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

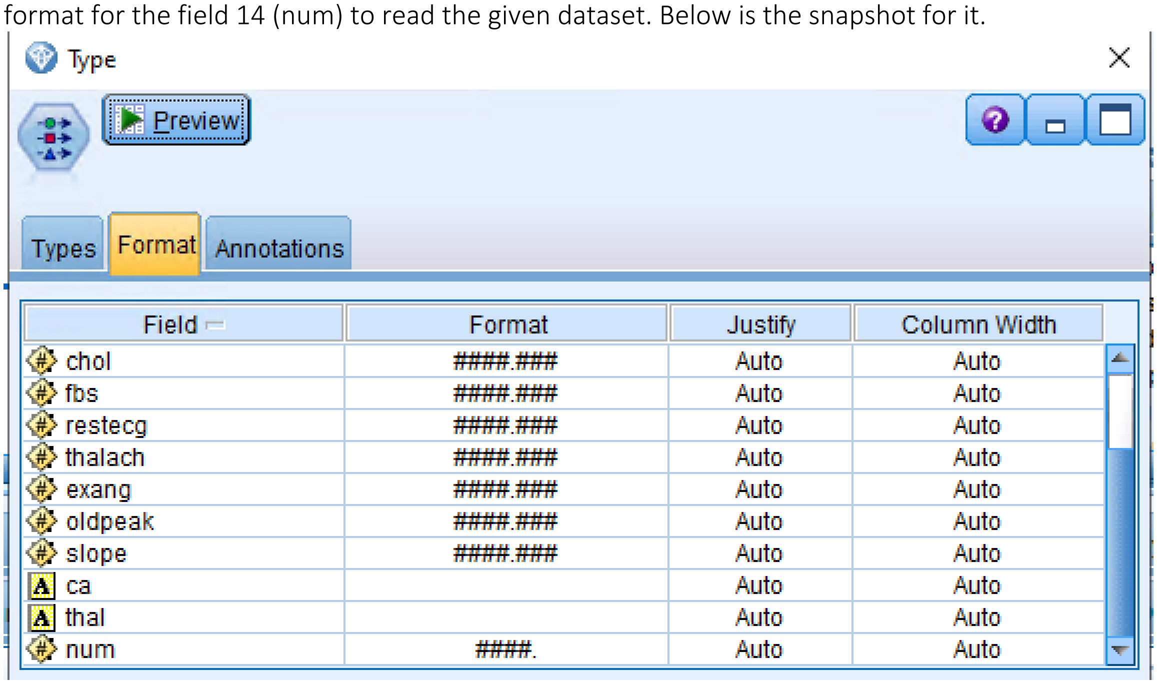
Euclidian distance between Record 2-3:
d (2-3) = √ [(1-3) ^2 + (3-3) ^2)]
d (2-3) = √ [(-2) ^2) + (0) ^2] = √ [(4)] = 2

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.

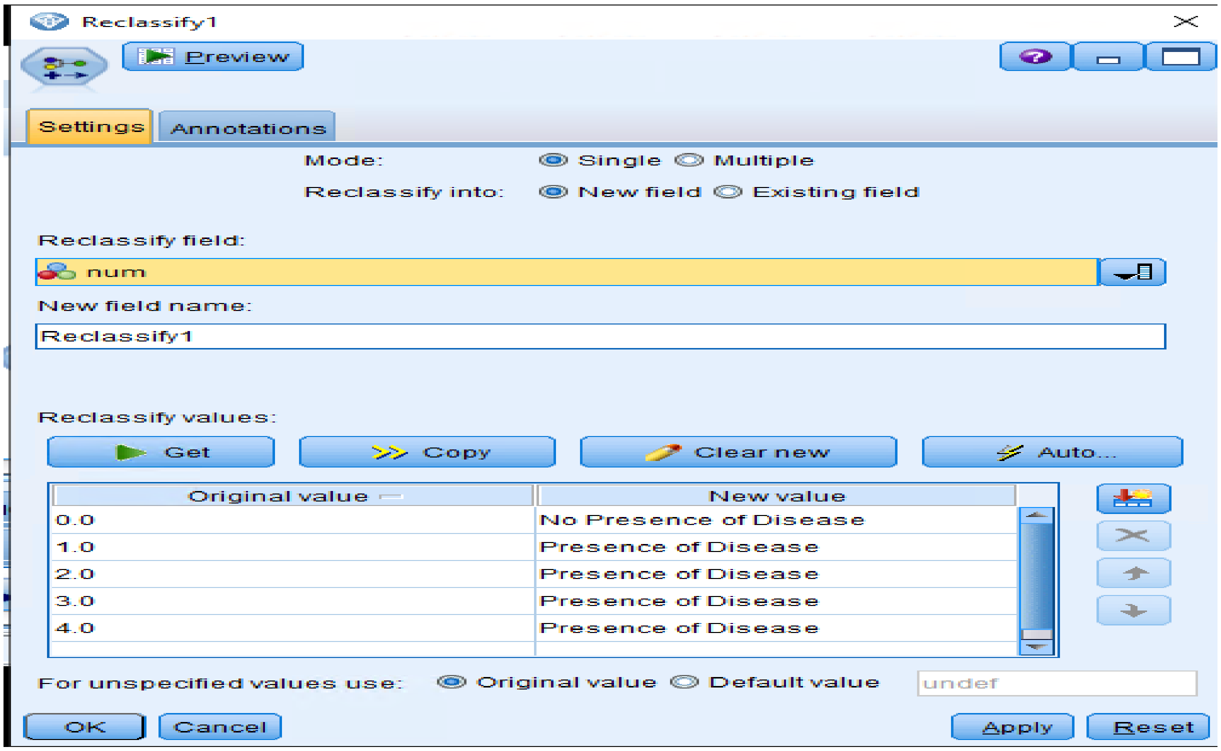
EXERCISE 2 – 65 pts

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.

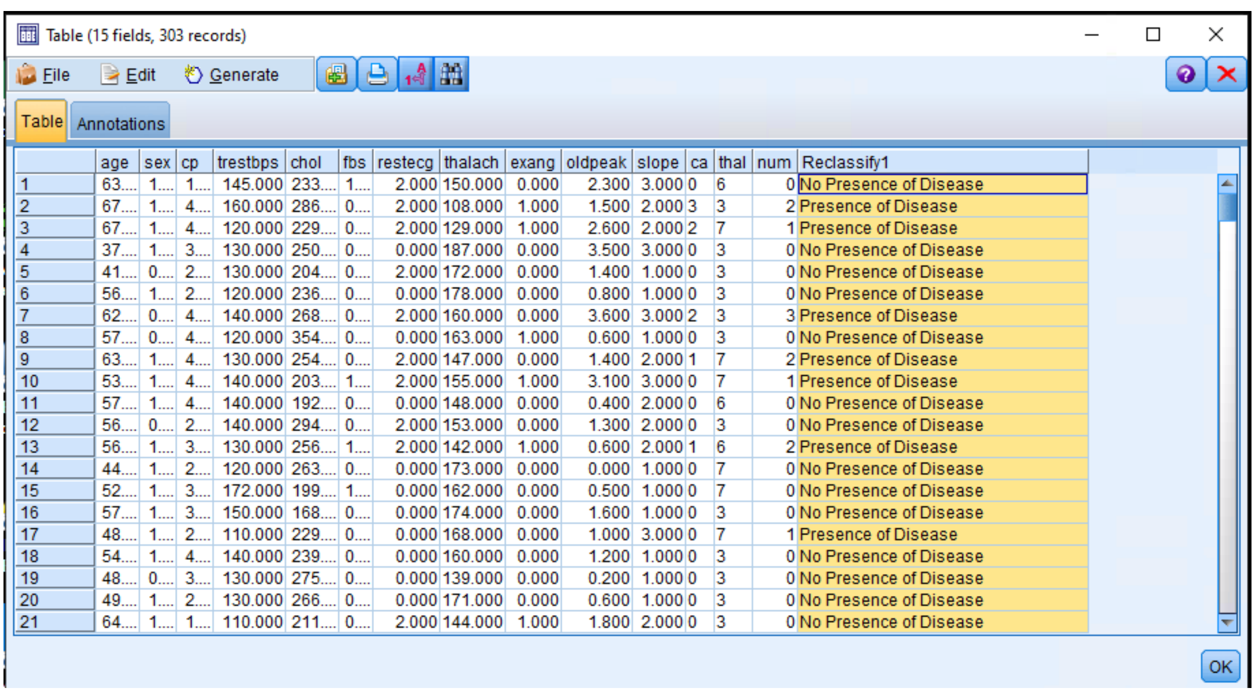
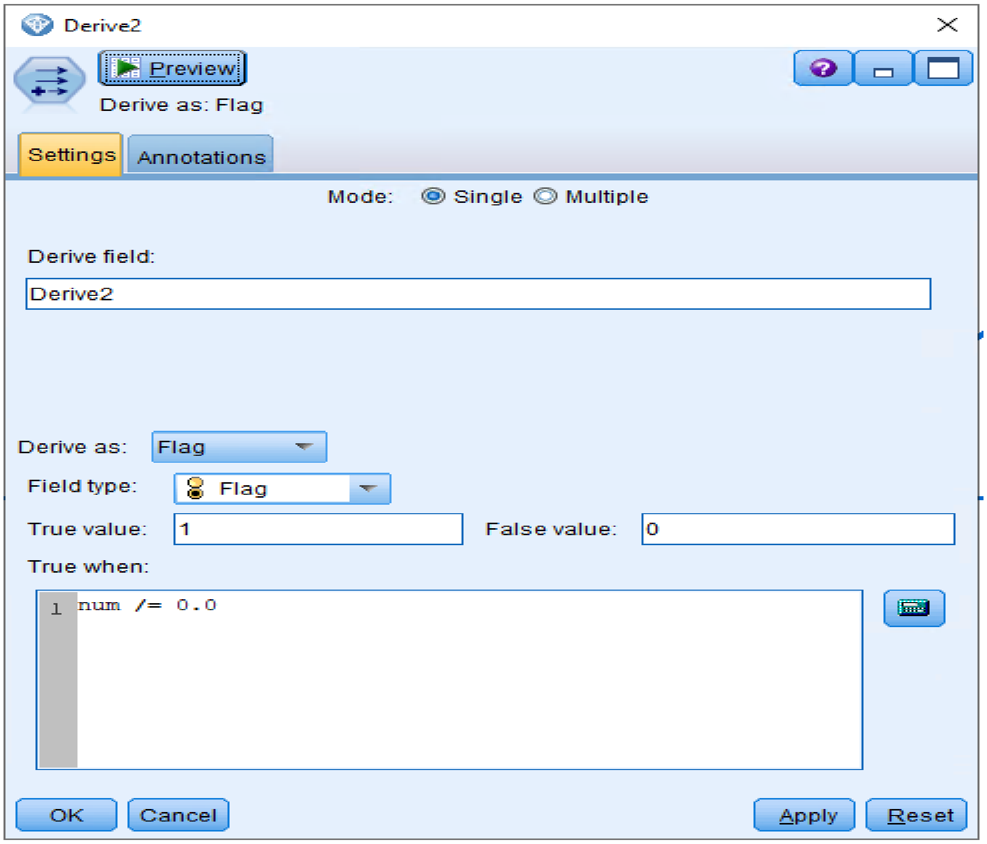


Table (15 fields, 303 records)

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num	Reclassify1
1	63....	1....	1....	145.000	233....	1....	2.000	150.000	0.000	2.300	3.000	0	6	0	No Presence of Disease
2	67....	1....	4....	160.000	286....	0....	2.000	108.000	1.000	1.500	2.000	3	3	2	Presence of Disease
3	67....	1....	4....	120.000	229....	0....	2.000	129.000	1.000	2.600	2.000	2	7	1	Presence of Disease
4	37....	1....	3....	130.000	250....	0....	0.000	187.000	0.000	3.500	3.000	0	3	0	No Presence of Disease
5	41....	0....	2....	130.000	204....	0....	2.000	172.000	0.000	1.400	1.000	0	3	0	No Presence of Disease
6	56....	1....	2....	120.000	236....	0....	0.000	178.000	0.000	0.800	1.000	0	3	0	No Presence of Disease
7	62....	0....	4....	140.000	268....	0....	2.000	160.000	0.000	3.600	3.000	2	3	3	Presence of Disease
8	57....	0....	4....	120.000	354....	0....	0.000	163.000	1.000	0.600	1.000	0	3	0	No Presence of Disease
9	63....	1....	4....	130.000	254....	0....	2.000	147.000	0.000	1.400	2.000	1	7	2	Presence of Disease
10	53....	1....	4....	140.000	203....	1....	2.000	155.000	1.000	3.100	3.000	0	7	1	Presence of Disease
11	57....	1....	4....	140.000	192....	0....	0.000	148.000	0.000	0.400	2.000	0	6	0	No Presence of Disease
12	56....	0....	2....	140.000	294....	0....	2.000	153.000	0.000	1.300	2.000	0	3	0	No Presence of Disease
13	56....	1....	3....	130.000	256....	1....	2.000	142.000	1.000	0.600	2.000	1	6	2	Presence of Disease
14	44....	1....	2....	120.000	263....	0....	0.000	173.000	0.000	0.000	1.000	0	7	0	No Presence of Disease
15	52....	1....	3....	172.000	199....	1....	0.000	162.000	0.000	0.500	1.000	0	7	0	No Presence of Disease
16	57....	1....	3....	150.000	168....	0....	0.000	174.000	0.000	1.600	1.000	0	3	0	No Presence of Disease
17	48....	1....	2....	110.000	229....	0....	0.000	168.000	0.000	1.000	3.000	0	7	1	Presence of Disease
18	54....	1....	4....	140.000	239....	0....	0.000	160.000	0.000	1.200	1.000	0	3	0	No Presence of Disease
19	48....	0....	3....	130.000	275....	0....	0.000	139.000	0.000	0.200	1.000	0	3	0	No Presence of Disease
20	49....	1....	2....	130.000	266....	0....	0.000	171.000	0.000	0.600	1.000	0	3	0	No Presence of Disease
21	64....	1....	1....	110.000	211....	0....	2.000	144.000	1.000	1.800	2.000	0	3	0	No Presence of Disease

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.

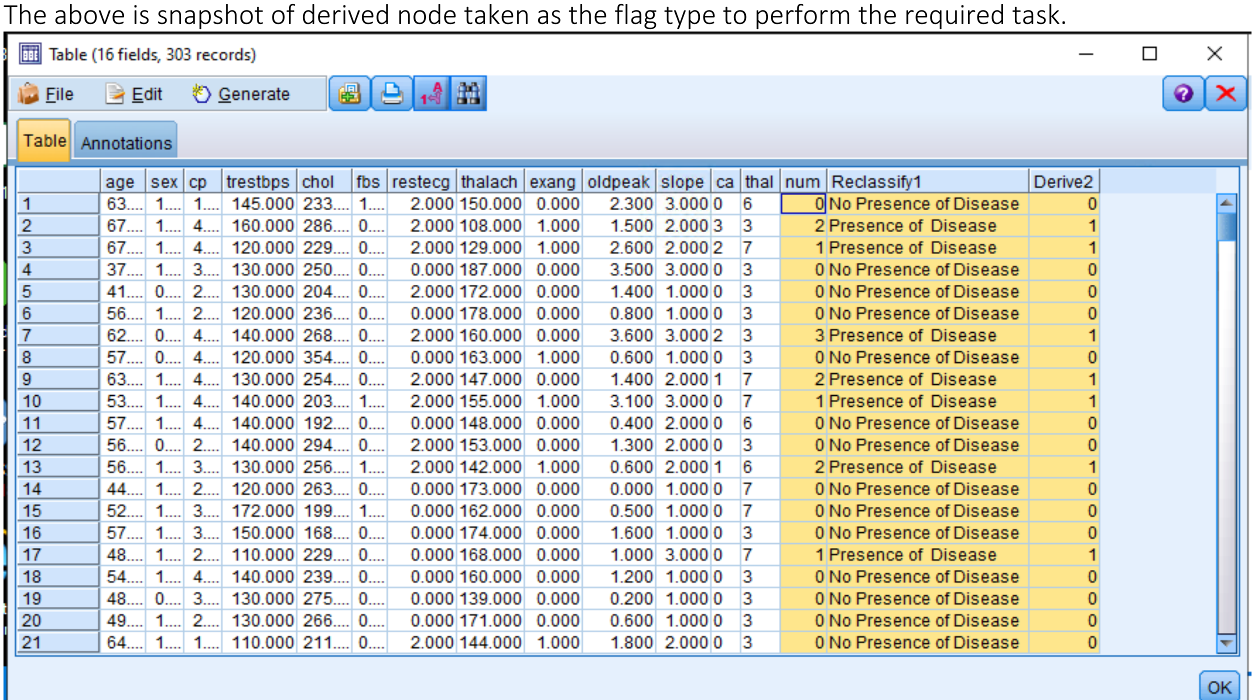


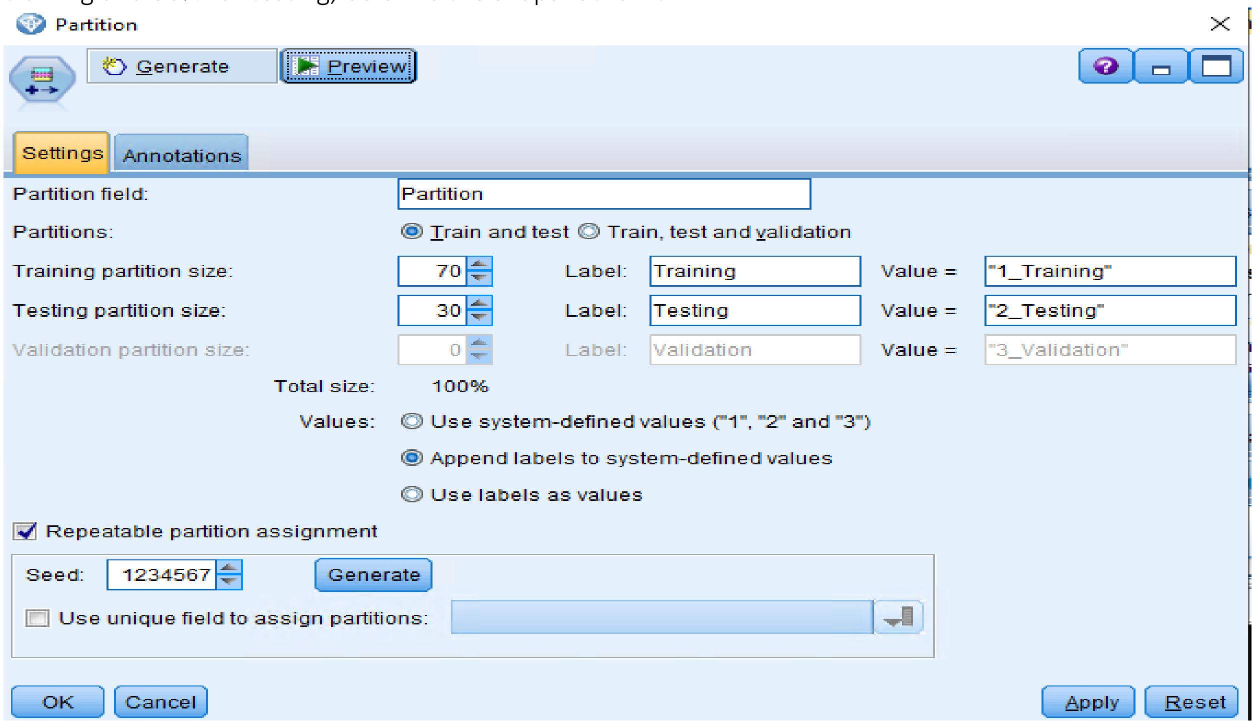
Table (16 fields, 303 records)

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num	Reclassify1	Derive2
1	63....	1....	1....	145.000	233....	1....	2.000	150.000	0.000	2.300	3.000	0	6	0	No Presence of Disease	0
2	67....	1....	4....	160.000	286....	0....	2.000	108.000	1.000	1.500	2.000	3	3	2	Presence of Disease	1
3	67....	1....	4....	120.000	229....	0....	2.000	129.000	1.000	2.600	2.000	2	7	1	Presence of Disease	1
4	37....	1....	3....	130.000	250....	0....	0.000	187.000	0.000	3.500	3.000	0	3	0	No Presence of Disease	0
5	41....	0....	2....	130.000	204....	0....	2.000	172.000	0.000	1.400	1.000	0	3	0	No Presence of Disease	0
6	56....	1....	2....	120.000	236....	0....	0.000	178.000	0.000	0.800	1.000	0	3	0	No Presence of Disease	0
7	62....	0....	4....	140.000	268....	0....	2.000	160.000	0.000	3.600	3.000	2	3	3	Presence of Disease	1
8	57....	0....	4....	120.000	354....	0....	0.000	163.000	1.000	0.600	1.000	0	3	0	No Presence of Disease	0
9	63....	1....	4....	130.000	254....	0....	2.000	147.000	0.000	1.400	2.000	1	7	2	Presence of Disease	1
10	53....	1....	4....	140.000	203....	1....	2.000	155.000	1.000	3.100	3.000	0	7	1	Presence of Disease	1
11	57....	1....	4....	140.000	192....	0....	0.000	148.000	0.000	0.400	2.000	0	6	0	No Presence of Disease	0
12	56....	0....	2....	140.000	294....	0....	2.000	153.000	0.000	1.300	2.000	0	3	0	No Presence of Disease	0
13	56....	1....	3....	130.000	256....	1....	2.000	142.000	1.000	0.600	2.000	1	6	2	Presence of Disease	1
14	44....	1....	2....	120.000	263....	0....	0.000	173.000	0.000	0.000	1.000	0	7	0	No Presence of Disease	0
15	52....	1....	3....	172.000	199....	1....	0.000	162.000	0.000	0.500	1.000	0	7	0	No Presence of Disease	0
16	57....	1....	3....	150.000	168....	0....	0.000	174.000	0.000	1.600	1.000	0	3	0	No Presence of Disease	0
17	48....	1....	2....	110.000	229....	0....	0.000	168.000	0.000	1.000	3.000	0	7	1	Presence of Disease	1
18	54....	1....	4....	140.000	239....	0....	0.000	160.000	0.000	1.200	1.000	0	3	0	No Presence of Disease	0
19	48....	0....	3....	130.000	275....	0....	0.000	139.000	0.000	0.200	1.000	0	3	0	No Presence of Disease	0
20	49....	1....	2....	130.000	266....	0....	0.000	171.000	0.000	0.600	1.000	0	3	0	No Presence of Disease	0
21	64....	1....	1....	110.000	211....	0....	2.000	144.000	1.000	1.800	2.000	0	3	0	No Presence of Disease	0

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.



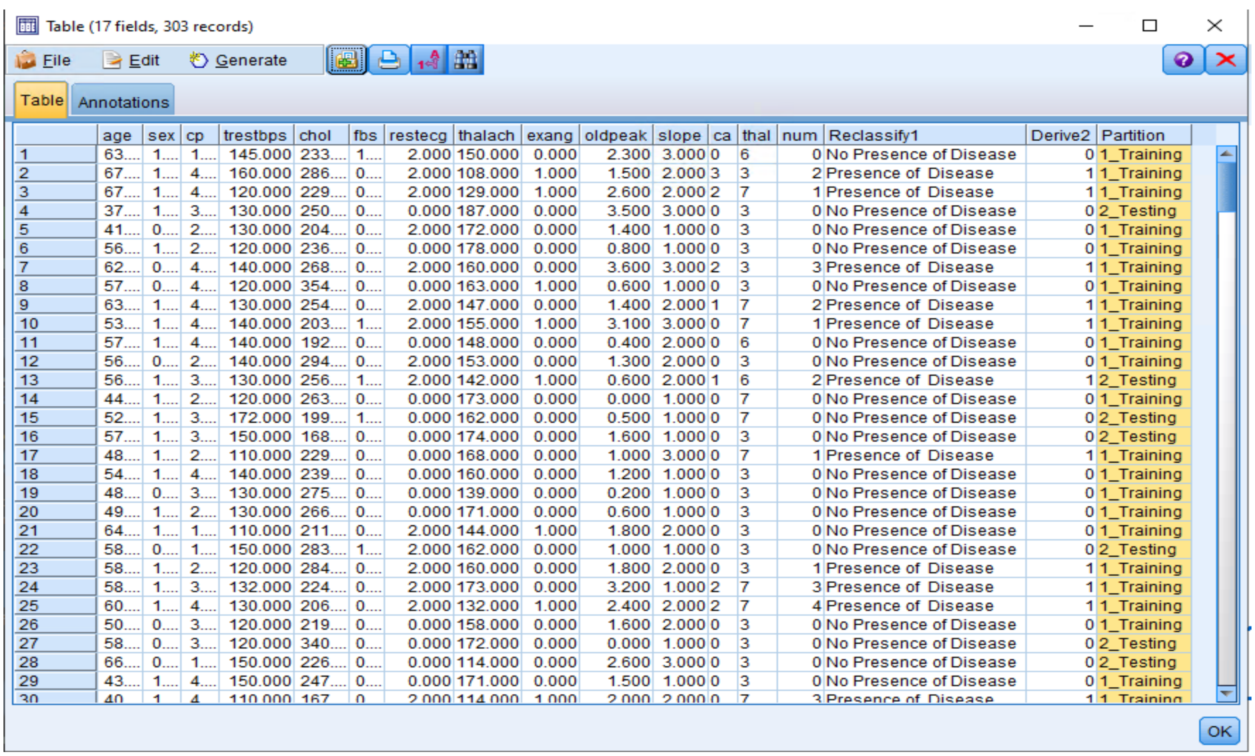
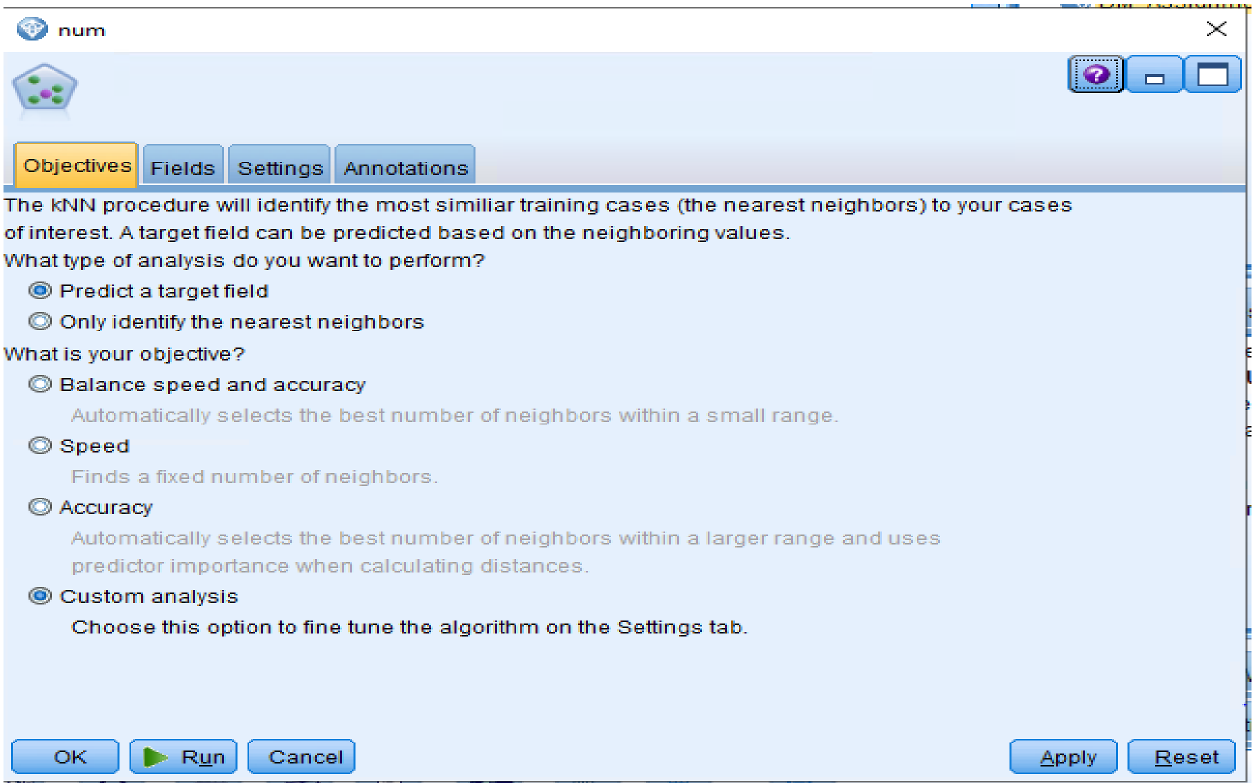


Table (17 fields, 303 records)

