Hands-on Exercise: Decision Tree with CreditSet Data

Loan Default Prediction

In this exercise, you will work with an excel data set that consist of information about individuals who have applied for a loan. The information included are client identification, income, age, loan amount they have applied for, and whether or not they have defaulted with a loan payment in the last 10 years. The process you go through and your final output is supposed to help you create a model that would be able to identify individuals who are most likely to pose a credit risk to the loan company.

We will use the RapidMiner Data mining software. **Make sure you install the correct version of RpaidMiner Studio on your laptop. Choose either a 32 bit or a 64 bit based on the specifications of the operating system on your computer. Include as many screenshot (of relevant steps) as possible.**

Particularly, you would use the Decision tree algorithm in the RapidMiner Machin learning software to build a model.

	A	B	C	D	E	F
1	clientid	income	age	loan	LTI	default_tenYear
2	1	66155.93	59.01702	8106.532	0.122537	0
3	2	34415.15	48.11715	6564.745	0.190752	0
4	3	57317.17	63.10805	8020.953	0.13994	0
5	4	42709.53	45.75197	6103.642	0.142911	0
6	5	66952.69	18.58434	8770.099	0.13099	1
7	6	24904.06	57.47161	15.4986	0.000622	0
8	7	48430.36	26.80913	5722.582	0.118161	0
9	8	24500.14	32.89755	2971.003	0.121265	1
10	9	40654.89	55.49685	4755.825	0.11698	0
11	10	25075.87	39.77638	1409.23	0.056199	0
12	11	64131.42	25.67958	4351.029	0.067846	0
13	12	59436.85	60.47194	9254.245	0.155699	0
14	13	61050.35	26.35504	5893.265	0.096531	0

Figure 1: Sample of Dataset

Exercise

The following are the main steps you would follow in building the Decision Tree (DT) model. (NB: check the DT videos uploaded on PILOT for more information and details for building a DT model).

Some steps have accompanying questions. Make sure your report includes a discussion of those questions as well. Include relevant snapshots and accompanying narratives in your report.

1. Create a new process in RapidMiner as shown in figure. Save your work at this point and remember to do it often throughout the process.

