Lab #4: Regression and Classification in IBM Modeler
Biostat 202: Opportunities and Challenges of "Big Data"

Due: Sunday, 8/20 by 11.55pm

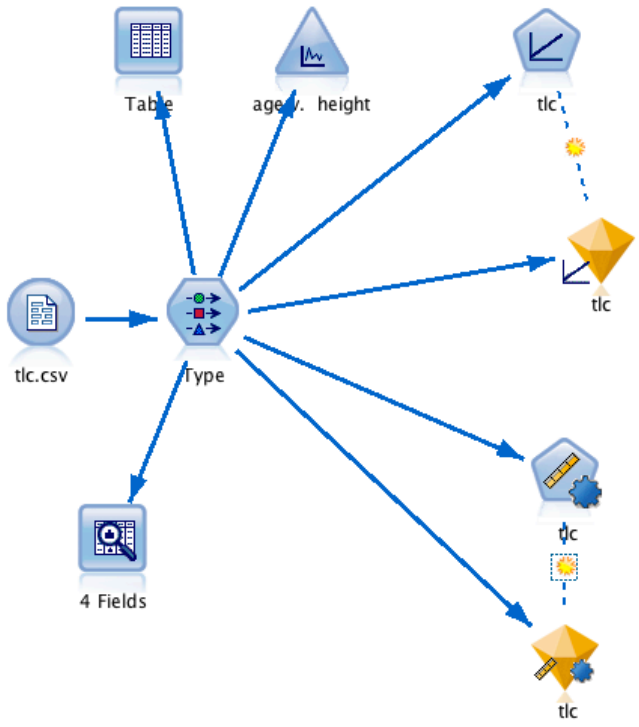
The purpose of this lab is to familiarize you to working with different regression and classification techniques and how to apply them in IBM SPSS Modeler.

Please respond to the questions below (marked in bold as **Q1** to **Q18**) and hand in your lab assignment by posting it to the CLE.

Regression: The **Total lung capacity dataset** gives the total lung capacity for 30 individuals along with age, sex and height (available as tlc.csv on the CLE). The outcome variable is tlc (total lung capacity) and predictors are age, sex, and height (in centimeters).

1. Download the Total lung capacity dataset from CLE ("tlc.csv") and read into IBM Modeler.
2. Use the Type Node to appropriately set the data given the description above.
3. Connect a Plot Node to the Type Node. Set the X field to age, the Y field to height and in the Overlay section set Color to sex. **Q1 What do you think is the label for females?**
4. Connect a Linear Node to the Type Node. Remove sex and height from the set of predictors under the *Fields Tab* so that age is the only predictor. Go to *Build Options*. In Objectives you want to *Build a new model* and *Create a standard model*. Click the Basics item. You will see that Modeler does some *Automatic Data Preparation* steps. Leave that checked for now (you can try running again without this checked later if you want to see whether it makes a difference). Go to *Model Selection*. Because we only have one predictor the Model selection method should not play a role (you can test this theory if you wish). Set the Criteria for entry/removal to *Adjusted R2*. Hit *Run*.
5. Open the Golden Nugget. Select the *Model* Tab. Look at the panels on the left. Look at the *Model Summary* top panel. **Q2 Do you think the model has good predictive power?**
6. Go back to the Linear Node and hit the Fields Tab. Select *Use predefined nodes*. You should find that all of age, sex and height are in the predictors. Click the Build Options Tab. Under Model selection method click *Best subsets* (there are not that many predictors so we may as well look at all combinations). The *Forward Stepwise Selection* area in the middle of the page will be grayed out since we are not using Stepwise. At the bottom in *Best Subsets Selection*, under *Criteria for entry/removal* – choose *Information Criterion (AICC)*. Hit Run.
7. Open the Golden Nugget. Select the Model Tab. Look at the panels on the left and click on the top *Model Summary* panel. **Q3 Do you think the selected model has better predictive power than the simple regression?**

8. Go down to the 3rd panel (*Predictor Importance*). Examine the relative importance of the variables that are in the best model. **Q4 Why do you think age is missing?**
9. Scan further down through the panels until you get to the *Coefficients* panel. **Q5 How does the predicted outcome change for increased height? How does the predicted outcome change for Males vs. Females?** Change the *Style* label at the bottom left from *Diagram* to *Table*. Click on the arrow in the *Coefficient* box on the table to expand. Do these coefficients confirm what you answered above?
10. Connect an Auto Numeric Node from the Type Node. Under *Fields* select *Use predefined roles*. Under the Models Tab select *Number of models to use* as 4. Note that *Rank models using Test partition* is selected by default; this is what we need to avoid overfitting. Also, leave the *Rank models by value* as *Correlation*. Under *Expert* select Linear, Linear-AS, C&R Tree, SVM, KNN Algorithm. Click *Model parameters* for the Linear-AS and select *Specify*. You will notice that *Consider two way interactions* is marked as *true*. This allows all possible interaction terms to be fit in the model. Click the Expert Tab inside the Model Parameters window. **Q6 What is the default Model selection method?** Click OK and then Run.
11. Open the golden nugget. Select the Model tab. **Q7 which algorithm did best with respect to the *Correlation* metric?** Was that also the best approach with respect to *Relative Error*? Click on the *Graph* Tab. Which predictor has the highest importance? Are you surprised by that?
12. Your stream should look something like the following when you are done:



Classification: The ***Iris* flower dataset** is a very famous dataset (at least in the world of statistics) introduced by Ronald Fisher (1890 – 1962). Ronald Fisher is arguably the most famous statistician of all time. The concept of randomization in clinical trials was his, and he was also famous for his work in Mendelian genetics and natural selection. However, he was not universally considered a "nice person"). The dataset is well described on its Wikipedia page: https://en.wikipedia.org/wiki/Iris_flower_data_set
The 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 (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.
3. Run an Audit Node and examine the graphs (histograms) of the individual predictors – **Q8 Which of the predictor(s) do you think are likely to be the most important in predicting Species based on looking at the graphs?**
4. Partition the data into 60% training, 20% validation, and 20% test
5. Add a *Discriminant Node* after the Partition Node and run it.
6. Look in the golden nugget and go to the Model Tab – **Q9 Do the Predictor Importance values match what you expected based on looking at the histograms in Q8?**
7. Add an analysis node after the golden nugget.
8. Add a *Logistic Node* after the partition. Set the Method to "Stepwise", and set the *base category* to setosa. Run.
9. Look in the golden nugget and go to the Model Tab. Click on the equation for setosa. It should just be a 0 because you have set setosa to be the base (or reference) category that the other groups are compared to. Now look at the virginica equation. This gives the log of the odds-ratio of virginica relative to setosa as an equation in terms of the predictors. Similarly take a look at versicolor. **Q10 Which variables have been retained in the model? Do the retained predictors match what you might expect based on looking at the histograms in Q8?**
10. Go to the *Advanced Tab*. Read the Warning. The problem here is (probably) that the logistic regression model is splitting the data up too perfectly and cannot fit a unique model. It is not ideal that IBM Modeler makes you work to find the warning.
11. Add an analysis node after the golden nugget.
12. Add a *C5.0 Node* after the partition and run it with the default settings.
13. Open up the golden nugget and look in the Model Tab. **Q11 Do the Predictor Importance values match what you expected based on looking at the histograms in Q8?**

14. Expand the line with Petal.Length > 1.900. **Q12 what would your prediction of the species of Iris be for a flower with petal length 2.75 based on the C5.0 fitted tree?**
15. Hit the "Viewer Tab" and check that your answer to Q12 still makes sense to you based on the tree.
16. Add an analysis node after the golden nugget.
17. Run each of the 3 model Analysis Nodes. **Q12 Rank the 3 models with how well they did with respect to classification accuracy in the *validation set*.** Note that in the literature outside of IBM Modeler this would be referred to as the ***test set***. Also, note that the validation set here is very small – 20% of 150 = 30. This is one of those situations where you would prefer to do *cross-validation*.
18. Your stream should look something like the following once you are done: