https://docs.google.com/document/d/1Ta0F6aGukIQZR1xih5Jx4QxZOzZR7EL7VIqS-XldMJs/pub

Lab #4: Regression and Classification in IBM Modeler

Biostat 202: Opportunities and Challenges of "Big Data"

Due: Sunday, 8/21 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.

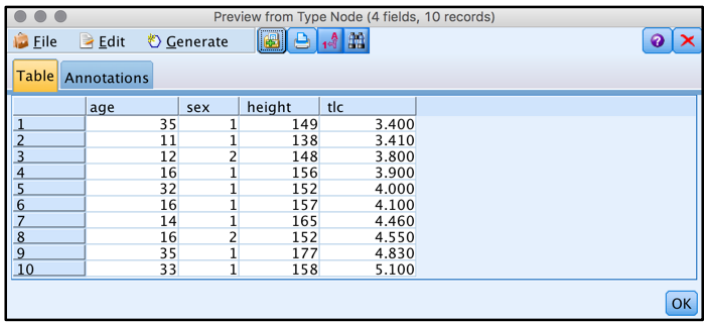
Please respond to the questions below (marked in bold as **Q1** to **Q*****) and hand in your lab assignment by posting it to the CLE.

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).

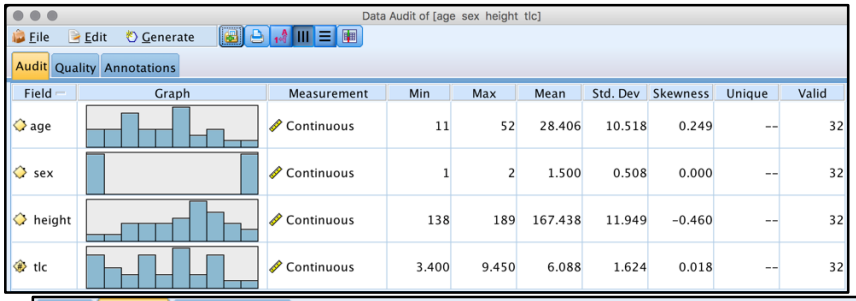
1. Download the Total lung capacity dataset from CLE ("tlc.csv") and read into IBM Modeler.

The tlc.csv has double quotes in the variable names, but the data look totally numerical.

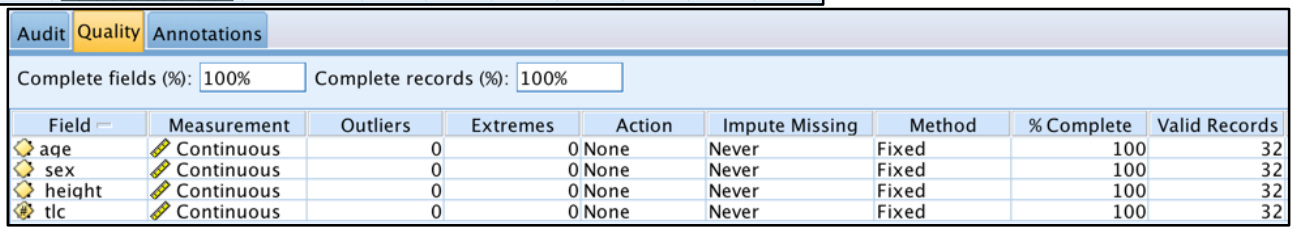
2. Use the Type Node to appropriately set the data given the description above.



	age	sex	height	tlc
1	35	1	149	3.400
2	11	1	138	3.410
3	12	2	148	3.800
4	16	1	156	3.900
5	32	1	152	4.000
6	16	1	157	4.100
7	14	1	165	4.460
8	16	2	152	4.550
9	35	1	177	4.830
10	33	1	158	5.100



Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
age		Continuous	11	52	28.406	10.518	0.249	--	32
sex		Continuous	1	2	1.500	0.508	0.000	--	32
height		Continuous	138	189	167.438	11.949	-0.460	--	32
tlc		Continuous	3.400	9.450	6.088	1.624	0.018	--	32



Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records
age	Continuous	0	0	None	Never	Fixed	100	32
sex	Continuous	0	0	None	Never	Fixed	100	32
height	Continuous	0	0	None	Never	Fixed	100	32
tlc	Continuous	0	0	None	Never	Fixed	100	32

3. Connect a Plot Node to the Type Node.

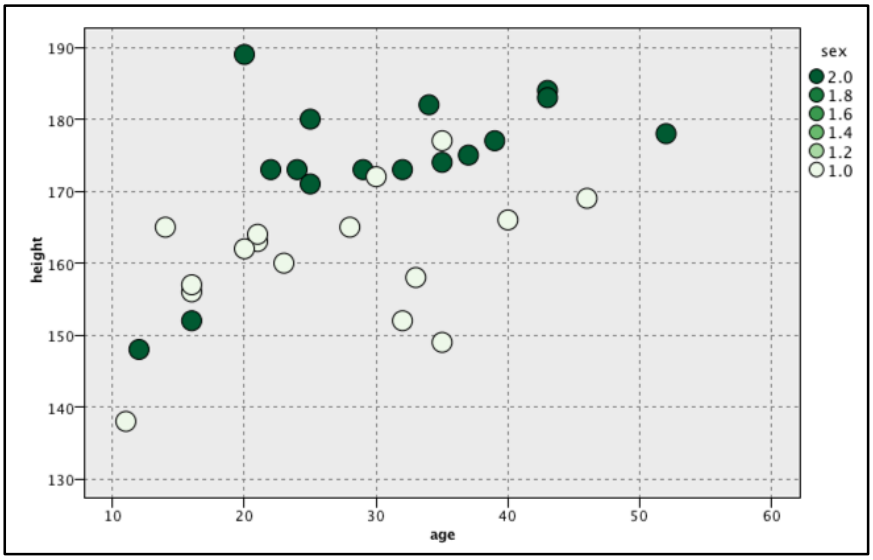
Set the **X field to age**,

the **Y field to height**

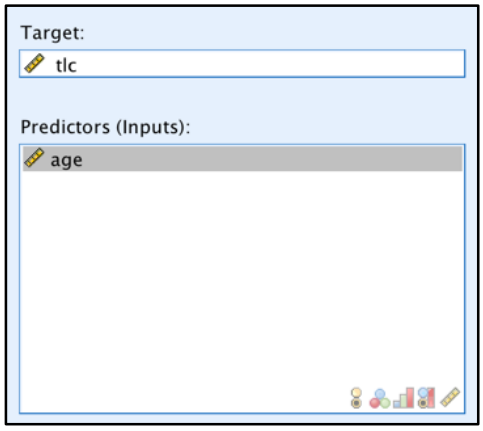
and in the **Overlay** section set **Color to sex**.

Q1 What do you think is the label for females?

I think the label for females is probably 1.



