# Translating Data into Information — an introduction to RapidMiner

### 8 May 2018

RapidMiner – Workshop

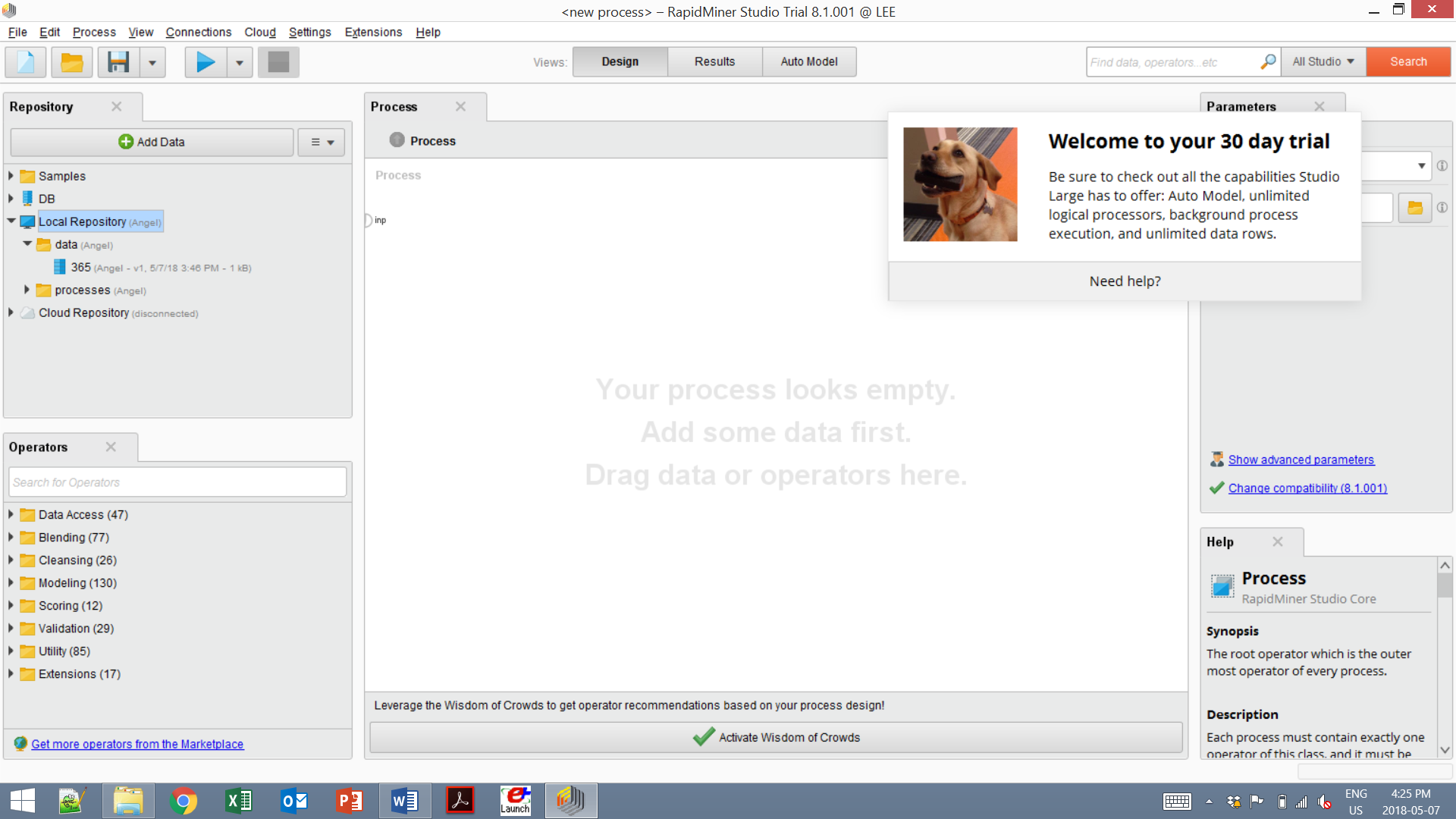
**Introduction**

In this training you will learn how to use the RapidMiner Studio 8.1 to execute data analysis.

**Exercise 1 – Import your database to Repository**

1. Download the excel file to your computer from the link.

<https://github.com/Brunoyylee/ESim2018_RapidMiner>

1. To import the downloaded data into RapidMiner, click in the Repository panel.
2. To store the excel file as **EX1** in your Local Repository, Folder “Data”.

