**EXERCISE 1 – 35 pts:**

The table below contains three training records with two numeric fields X (range: 1-10), and Y (range: 1-5), and a categorical variable (values A and B). Class is the categorical target variable. Find the Euclidean distance between each pair of points. Which records are closest?

Euclidean Distance = **((x2-x1) ^2 + (y2-y1) ^2) ^0.5** where points are (x1, y1) and (x2, y2).

Here we have 3 records Record 1(3,1) Record2(3,3) and record 3(1,3).

If we calculate the Euclidian distance between Record 1-2:

d (1-2) = √ [ (3-3) ^2 + (3-1) ^2)]

d (1-2) = √ [ (0) ^2 + (-2) ^2)] = √ [ (4)] = 2

Euclidian distance between Record 2-3:

d (2-3) = √ [ (1-3) ^2 + (3-3) ^2)]

d (2-3) = √ [(-2) ^2) + (0) ^2] = √ [ (4)] = 2

Euclidean between record 3 and 1:

d (3-1) = √ [ (3-1) ^2 + (1-3) ^2)]

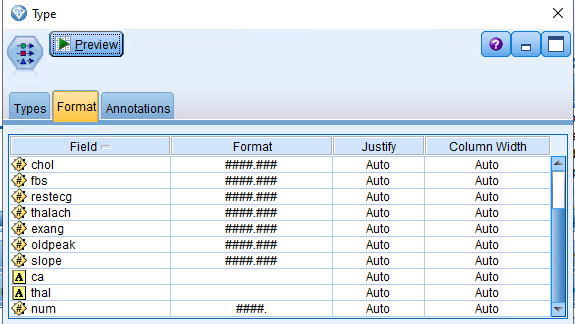
d (3-1) = √ [(2) ^2) + (-2) ^2] = √ [ (8)] = 2√2.

From the above we can say that the record 1 -2 and record 2-3 are the closest.

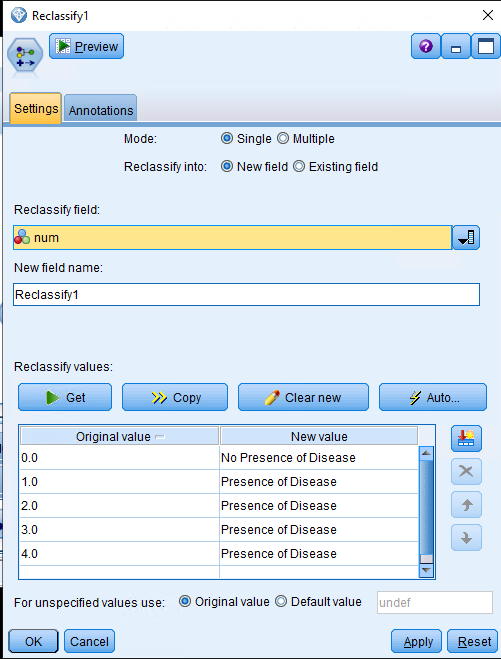
**EXERCISE 2 – 65 pts**

a) Derive a binary target field from field 14 (num) where 0 implies no disease and 1 implies presence of disease, and choose appropriate data types for the predictors when reading the data.

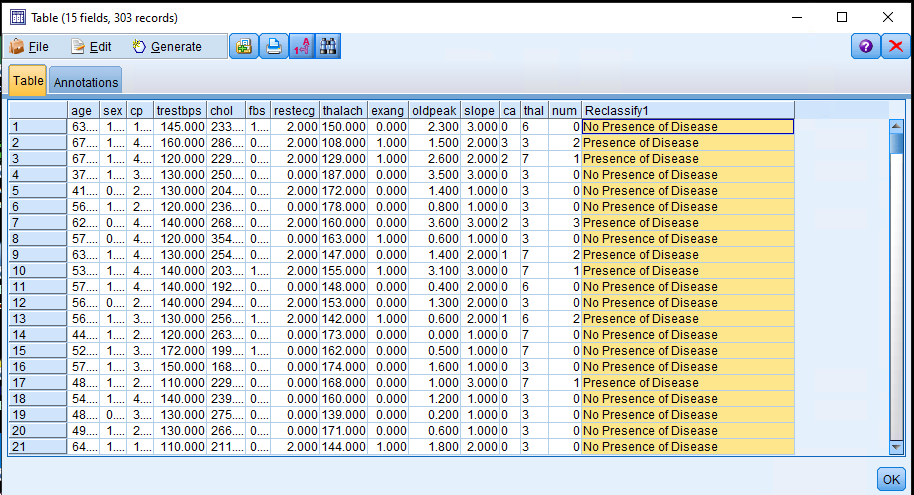
In order to derive the binary target field as required, we first have to use the type node to change the format for the field 14 (num) to read the given dataset. Below is the snapshot for it.



Then we have used the reclassify node to achieve the desired task, we have changed the label as required, we have created a new value for 0 to “No Presence of Disease” and the rest (1,2,3,4) to Presence of Disease.

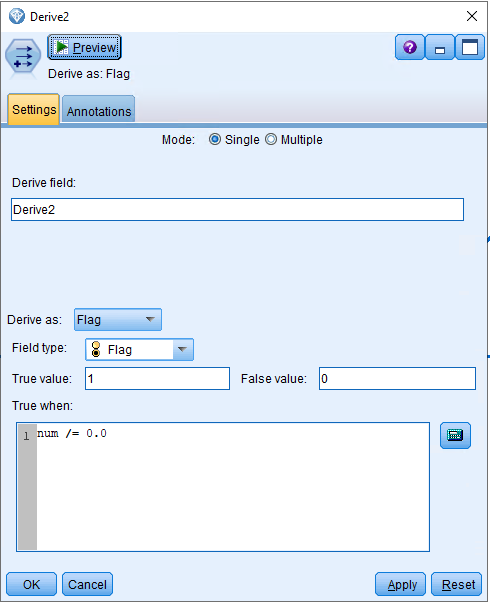


Which can be seen in above snapshot.

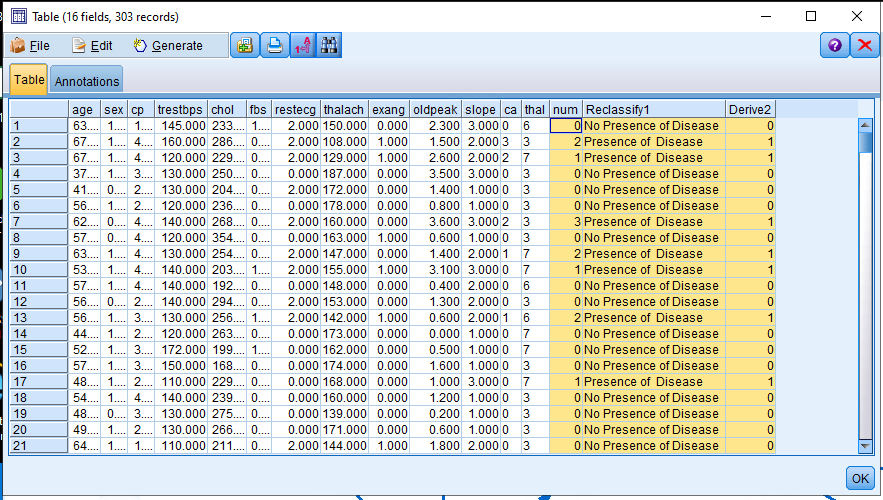


The above is the snapshot with the updated table view as we have derived a binary target field from field 14 (num).

But the above reclassified node is not performing the required task as we need binary target for entire field 14 (num).

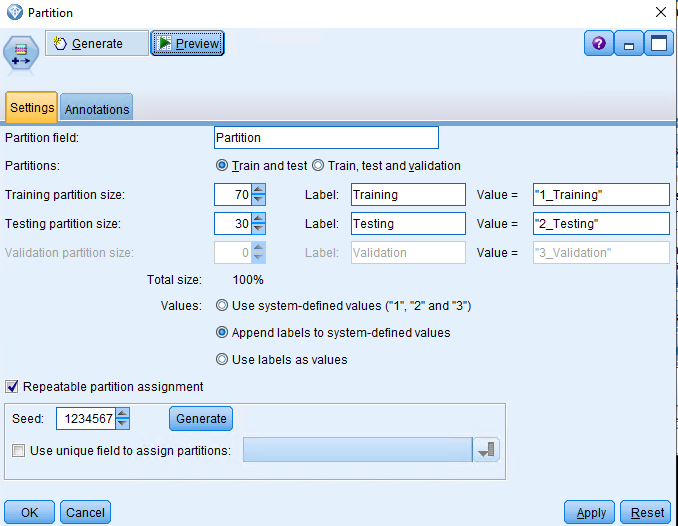


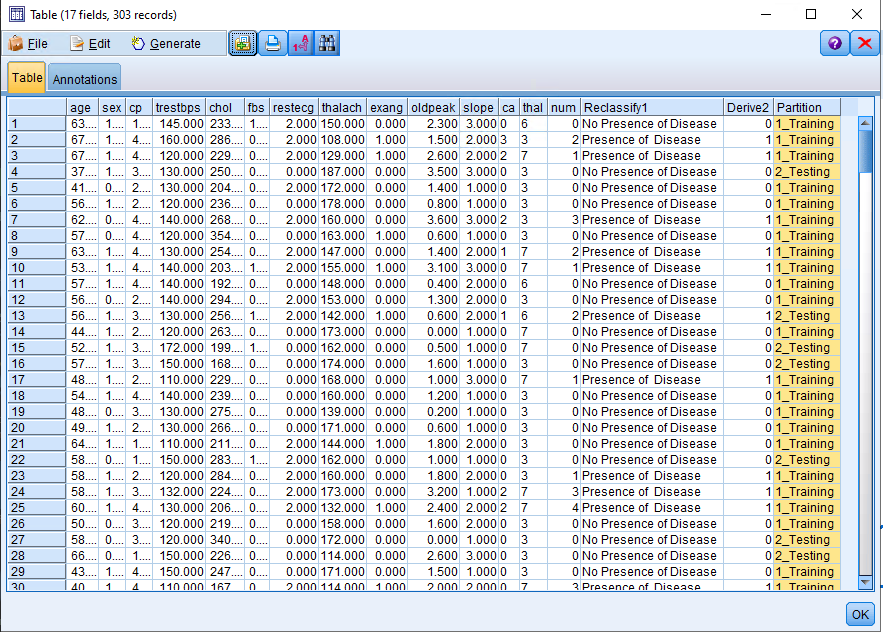
The above is snapshot of derived node taken as the flag type to perform the required task.



If we connect a table to see the updated dataset, we can see that the we are now having the required binary field.

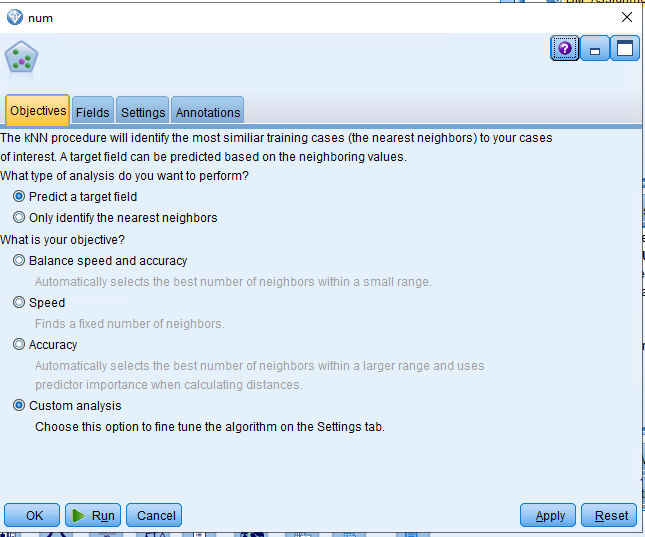
**b) Partition the data into training (70%) and test (30%) sets**.

In order to complete this step, we have first used the partition node and divided the data set 70% for training and 30% for testing, below is the snapshot for it.