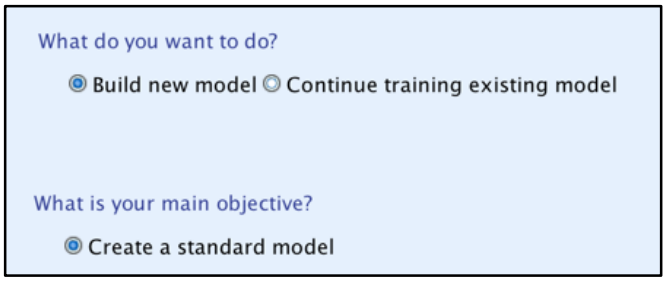
4. Connect a **Linear Node** to the Type Node.



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**.

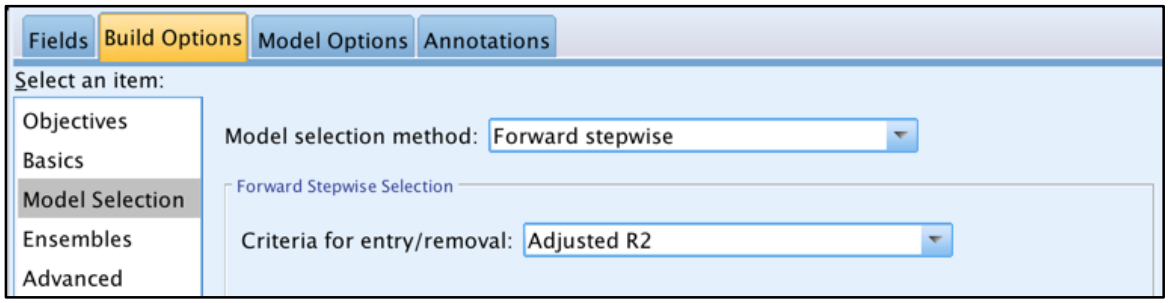


Click the **Basics** item.

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).

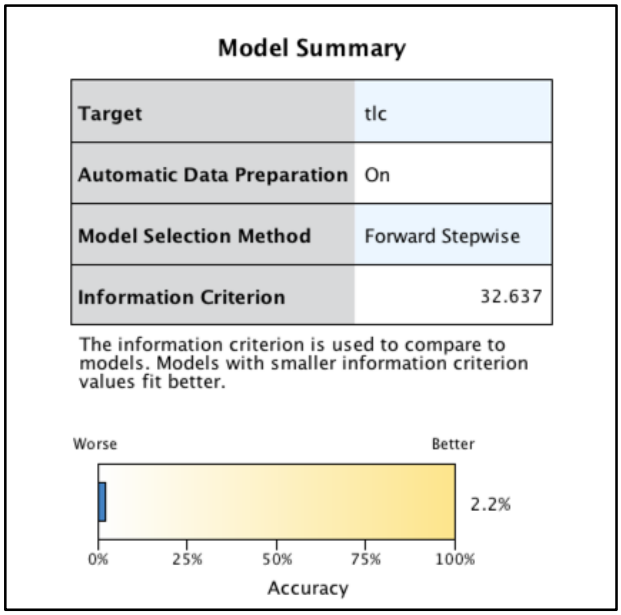
Go to **Model Selection**. Because we only have one predictor the Model selection method should not play a role (you can test this theory if you wish). Set the Criteria for entry/removal to **Adjusted R2.**



Hit *Run*.

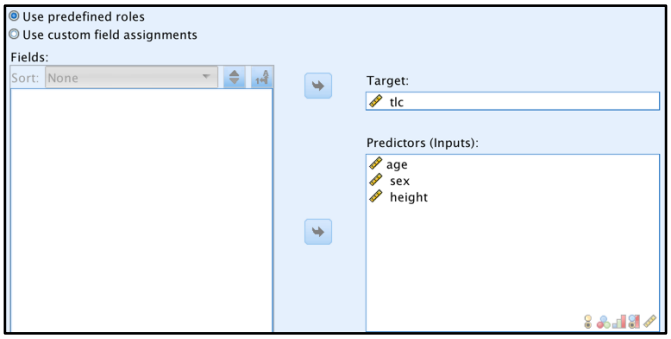
5. Open the Golden Nugget. Select the *Model* Tab. Look at the panels on the left. Look at the *Model Summary* top panel.

Q2 Do you think the model has good predictive power?



Provided the terms, "worse," "better," and "accuracy" still mean what I think they do, then no. This model did not have good predictive power. The

6. Go back to the **Linear Node** and hit the **Fields** Tab.

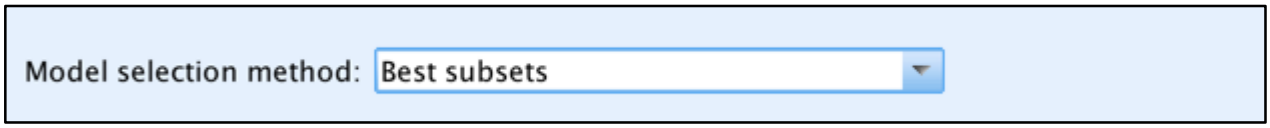


Select **Use predefined nodes**.

You should find that all of age, sex and height are in the predictors.

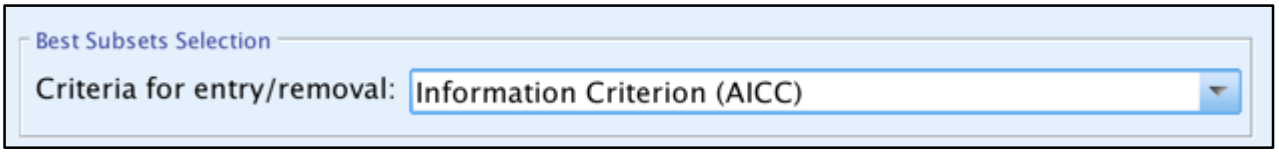
Click the **Build Options** Tab.

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)*.

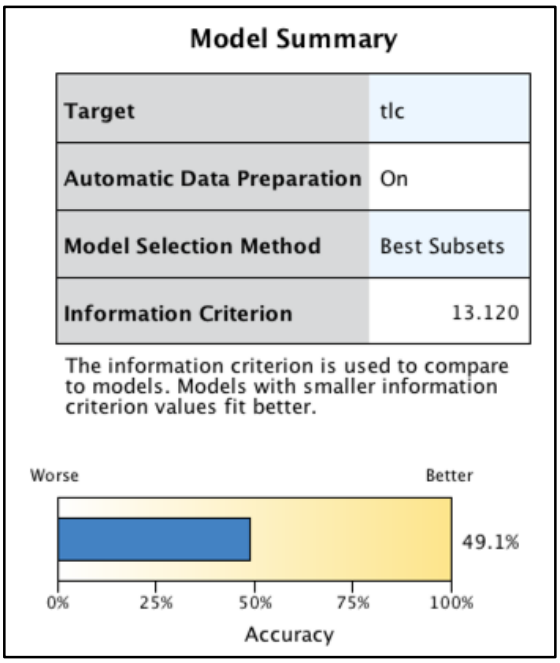


Hit Run.

