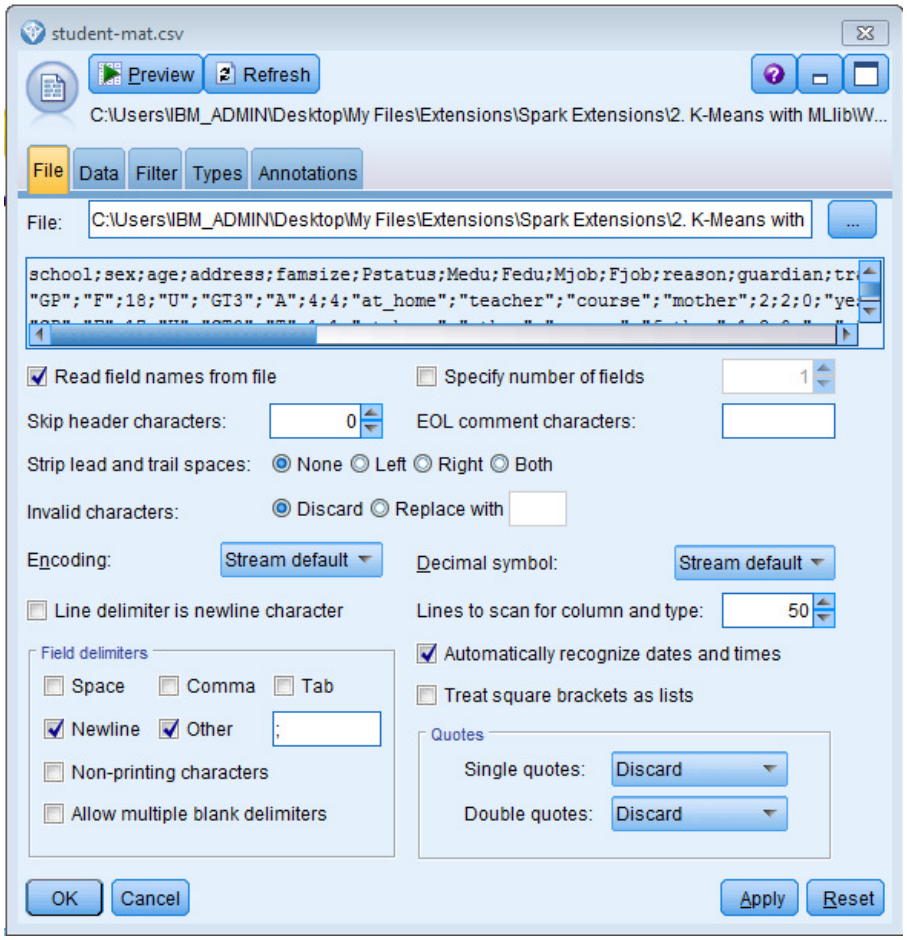
Step by Step Example – K-Means Clustering

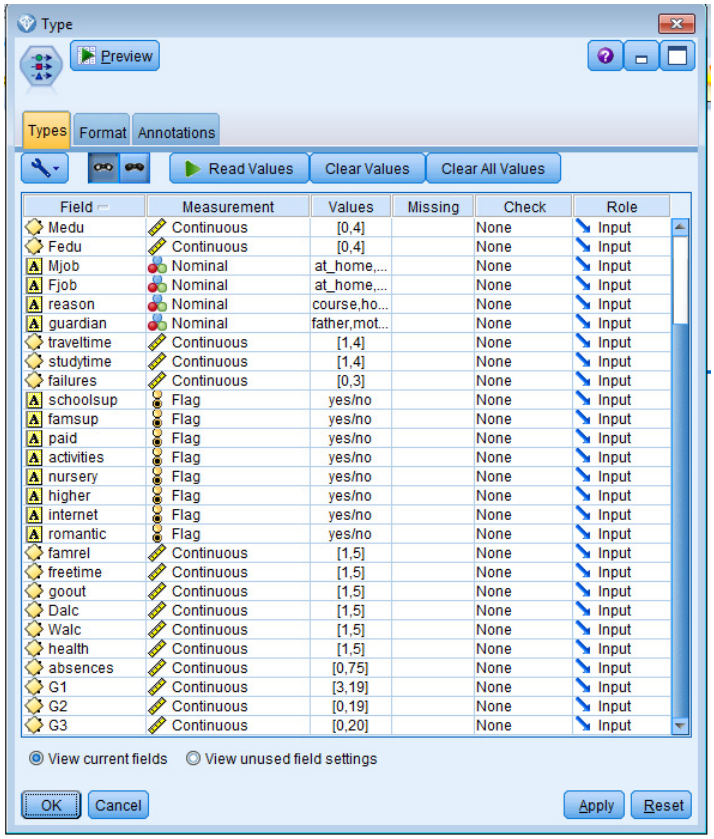
For this example, we will cluster students on a number of variables. The dataset used for this example can be found at the UCI Machine Learning Repository [1]. This dataset can be found at http://archive.ics.uci.edu/ml/datasets/Student+Performance. The data was originally used by Paulo Cortez [2].

