

[Martin Frigaard's Lab 5 comments](#)

Lab #5: Regression and Classification in IBM Modeler Biostat 202: Opportunities and Challenges of “Big Data”

Due: Sunday, 8/28 by 11.55pm

The purpose of this lab is to familiarize you to working with different clustering techniques, to reinforce regression and classification techniques including neural nets and Bayes nets in IBM SPSS Modeler.

Please respond to the questions below (marked in bold as **Q1** to **Q12** and hand in your lab assignment by posting it to the CLE.

Clustering: We will revisit the **Iris flower dataset** from lab 4. Just to remind you, the iris dataset gives the measurements in centimeters of the variables sepal length and width and petal length and width (the 4 predictors), respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*.

1. Download the Iris flower dataset from CLE (“iris.csv”) and read into IBM Modeler.

We already know a lot about this dataset.

"This famous (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are *Iris setosa*, *versicolor*, and *virginica*." from [Edgar Anderson's Iris Data](#)

pair and discard the double quotes:

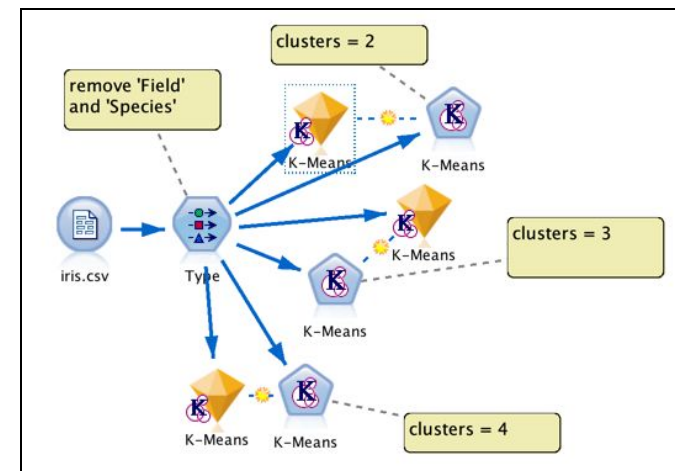
Table		Annotations				
	field1	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	1	5.100	3.500	1.400	0.200	setosa
2	2	4.900	3.000	1.400	0.200	setosa
3	3	4.700	3.200	1.300	0.200	setosa
4	4	4.600	3.100	1.500	0.200	setosa
5	5	5.000	3.600	1.400	0.200	setosa
6	6	5.400	3.900	1.700	0.400	setosa
7	7	4.600	3.400	1.400	0.300	setosa
8	8	5.000	3.400	1.500	0.200	setosa
9	9	4.400	2.900	1.400	0.200	setosa
10	10	4.900	3.100	1.500	0.100	setosa

Looks good!

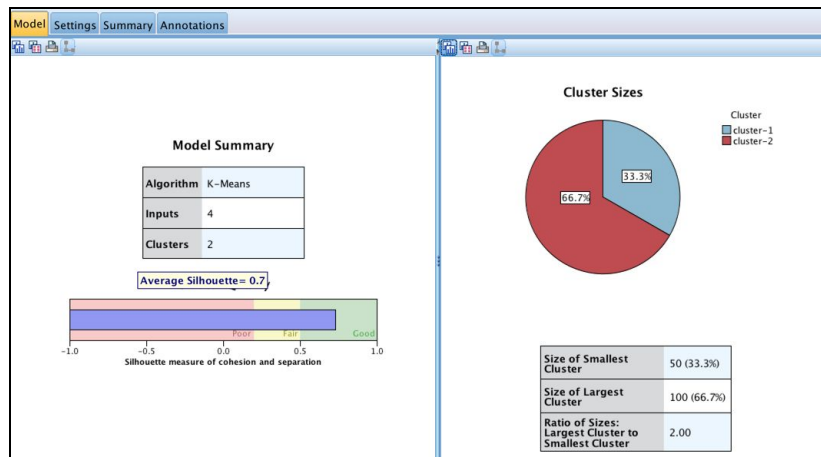
- Use the Type Node to appropriately set the data given the description above. Make sure Species and field1 are set to have the role “none”.

Types					
Format					
Annotations					
Read Values Clear Values Clear All Values					
Field	Measurement	Values	Missing	Check	Role
field1	Continuous	<Read>		None	None
Sepal.Length	Continuous	<Read>		None	Input
Sepal.Width	Continuous	<Read>		None	Input
Petal.Length	Continuous	<Read>		None	Input
Petal.Width	Continuous	<Read>		None	Input
Species	Categorical	<Read>		None	None

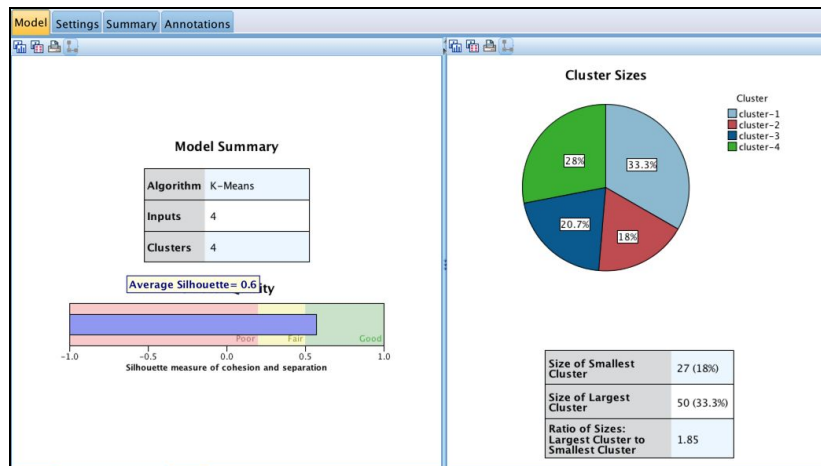
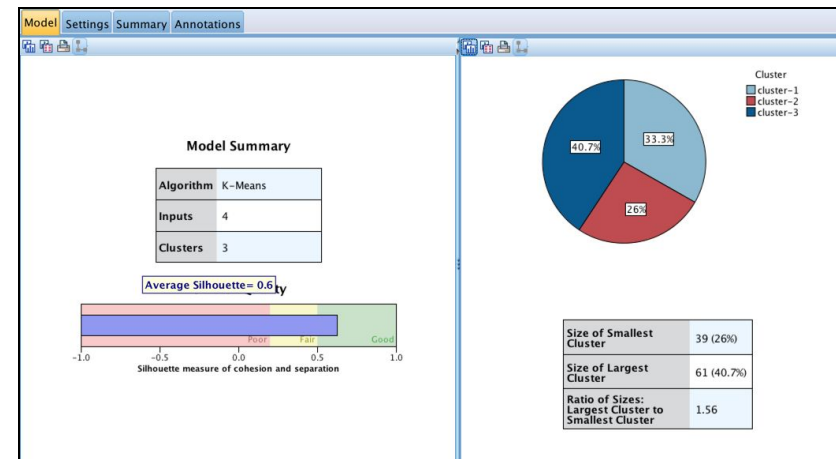
- Connect a K-Means Node  after the Type Node and use predefined roles. Run it for K = 2, 3, and 4 (see Model Tab). Look inside the golden nugget.



Q1 What value of K performed best in terms of the Silhouette measure?



To find the optimal number of clusters, a different rule of thumb exists. The SPSS Modeler provides the **Silhouette chart** as a measure of cohesion in the cluster, and a measure of separation between the clusters.



To measure the goodness of a classification, the silhouette value S can be calculated:

1. Calculation of average distance from the object to all objects in the nearest cluster.
2. Calculation of average distance from the object to all objects in the same cluster the object belongs to.
3. Calculation of the difference between the average distances (1)-(2).
4. This difference is then divided by the maximum of both those average distances. This standardizes the result.

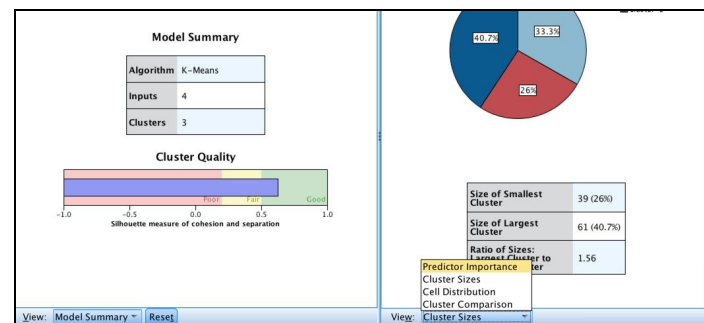
$$\frac{\text{avg. dist. to nearest cluster} - \text{avg. dist. to objects in same cluster}}{\text{maximum of both avg. above}}$$

More details can be found in Struyf et al. (1997), p. 5–7.

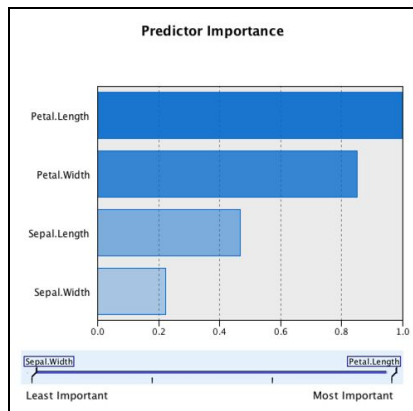
S can take on values between -1 and +1. If $S = -1$, the object is not well classified. If $S = 0$; the object lies between two clusters, and if $S = +1$ the object is well classified. The average of the Silhouette values of all objects represents a measure of goodness for the overall classification.

As outlined in IBM (2015b), p. 77 and IBM (2015b), p. 209, the IBM SPSS Modeler calculates a Silhouette Ranking Measure based on the Silhouette value. It is a measure of cohesion in the cluster and separation between the clusters. Additionally, it provides thresholds for poor (up to +0.25), fair (+0.5), and good (above +0.5) models.

4. Rerun the K-means clustering for $K = 3$ (we know 3 is the number of species). View the Predictor Importance in the golden nugget.



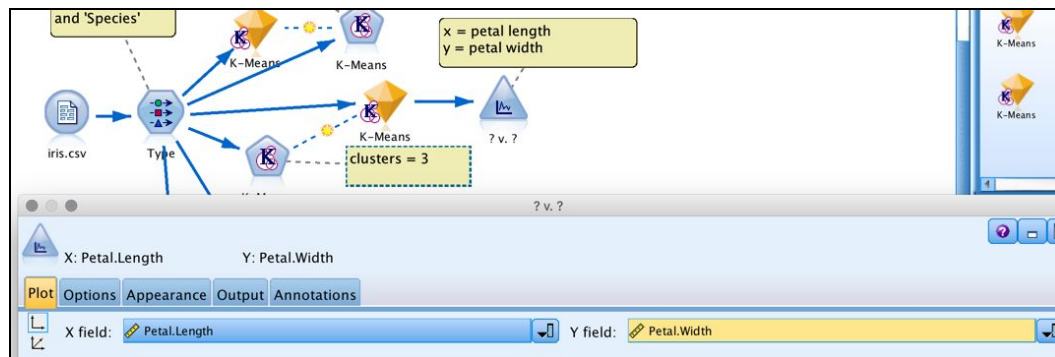
Q2. Which are the most important 2 variables for clustering?



It looks like the petal length and petal width are the most important variables for clustering

5. Add a plot node to the golden nugget.

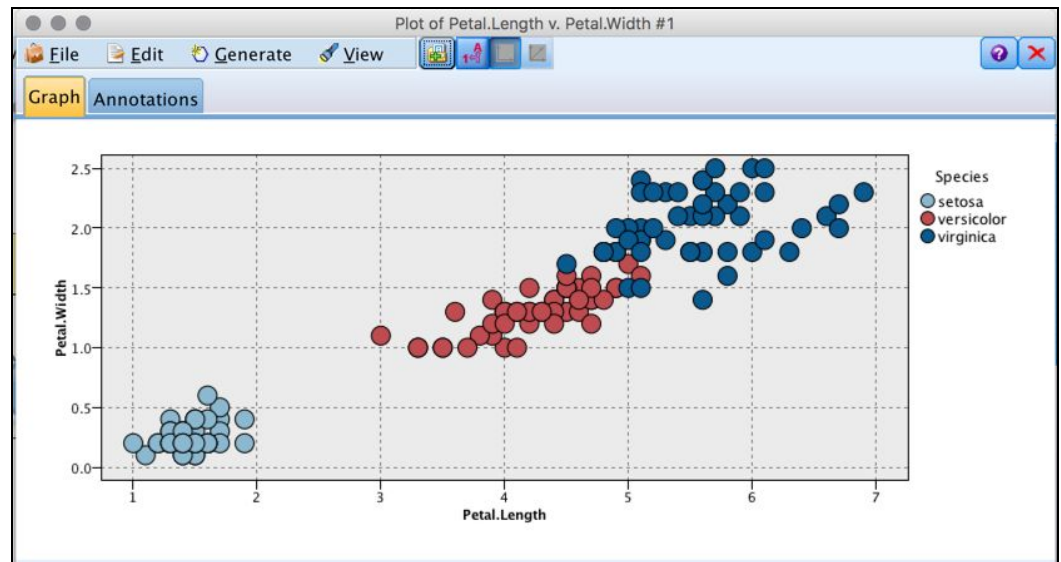
Put the most important clustering variable in the X field and 2nd most important in the Y field.



Determine whether the clustering algorithm did well at recreating the original Species classes.

Where did it tend to break down?

As we can see from the graph, the petal length and width did a great job differentiating setosa from versicolor and virginica. It didn't do as well separating wider versicolors from less-wide virginica, and shorter virginica from longer versicolors.



Q3. Which of the Species was most easily put into a separate cluster by K-means?

setosa



6. Add an Auto Cluster Node after the Type Node.

Again you just need to use predefined roles (assuming they are correctly set up in the type node).

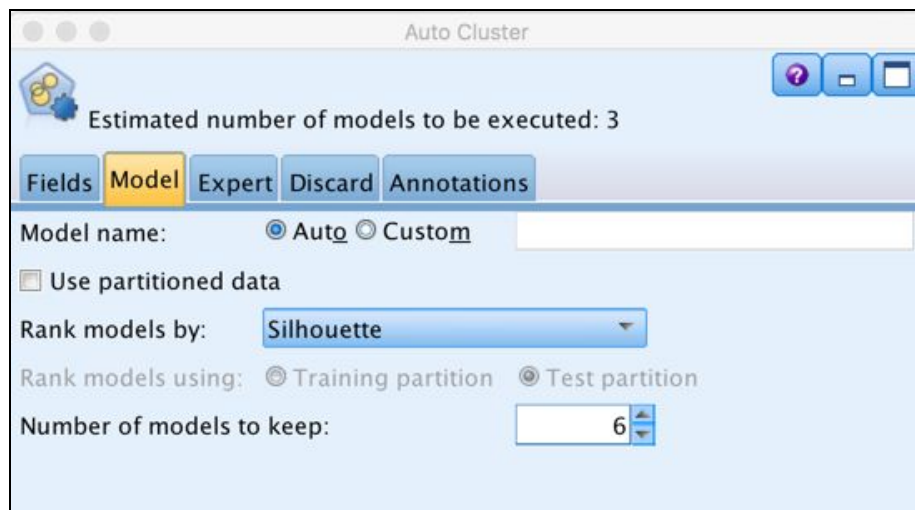
Under the Model Tab leave as defaults.

Uncheck the “Use partition data”.

It actually doesn't seem to play any role which makes sense since you haven't created any partition, but better to be safe (the dataset is really too small to seriously do any partitioning).

Leave “Rank models by” on the default of “Silhouette”.

Set the number of models to keep as 6.

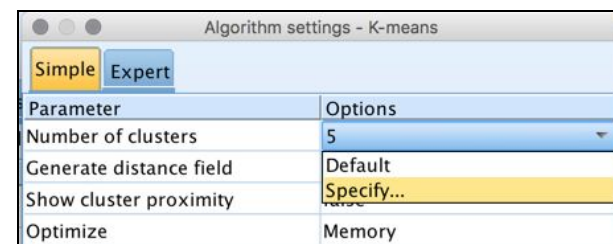


7. Under the Expert Tab make sure all 3 Model types are checked.

Go into the Model parameters for K-means changing “Default” to “Specify”.

Fields	Model	Expert	Discard	Annotations
Models used:				
Use?	Model type	Model parameters	No of models	
☒	Kohonen	Default	1	
☒	K-means	Default	1	
☒	TwoStep	Default Specify...	1	

Under the Simple Tab click on the number 5 under options and hit specify.