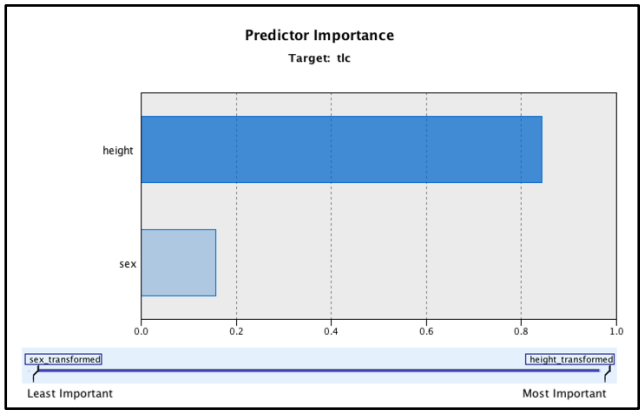
7. Open the Golden Nugget. Select the Model Tab. Look at the panels on the left and click on the top *Model Summary* panel.

Q3 Do you think the selected model has better predictive power than the simple regression?

Yes, this 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?

It did not meet the Criteria for entry/removal criteria. When I chose the **Best Subsets** criteria, it chooses "the subsets of variables that gives the best criteria value as the input variable for the regression model." Age was not one of them.

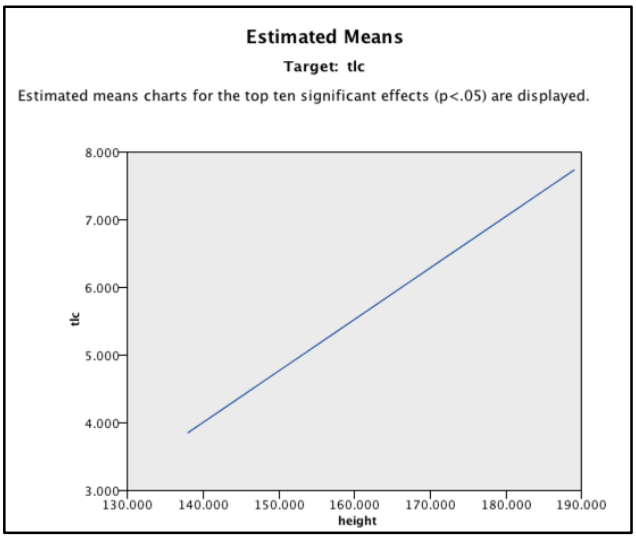
Table 5.2 Variable selection methods for linear regression in the Linear and Regression node		
Method	**Description**	**Linear node**

Include all predictors/ Enter	This is the naive approach, where all input variables are included.	X
(Forward) Stepwise	Like the Forwards method, but with the adding and removing of variables in each step.	X
Best subset	The subset of variables that gives the best criteria value is selected as the input variable for the regression model.	X

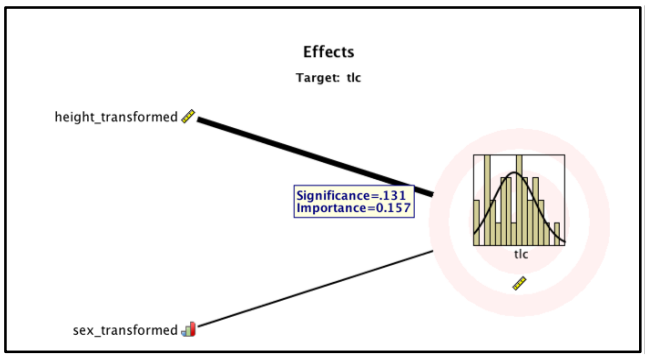
9. Scan further down through the panels until you get to the **Coefficients** panel.

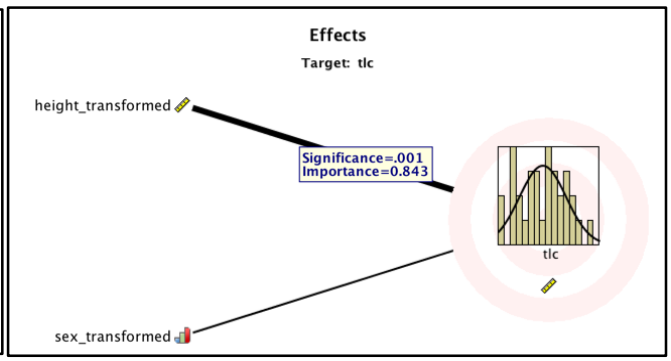
Q5 How does the predicted outcome change for increased height?

Predicted tlc goes up with increases in height.



How does the predicted outcome change for Males vs. Females?





It doesn't change much for age, and it certainly isn't significant.

Change the *Style* label at the bottom left from *Diagram* to *Table*.

Coefficients

Target: tlc

Model Term	Coefficient ▼	Std.Error	t	Sig.	95% Confidence Interval		Importance
					Lower	Upper	
Intercept	-6.264	3.680	-1.702	.099	-13.790	1.263	
height_transformed	0.076	0.021	3.609	.001	0.033	0.119	0.843
sex_transformed=1	-0.771	0.496	-1.555	.131	-1.785	0.243	0.157
sex_transformed=2	0.000						0.157

Click on the arrow in the *Coefficient* box on the table to expand. **Do these coefficients confirm what you answered above?**

Yes. The regression equation could be written as,

y = -6.264(intercept) + (0.076*height) + (-0.771*male)

Effects

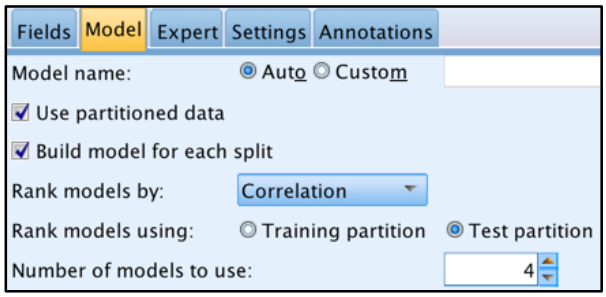
Target: tlc

Source	Sum of Squares	df	Mean Square	F	Sig.
Corrected Model ▶	42.794	2	21.397	15.944	.000
Residual	38.918	29	1.342		
Corrected Total	81.712	31			

The overall model is significant. I can see from the ANOVA table that all three variables are included in this model (If there were *k* groups being compared, then there are ***k-1*** degrees of freedom associated with the outcome of interest).

10. Connect an Auto Numeric Node 

from the Type Node.