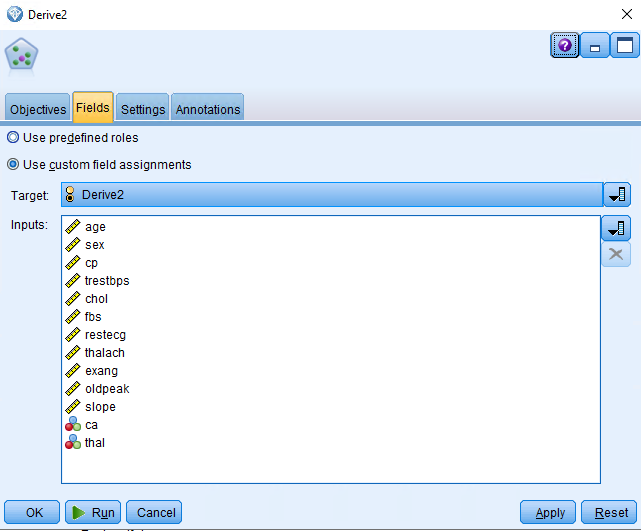


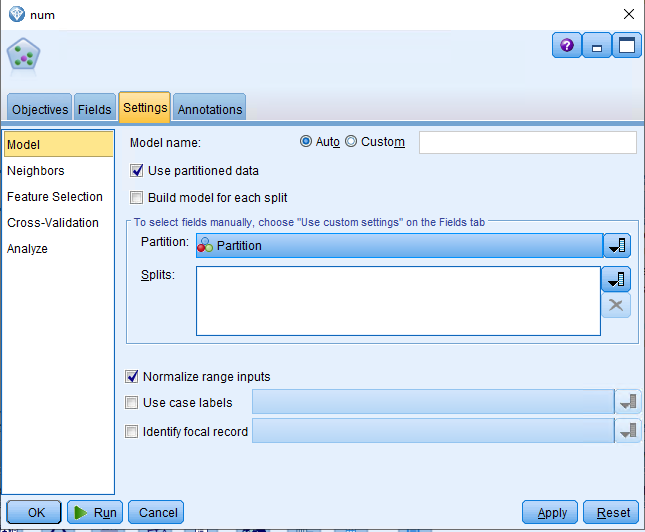
In the above snapshot we can see that the data set is randomly divided into training and testing as in required ratio.

**c) Perform a k-NN classification with all 13 predictors. Choose automatic k selection (range of k: 2 to 10). Make sure to check the box to normalize input data. Use Euclidean distance and weight features by importance when computing distances. Perform cross-validation and no feature selection.**

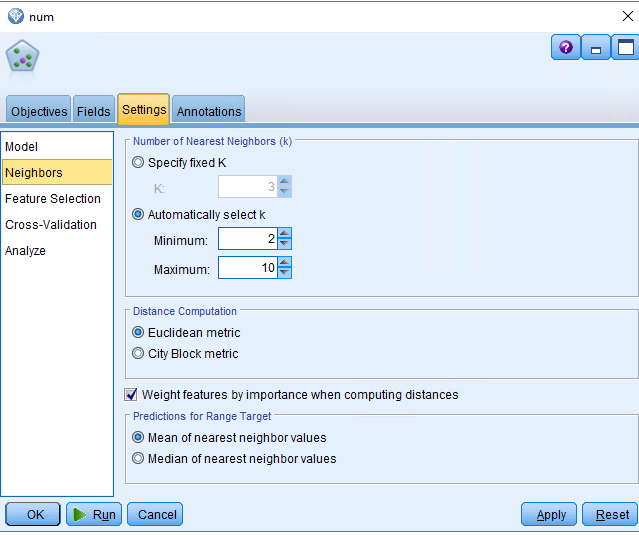


For this step of the question, we have used the KNN node under the model option in SPSS, and selected the custom analysis option, as shown above.