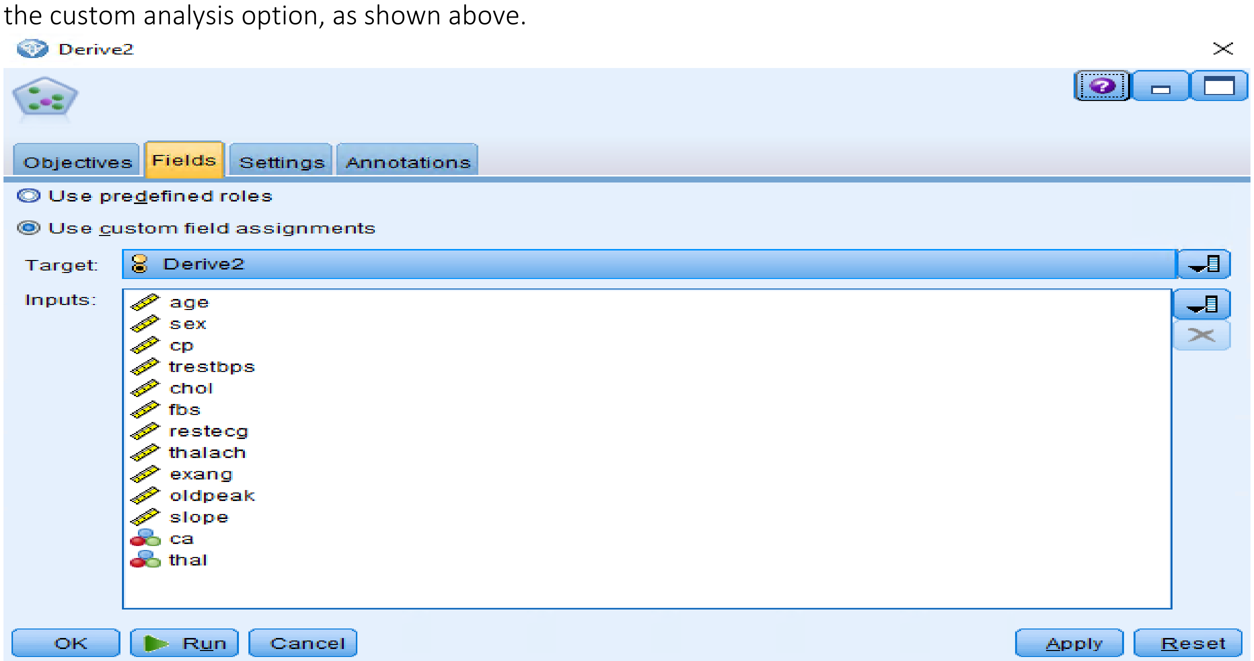
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num	Reclassify1	Derive2	Partition
1	63....	1....	1....	145.000	233....	1....	2.000	150.000	0.000	2.300	3.000	0	6	0	No Presence of Disease	0	1_Training
2	67....	1....	4....	160.000	286....	0....	2.000	108.000	1.000	1.500	2.000	3	3	2	Presence of Disease	1	1_Training
3	67....	1....	4....	120.000	229....	0....	2.000	129.000	1.000	2.600	2.000	2	7	1	Presence of Disease	1	1_Training
4	37....	1....	3....	130.000	250....	0....	0.000	187.000	0.000	3.500	3.000	0	3	0	No Presence of Disease	0	2_Testing
5	41....	0....	2....	130.000	204....	0....	2.000	172.000	0.000	1.400	1.000	0	3	0	No Presence of Disease	0	1_Training
6	56....	1....	2....	120.000	236....	0....	0.000	178.000	0.000	0.800	1.000	0	3	0	No Presence of Disease	0	1_Training
7	62....	0....	4....	140.000	268....	0....	2.000	160.000	0.000	3.600	3.000	2	3	3	Presence of Disease	1	1_Training
8	57....	0....	4....	120.000	354....	0....	0.000	163.000	1.000	0.600	1.000	0	3	0	No Presence of Disease	0	1_Training
9	63....	1....	4....	130.000	254....	0....	2.000	147.000	0.000	1.400	2.000	1	7	2	Presence of Disease	1	1_Training
10	53....	1....	4....	140.000	203....	1....	2.000	155.000	1.000	3.100	3.000	0	7	1	Presence of Disease	1	1_Training
11	57....	1....	4....	140.000	192....	0....	0.000	148.000	0.000	0.400	2.000	0	6	0	No Presence of Disease	0	1_Training
12	56....	0....	2....	140.000	294....	0....	2.000	153.000	0.000	1.300	2.000	0	3	0	No Presence of Disease	0	1_Training
13	56....	1....	3....	130.000	256....	1....	2.000	142.000	1.000	0.600	2.000	1	6	2	Presence of Disease	1	2_Testing
14	44....	1....	2....	120.000	263....	0....	0.000	173.000	0.000	0.000	1.000	0	7	0	No Presence of Disease	0	1_Training
15	52....	1....	3....	172.000	199....	1....	0.000	162.000	0.000	0.500	1.000	0	7	0	No Presence of Disease	0	2_Testing
16	57....	1....	3....	150.000	168....	0....	0.000	174.000	0.000	1.600	1.000	0	3	0	No Presence of Disease	0	2_Testing
17	48....	1....	2....	110.000	229....	0....	0.000	168.000	0.000	1.000	3.000	0	7	1	Presence of Disease	1	1_Training
18	54....	1....	4....	140.000	239....	0....	0.000	160.000	0.000	1.200	1.000	0	3	0	No Presence of Disease	0	1_Training
19	48....	0....	3....	130.000	275....	0....	0.000	139.000	0.000	0.200	1.000	0	3	0	No Presence of Disease	0	1_Training
20	49....	1....	2....	130.000	266....	0....	0.000	171.000	0.000	0.600	1.000	0	3	0	No Presence of Disease	0	1_Training
21	64....	1....	1....	110.000	211....	0....	2.000	144.000	1.000	1.800	2.000	0	3	0	No Presence of Disease	0	1_Training
22	58....	0....	1....	150.000	283....	1....	2.000	162.000	0.000	1.000	1.000	0	3	0	No Presence of Disease	0	2_Testing
23	58....	1....	2....	120.000	284....	0....	2.000	160.000	0.000	1.800	2.000	0	3	1	Presence of Disease	1	1_Training
24	58....	1....	3....	132.000	224....	0....	2.000	173.000	0.000	3.200	1.000	2	7	3	Presence of Disease	1	1_Training
25	60....	1....	4....	130.000	206....	0....	2.000	132.000	1.000	2.400	2.000	2	7	4	Presence of Disease	1	1_Training
26	50....	0....	3....	120.000	219....	0....	0.000	158.000	0.000	1.600	2.000	0	3	0	No Presence of Disease	0	1_Training
27	58....	0....	3....	120.000	340....	0....	0.000	172.000	0.000	0.000	1.000	0	3	0	No Presence of Disease	0	2_Testing
28	66....	0....	1....	150.000	226....	0....	0.000	114.000	0.000	2.600	3.000	0	3	0	No Presence of Disease	0	2_Testing
29	43....	1....	4....	150.000	247....	0....	0.000	171.000	0.000	1.500	1.000	0	3	0	No Presence of Disease	0	1_Training
30	40....	1....	4....	110.000	167....	0....	2.000	114.000	1.000	2.000	2.000	0	7	3	Presence of Disease	1	1_Training

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

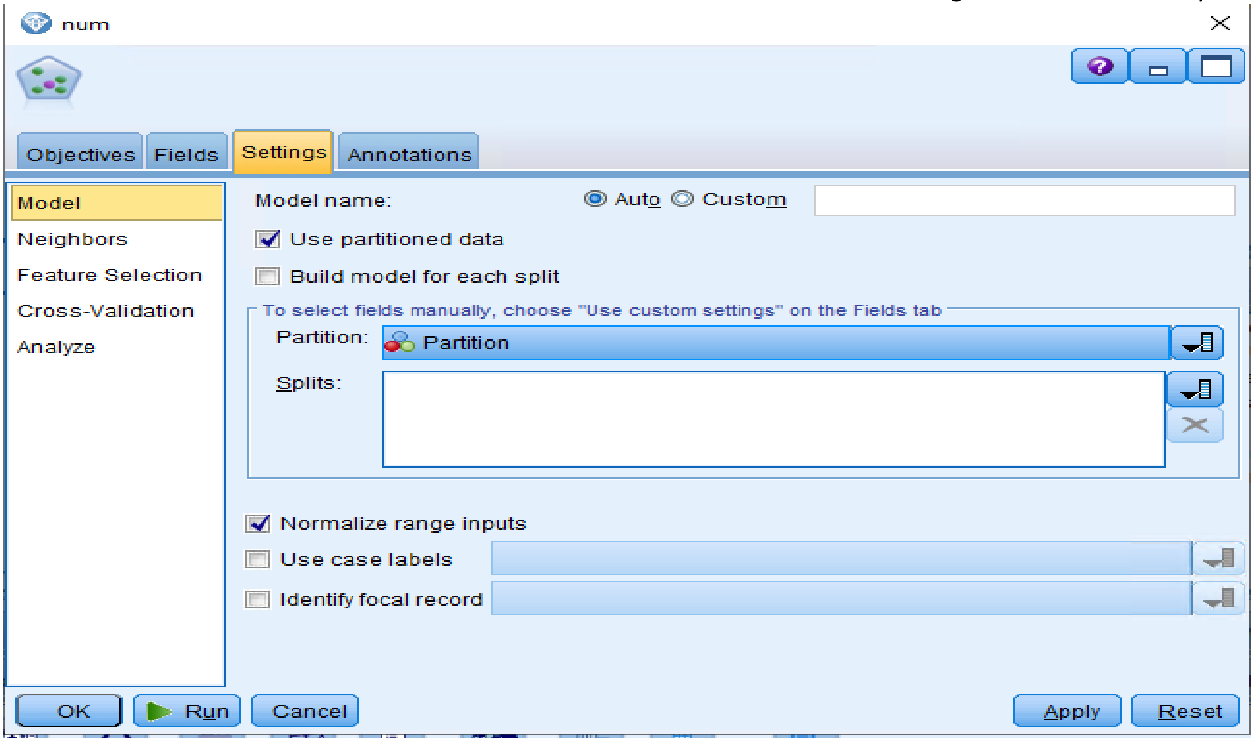
c) Perform a k-NN classification with all 13 predictors. Choose automatic k selection (range of k: 2 to 10). Make sure to check the box to normalize input data. Use Euclidean distance and weight features by importance when computing distances. Perform cross-validation and no feature selection.



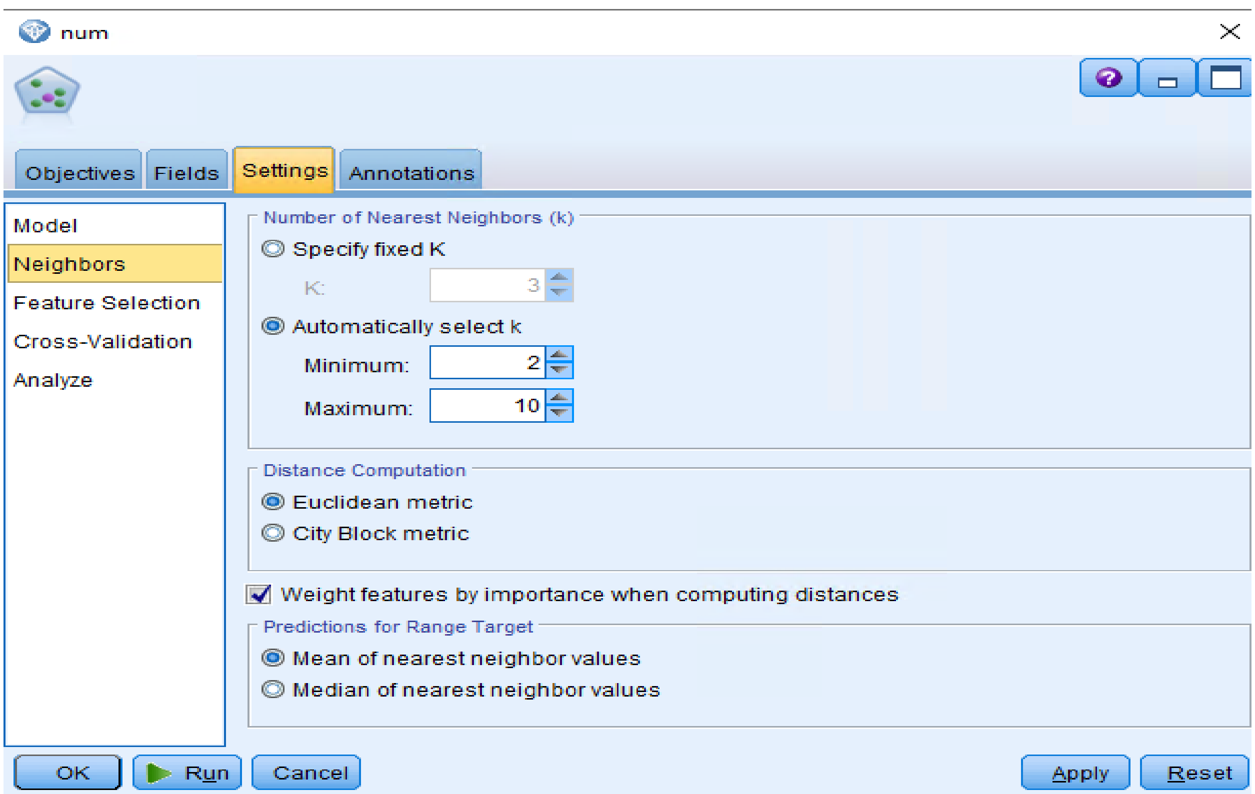
For this step of the question, we have used the KNN node under the model option in SPSS, and selected the custom analysis option, as shown above.



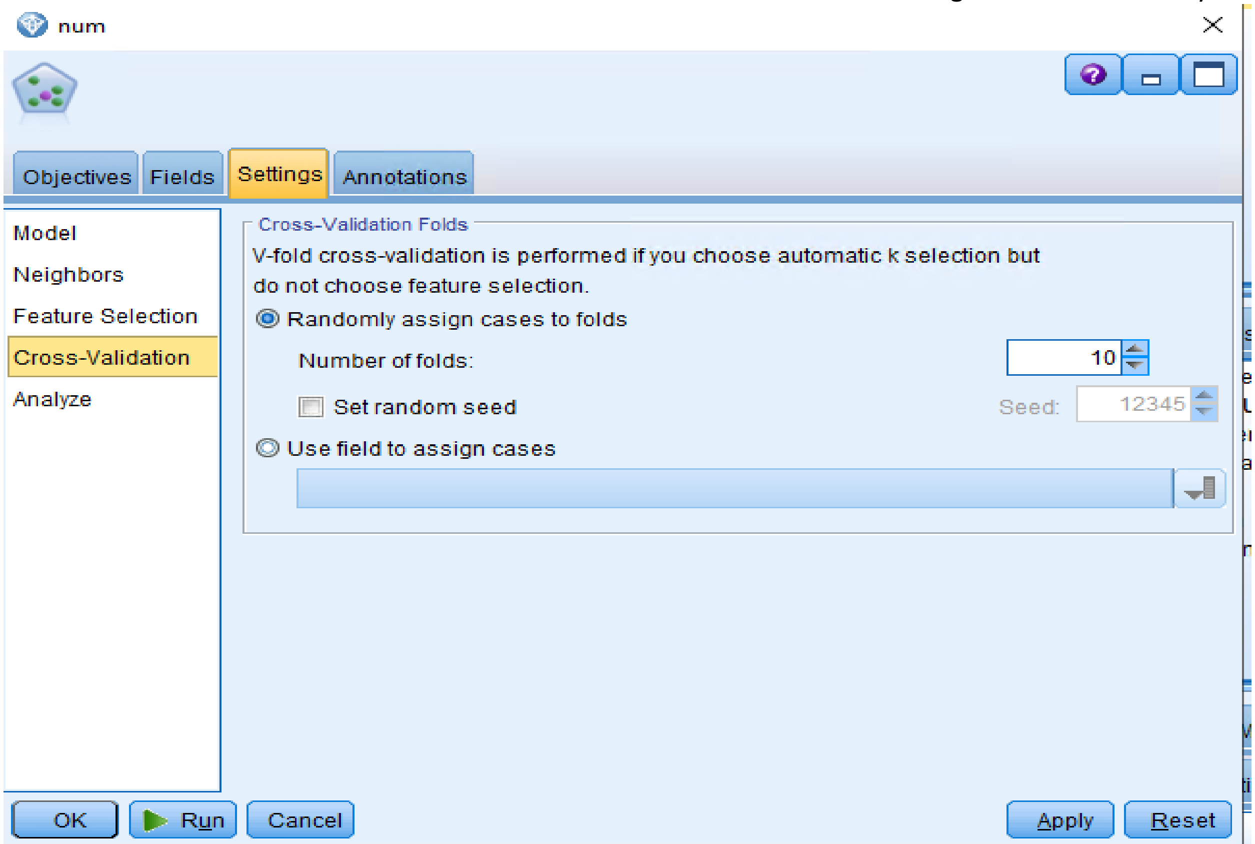
As required, we have used the 13 fields as the input and derived as our target field, this is very important.