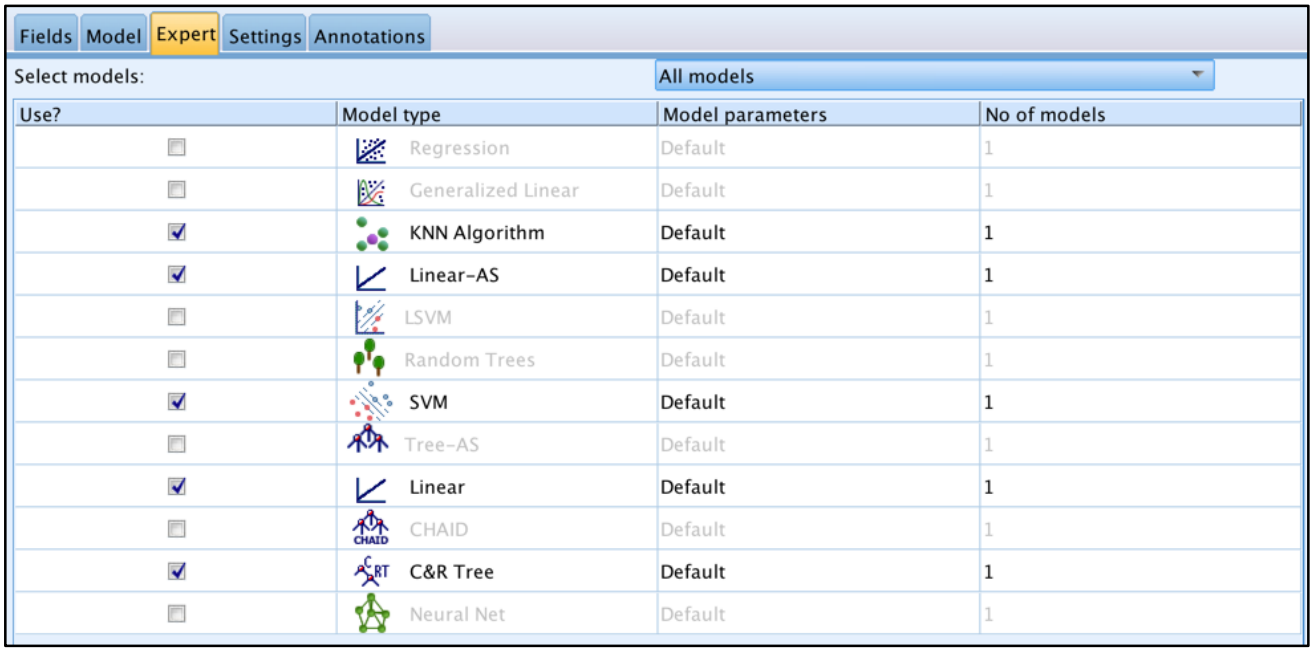


Under **Fields** select **Use predefined roles.**

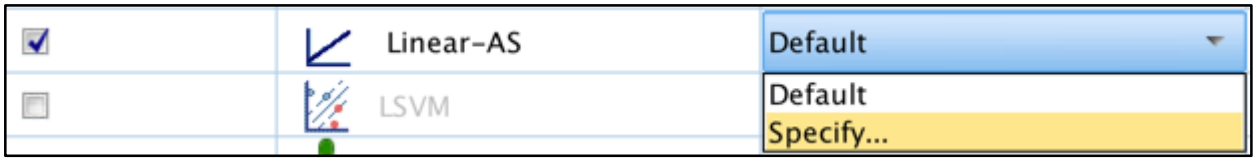
Under the Models Tab select **Number of models to use** as 4.

Note that **Rank models using Test partition** is selected by default; this is what we need to avoid overfitting.

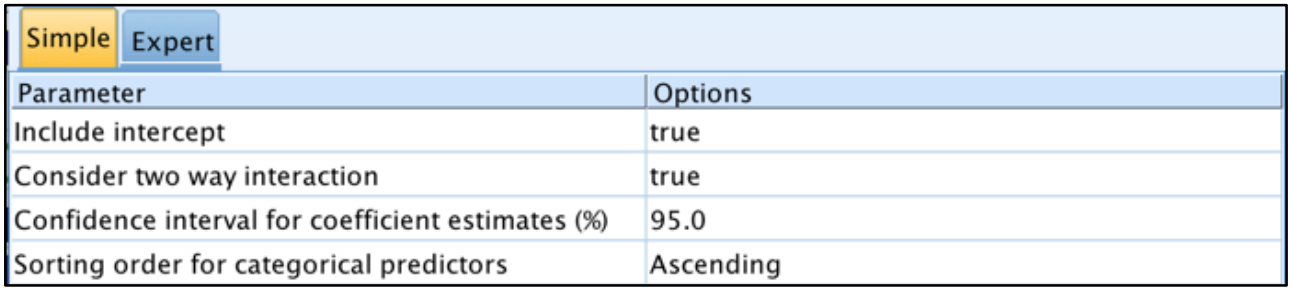
Also, leave the *Rank models by value* as *Correlation*.



Under *Expert* select Linear, Linear-AS, C&R Tree, SVM, KNN Algorithm.

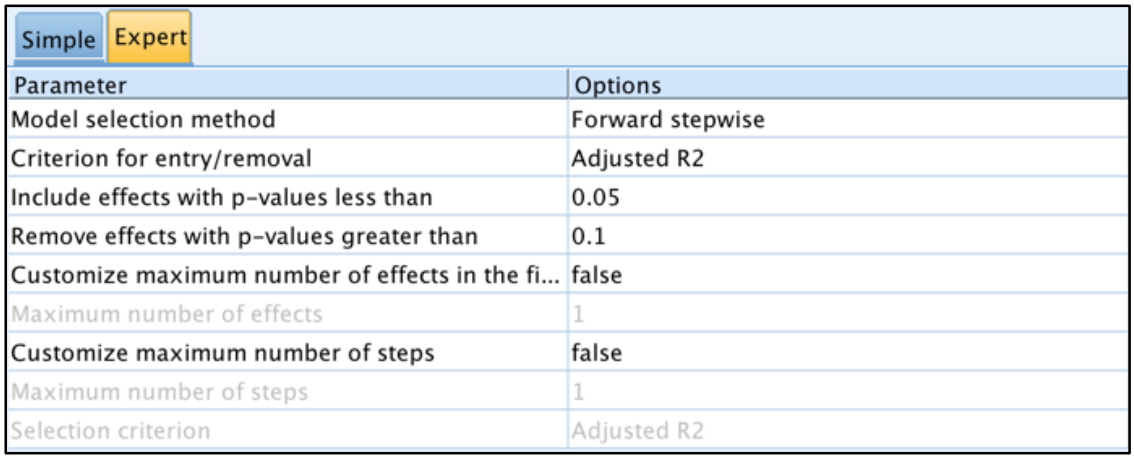


Click *Model parameters* for the Linear-AS and select *Specify*.



You will notice that *Consider two way interaction s* is marked as *true*.

This allows all possible interaction terms to be fit in the model.