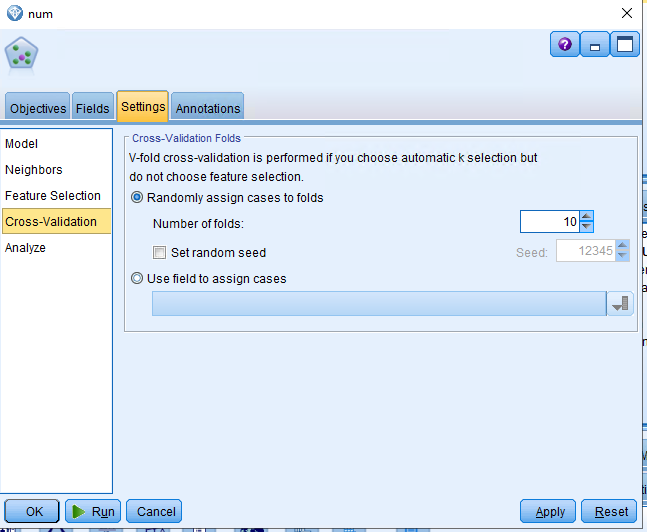
As required, we have used the 13 fields as the input and derived as our target field, this is very important.

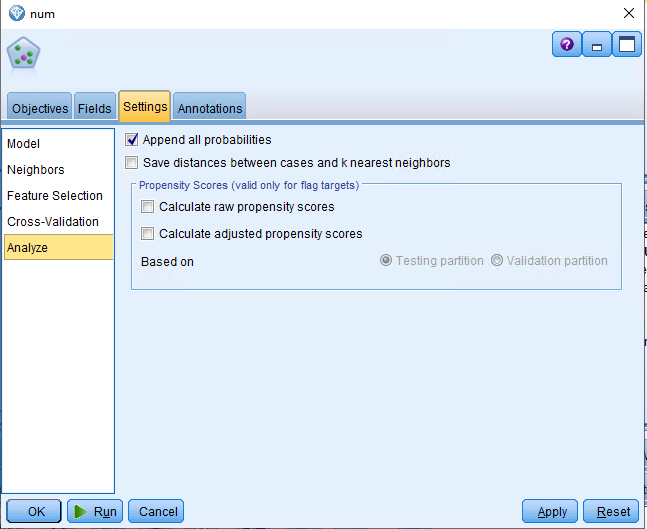


We have chosen the partition data, normalized range inputs under the model settings.

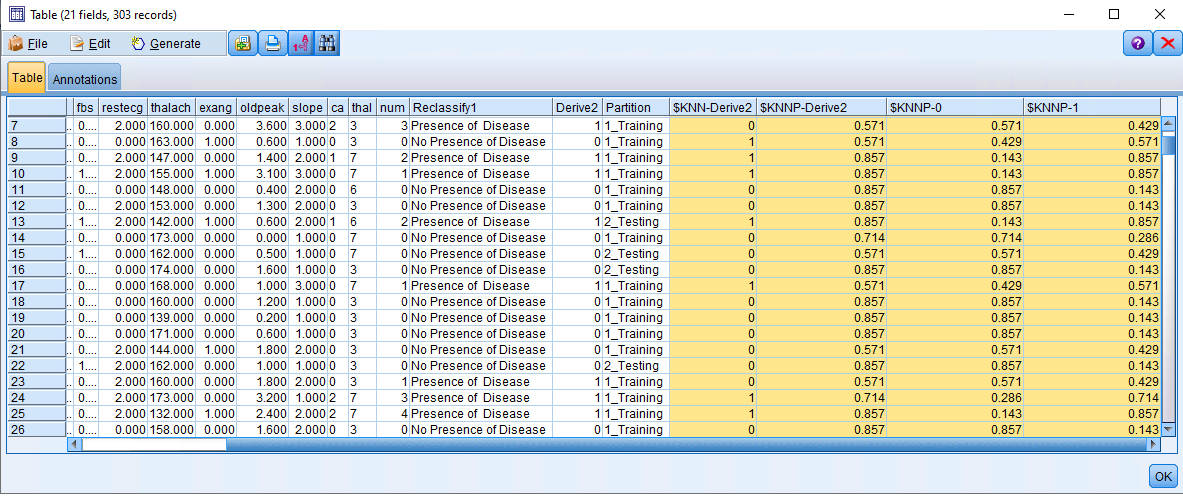


Choose automatic k selection (range of k: 2 to 10) and selected Euclidean distance and weight features by importance when computing distances as shown from the above snapshot.