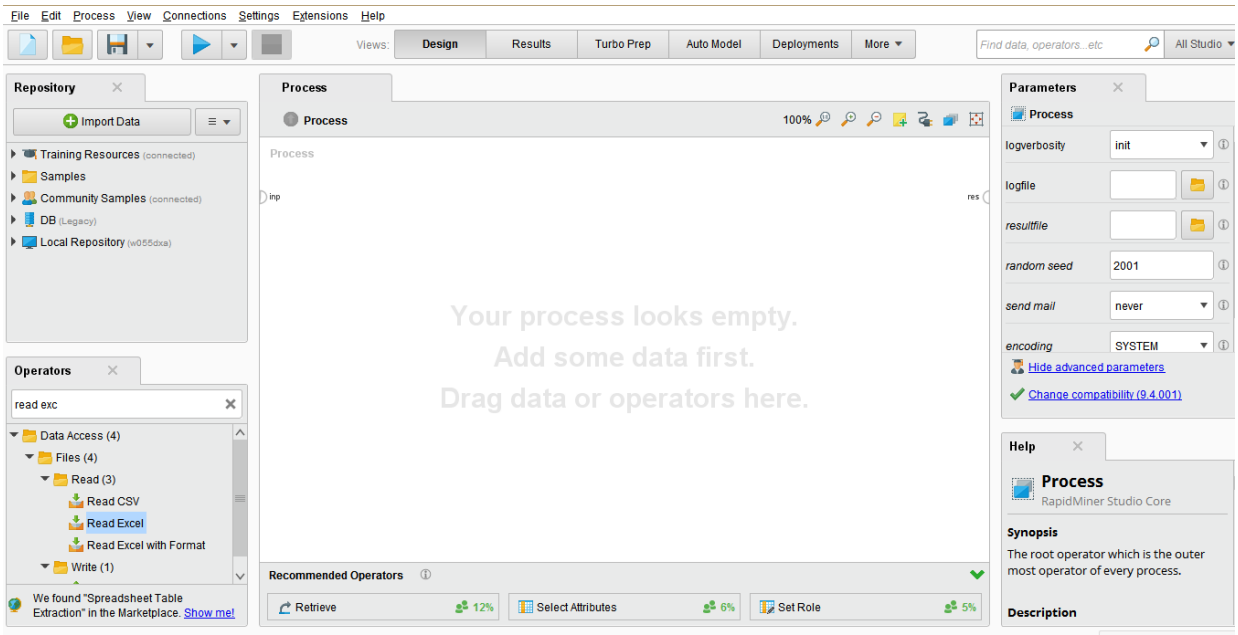

Figure 2: New Process

2. Import the excel Creditset data set to be used for building the model. Perform this task with a ***Read Excel*** operator. Type Read Excel into the Operator search box in the top-left corner. Drag the operator into the process area.

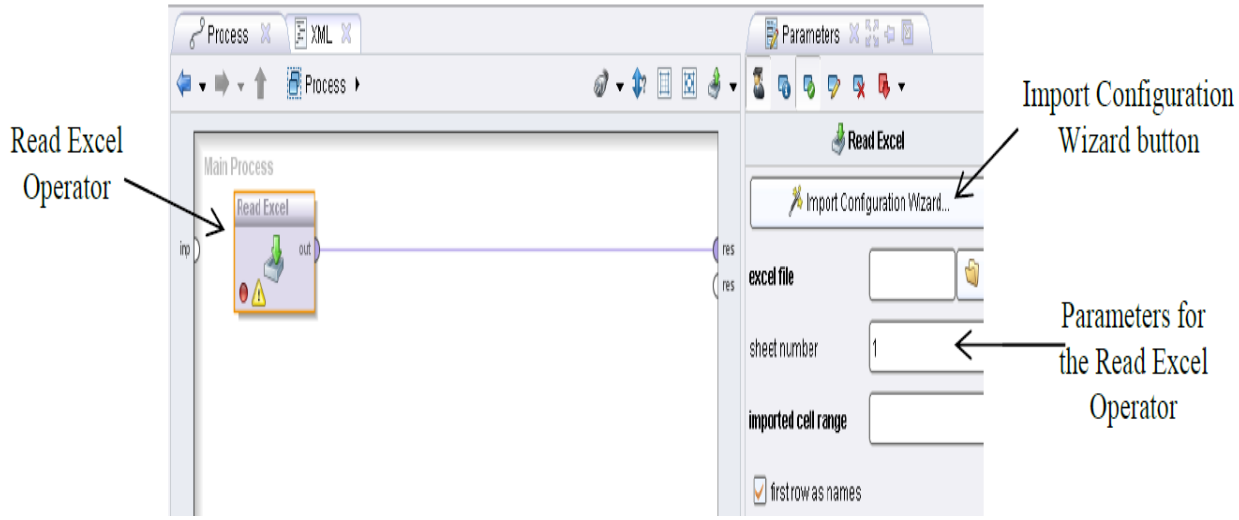

Figure 3: Read Excel operator

3. In the parameters area, click on "***Import Configuration Wizard***" button and navigate to the location of the CreditSet data on your computer. Follow the following steps to load the data correctly.