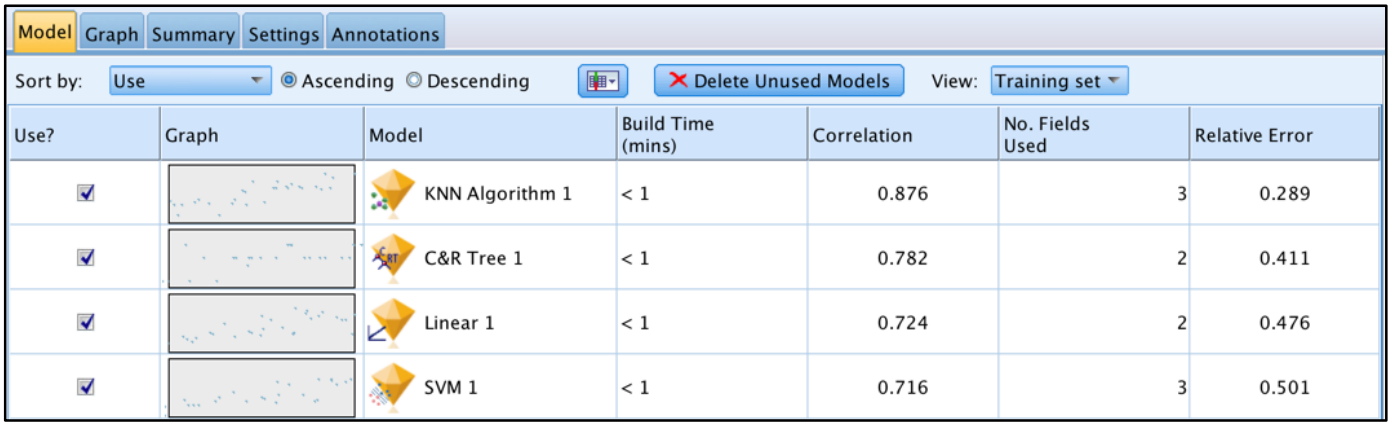


Click the Expert Tab inside the Model Parameters window. **Q6 What is the default Model selection method?**

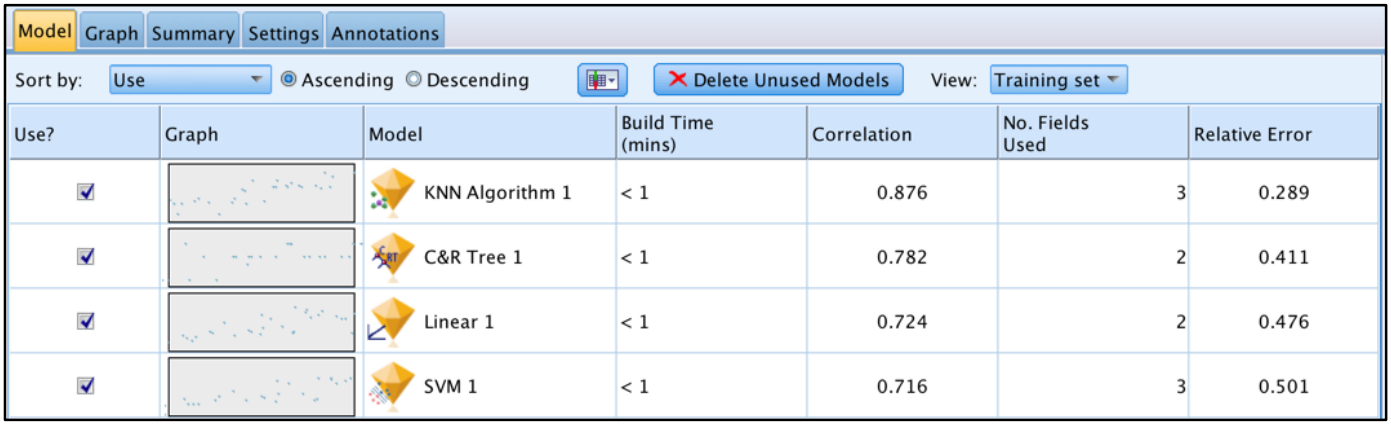
Click OK and then Run.

11. Open the golden nugget. Select the Model tab.

Q7 which algorithm did best with respect to the *Correlation* metric?



Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
☑		KNN Algorithm 1	< 1	0.876	3	0.289
☑		C&R Tree 1	< 1	0.782	2	0.411
☑		Linear 1	< 1	0.724	2	0.476
☑		SVM 1	< 1	0.716	3	0.501

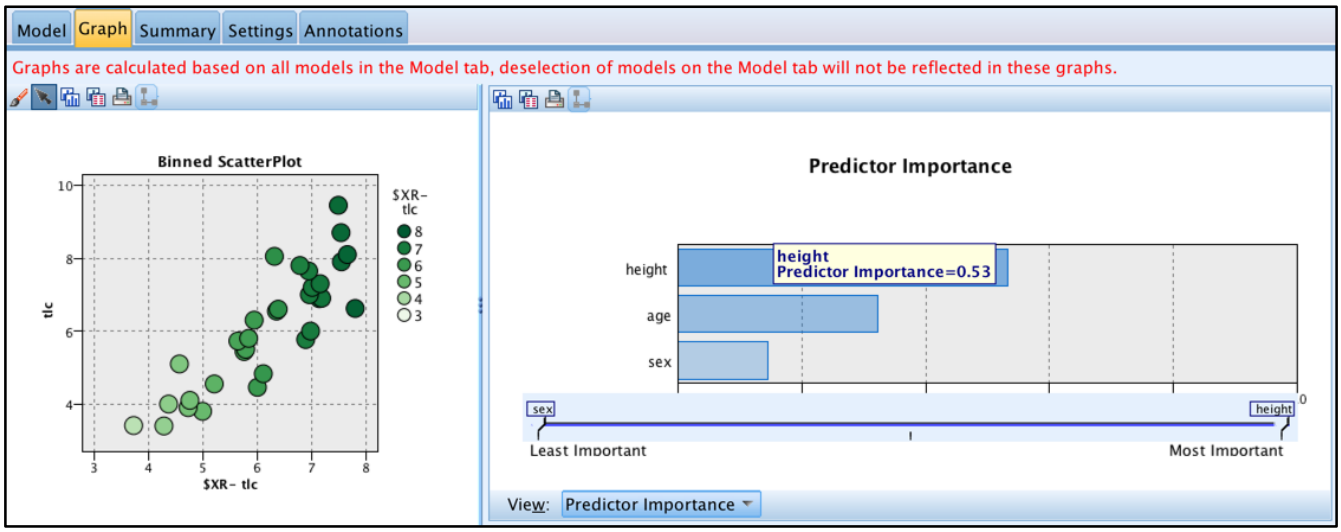


Use?	Graph	Model	Build Time (mins)	Correlation	No. Fields Used	Relative Error
☑		KNN Algorithm 1	< 1	0.876	3	0.289
☑		C&R Tree 1	< 1	0.782	2	0.411
☑		Linear 1	< 1	0.724	2	0.476
☑		SVM 1	< 1	0.716	3	0.501

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? No.

