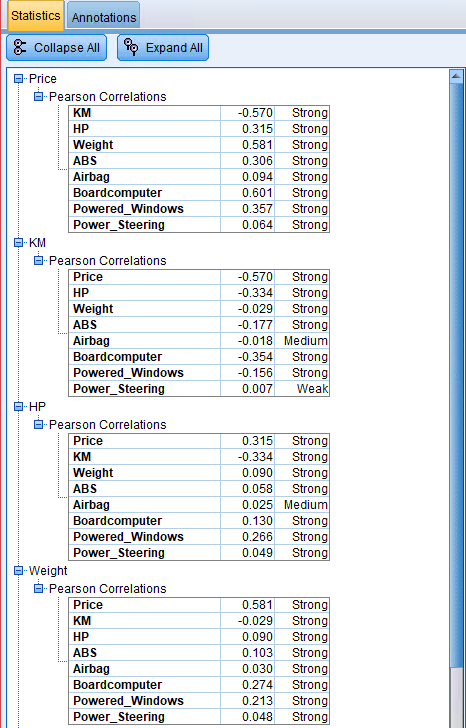
1. First, we have uploaded the dataset into SPSS modeler.
2. To have the visual of the given data, we have connected a table node, below is the attachment from table node. 
3. Then do check the quality of the data, we have connected the data audit node to see if there is any missing data from the given data set, there is no missing data, below is the attachment from data audit node.

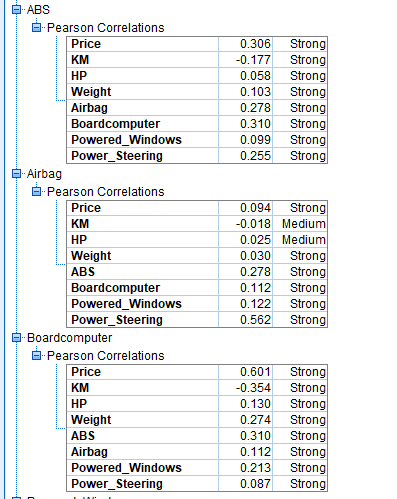


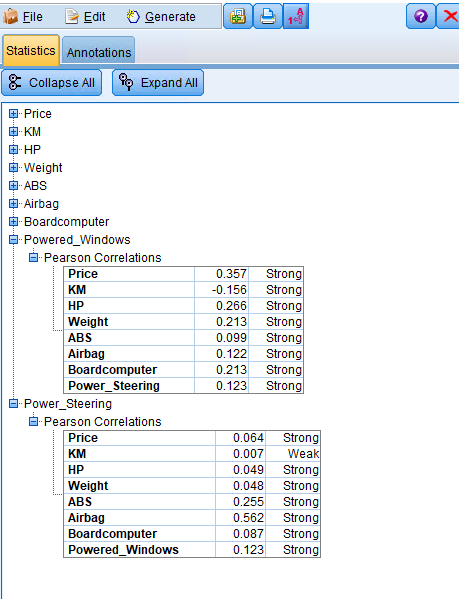
1. To have a visual of the data numerical as well as the categorical fields, below is the graph visuals:



1. Then we have connected a statistics node to the excel file node to see the correlation from the given data set. We have selected all the node except the first field the ID

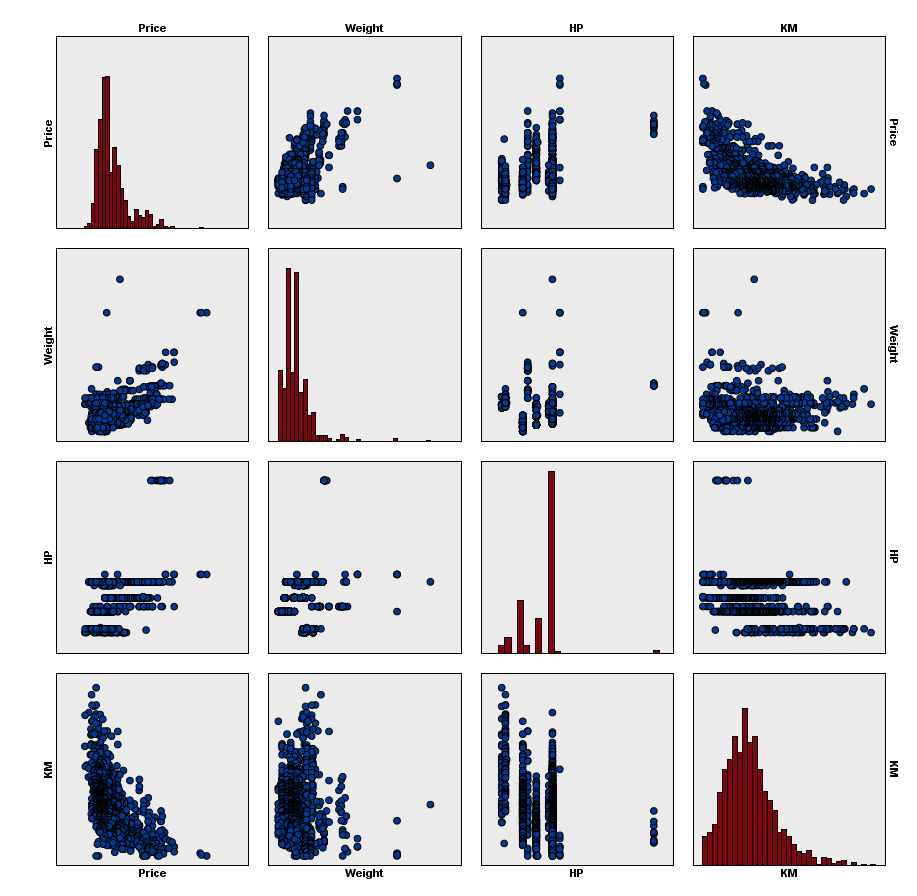


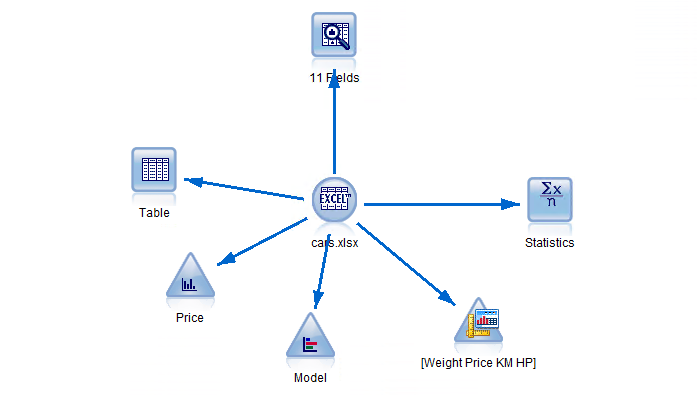




As we can see from the above that there are some strong correlations.

1. Finally, we have a scatter matrix plot for the Price, Weight, HP and KM



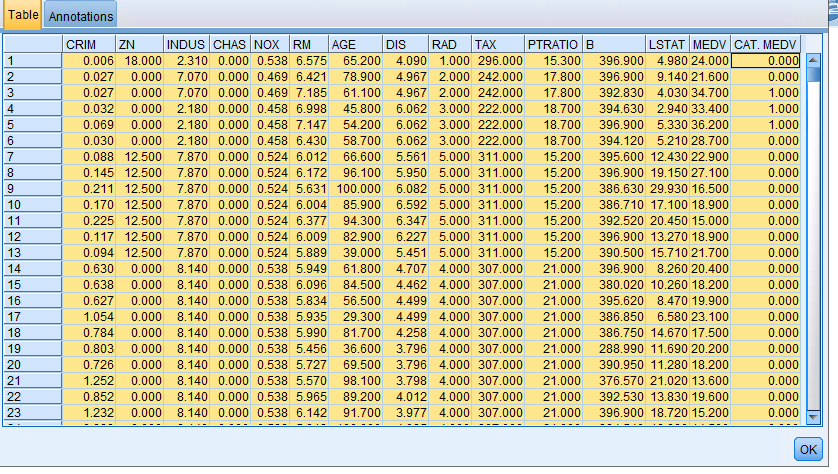
As we can see from the graph above, there are some strong correlations. Below is the complete stream visual. 

For more details please refer to attached stream file from submitted zip folder.

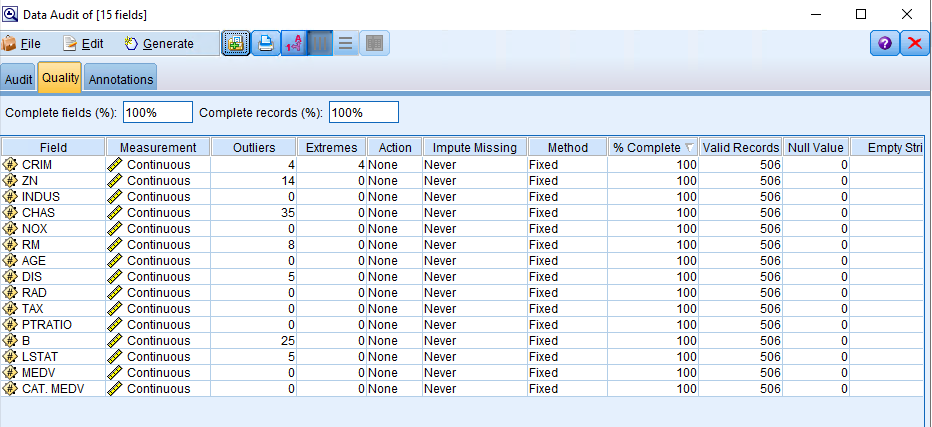
**Q4. The file BostonHousing.xlsx contains information over 500 census tracts in Boston, where for each tract 14 variables are recorded. Attribute MEDV represents the median value of a tract given the information of the other 13 attributes. The last attribute (CAT.MEDV) is a discrete recoding of MEDV such that it carries value 1 if MEDV >30 and 0 otherwise**

**Build kNN classification models of the median value of a tract (CAT.MEDV) using the attached dataset, with a 70%- 30% partition, and varying k between 1 and 5. Report the predictive performance of the models. What k would you choose? For one of the k values the training error rate is zero. Why would you say that the error is zero? (Use SPSS Modeler to complete this question)**

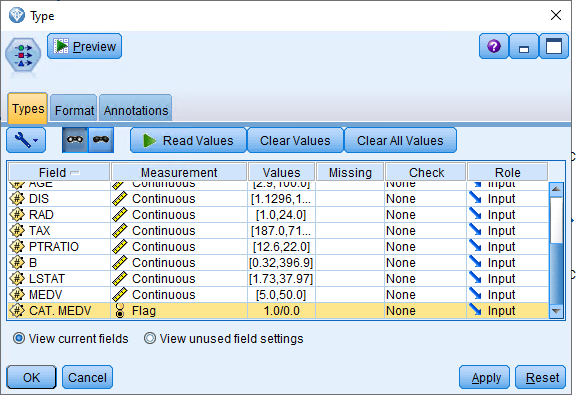
1. First, we have uploaded the dataset into SPSS modeler.
2. To have the visual of the given data, we have connected a table node, below is the attachment from table node.



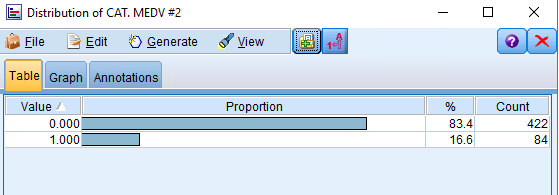
1. Then we have check quality of the data, by connecting the audit node to Excel node, there are no missing data, below is the attached snapshot.



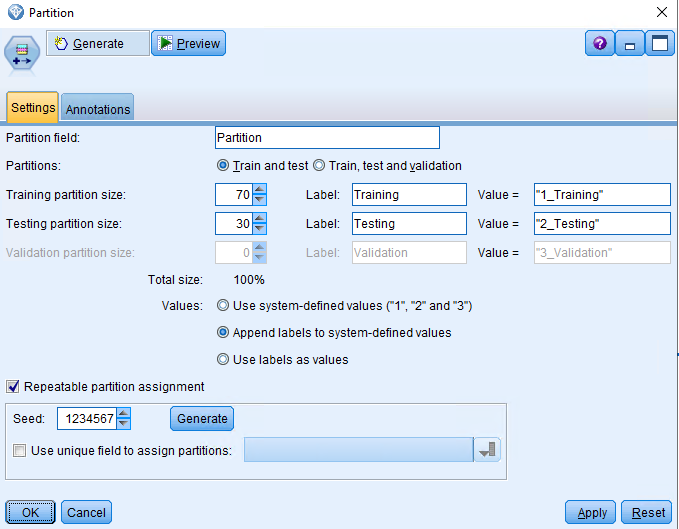
1. As we have built kNN classification models of the median value of a tract (CAT.MEDV), we first check the type of the CAT.MEDV, here we have to change the measurement to flag, which can be seen from the snapshot below:



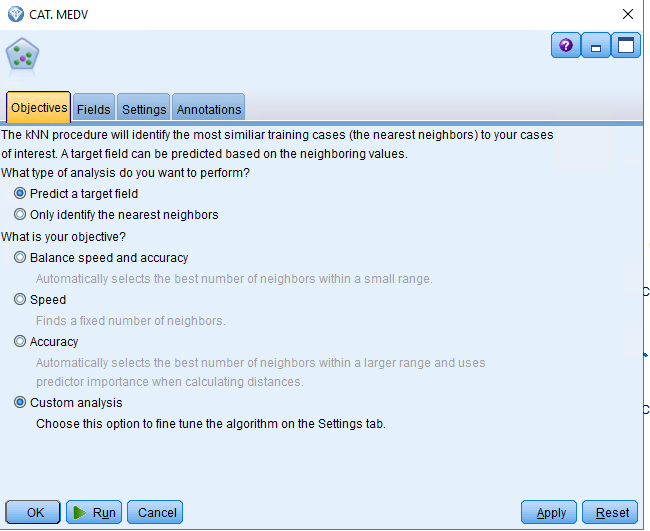
1. We have connected distribution node to have a visual of the CAT.MEDV attribute, below is the snapshot.

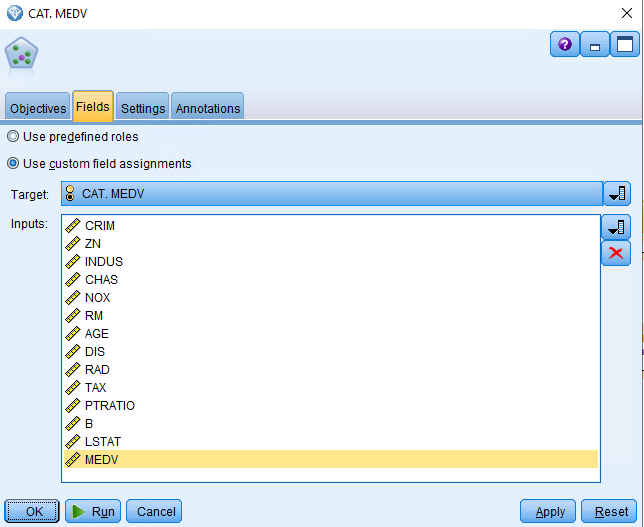


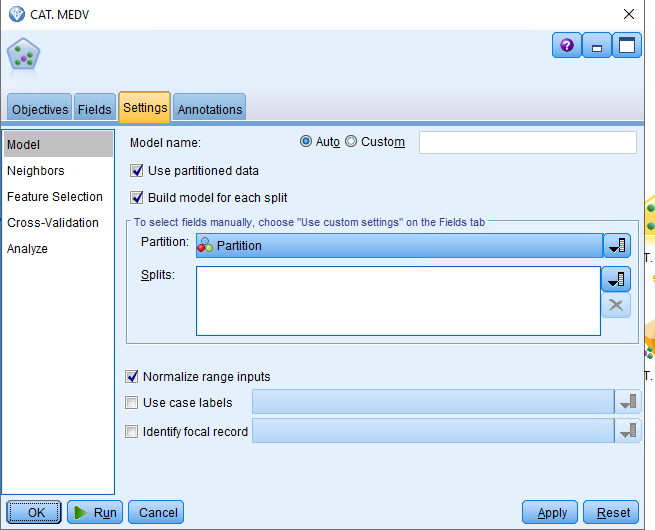
1. The next step is the partition the dataset in required ratio (70: 30).

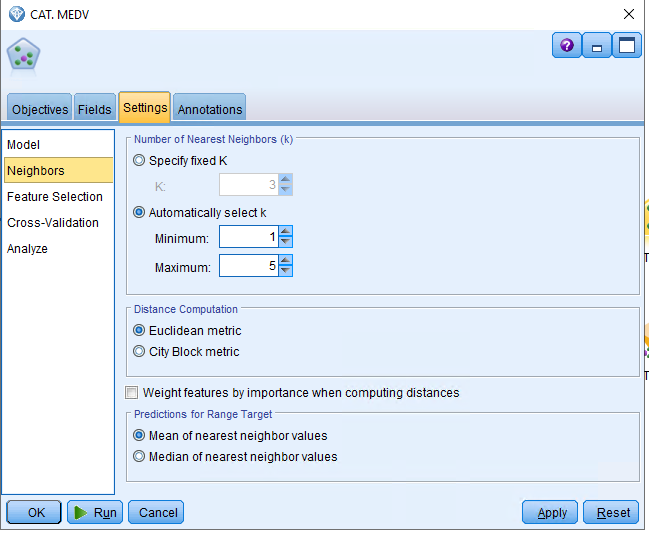


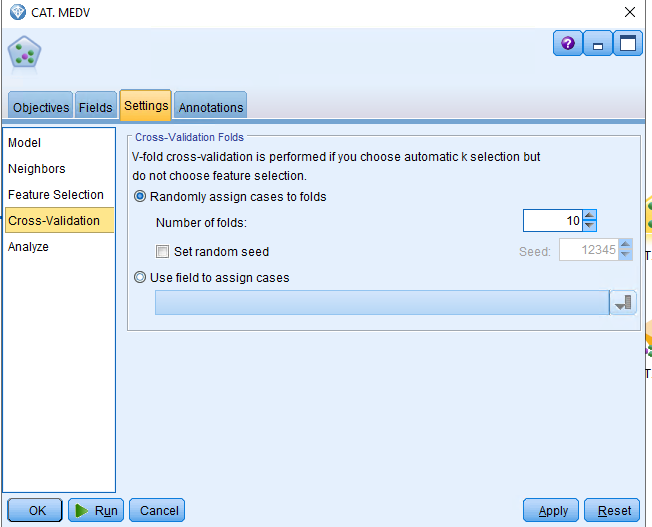
1. Using the KNN node from modeling we have implement the KNN with the following features are asked:

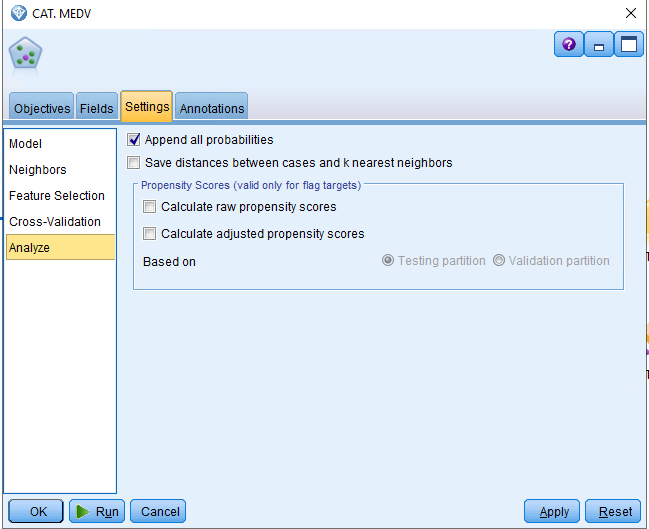




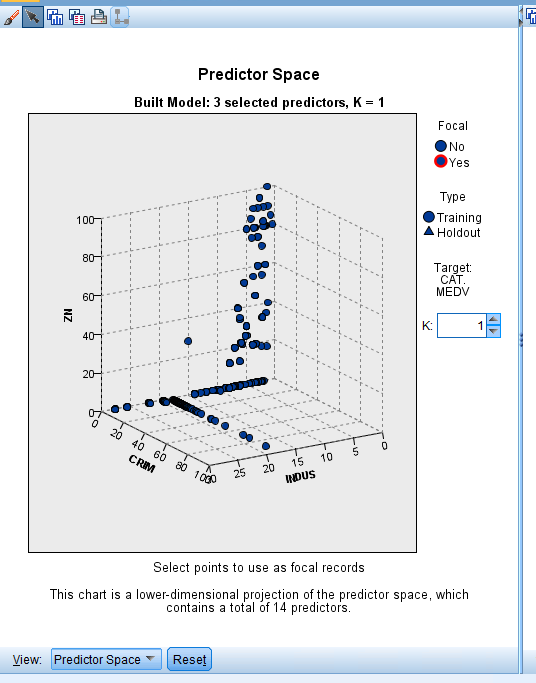




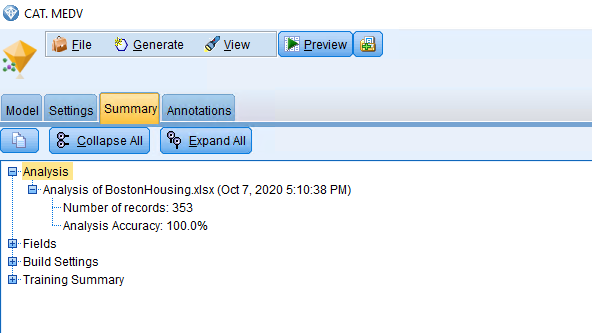




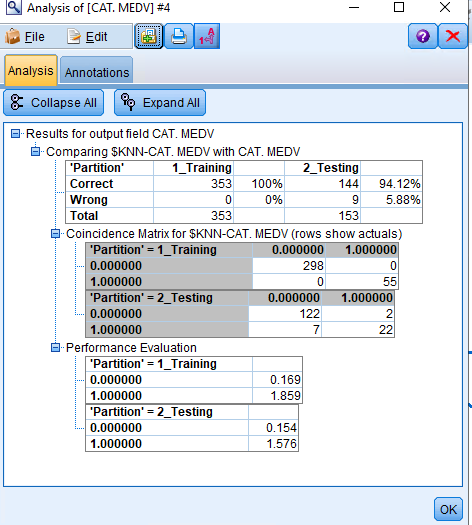
After creating the implementing the KNN with above specification, we have the diamond where the computed K value is 1:



When we had the view for summary tab from our diamond we can see the analysis accuracy is 100%



To have a better understanding, we have added a Analysis node from our KNN model, break down analysis field of KNN\_CAT.MEDV, we get:



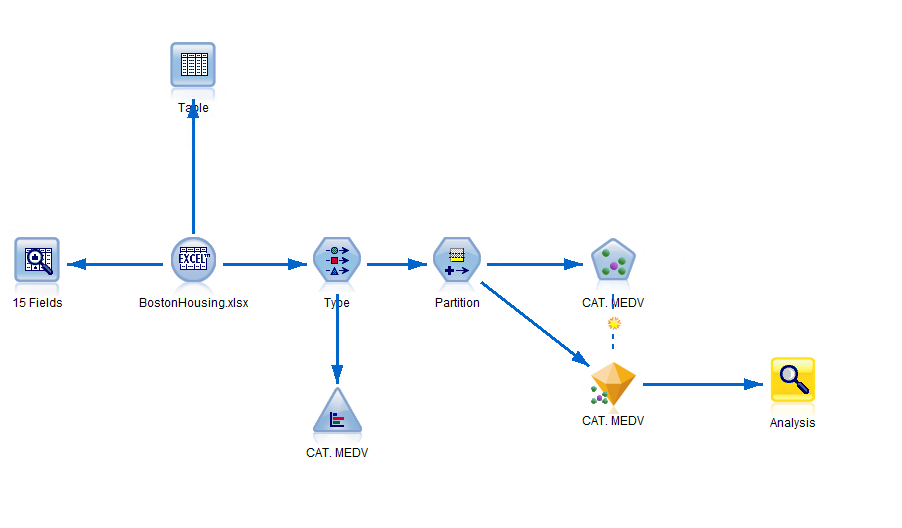
From the above image,

Predictive accuracy of Training set = (298+55)/ (298+0+55+0) = 100%

Predictive accuracy of Testing set = (122+22)/ (122+2+7+22) = 94.12%

As we can see from the above that the given data set is skewed and we are not able to predict the true prediction.

Below is the complete stream file:



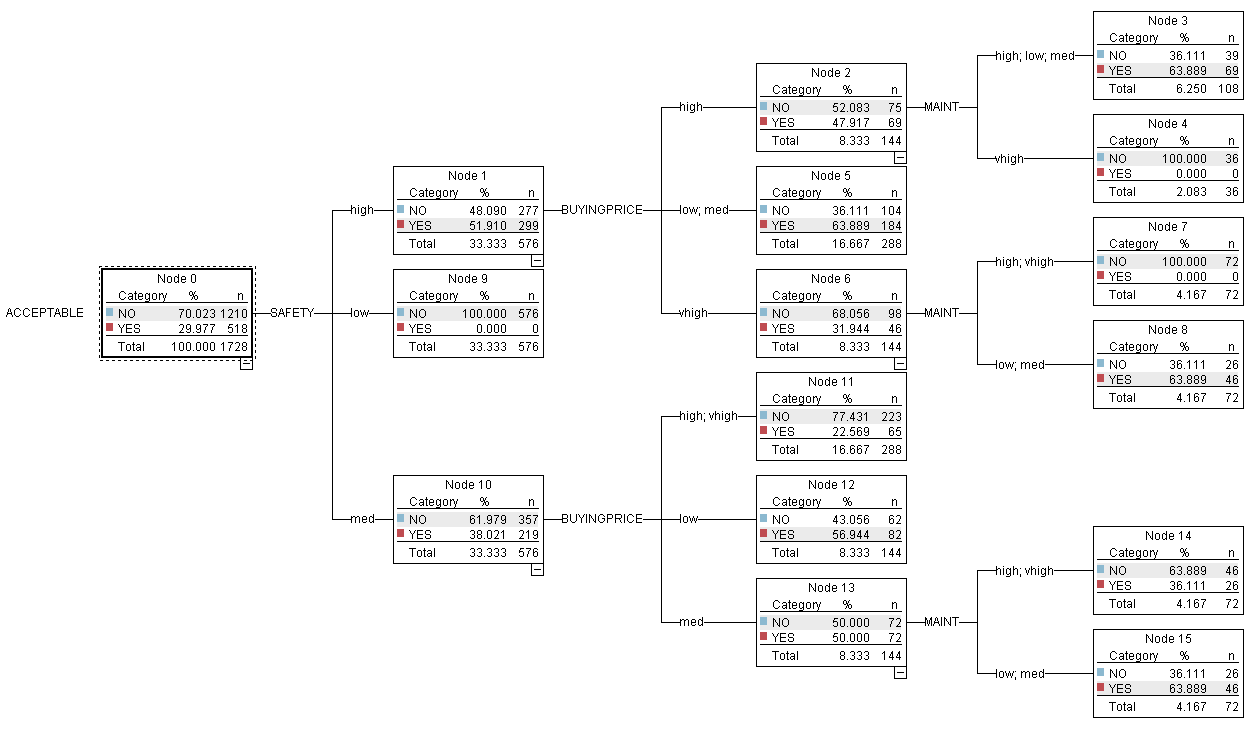
**Q5. This is a modified version of the Car Evaluation Database was derived from a simple hierarchical decision model originally developed for the demonstration of DEX, M. Bohanec, V. Rajkovic: Expert system for decision making. Sistemica 1(1), pp. 145-157, 1990.).**

**The Car Evaluation Database contains examples with the structural information removed, i.e., directly relates CAR to three input attributes: buyingprice, maintenance, safety.**

Attribute Information**:**

Class attribute Acceptable: YES, NO   
  
Attributes:   
buyingprice: vhigh, high, med, low.   
maint: vhigh, high, med, low.   
safety: low, med, high.

A C5.0 algorithm is applied on the whole data set, yielding the decision tree below.



Write the rules for car acceptability that can be derived from the decision tree, together with the support and the confidence of each rule (hint: check my notes and the textbook to calculate confidence and

**Answer**: The rules that can be derived from the decision tree together is given as below:

|  |  |  |  |
| --- | --- | --- | --- |
| Antecedent | Consequence | Support | Confidence |
| If (safety = high) and (buying price = high) and (Maintenance = high, low, med) | Then Acceptable = Yes | 69/1728 | 69/108 = 63.889% |
| If (safety = high) and (buying price = high) and (Maintenance = very high) | Then Acceptable = No | 36/1728 | 36/36 = 100% |
| If (safety = high) and (buying price = very high) and (Maintenance = high, very high) | Then Acceptable = No | 72/1728 | 72/72 = 100% |
| If (safety = high) and (buying price = very high) and (Maintenance = low, med) | Then Acceptable = Yes | 46/1728 | 46/72 = 63.889% |
| If (safety = med) and (buying price = med) and (Maintenance = high, very high) | Then Acceptable = No | 46/1728 | 46/72 = 63.889% |
| If (safety = med) and (buying price = med) and (Maintenance = low, med) | Then Acceptable = Yes | 46/1728 | 46/72 = 63.889% |
| If (safety = high) and (buying price = low, med) | Then Acceptable = No | 184/1728 | 184/288 = 63.889% |
| If (safety = med) and (buying price = high, very high) | Then Acceptable = No | 233/1728 | 223/288 = 77.43% |
| If (safety = med) and (buying price = low) | Then Acceptable = Yes | 82/1728 | 82/144 = 56.944% |
| If (Safety = low) | Then Acceptable = No | 576/1728 | 576/576 = 100% |