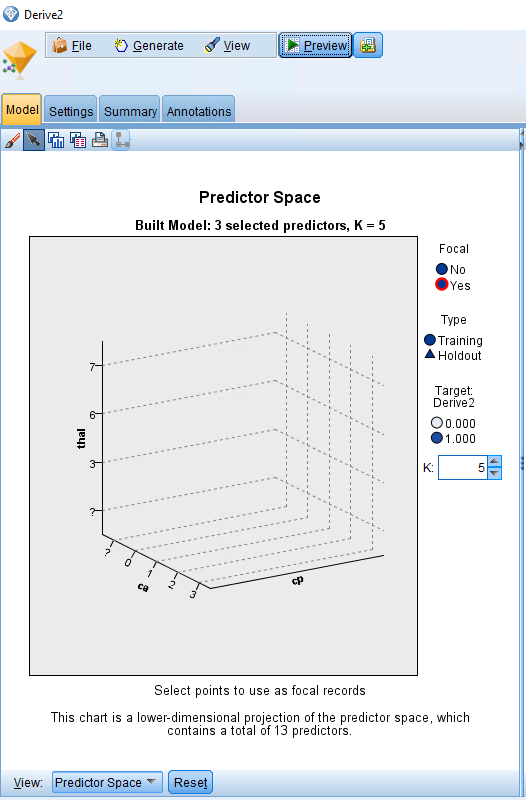


The above two snapshot are the cross validation and analyze features selected under the model, we have chosen “append all probabilities” while analyze the final data.



In the above snapshot we can see all the probabilities using KNN algorithm has been created, this can snapshot from the table connected to the nugget created after running the KNN node.

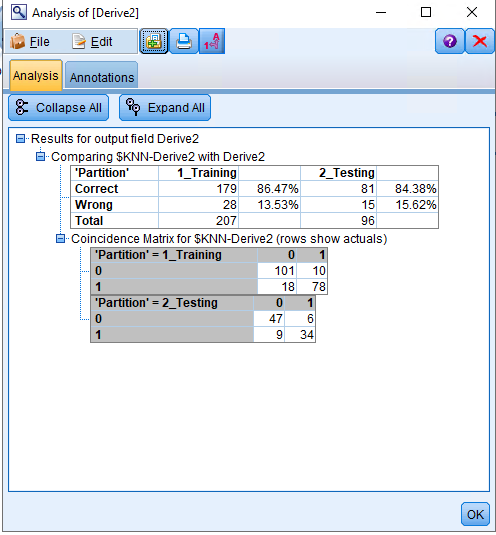
**d)** **What is the best value of K chosen by SPSS Modeler?**



We can see from the above snapshot that we are getting the k= 5 value from SPSS modeler.

**e) Perform an evaluation analysis of the classification task, reporting accuracy, precision, TP rate (aka recall) and FP rate.**

In order perform the evaluation analysis of the classification task we first need the confusion matrix, below is the snapshot from the analysis node connected to the diamond created after running the KNN model node.



* Reporting accuracy for the test data is: (47+34)/47+34+6+9) = 84.38%.
* We have this value slightly smaller than training accuracy but data set is balanced about heart disease diagnosis (sign of disease to non-disease).
* Precision = 34/ (34 + 6) = 85%
* TP rate (aka recall) = 34/ (9+34) = 79.06%
* FP rate can be calculated by: (1 – Specificity) = 1 – [47 / (47 + 6)] = 11.3%

From the above predictive performance outcome, we can say that from the given 14 attributes:

* We are able to capture 84.38% of true smokers who smokes 1,2,3 or 4 cigarettes a day.
* 85% of the prospects diagnosed predicted were actual smokers or who smokes 1 or more than 1 cigarette a day.
* 11.3% of the diagnosed were non smoker or who smokes 0 cigarette a day.

Below is the snapshot of complete stream file.