12. Your stream should look something like the following when you are done:

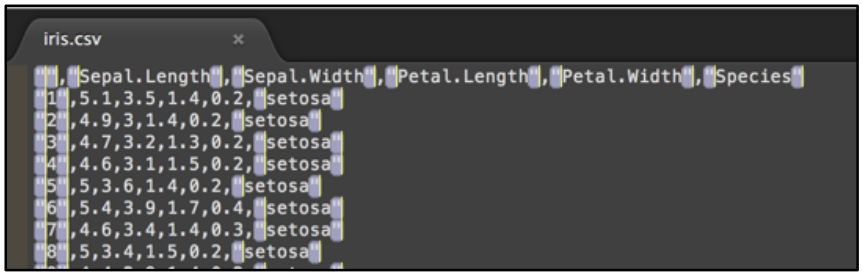
Classification: The ***Iris* flower dataset** is a very famous dataset (at least in the world of statistics) introduced by Ronald Fisher (1890 – 1962). Ronald Fisher is arguably the most famous statistician of all time. The concept of randomization in clinical trials was his, and he was also famous for 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: h

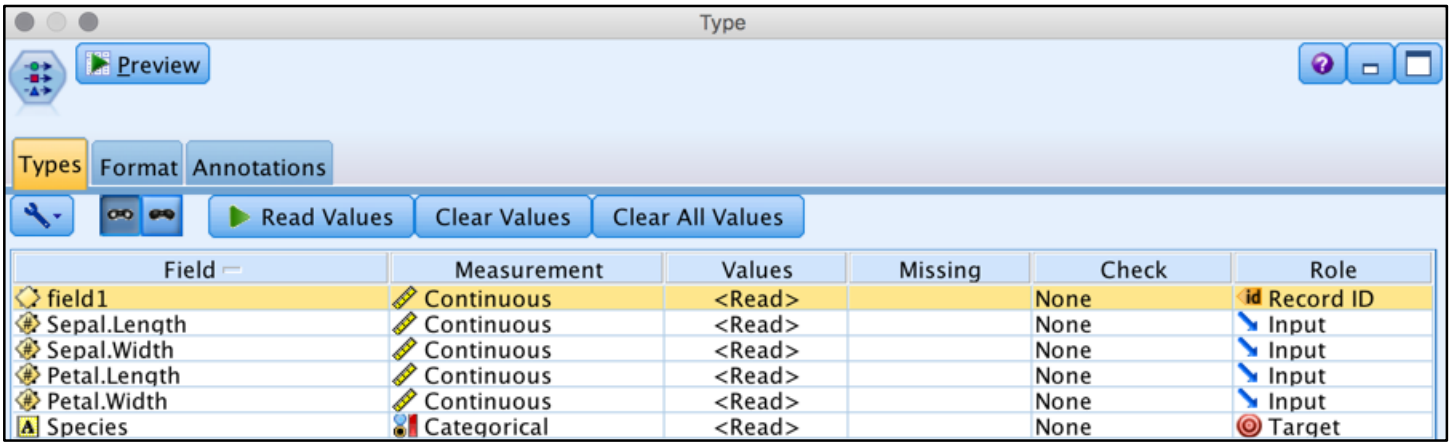
ttps://en.wikipedia.org/wiki/Iris_flower_data_set

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

The iris.csv dataset has double-quotes that need to be paired & discarded. It looks like one of the numbers is stored as text (probably an ID number?).





Field	Measurement	Values	Missing	Check	Role
field1	Continuous	<Read>		None	Record ID
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	Target

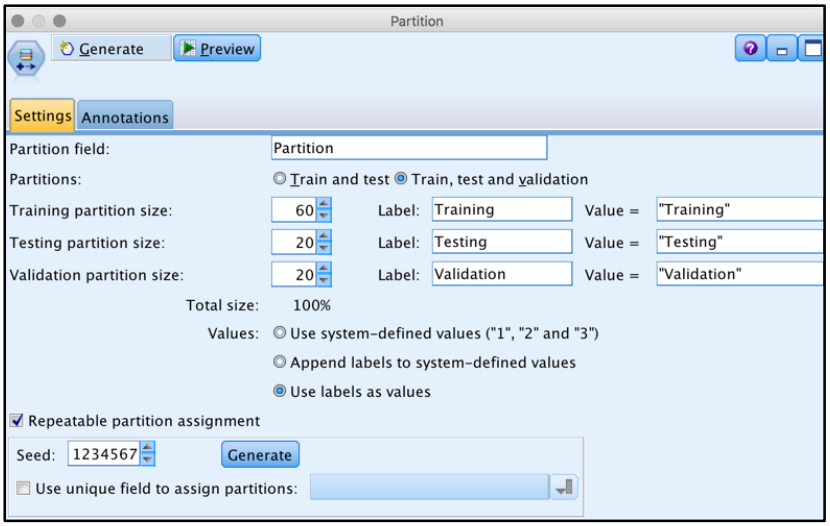
2. Use the Type Node to appropriately set the data given the description above.

3. Run an Audit Node and examine the graphs (histograms) of the individual predictors –

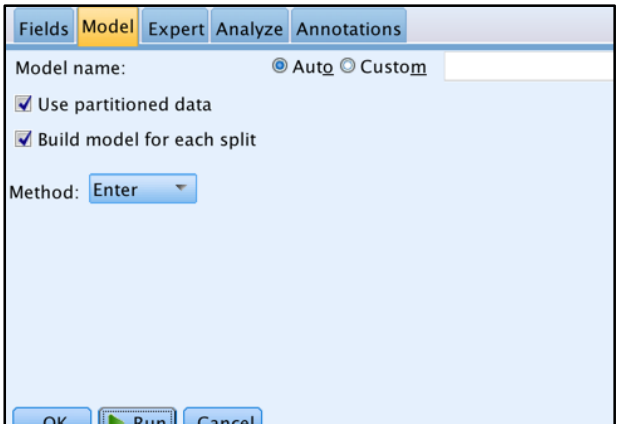
Q8 Which of the predictor(s) do you think are likely to be the most important in predicting Species based on looking at the graphs? petal length & petal width

Audit | Quality | Annotations

Field	Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
field1		Continuous	1	150	75.500	43.445	0	--	150
Sepal.Length		Continuous	4.300	7.900	5.843	0.828	0.315	--	150
Sepal.Width		Continuous	2.000	4.400	3.057	0.436	0.319	--	150
Petal.Length		Continuous	1.000	6.900	3.758	1.765	-0.275	--	150
Petal.Width		Continuous	0.100	2.500	1.199	0.762	-0.103	--	150
Species		Nominal	--	--	--	--	--	3	150

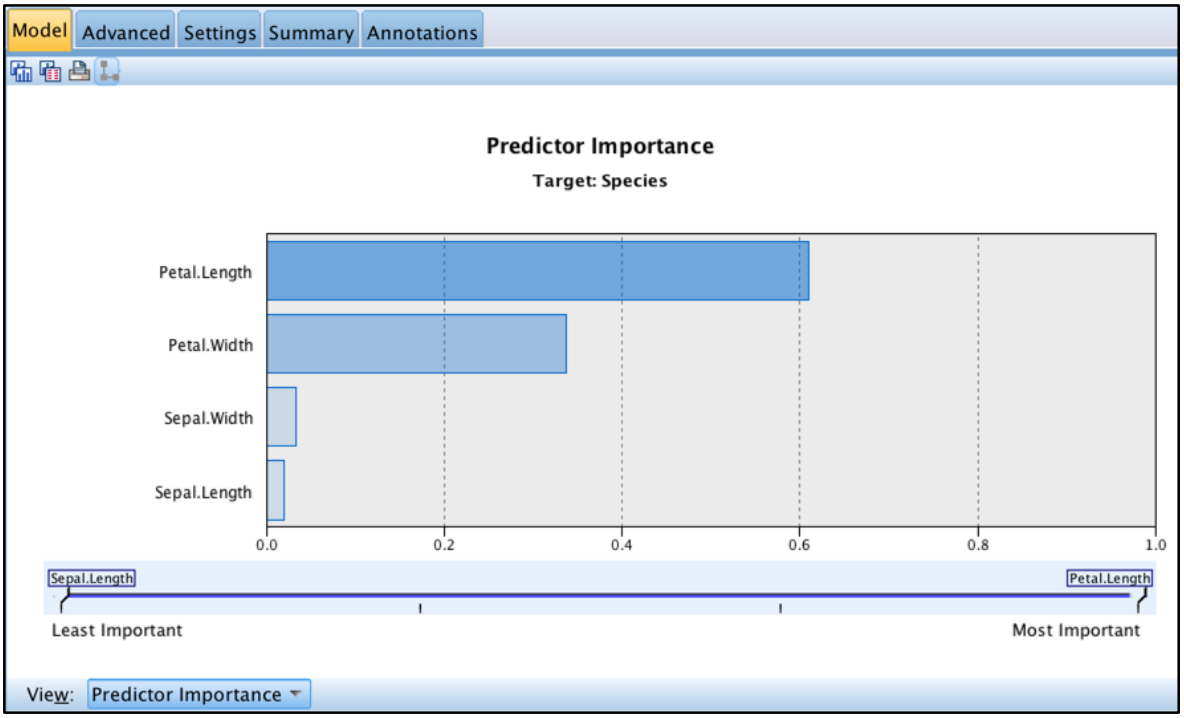


4. Partition the data into 60% training, 20% validation, and 20% test



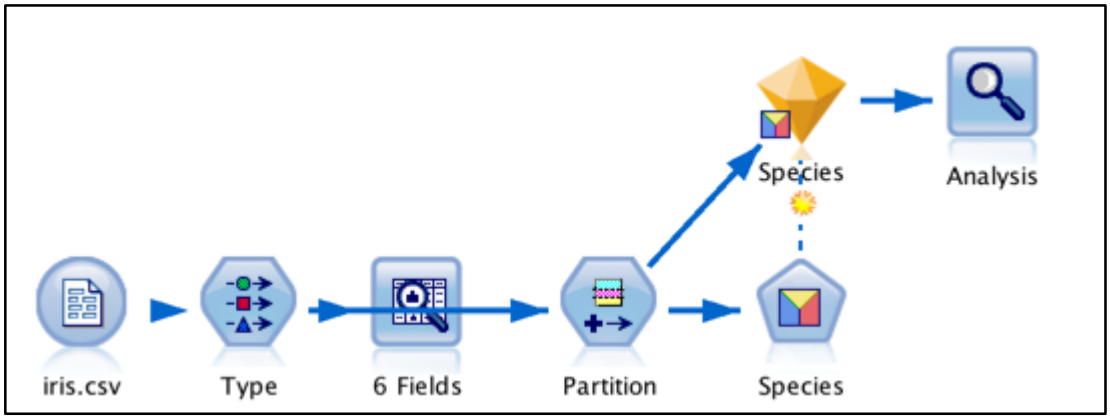
5. Add a *Discriminant Node* after the Partition Node and run it.

6. Look in the golden nugget and go to the Model Tab –

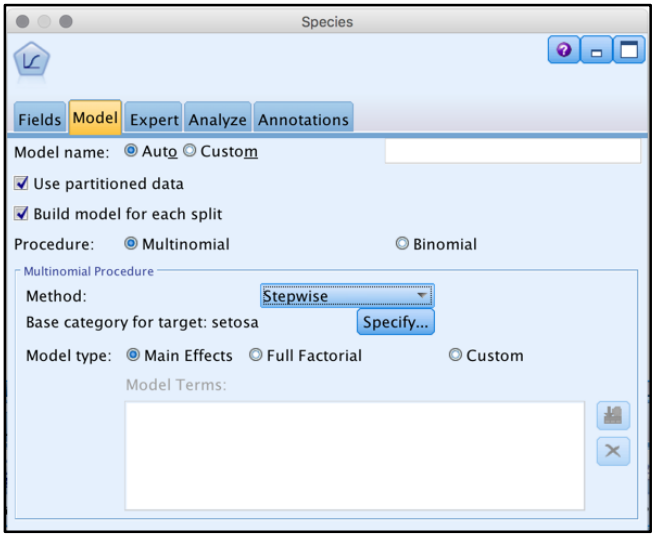


Q9 Do the Predictor Importance values match what you expected based on looking at the histograms in Q8?

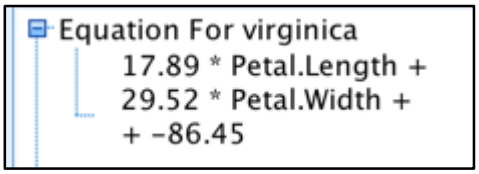
7. Add an analysis node after the golden nugget.

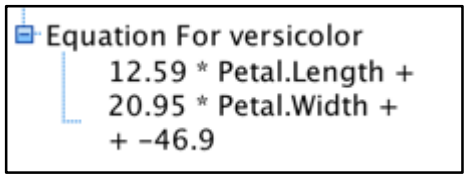


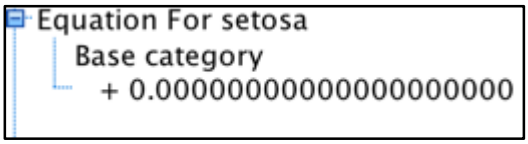
8. Add a *Logistic Node* after the partition. Set the Method to "Stepwise", and set the *base category* to setosa. Run.



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.