This is a toy example to demonstrate how this extension can be used locally to use K-Means Clustering for smaller datasets, as well as in a Spark environment using Analytic Server.

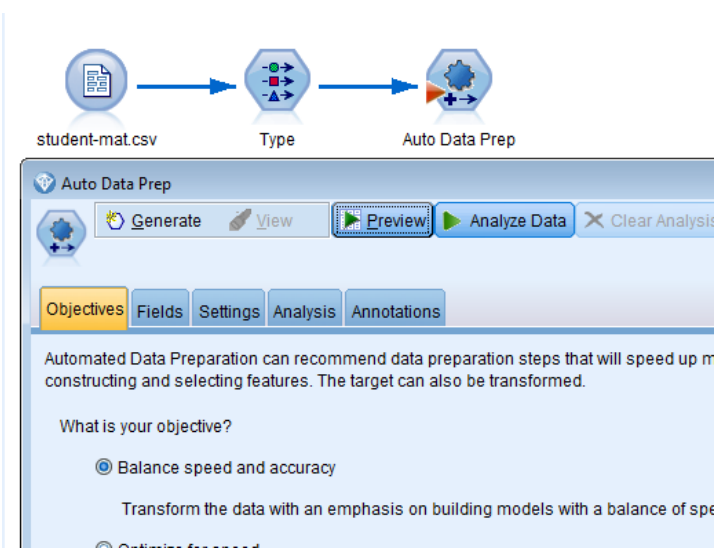
1. Download data and open in Modeler using a Var File node. This file is ';' (semi-colon) separated, so you will need to mark other for type of delimiter and type in a semi-colon.



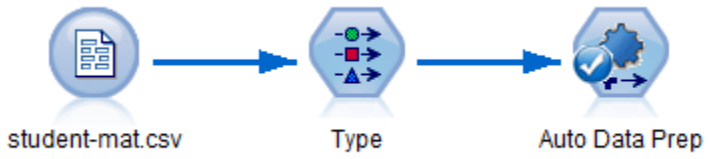
2. Next we can add a Type node, to read values to make sure all the meta-data is correct.



3. Before we cluster our data, we should do some pre-processing to make sure our continuous variables are on the same scale. This is easily done with the Auto Data Prep node in Modeler
    a. Add the Auto Data Prep node from the Field Ops palette
    b. Open the node by double clicking and click the button to "Analyze Data"

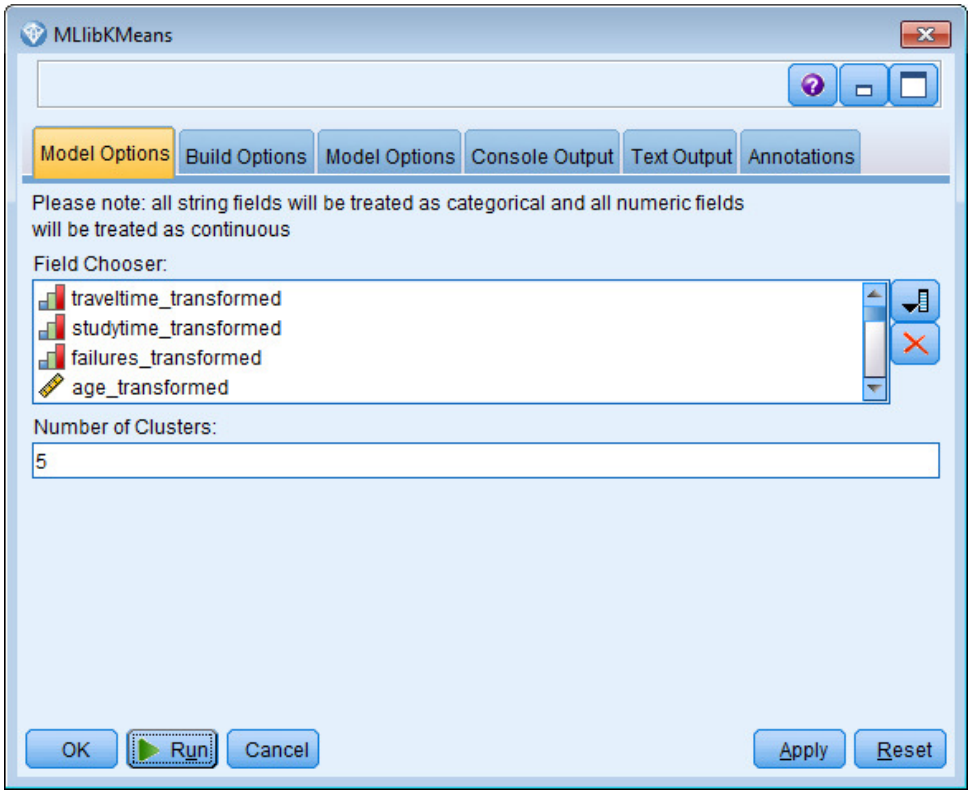


c. This will create dummy variables for categorical variables and will standardize numeric variables
d. You will know this step is complete because a check mark will appear on the node

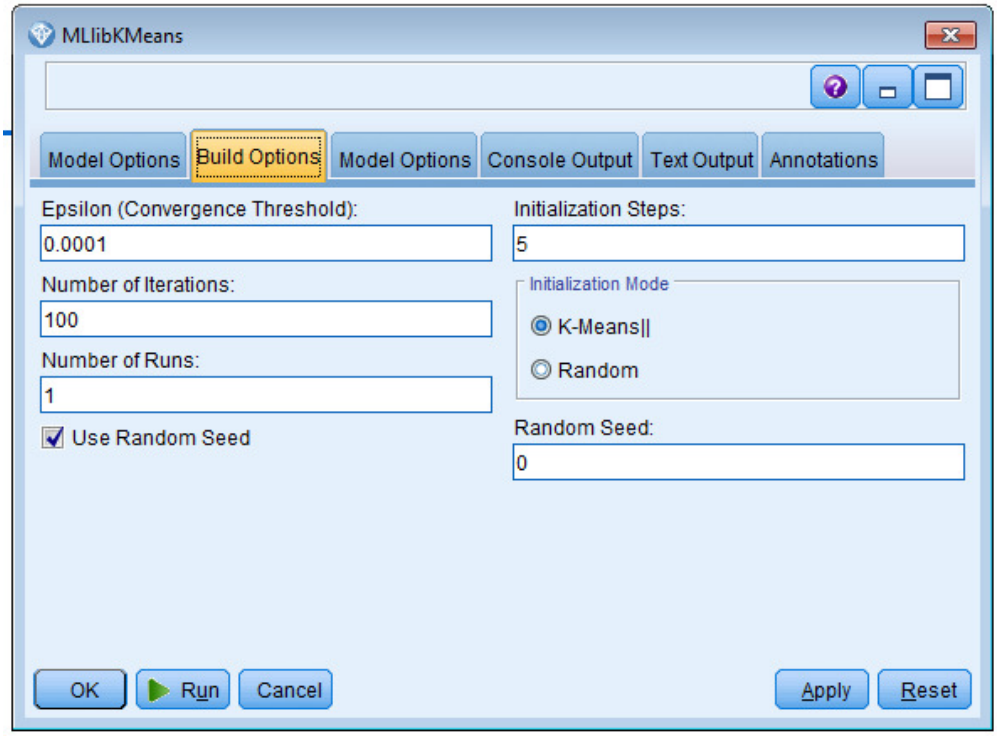


4. Now, let's add our K-Means node so our stream will look like this

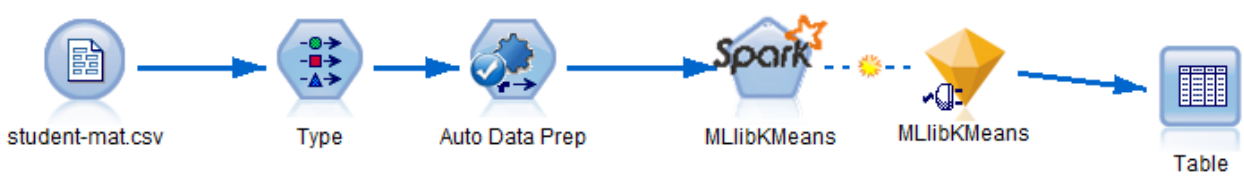
5. Open the MLlib K-Means dialog
   a. Add all the Fields
   b. Since we do not have more knowledge of the dataset , let's use the default 5 clusters



c. Let's also use the default options for the algorithm; these can be changed on the Build Options tab.



6. Run this stream to create the model.
7. Now add a table node connected to the model nugget from the Output palette
8. Right click the table node and run to get the following stream:



9. The table produced will include the cluster assignment for each student:

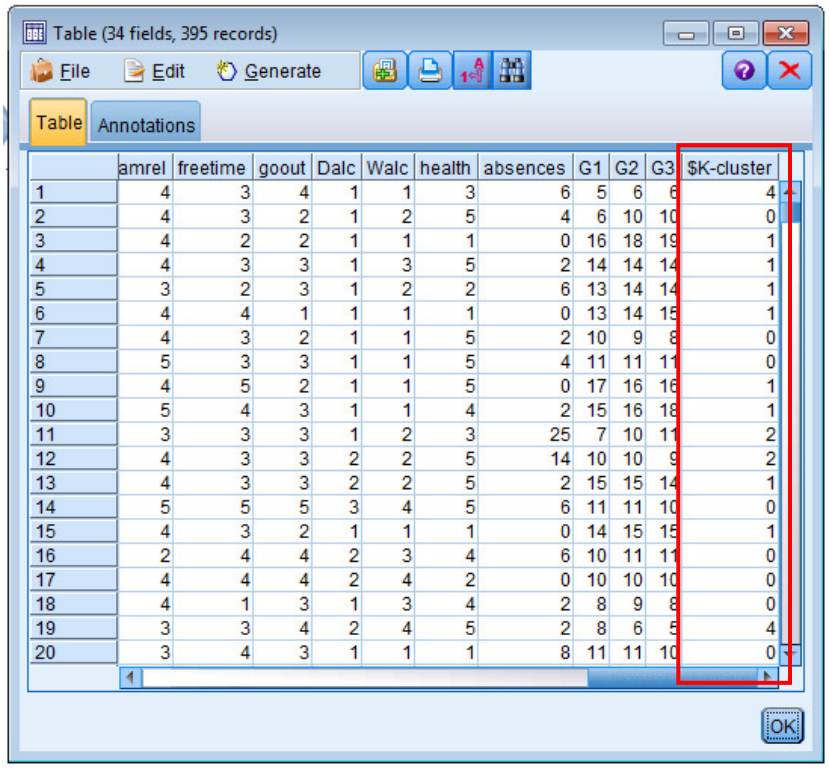


Table (34 fields, 395 records)

File Edit Generate

Table Annotations

	amrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3	$K-cluster
1	4	3	4	1	1	3	6	5	6	6	4
2	4	3	2	1	2	5	4	6	10	10	0
3	4	2	2	1	1	1	0	16	18	19	1
4	4	3	3	1	3	5	2	14	14	14	1
5	3	2	3	1	2	2	6	13	14	14	1
6	4	4	1	1	1	1	0	13	14	15	1
7	4	3	2	1	1	5	2	10	9	8	0
8	5	3	3	1	1	5	4	11	11	11	0
9	4	5	2	1	1	5	0	17	16	16	1
10	5	4	3	1	1	4	2	15	16	18	1
11	3	3	3	1	2	3	25	7	10	11	2
12	4	3	3	2	2	5	14	10	10	9	2
13	4	3	3	2	2	5	2	15	15	14	1
14	5	5	5	3	4	5	6	11	11	10	0
15	4	3	2	1	1	1	0	14	15	15	1
16	2	4	4	2	3	4	6	10	11	11	0
17	4	4	4	2	4	2	0	10	10	10	0
18	4	1	3	1	3	4	2	8	9	8	0
19	3	3	4	2	4	5	2	8	6	5	4
20	3	4	3	1	1	1	8	11	11	10	0

OK

10. This clustering assignment can now be used for other steps in our Modeler workflow.

References:

[1] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.