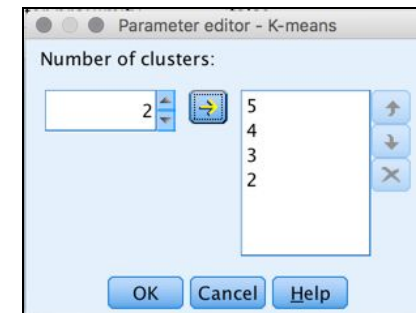
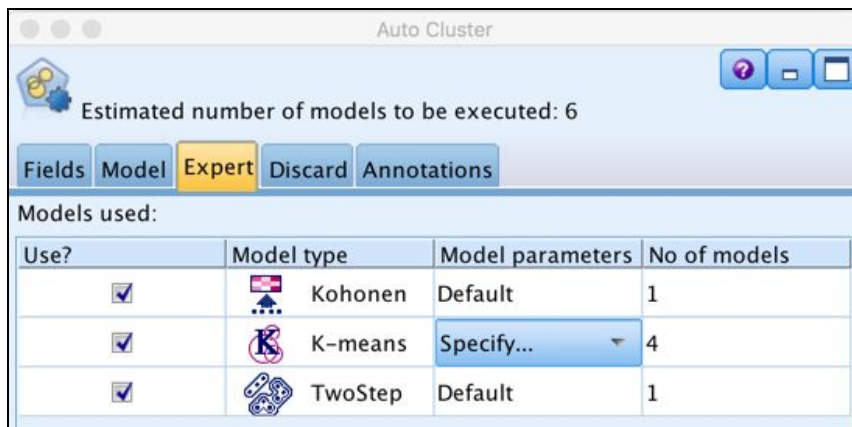


Add 4, 3, and 2 as number of clusters on the right hand side.

This will make the auto-classifier run K-means models for all of K = 2, 3, 4, and 5.

Hit OK and OK again to go back to the main screen.

Leave the Kohonen and TwoStep clustering approaches on the defaults; they are both methods that optimize over cluster size.



Hit Run.

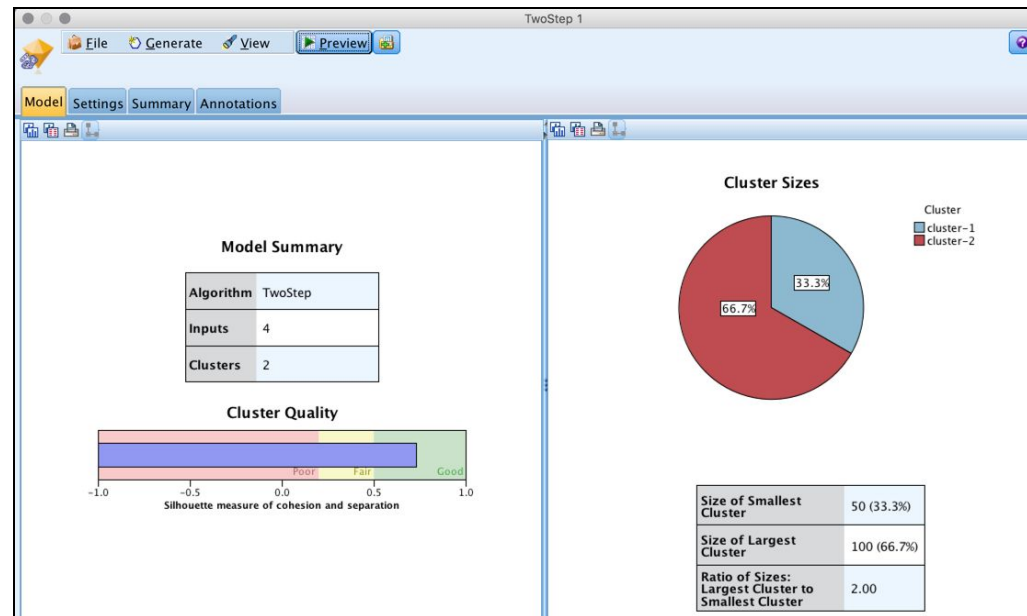
8. Look in the Model Tab of the golden nugget for the Auto Cluster Node.

Q4. Which methods performed best with respect to Silhouette measure?

The K-means with 2 clusters.

Model Summary Annotations										
Sort by: Use Ascending Descending Delete Unused Models View: Training set										
Use?	Graph	Model	Build Time (mins)	Silhouette	Number of Clusters	Smallest Cluster (N)	Smallest Cluster (%)	Largest Cluster (N)	Largest Cluster (%)	Smallest/Largest
☒		K-m...	< 1	0.733	2	50	33	100	66	0.5
☐		Two...	< 1	0.733	2	50	33	100	66	0.5
☐		K-m...	< 1	0.626	3	39	26	61	40	0.639
☐		K-m...	< 1	0.574	4	27	18	50	33	0.54
☐		K-m...	< 1	0.504	5	21	14	42	28	0.5
☐		Koh...	< 1	0.401	10	1	0	50	33	0.02

9. In the model column of the data click on the TwoStep golden nugget.



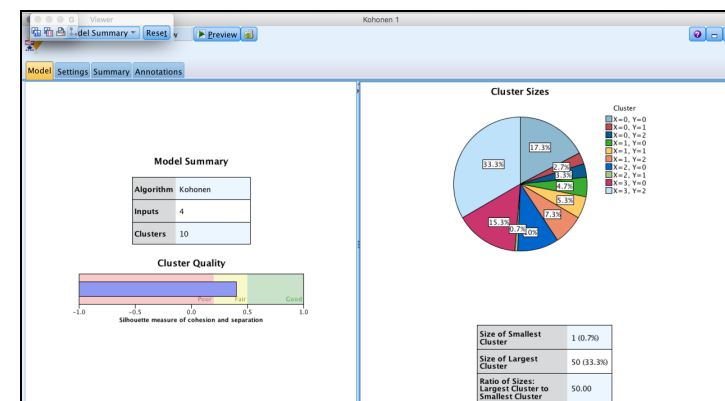
Q5. How many clusters did the TwoStep clustering procedure determine to be optimal?

Two clusters.

10. In the model column of the data click on the Kohonen golden nugget.

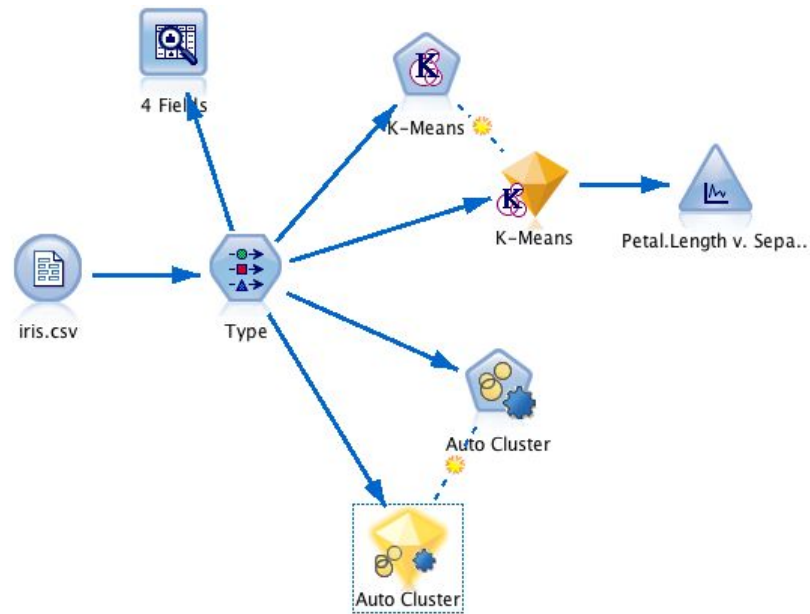
Q6. How many clusters did the Kohonen clustering procedure determine to be optimal?

10. Wayyyyy too many.



It seems that this procedure has overfitted the number of clusters. Perhaps not surprising since it is based on a complex neural network-related model and the dataset is quite small.

11. At this point your stream should look something like the following:



Clustering, Regression and Classification: We will now turn to the Los Angeles heart dataset (laheart.xlsx) available on the CLE. The dataset includes the following variables:

Var	Description	Units	Scale	Comments
ID	ID	none	nominal	Patients numbered sequentially
AGE_50	Age in 1950	yr	ratio	Age at last birthday (in 1950)
MD_50	Examining M.D. (1950)	none	nominal	Coded from 1-4
SBP_50	Systolic BP (1950)	mm Hg	ratio	Recorded to nearest integer
DBP_50	Diastolic BP (1950)	mm Hg	ratio	Recorded to nearest integer
HT_50	Height in 1950	in	ratio	Recorded to nearest inch
WT_50	Weight in 1950	lb	ratio	Recorded to nearest pound
CHOL_50	Serum cholesterol(1950)	mg%	ratio	Recorded to nearest integer
SES	Socio-economic status	none	Ordinal	1=high,...,5=low
CL_STATUS	Clinical status	none	Nominal	0=other heart disease(hd) 1=Coronary hd 2=Coronary and hypertensive hd 3=Hypertensive hd 4=Hypertensive and rheumatic hd 5=Rheumatic hd 6=Possible/potential hd 7=Hypertension wout/hd 8=Normal
MD_62	Examining MD (1962)	none	nominal	Coded from 1-4
SBP_62	Systolic BP (1962)	mm Hg	ratio	Recorded to nearest integer
DBP_62	Diastolic BP (1962)	mm Hg	ratio	Recorded to nearest integer
CHOL_62	Serum cholesterol(1962)	mg%	ratio	Recorded to nearest integer
WT_62	Weight (1962)	lb	ratio	Recorded to nearest pound
IHD_DX	Ischemic hd diagnosis	none	nominal	0=Not known 1-3=Myocardial infarction 4-7=Angina pectoris 8-9=Other
DEATH_YR	Year of death	none	interval	0=Not Dead as of 1968 Otherwise year of death
DEATH	Death	none	nominal	0=Alive 1=Dead

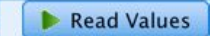
12. Download the LA Heart dataset from CLE (“laheart.csv”) and read into IBM Modeler.

Table	Annotations																	
	ID	AGE_50	MD_50	SBP_50	DBP_50	HT_50	WT_50	CHOL_50	SES	CL_STATUS	MD_62	SBP_62	DBP_62	CHOL_62	WT_62	IHD_DX	DEATH_YR	DEATH
1	1.000	42.000	1.000	110.000	65.000	64.000	147.000	291.000	2.000	8.000	4.000	120.000	78.000	271.000	146.000	2.000	68.000	1.000
2	2.000	53.000	1.000	130.000	72.000	69.000	167.000	278.000	1.000	6.000	2.000	122.000	68.000	250.000	165.000	9.000	67.000	1.000
3	3.000	53.000	2.000	120.000	90.000	70.000	222.000	342.000	4.000	8.000	1.000	132.000	90.000	304.000	223.000	2.000	64.000	1.000
4	4.000	48.000	4.000	120.000	80.000	72.000	229.000	239.000	4.000	8.000	2.000	118.000	68.000	209.000	227.000	3.000	66.000	1.000
5	5.000	53.000	3.000	118.000	74.000	66.000	134.000	243.000	3.000	8.000	5.000	118.000	56.000	261.000	138.000	2.000	66.000	1.000
6	6.000	58.000	2.000	122.000	72.000	69.000	135.000	210.000	3.000	8.000	4.000	130.000	72.000	245.000	136.000	2.000	64.000	1.000
7	7.000	48.000	4.000	130.000	90.000	67.000	165.000	219.000	3.000	8.000	4.000	138.000	86.000	275.000	166.000	2.000	63.000	1.000
8	8.000	60.000	1.000	124.000	80.000	74.000	235.000	203.000	3.000	8.000	1.000	160.000	90.000	271.000	226.000	3.000	65.000	1.000
9	9.000	59.000	4.000	160.000	100.0...	72.000	206.000	269.000	5.000	8.000	3.000	150.000	100.0...	291.000	198.000	3.000	67.000	1.000
10	10.000	40.000	3.000	120.000	80.000	69.000	148.000	185.000	3.000	8.000	3.000	110.000	64.000	241.000	152.000	2.000	66.000	1.000

^this is obviously an .xlsx file (see the 3 trailing zeros?), not a .csv

13. Use the Type Node to appropriately set the data given the description above.

- Set ID to the Role of Record ID.
- DEATH_YR and DEATH should be set to have the role “None”.

Types		Format		Annotations					
									
Field	Measurement	Values	Missing	Check	Role				
 ID	 Continuous	<Read>		None	 Record ID				
 AGE 50	 Continuous	<Read>		None	 Input				
 MD 50	 Continuous	<Read>		None	 Input				
 SBP 50	 Continuous	<Read>		None	 Input				
 DBP 50	 Continuous	<Read>		None	 Input				
 HT 50	 Continuous	<Read>		None	 Input				
 WT 50	 Continuous	<Read>		None	 Input				
 CHOL 50	 Continuous	<Read>		None	 Input				
 SES	 Continuous	<Read>		None	 Input				
 CL STATUS	 Continuous	<Read>		None	 Input				
 MD 62	 Continuous	<Read>		None	 Input				
 SBP 62	 Continuous	<Read>		None	 Input				
 DBP 62	 Continuous	<Read>		None	 Input				
 CHOL 62	 Continuous	<Read>		None	 Input				
 WT 62	 Continuous	<Read>		None	 Input				
 IHD DX	 Continuous	<Read>		None	 Input				
 DEATH YR	 Continuous	<Read>		None	 None				
 DEATH	 Continuous	<Read>		None	 None				

14. Use the Reclassify Node to create a new variable based on the IHD_DX with category names “Not known”, “Myocardial infarction”, “Angina pectoris”, “Other” as prescribed in the variable description.

IHD_DX	Ischemic hd diagnosis	none	nominal	0=Not known
		1-3=Myocardial infarction		
		4-7=Angina pectoris		
		8-9=Other		

Tried this but got an error (see below) > Be sure to make the variable nominal!

Reclassify1

Preview

Settings Annotations

Mode: ☒ Single ☐ Multiple

Reclassify into: ☒ New field ☐ Existing field

Reclassify field:

New field name:

Reclassify values:

☒ Get ☐ Copy ☐ Clear new ☐ Auto...

No values were found - reclassify field(s) may not be selected or values may not be available at type node.

OK

For unrecategorized values use: ☒ Original value ☐ Default value

OK Cancel Apply Reset

[illegible]

15. Connect another Type Node from the Reclassify Node. Set the Role of IHD_DX to None.

The screenshot shows the Orange3 software interface. On the left, a workflow diagram includes nodes for 'laheart.csv', 'laheart.xlsx', 'Type', 'IHD_DX_5_cat', 'Reclassify3', and '17 Fields'. The 'Type' node is connected to 'IHD_DX_5_cat', which is then connected to 'Reclassify3'. The '17 Fields' node is also connected to 'Reclassify3'. On the right, the 'Type' node's 'Preview' window is open, displaying a table of data fields and their roles.

Field	Measurement	Values	Missing	Check	Role
ID	Continuous	[1.0,200...	None		Record ...
AGE_50	Continuous	[20.0,69...	None		Input
MD_50	Nominal	1.0,2.0...	None		Input
SBP_50	Continuous	[88.0,21...	None		Input
DBP_50	Continuous	[38.0,14...	None		Input
HT_50	Continuous	[61.0,75...	None		Input
WT_50	Continuous	[109.0,2...	None		Input
CHOL_50	Continuous	[120.0,5...	None		Input
SES	Ordinal	1.0,2.0...	None		Input
CL STATUS	Nominal	0.0,3.0...	None		Input
MD_62	Nominal	1.0,2.0...	None		Input
SBP_62	Continuous	[70.0,23...	None		Input
DBP_62	Continuous	[56.0,12...	None		Input
CHOL_62	Continuous	[139.0,3...	None		Input
WT_62	Continuous	[108.0,2...	None		Input
IHD_DX	Nominal	0.0,1.0...	None		None
DEATH_YR	Ordinal	0.0,63.0...	None		None
DEATH	Flag	1.0/0.0	None		None
IHD_DX_5 ...	Nominal	*Not kno...	None		Input

I created a .csv file just in case the formatting gets in the way.

16. Connect a Partition Node to the Type Node.

Open the Partition Node.

Hit the Train and test Partition radio button.

Set the Training partition size to 60 and the Testing partition size to 40.

Hit Apply and OK.

The screenshot shows the 'Partition' node settings window in Orange3. The 'Settings' tab is active. The 'Partition field' is set to 'Partition'. The 'Partitions' section has the 'Train and test' radio button selected. The 'Training partition size' is set to 60, and the 'Testing partition size' is set to 40. The 'Validation partition size' is set to 0. The 'Label' and 'Value' fields are populated with 'Training', 'Testing', and 'Validation' respectively. The 'Total size' is 100%. The 'Values' section has 'Append labels to system-defined values' selected. The 'Repeatable partition assignment' checkbox is checked. The 'Seed' is set to 1234567. The 'Use unique field to assign partitions' checkbox is unchecked. The 'Apply' and 'OK' buttons are visible at the bottom.

17. Connect an Auto Cluster Node after the Type Node.

Again you just need to use predefined roles (assuming they are correctly set up in the type node).

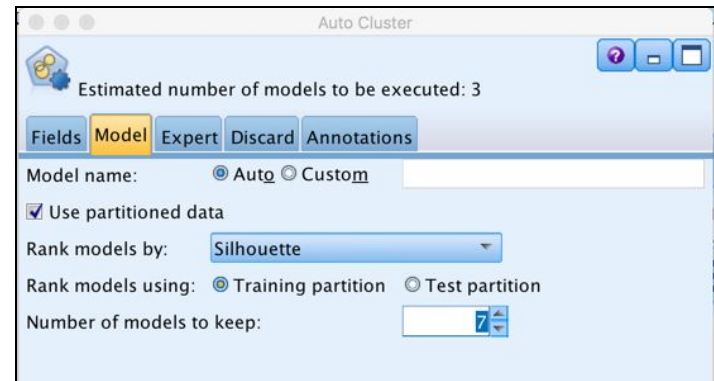
Under the Model Tab leave as defaults.

Check the “Use partition data”.

Leave “Rank models by” on the default of “Silhouette”.

By “Rank models using” hit the “Training partition” radio button.

Set the number of models to 7.



18. Under the Expert Tab make sure all 3 Model types are checked.

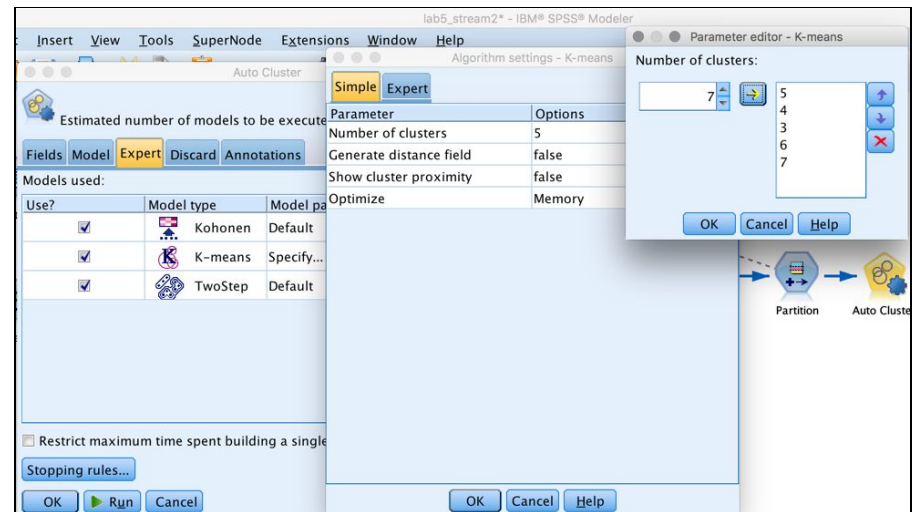
Go into the Model parameters for K-means changing “Default” to “Specify”.

Under the Simple Tab click on the number 5 under options and hit specify.

Add 3, 4, 6, and 7 as number of clusters on the right hand side.

Hit OK and OK again to go back to the main screen.

Leave the Kohonen and TwoStep clustering approaches on the defaults. Hit Run.



19. Look in the Model Tab of the golden nugget for the Auto Cluster Node.

Q7. Which method performed best with respect to Silhouette measure and note the associated value of the measure?

Model Summary Annotations											
Sort by: Use Ascending Descending Delete Unused Models View: Testing set											
Use?	Graph	Model	Build Time (mins)	Silhouette	Number of Clusters	Smallest Cluster (N)	Smallest Cluster (%)	Largest Cluster (N)	Largest Cluster (%)	Smallest/Largest	Importance
☒		TwoStep 1	< 1	0.126	2	29	33	58	66	0.5	0.0
☐		Kohonen 1	< 1	0.101	12	1	1	18	20	0.056	0.0
☐		K-means 4	< 1	0.07	6	4	4	42	48	0.095	0.0
☐		K-means 5	< 1	0.066	7	4	4	40	45	0.1	0.0
☐		K-means 3	< 1	0.065	3	20	22	39	44	0.513	0.0
☐		K-means 2	< 1	0.047	4	12	13	37	42	0.324	0.0
☐		K-means 1	< 1	0.046	5	5	5	32	36	0.156	0.0

The TwoStep model has the highest Silhouette value (0.126)

20. At the top right of the table go to the View button and change the value to Test set. - I am assuming you mean training set?

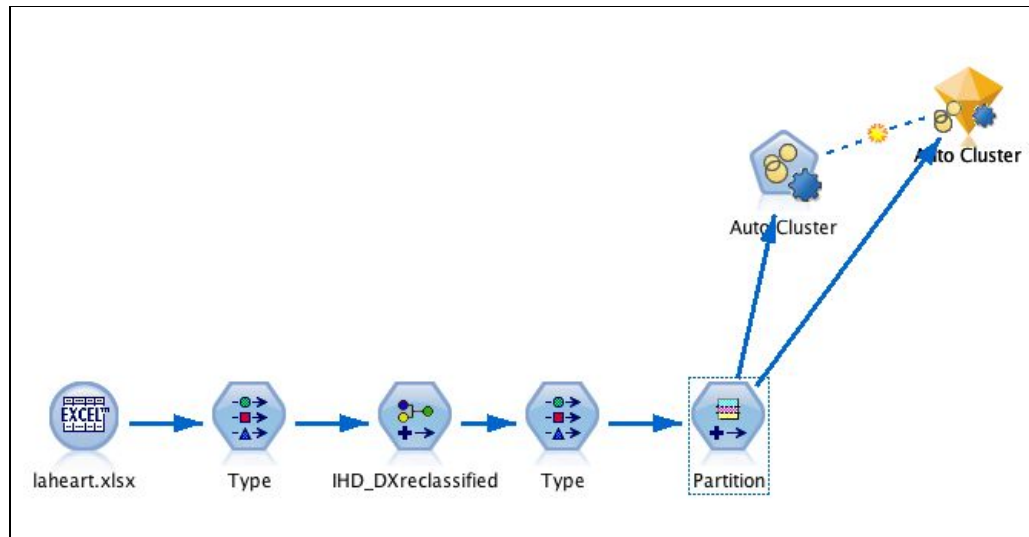
Q8. Is the ranking of the methods still the same? TwoStep is still at the top, but the Silhouette value has changed (0.153)

Model Summary Annotations											
Sort by: Use Ascending Descending Delete Unused Models View: Training set											
Use?	Graph	Model	Build Time (mins)	Silhouette	Number of Clusters	Smallest Cluster (N)	Smallest Cluster (%)	Largest Cluster (N)	Largest Cluster (%)	Smallest/Largest	Importance
☒		TwoStep 1	< 1	0.153	2	43	38	70	61	0.614	0.0
☐		K-means 2	< 1	0.082	4	16	14	40	35	0.4	0.0
☐		K-means 3	< 1	0.075	3	30	26	44	38	0.682	0.0
☐		Kohonen 1	< 1	0.069	12	1	0	24	21	0.042	0.0
☐		K-means 5	< 1	0.067	7	7	6	43	38	0.163	0.0
☐		K-means 1	< 1	0.064	5	13	11	30	26	0.433	0.0
☐		K-means 4	< 1	0.062	6	9	7	45	39	0.2	0.0

Does the best Test set model have higher or lower Silhouette score than the Training set?

My testing set was slightly lower than my training set (0.126 vs. 0.153).

21. At this point your stream should look something like the following:



22. Go back into the 2nd Type Node.

Set CHOL_62 to be the Target Node and set all the other variables with _62 in the name to the Role of “None”.

Types

Preview

Types Format Annotations

Read Values Clear Values Clear All Values

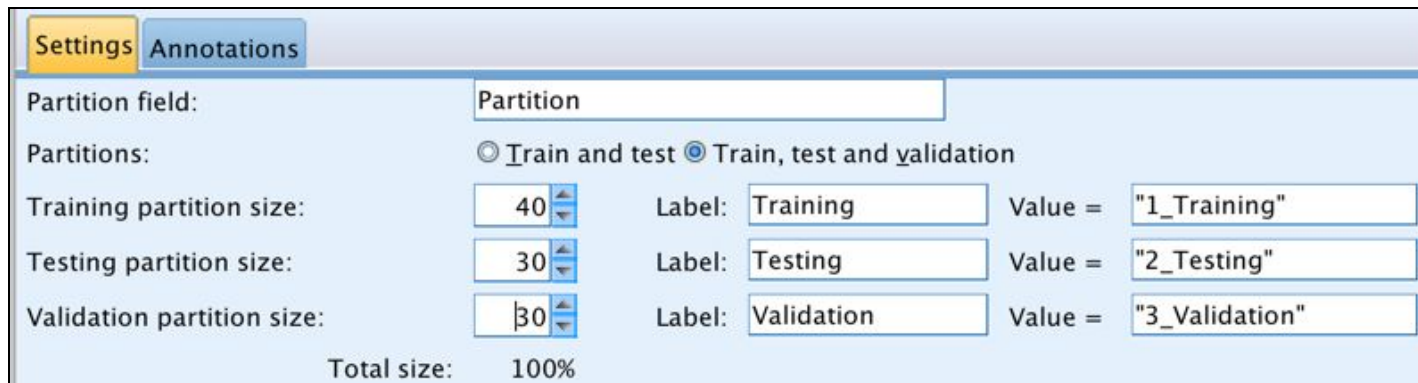
Field	Measurement	Values	Missing	Check	Role
ID	Continuous	{1.0,200...		None	Record ...
AGE_50	Continuous	{20.0,69...		None	Input
MD_50	Nominal	{1.0,2.0...		None	Input
SBP_50	Continuous	{88.0,21...		None	Input
DBP_50	Continuous	{38.0,14...		None	Input
HT_50	Continuous	{61.0,75...		None	Input
WT_50	Continuous	{109.0,2...		None	Input
CHOL_50	Continuous	{120.0,5...		None	Input
SES	Ordinal	{1.0,2.0...		None	Input
CL STATUS	Nominal	{0.0,3.0...		None	Input
MD_62	Nominal	{1.0,2.0...		None	None
SBP_62	Continuous	{70.0,23...		None	None
DBP_62	Continuous	{56.0,12...		None	None
CHOL_62	Continuous	{139.0,3...		None	Target
WT_62	Continuous	{108.0,2...		None	None
IHD_DX	Nominal	{0.0,1.0...		None	None
DEATH_YR	Ordinal	{0.0,63.0...		None	None
DEATH	Flag	{1.0/0.0}		None	None
IHD_DX_5 ...	Nominal	"Not kno...		None	Input

View current fields View unused field settings

We want to restrict the prediction to just the variables known at Age 50 so that we can attempt to predict how they will do further in the future at Age 62, and thereby see if we can flag patients early that will benefit from preventative measures.

23. Go into the Partition Node on the Settings Tab and set Partitions to “Train, test and validation”.

24. Set the split to 40 Training, 30 Testing and 30 Validation.



The screenshot shows the 'Settings' tab of a Partition Node. The 'Partition field' is set to 'Partition'. Under 'Partitions', the 'Train, test and validation' radio button is selected. The 'Training partition size' is 40, 'Testing partition size' is 30, and 'Validation partition size' is 30. Each size has a corresponding 'Label' and 'Value' field. The 'Total size' is 100%.

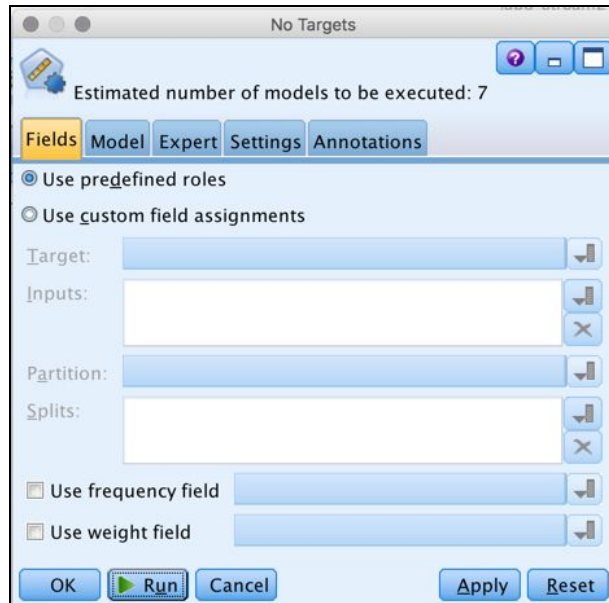
Partition	Label	Value
40	Training	"1_Training"
30	Testing	"2_Testing"
30	Validation	"3_Validation"

Total size: 100%

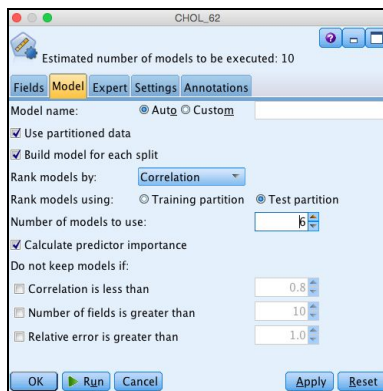


25. Add an Auto Numeric Node

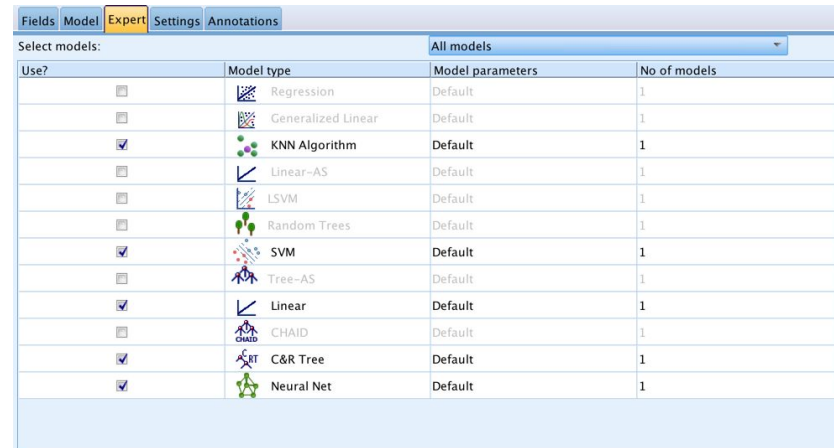
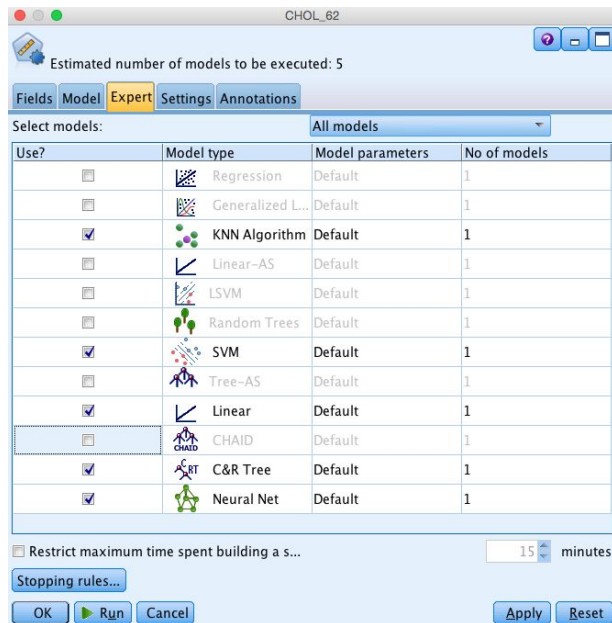
26. Use the predefined roles under the Fields Tab



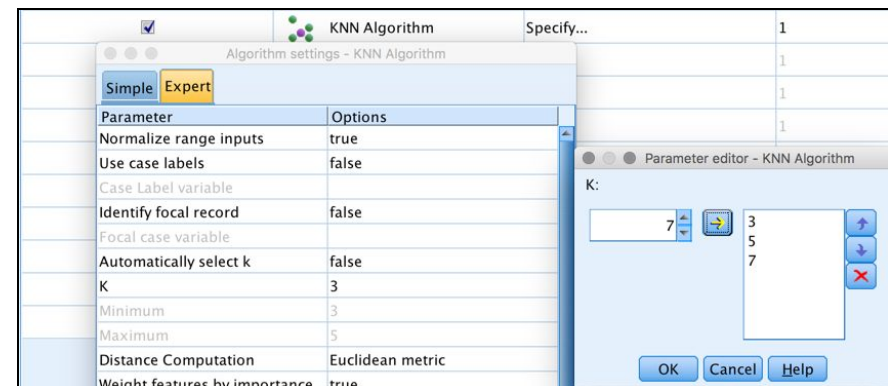
27. Under the Model Tab use the following settings (you need to change Number of models to use to 6).



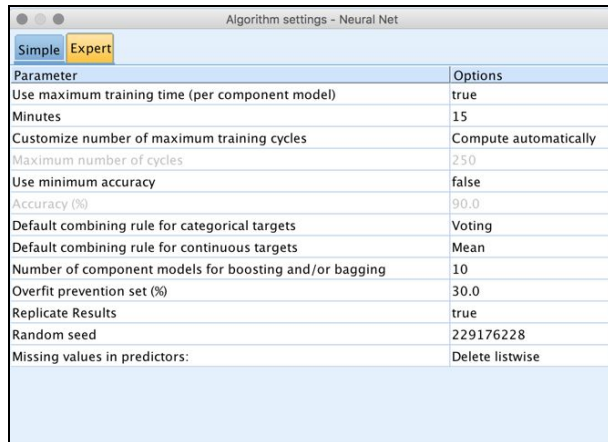
28. In the Expert Tab check the models as below



29. Click the cell under Model parameters for the KNN Algorithm to Specify and click to the Expert Tab.
Click the cell under Options for K and hit Specify.
Set so that K=3, 5, and 7 are considered.
Hit OK.



30. Open up the Model Parameters for Neural Net.
Go to the Expert Tab.

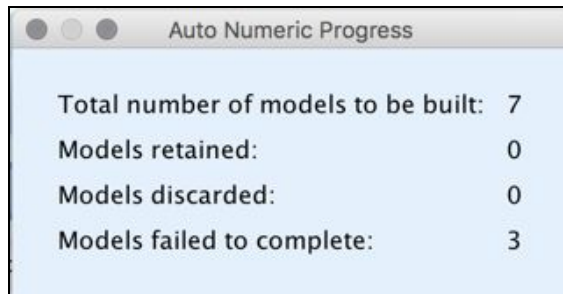


The screenshot shows a window titled "Algorithm settings - Neural Net". It has two tabs: "Simple" and "Expert", with "Expert" being the active tab. Below the tabs is a table with two columns: "Parameter" and "Options".

Parameter	Options
Use maximum training time (per component model)	true
Minutes	15
Customize number of maximum training cycles	Compute automatically
Maximum number of cycles	250
Use minimum accuracy	false
Accuracy (%)	90.0
Default combining rule for categorical targets	Voting
Default combining rule for continuous targets	Mean
Number of component models for boosting and/or bagging	10
Overfit prevention set (%)	30.0
Replicate Results	true
Random seed	229176228
Missing values in predictors:	Delete listwise

Q9. What is the default Number of component models? 10

31. Back out and run the Auto Numeric Node.



The screenshot shows a window titled "Auto Numeric Progress". It contains a list of statistics with their corresponding values.

Total number of models to be built:	7
Models retained:	0
Models discarded:	0
Models failed to complete:	3

The three KNN models failed to complete.

Q10. Which regression method has the highest Correlation value and what is the associated number of fields used?

Model Graph Summary Settings Annotations						
Sort by: Use		Ascending Descending		Delete Unused Models		View: Testing set
Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
☒		Linear 1	< 1	0.363	4	1.012
☒		Neural Net 1	< 1	0.255	10	1.074
☒		SVM 1	< 1	0.156	10	1.056
☒		C&R Tree 1	< 1	0.0	10	1.147

The Linear model did (0.401) and it used 7 fields. I had selected KNN, SVM, Linear, C&R Tree, and Neural Net, but I am only getting Linear, Neural Net, C&R Tree, and SVM back?

32. Go back into the 2nd Type Node. Set DEATH as the Target variable and set the Role of CHOL_62 to "None".

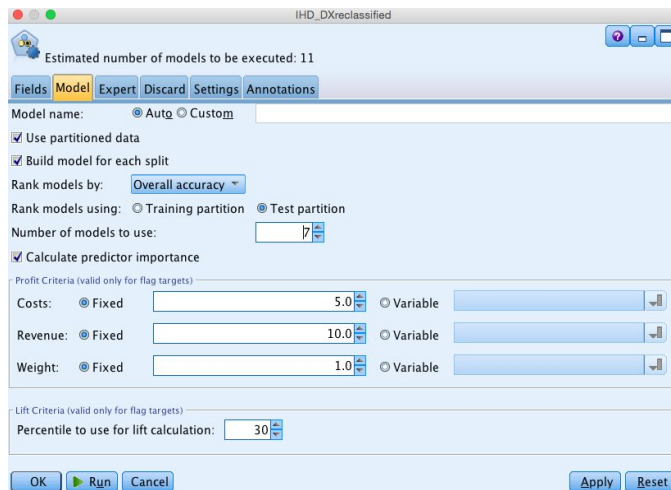
Types Format Annotations					
Read Values		Clear Values		Clear All Values	
Field	Measurement	Values	Missing	Check	Role
ID	Continuous	[1.0,200.0]		None	Record ID
AGE_50	Continuous	[20.0,69.0]		None	Input
MD_50	Nominal	1.0,2.0,3...		None	Input
SBP_50	Continuous	[88.0,210...		None	Input
DBP_50	Continuous	[38.0,140...		None	Input
HT_50	Continuous	[61.0,75.0]		None	Input
WT_50	Continuous	[109.0,24...		None	Input
CHOL_50	Continuous	[120.0,53...		None	Input
SES	Ordinal	1.0,2.0,3...		None	Input
CL_STATUS	Nominal	0.0,3.0,4...		None	Input
MD_62	Nominal	1.0,2.0,3...		None	Input
SBP_62	Continuous	[70.0,230...		None	Input
DBP_62	Continuous	[56.0,125...		None	Input
CHOL_62	Continuous	[139.0,36...		None	None
WT_62	Continuous	[108.0,24...		None	Input
IHD_DX	Nominal	0.0,1.0,2...		None	Input
DEATH_YR	Ordinal	0.0,63.0,6...		None	None
DEATH	Flag	1.0/0.0		None	Target
IHD_DX_5_cat	Nominal	"Not know...		None	Input



33. Attach an Auto Classifier Node to the Partition Node

34. Use predefined roles for in the Fields Tab

35. Use the following settings under the Model Tab:



Estimated number of models to be executed: 11

Fields | **Model** | Expert | Discard | Settings | Annotations

Model name: ☐ Auto ☐ Custom

☒ Use partitioned data

☒ Build model for each split

Rank models by: Overall accuracy

Rank models using: ☐ Training partition ☒ Test partition

Number of models to use:

☒ Calculate predictor importance

Profit Criteria (valid only for flag targets)

Costs: ☒ Fixed ☐ Variable

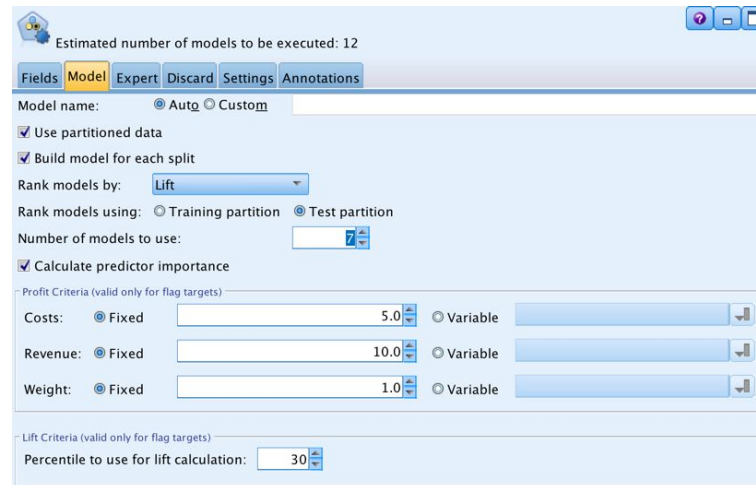
Revenue: ☒ Fixed ☐ Variable

Weight: ☒ Fixed ☐ Variable

Lift Criteria (valid only for flag targets)

Percentile to use for lift calculation:

OK Run Cancel Apply Reset



Estimated number of models to be executed: 12

Fields | **Model** | Expert | Discard | Settings | Annotations

Model name: ☐ Auto ☐ Custom

☒ Use partitioned data

☒ Build model for each split

Rank models by: Lift

Rank models using: ☐ Training partition ☒ Test partition

Number of models to use:

☒ Calculate predictor importance

Profit Criteria (valid only for flag targets)

Costs: ☒ Fixed ☐ Variable

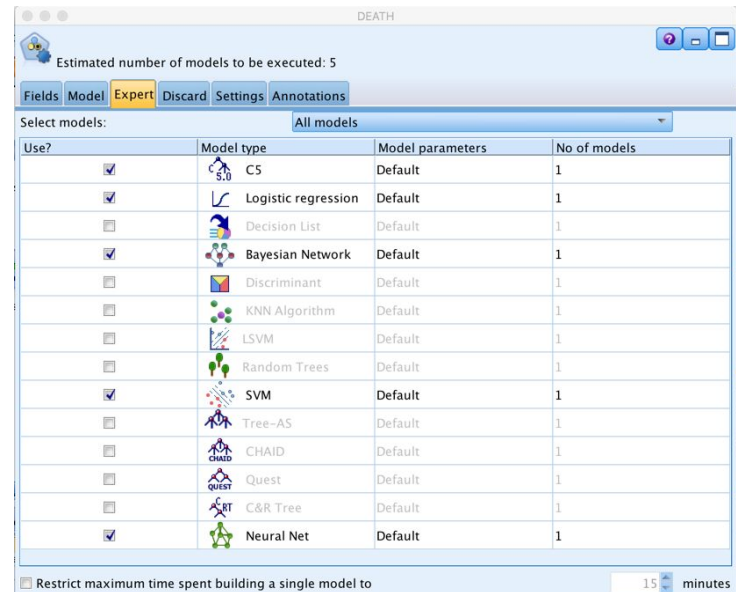
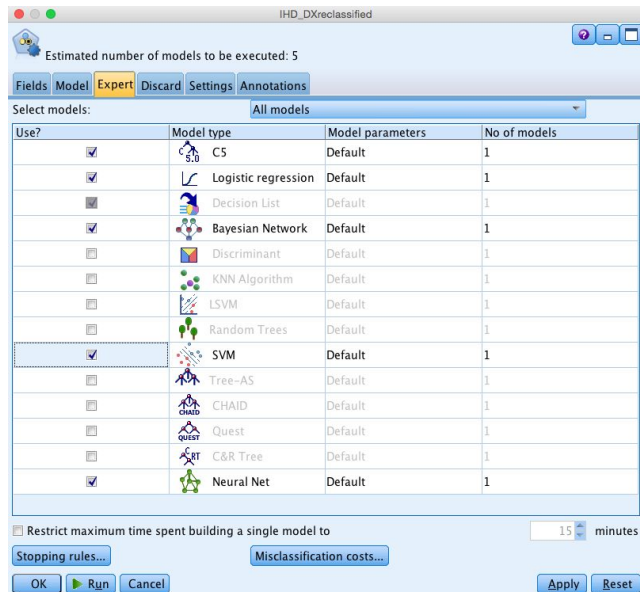
Revenue: ☒ Fixed ☐ Variable

Weight: ☒ Fixed ☐ Variable

Lift Criteria (valid only for flag targets)

Percentile to use for lift calculation:

36. Under Expert select the following models:



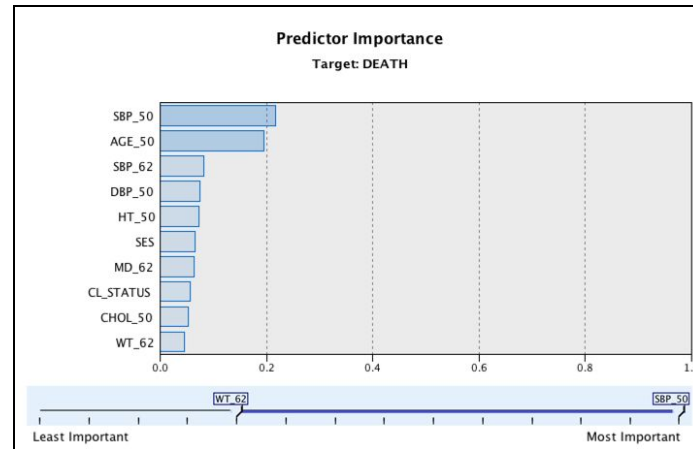
37. Hit Run

38. Open the golden nugget

39. Under the Model Tab hit the Model for Neural Net.

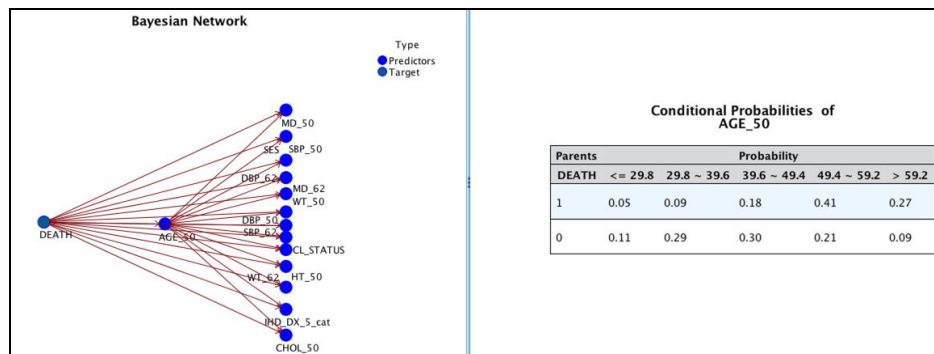
Q11. Which is the most important predictor?

The most important predictor is SBP_50



40. Now look in Model for the Bayesian Network.

Q12. Looking at the Network, do you think the role attached to the AGE_50 node makes sense? (AGE_50 is the Age of the subject in 1950.)

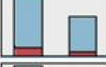


Yes, this makes sense based on the Bayesian Network it makes sense that AGE_50

41. Go back to the Model Tab. In the Use? Column check only C5 1. Hit Apply and OK.

42. Add an Analysis Node after the Auto Classifier golden nugget and hit Run with the default settings.

43. Record the Accuracies in each of the Training, Testing and Validation sets.

Use?	Graph	Model
☐		 Neural Net 1
☐		 Logistic regression 1
☐		 Bayesian Network 1
☒		 C5 1

Analysis of [DEATH] #4

File Edit

Analysis Annotations

Collapse All Expand All

Results for output field DEATH

Comparing \$XF-DEATH with DEATH

'Partition'	1_Training		2_Testing		3_Validation	
Correct	61	78.21%	31	56.36%	44	65.67%
Wrong	17	21.79%	24	43.64%	23	34.33%
Total	78		55		67	

44. Now go back in the golden nugget and select Bayesian Networks in the Use? Column. Hit Apply and OK.

Use?	Graph	Model
☒		 Bayesian Network 1
☐		 Neural Net 1
☐		 Logistic regression 1
☐		 C5 1
☐		 SVM 1

45. Rerun the Analysis Node and compare the Bayesian Networks accuracies with those of C5.

Analysis of [DEATH] #5

File Edit

Analysis Annotations

Collapse All Expand All

Results for output field DEATH

Comparing \$XF-DEATH with DEATH

'Partition'	1_Training	2_Testing	3_Validation
Correct	74 94.87%	23 41.82%	27 40.3%
Wrong	4 5.13%	32 58.18%	40 59.7%
Total	78	55	67

46. Here is what your stream should look something like when you're done.