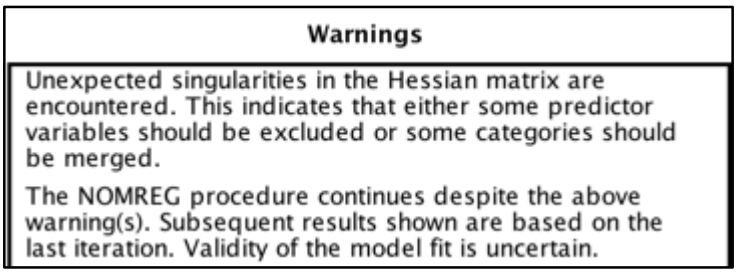
Q10 Which variables have been retained in the model?

Petal length and petal width.

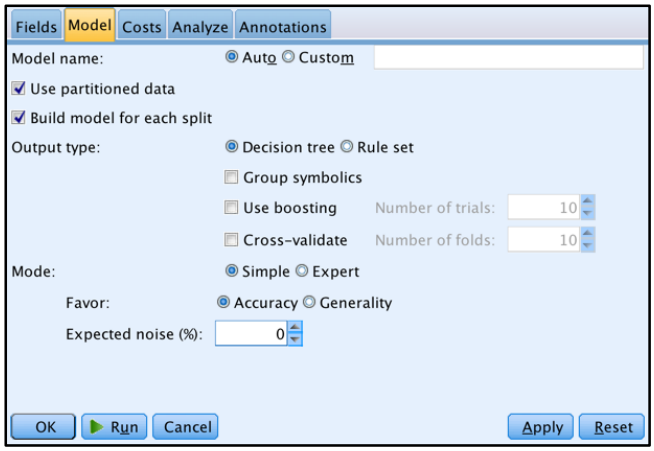
Do the retained predictors match what you might expect based on looking at the histograms in Q8?

Yes.

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.

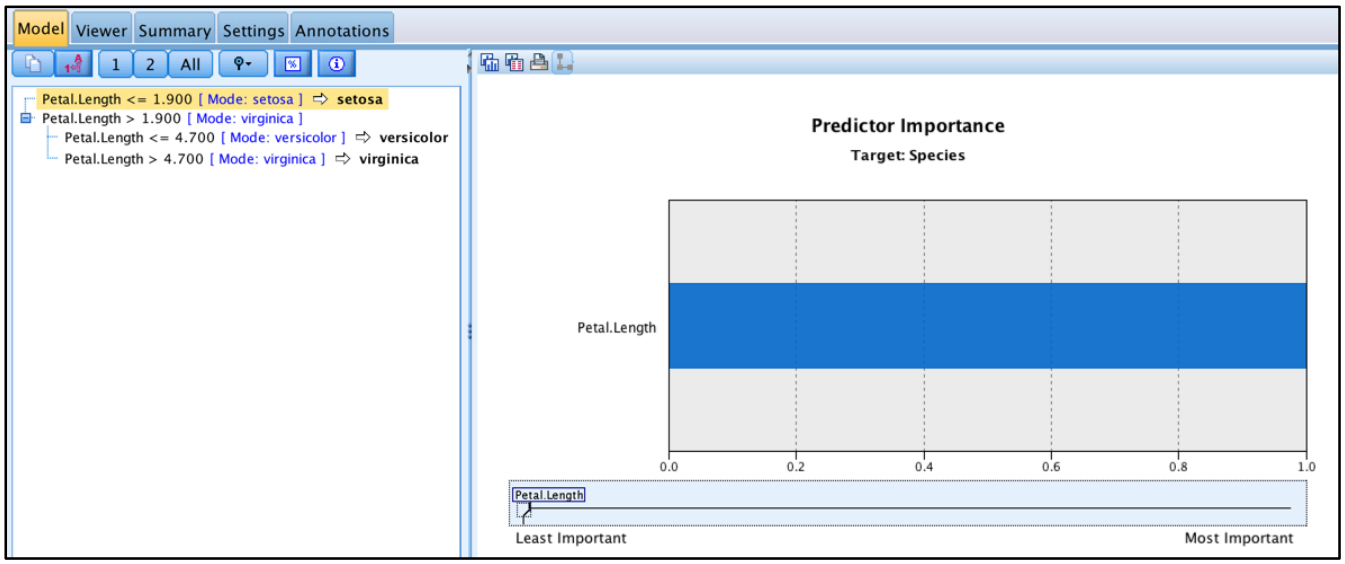


11. Add an analysis node after the golden nugget.



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.

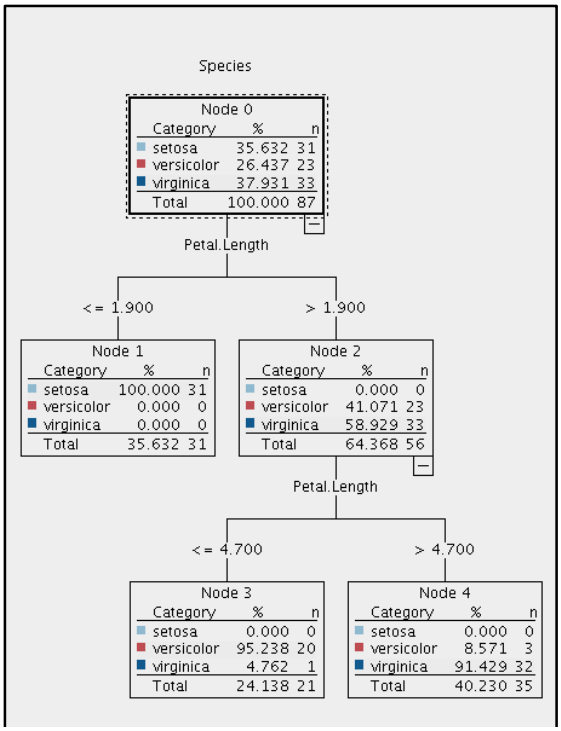


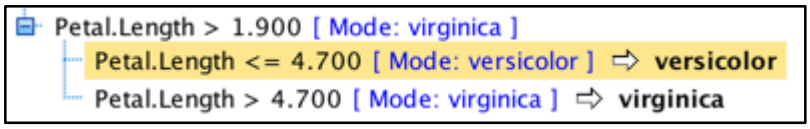
Q11 Do the Predictor Importance values match what you expected based on looking at the histograms in Q8?

No.

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.

Results for output field Species
Comparing $D-Species with Species

'Partition'	Testing		Training		Validation	
Correct	25	100%	84	96.55%	38	100%
Wrong	0	0%	3	3.45%	0	0%
Total	25		87		38	

Discriminant

Results for output field Species
Comparing $L-Species with Species

'Partition'	Testing		Training		Validation	
Correct	25	100%	83	95.4%	37	97.37%
Wrong	0	0%	4	4.6%	1	2.63%
Total	25		87		38	

Logistic

Results for output field Species
Comparing $C-Species with Species

'Partition'	Testing		Training		Validation	
Correct	24	96%	83	95.4%	36	94.74%
Wrong	1	4%	4	4.6%	2	5.26%
Total	25		87		38	

C 5.0

Q12 Rank the 3 models with how well they did with respect to classification accuracy in the *validation set*.

Discriminant > Logistic > C 5.0

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: