Lab #5: Regression and Classification in IBM Modeler
Biostat 202: Opportunities and Challenges of "Big Data"
Due: Sunday, 8/27 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.

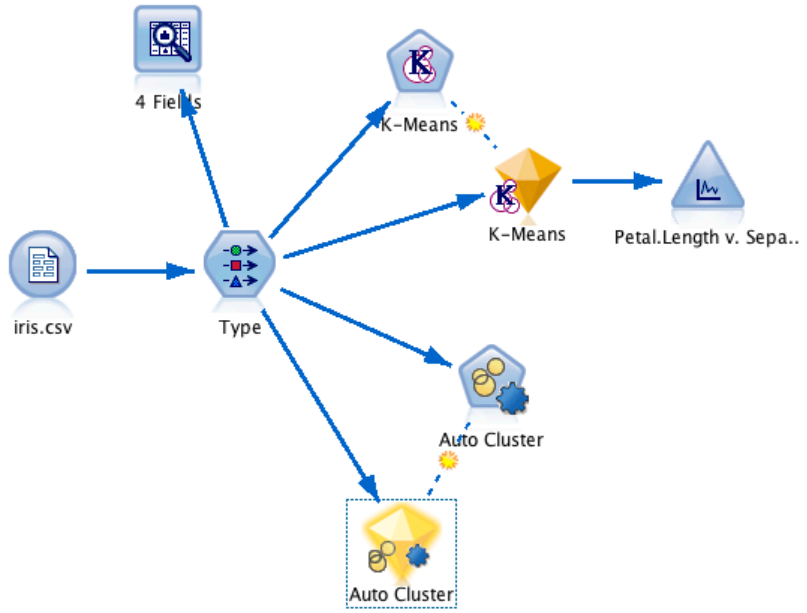
Please respond to the questions below (marked in bold as **Q1** to **Q12** and hand in your lab assignment by posting it to the CLE.

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.
2. 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".
3. 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?**
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?**
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? **Q3. Which of the Species was most easily put into a separate cluster by K-means?**
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". 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?**
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?**
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?** It seems that this procedure has overfitted the number of clusters. Perhaps not surprising since it is based on a complex nueral network-related model and the dataset is quite small.
11. At this point your stream should look something like the following:



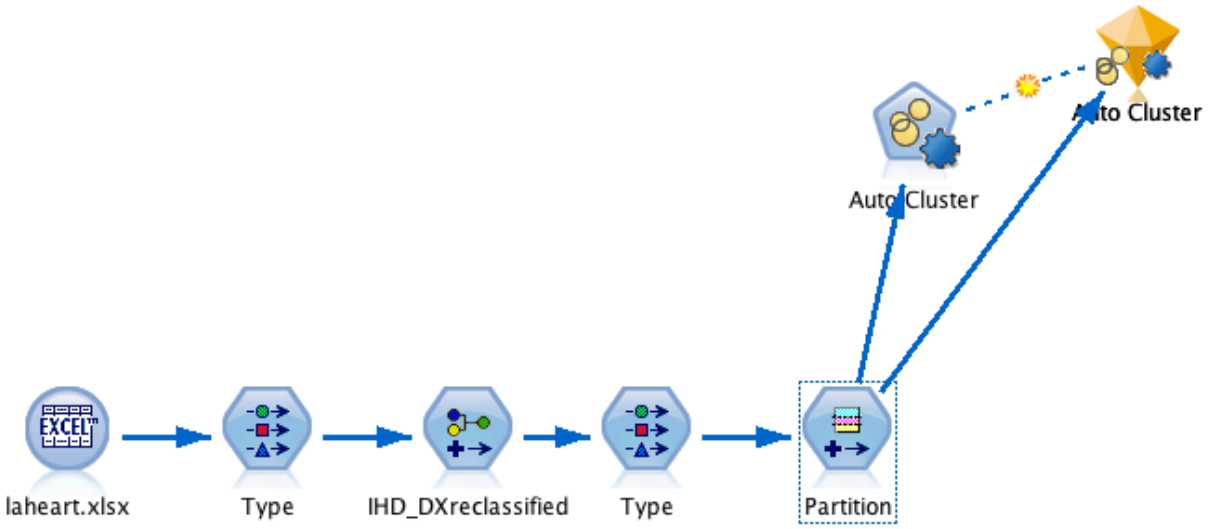
<u>Clustering, Regression and Classification</u>: We will now turn to the Los Angeles heart dataset (laheart.xlsx) available on the CLE. The dataset includes the following variables:

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

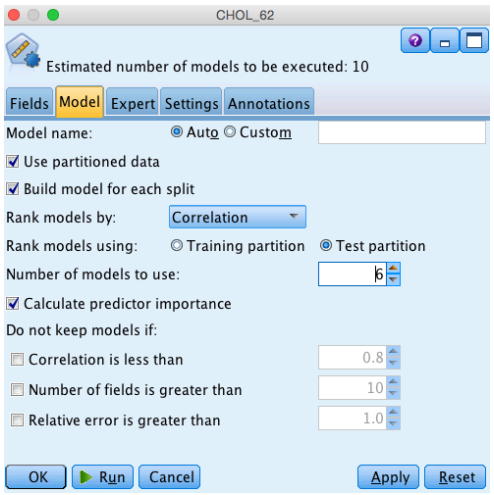
```
CL_STATUS     Clinical status            none   Nominal
                                         0=other heart disease(hd)
                                         1=Coronary hd
                                         2=Coronary and hypertensive hd
                                         3=Hypertensive 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.
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”.
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.
15. Connect another Type Node from the Reclassify Node. Set the Role of IHD_DX to None.
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.
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?**

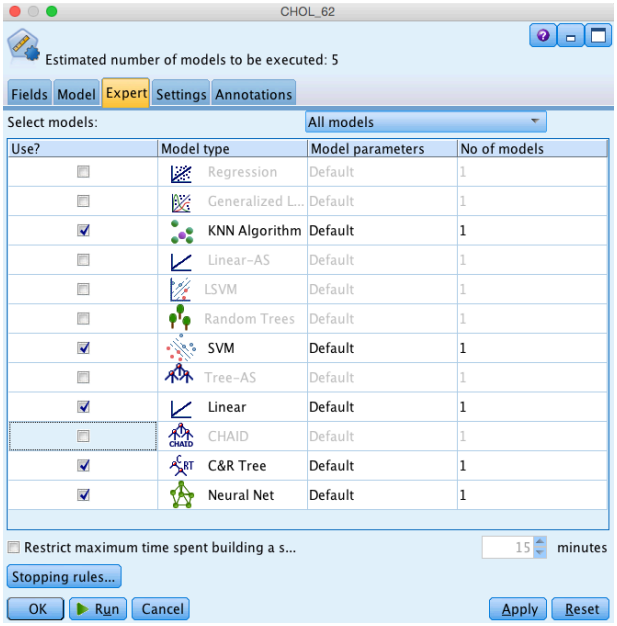
20. At the top right of the table go to the View button and change the value to Test set. **Q8. Is the ranking of the methods still the same? Does the best Test set model have higher or lower Silhouette score than the Training set?**
21. At this point your stream should look something like the following:



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". 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.
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

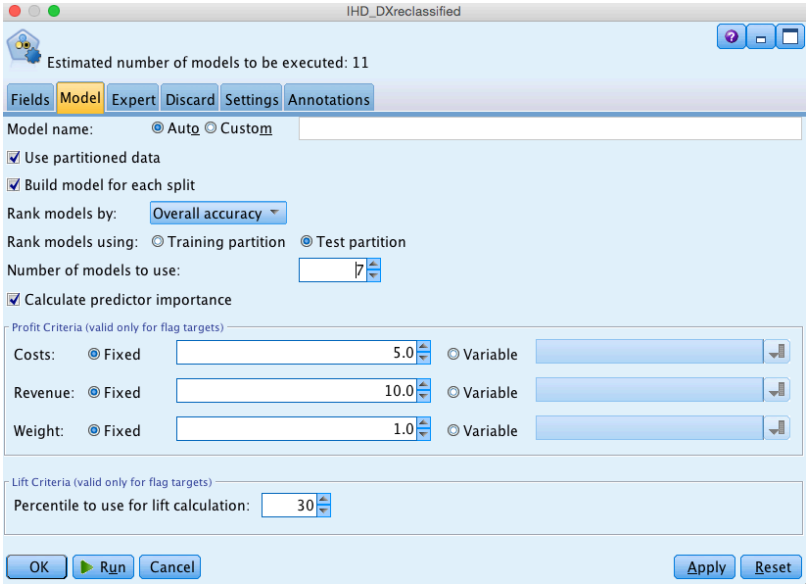


29. Click the cell under Model parameters for the KNN Algorithm to Specify and click to the Expert Tab. Click the cell under Options for K and hit Specify. Set so that K=3, 5, and 7 are considered. Hit OK.
30. Open up the Model Parameters for Nueral Net. Go to the Expert Tab. **Q9. What is the default Number of component models?**
31. Back out and run the Auto Numeric Node. **Q10. Which regression method has the highest Correlation value and what is the associated number of fields used?**
32. Go back into the 2nd Type Node. Set DEATH as the Target variable and set the Role of CHOL_62 to "None".

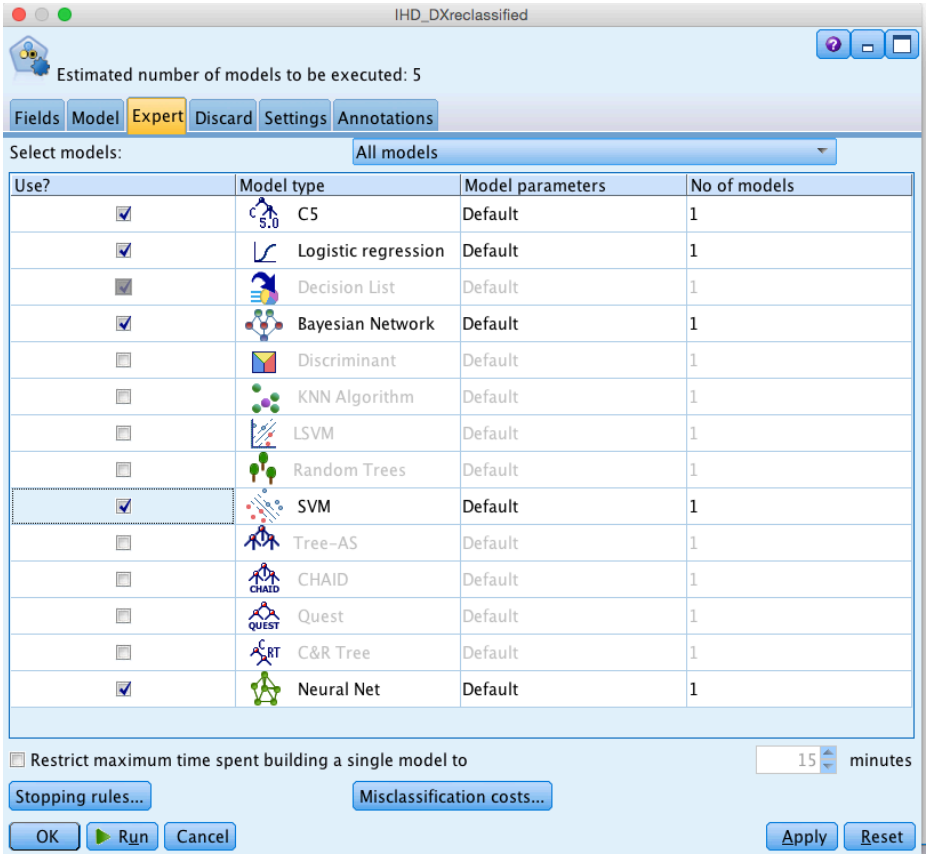
33. Attach an Auto Classifier Node to the Partition Node

34. Use predefined roles for in the Fields Tab

35. Use the following settings under the Model Tab:



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?**
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.)
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.
44. Now go back in the golden nugget and select Bayesian Networks in the Use? Column. Hit Apply and OK.
45. Rerun the Analysis Node and compare the Bayesian Networks accuracies with those of C5.
46. Here is what your stream should look something like when you're done.