   3.1 Cells selected. Click on Next.

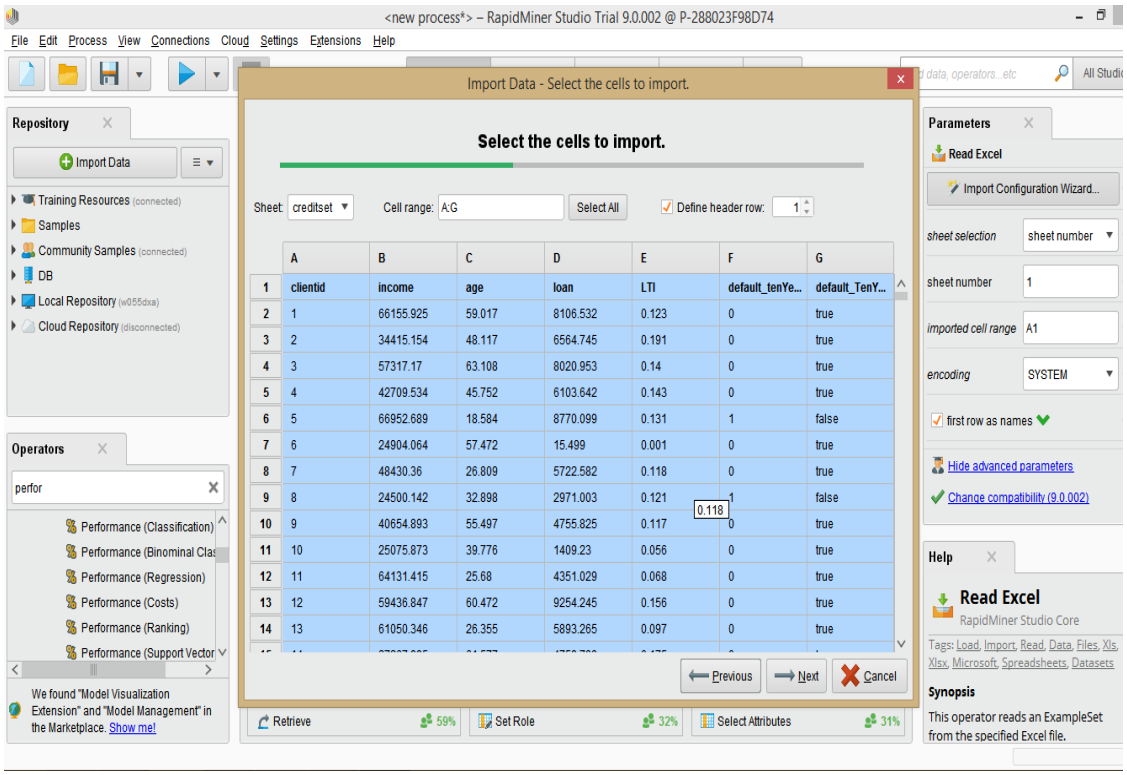

Figure 4.1: Variable selection

3.2 Click on drop down arrow at the top right of default_TenYear and choose “change Role”

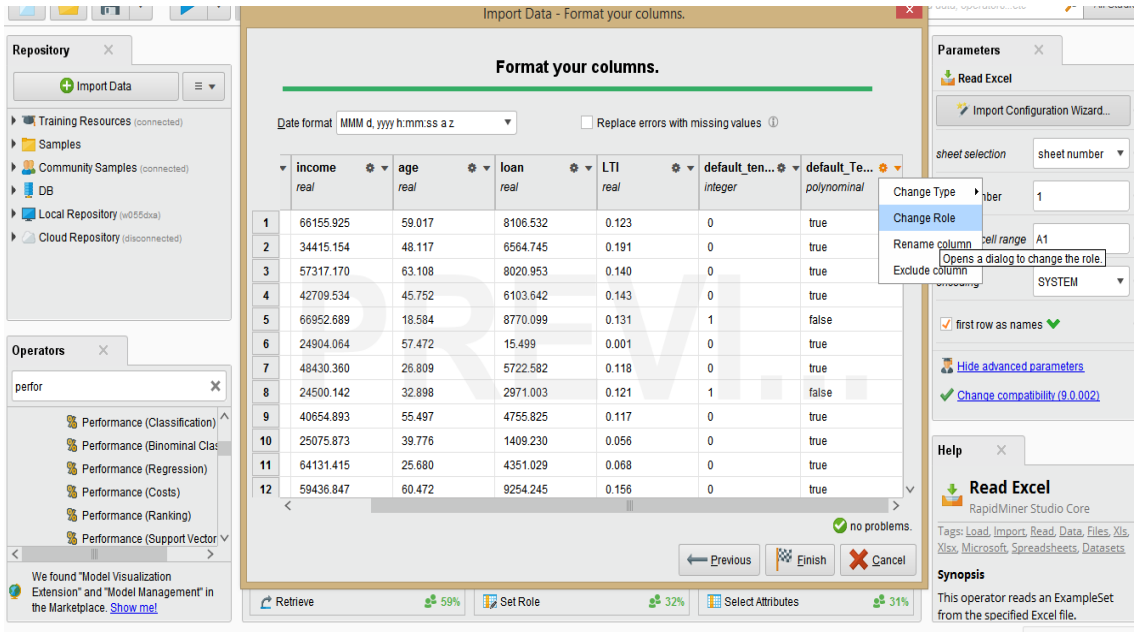

Figure 4.2: Change variable parameters

3.3 In the next screen, change role of default_TenYear to label

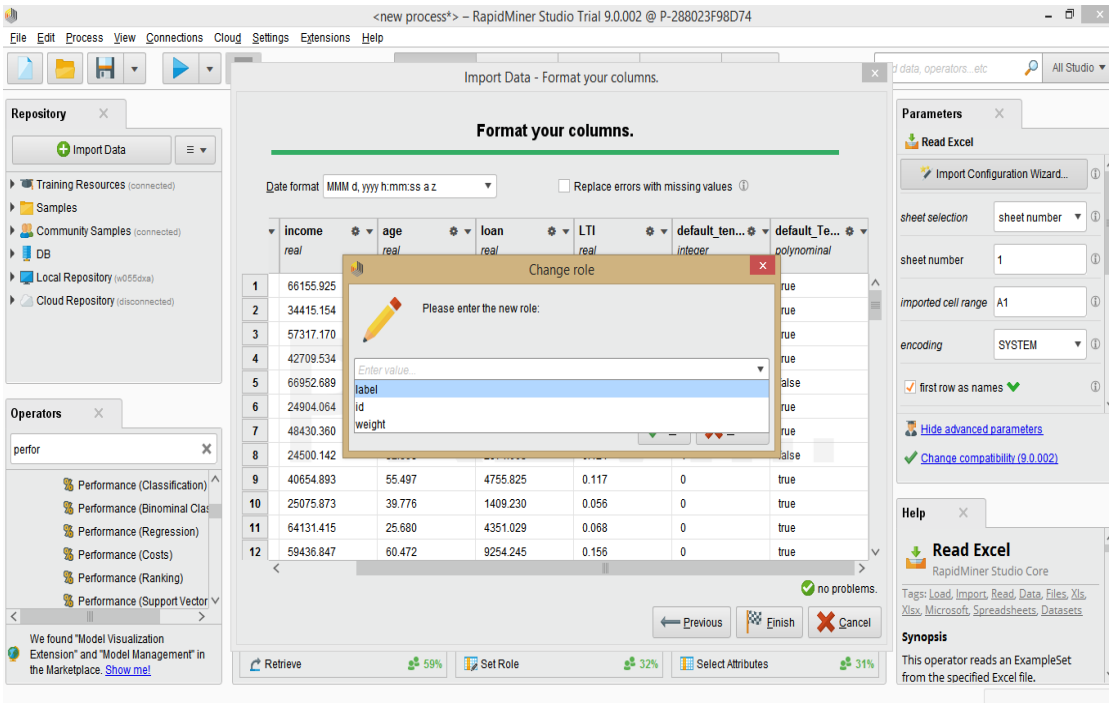

Figure 4.3: Change variable parameters

3.4 For subsequent steps, exclude default_tenYear, clientid and LTI attributes.

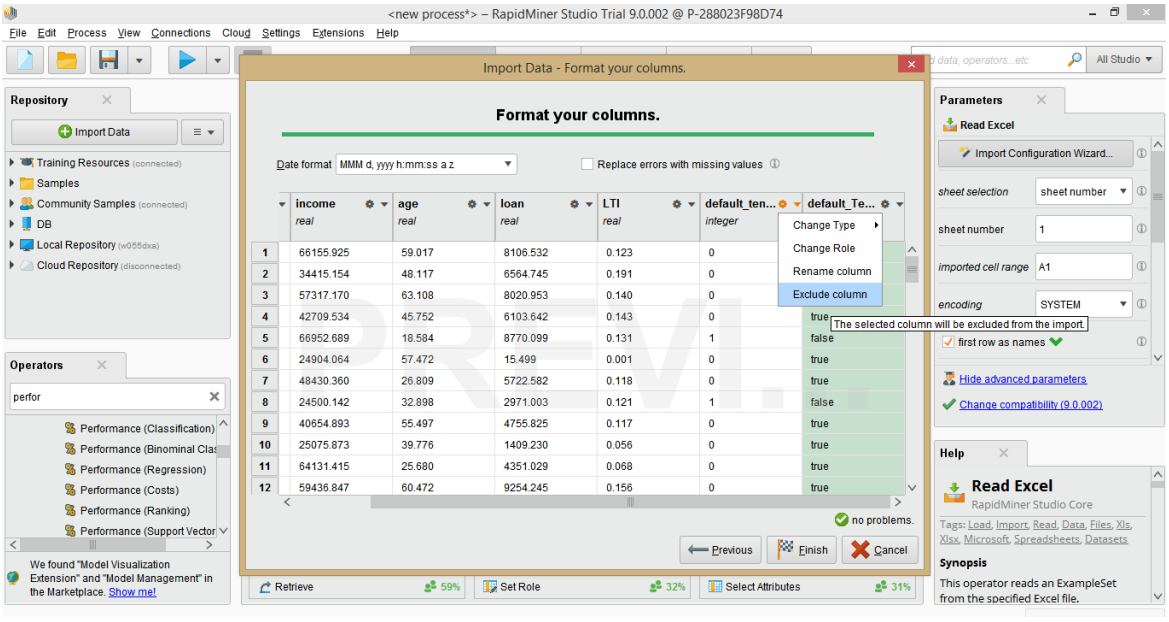

Figure 4.4: Change variable parameters

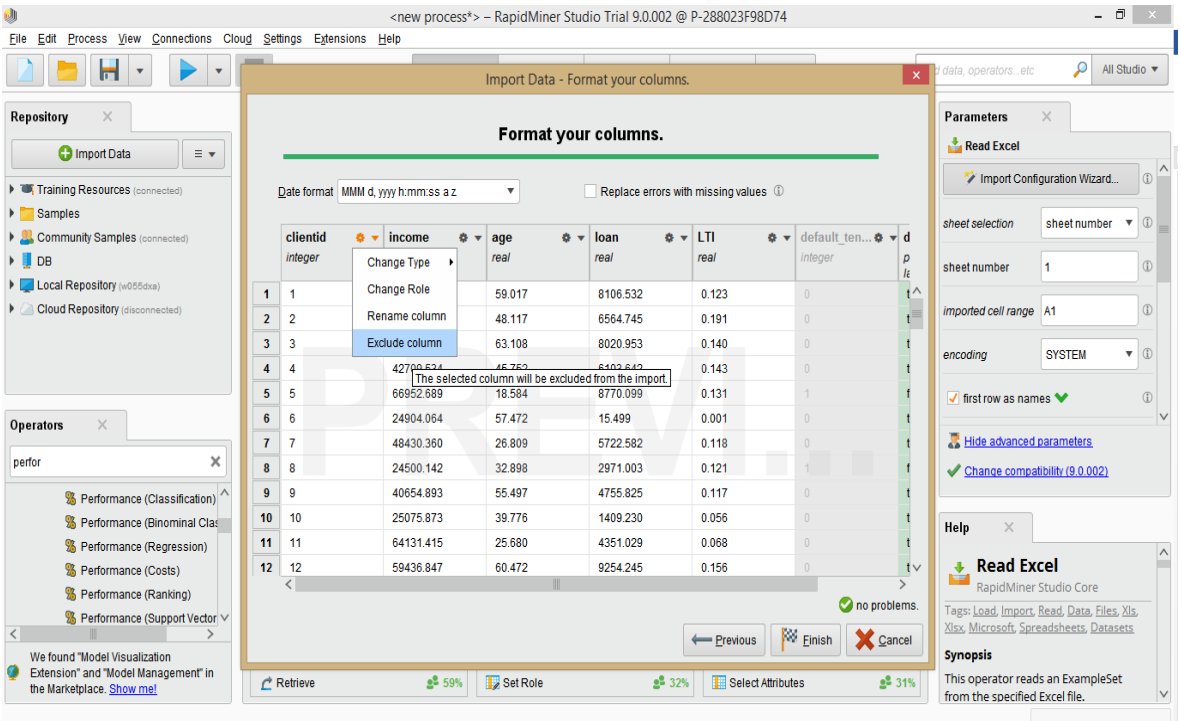

Figure 4.5: Change variable parameters

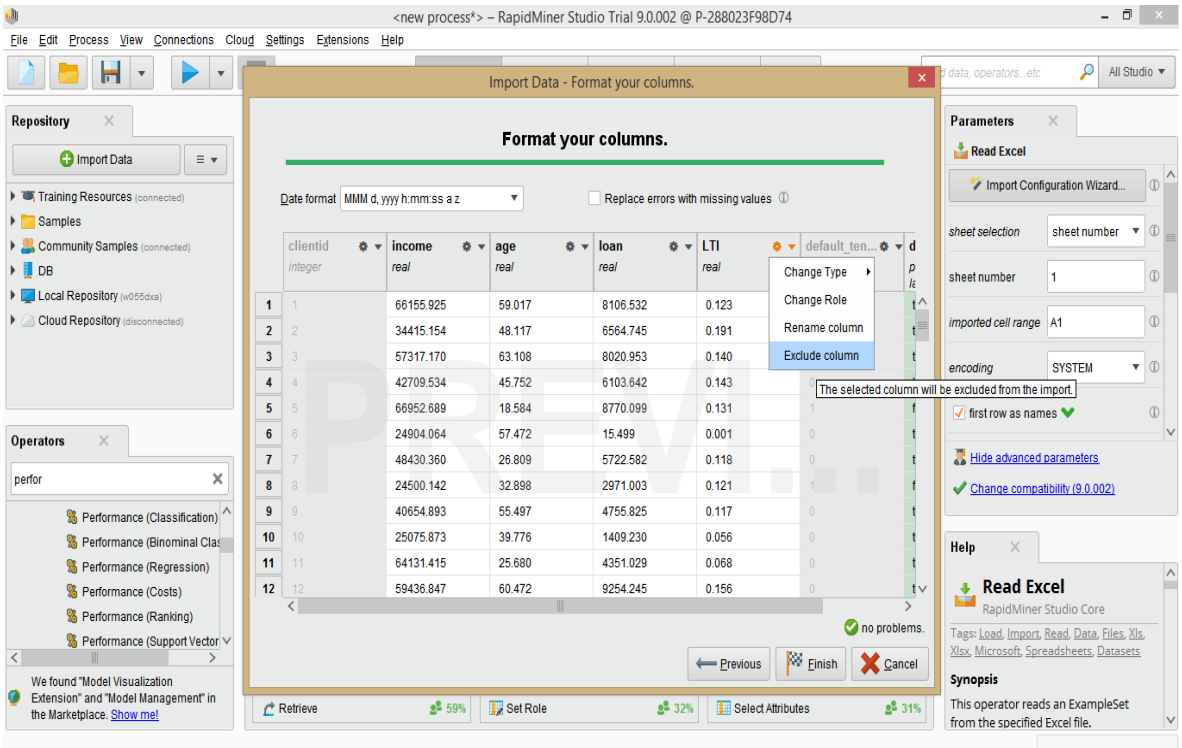

Figure 4.6: Change variable parameters

4. Search and drag the **Split Validation** operator into the process area. Use a relative split and a split ratio of 0.7 in the parameters areas. Explain the meaning of **Split Ratio**. Discuss how it impacts data validation in data mining? Connect the operators together as shown in figure 5.

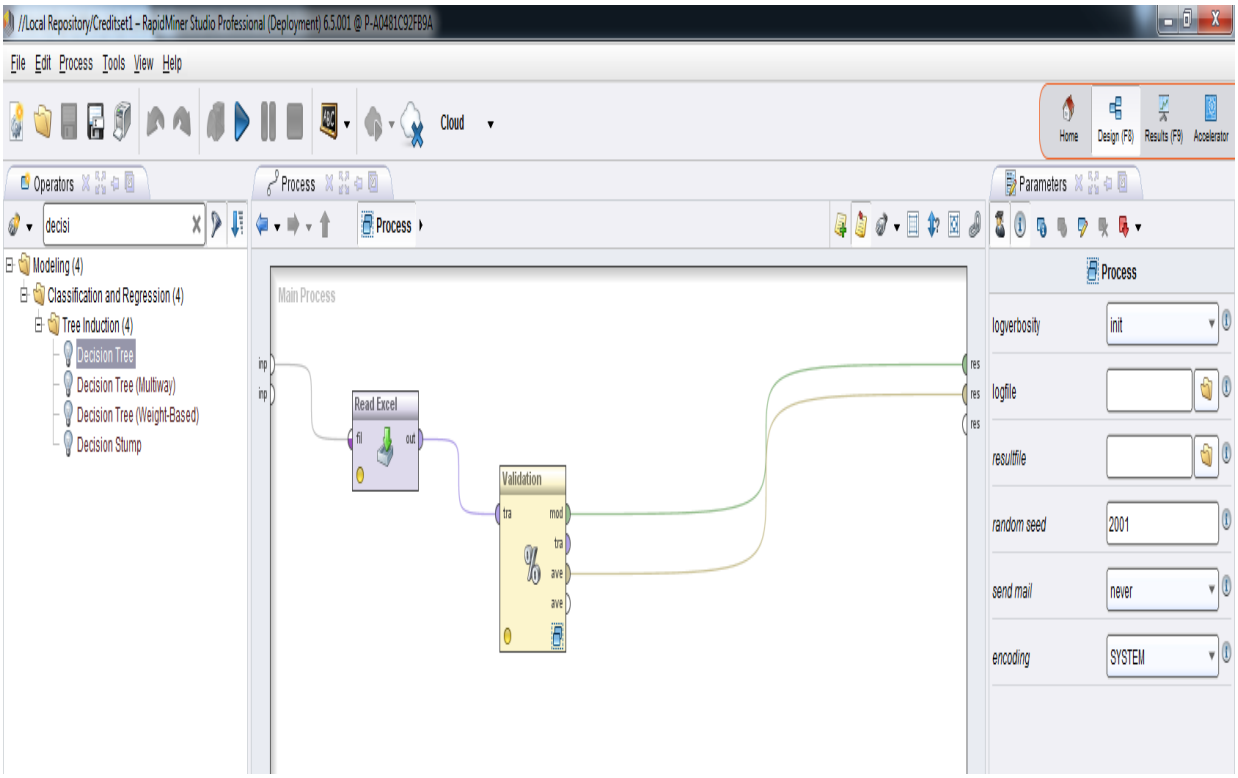

Figure 5: Validation operator

5. Double-click to open the Split Validation nested operator. Two sub-processes are presented, a Training process and a Testing process. At this point, you will build your model in the Training process and test it in the Testing process.
6. At this point, you are ready to build your DT model. Search and drag the Decision Tree, Apply Model and Performance operators into the process areas. Connect the operators as shown figure 6. In the DT operator, choose a maximal depth of 10 and gain_ratio criterion. Leave all other DT parameters default. Mention and explain 1 other criterion parameter.

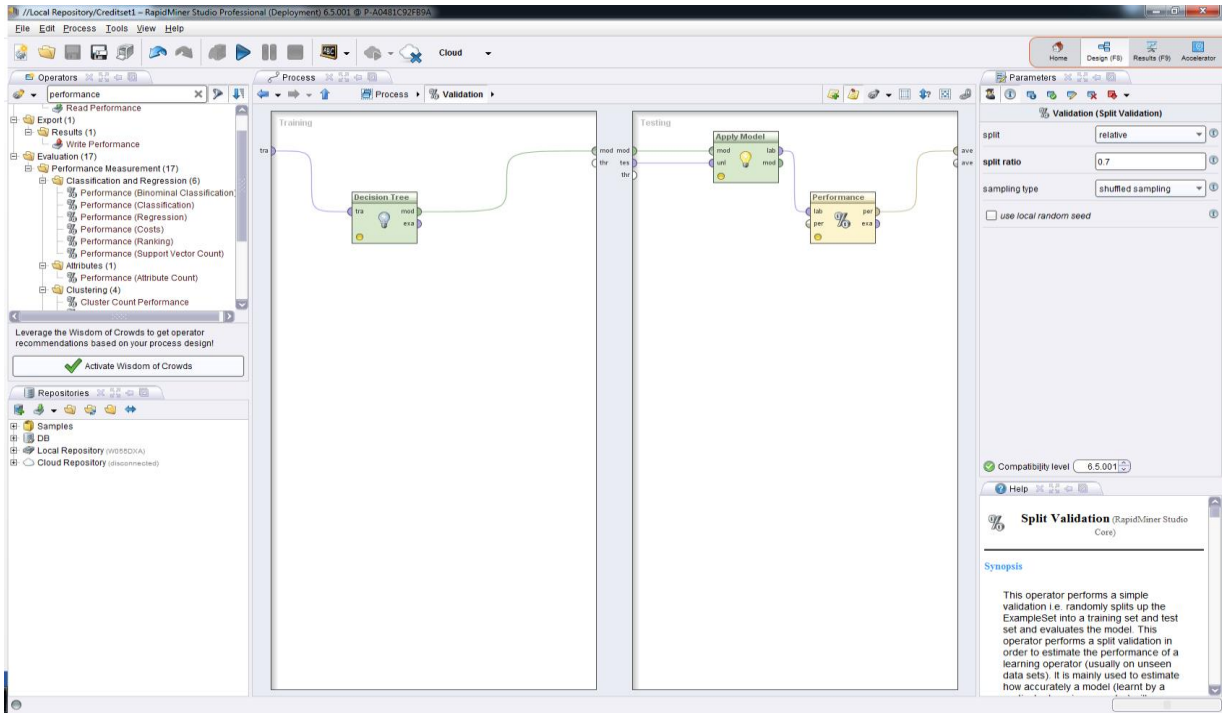

Figure 6: Validation operator sub-processes

7. Click on the Run button.
8. Include a snapshot of the **Accuracy criterion** (in the Performance tab) in your report. Explain the parameters and how it relates to prediction of individuals who are likely to default on their loan payments.
9. Include a snapshot of the **DT graph** and **Description** in your report. Explain the graph and how it relates to prediction of individuals who are likely to default on their loan payments.
10. Assume you would want to prune the Tree to a depth of 5. Edit the necessary parameter(s) in the DT operator and run the model again.
11. Include a snapshot of your new DT graph and explain how it relates to prediction of individuals who are likely to default on their loan payments.