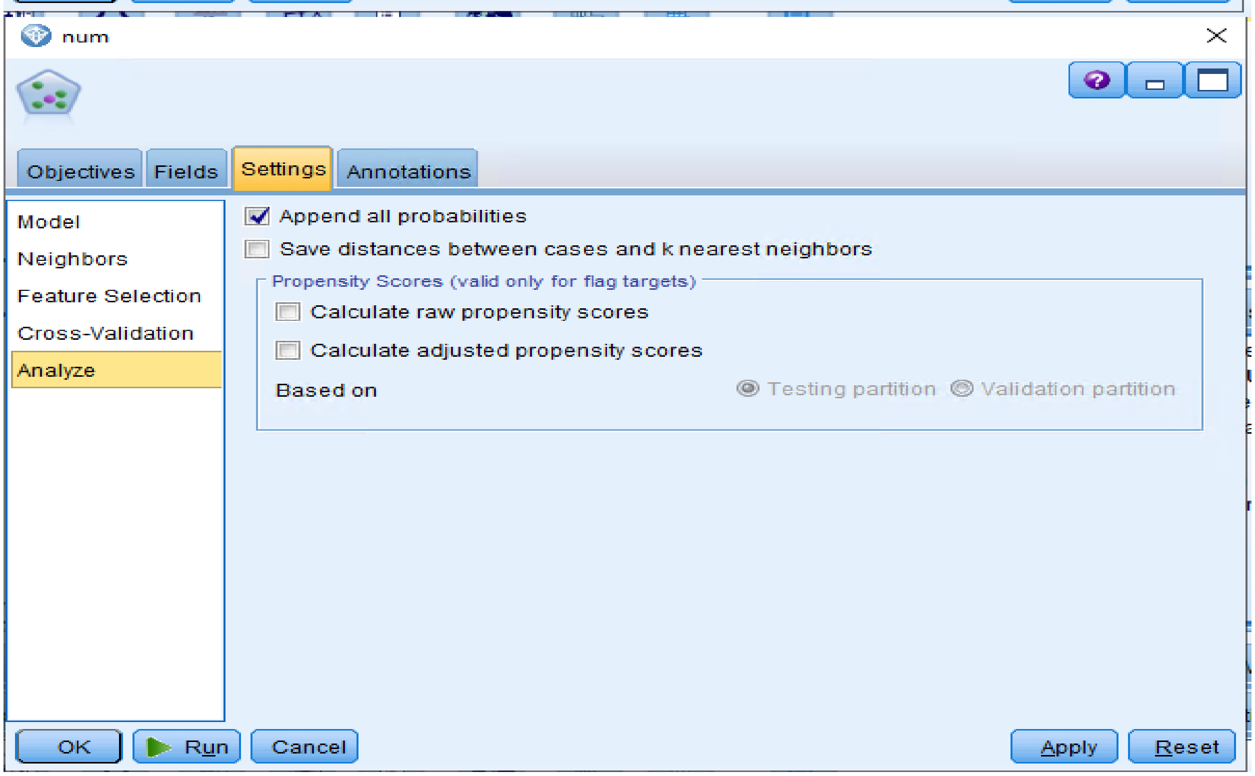


We have chosen the partition data, normalized range inputs under the model settings.



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.

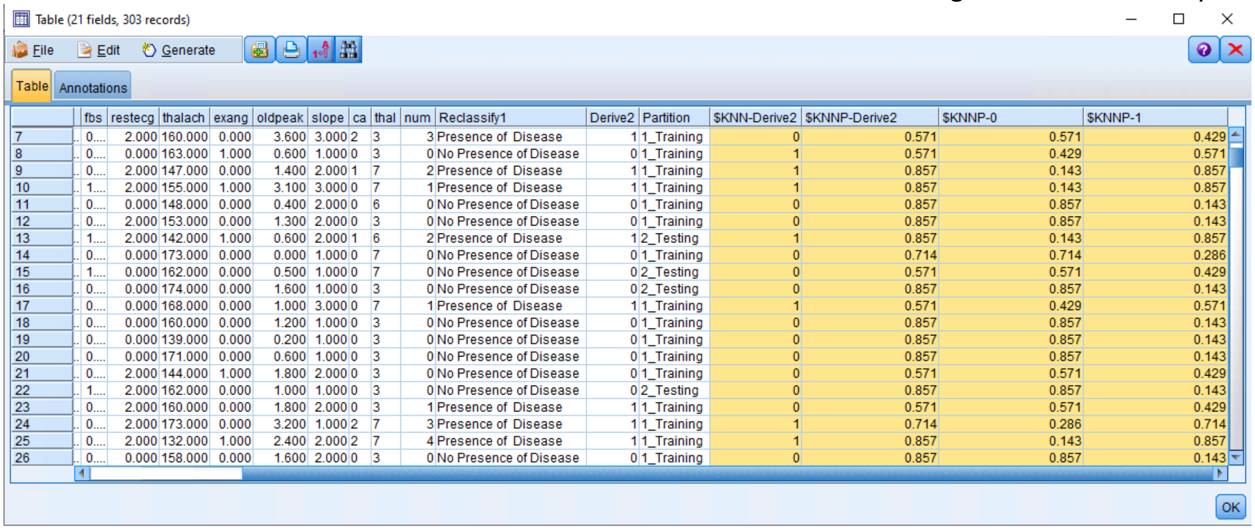
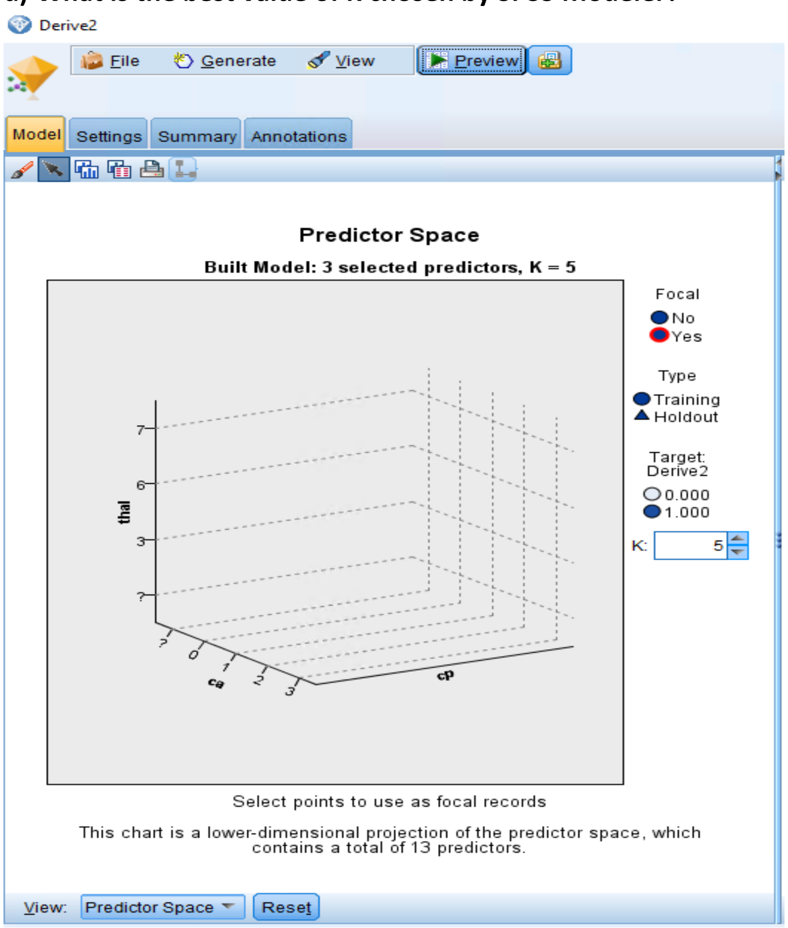


Table (21 fields, 303 records)

		fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num	Reclassify1	Derive2	Partition	$KNN-Derive2	$KNNP-Derive2	$KNNP-0	$KNNP-1
7	..	0....	2.000	160.000	0.000	3.600	3.000	2	3	3	Presence of Disease	1	1_Training	0	0.571	0.571	0.429
8	..	0....	0.000	163.000	1.000	0.600	1.000	0	3	0	No Presence of Disease	0	1_Training	1	0.571	0.429	0.571
9	..	0....	2.000	147.000	0.000	1.400	2.000	1	7	2	Presence of Disease	1	1_Training	1	0.857	0.143	0.857
10	..	1....	2.000	155.000	1.000	3.100	3.000	0	7	1	Presence of Disease	1	1_Training	1	0.857	0.143	0.857
11	..	0....	0.000	148.000	0.000	0.400	2.000	0	6	0	No Presence of Disease	0	1_Training	0	0.857	0.857	0.143
12	..	0....	2.000	153.000	0.000	1.300	2.000	0	3	0	No Presence of Disease	0	1_Training	0	0.857	0.857	0.143
13	..	1....	2.000	142.000	1.000	0.600	2.000	1	6	2	Presence of Disease	1	2_Testing	1	0.857	0.143	0.857
14	..	0....	0.000	173.000	0.000	0.000	1.000	0	7	0	No Presence of Disease	0	1_Training	0	0.714	0.714	0.286
15	..	1....	0.000	162.000	0.000	0.500	1.000	0	7	0	No Presence of Disease	0	2_Testing	0	0.571	0.571	0.429
16	..	0....	0.000	174.000	0.000	1.600	1.000	0	3	0	No Presence of Disease	0	2_Testing	0	0.857	0.857	0.143
17	..	0....	0.000	168.000	0.000	1.000	3.000	0	7	1	Presence of Disease	1	1_Training	1	0.571	0.429	0.571
18	..	0....	0.000	160.000	0.000	1.200	1.000	0	3	0	No Presence of Disease	0	1_Training	0	0.857	0.857	0.143
19	..	0....	0.000	139.000	0.000	0.200	1.000	0	3	0	No Presence of Disease	0	1_Training	0	0.857	0.857	0.143
20	..	0....	0.000	171.000	0.000	0.600	1.000	0	3	0	No Presence of Disease	0	1_Training	0	0.857	0.857	0.143
21	..	0....	2.000	144.000	1.000	1.800	2.000	0	3	0	No Presence of Disease	0	1_Training	0	0.571	0.571	0.429
22	..	1....	2.000	162.000	0.000	1.000	1.000	0	3	0	No Presence of Disease	0	2_Testing	0	0.857	0.857	0.143
23	..	0....	2.000	160.000	0.000	1.800	2.000	0	3	1	Presence of Disease	1	1_Training	0	0.571	0.571	0.429
24	..	0....	2.000	173.000	0.000	3.200	1.000	2	7	3	Presence of Disease	1	1_Training	1	0.714	0.286	0.714
25	..	0....	2.000	132.000	1.000	2.400	2.000	2	7	4	Presence of Disease	1	1_Training	1	0.857	0.143	0.857
26	..	0....	0.000	158.000	0.000	1.600	2.000	0	3	0	No Presence of Disease	0	1_Training	0	0.857	0.857	0.143

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.

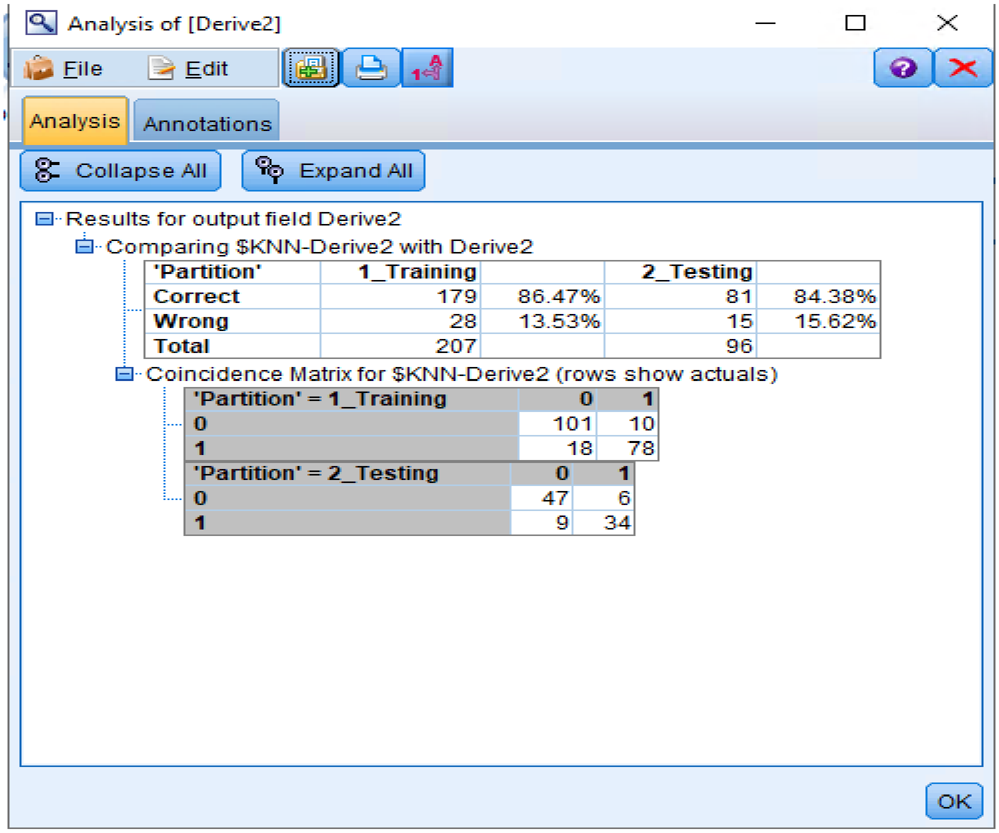
d) What is the best value of K chosen by SPSS Modeler?



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.