What is **Repository?**

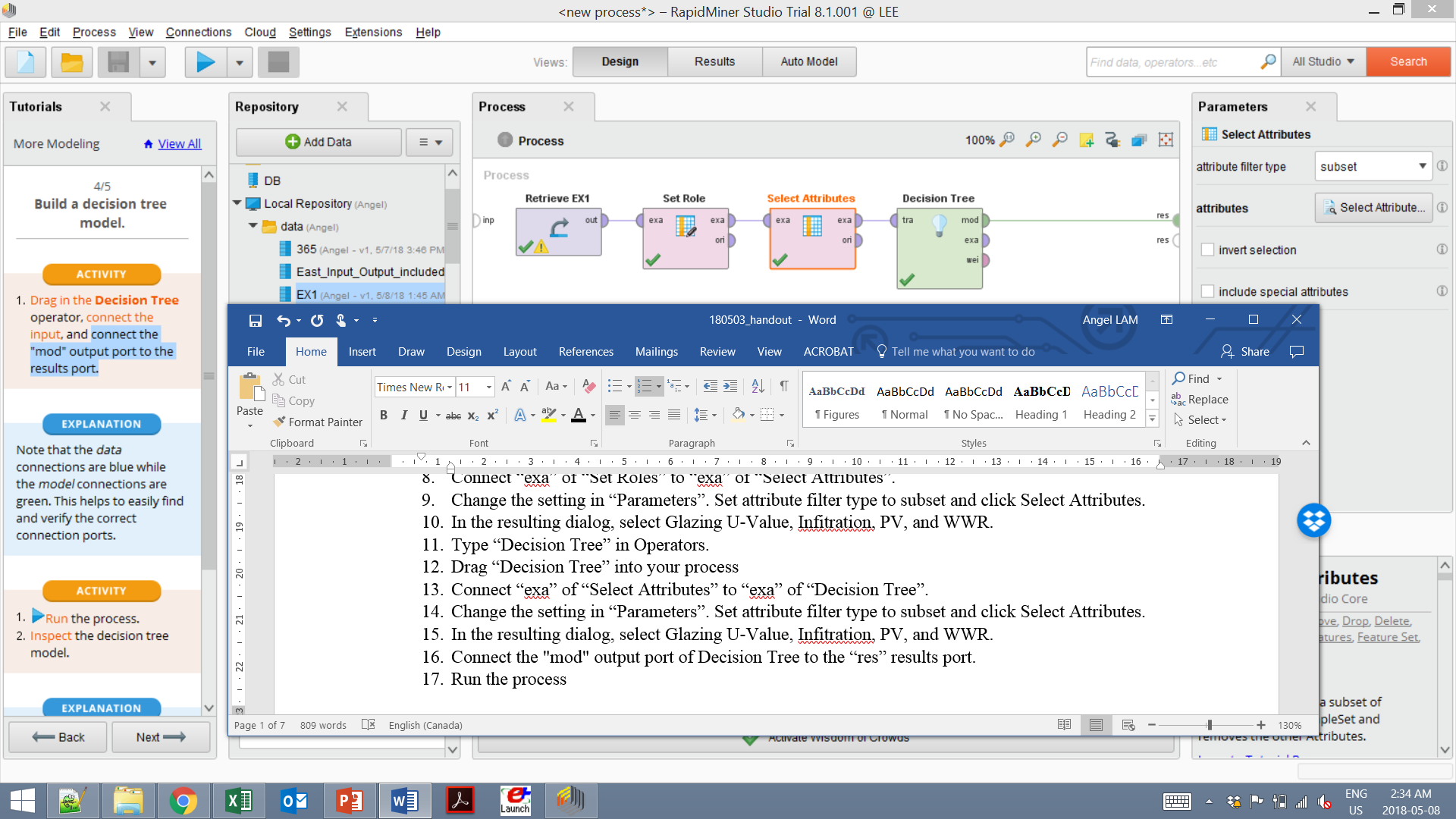
* Store all your data, processes, and results
* Simplify the design of analytical processes

**Exercise 2 – To build predictive models – Decision Tree**

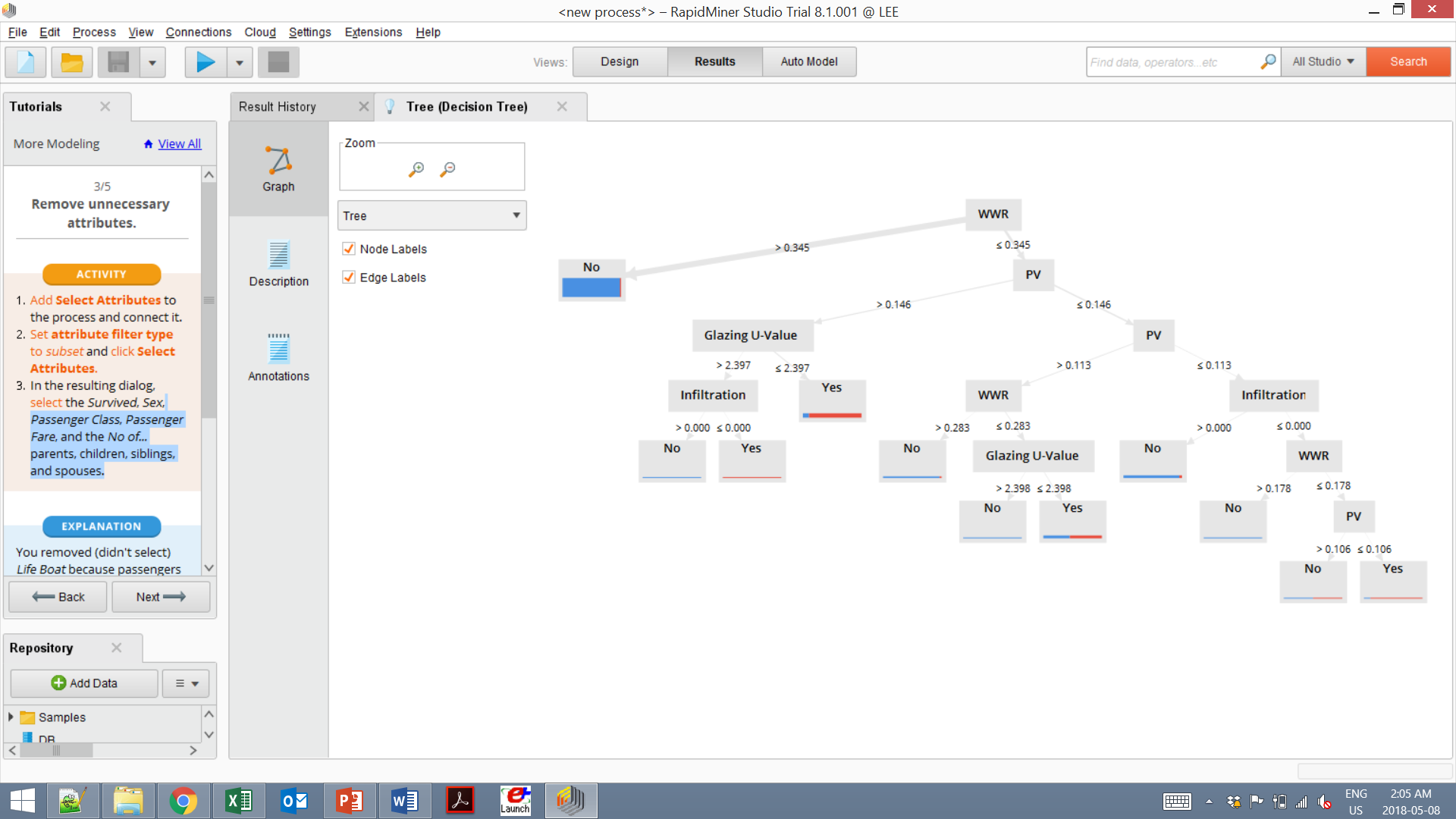
In this exercise, you will use decision tree to predict which building designs are net-zero buildings.

1. Drag the EX1 into your process.
2. Type “Set Role” in Operators.
3. Drag “Set Role” into your process
4. Connect “out” of “retrieve EXX1” to “exa” of “Set Roles”.
5. Change the setting in “Parameters”. Select “Net-Zero” in “attribute name” and “target role” into “label”

The attribute with role *label* is the one you want to predict. It is important to set the label, because there are machine learning methods, like the decision tree algorithm, that use existing data with known label values (a training set) to find hidden patterns. It then creates predictions from those patterns and applies them to new data without known labels (the testing set).

1. Type “Select Attributes” in Operators.
2. Drag “Select Attributes” into your process
3. Connect “exa” of “Set Roles” to “exa” of “Select Attributes”.
4. Change the setting in “Parameters”. Set attribute filter type to subset and click Select Attributes.
5. In the resulting dialog, select Glazing U-Value, Infitration, PV, and WWR.
6. Type “Decision Tree” in Operators.
7. Drag “Decision Tree” into your process
8. Connect “exa” of “Select Attributes” to “exa” of “Decision Tree”.
9. Change the setting in “Parameters”. Set attribute filter type to subset and click Select Attributes.
10. In the resulting dialog, select Glazing U-Value, Infitration, PV, and WWR.
11. Connect the "mod" output port of Decision Tree to the “res” results port.
12. Run  the process

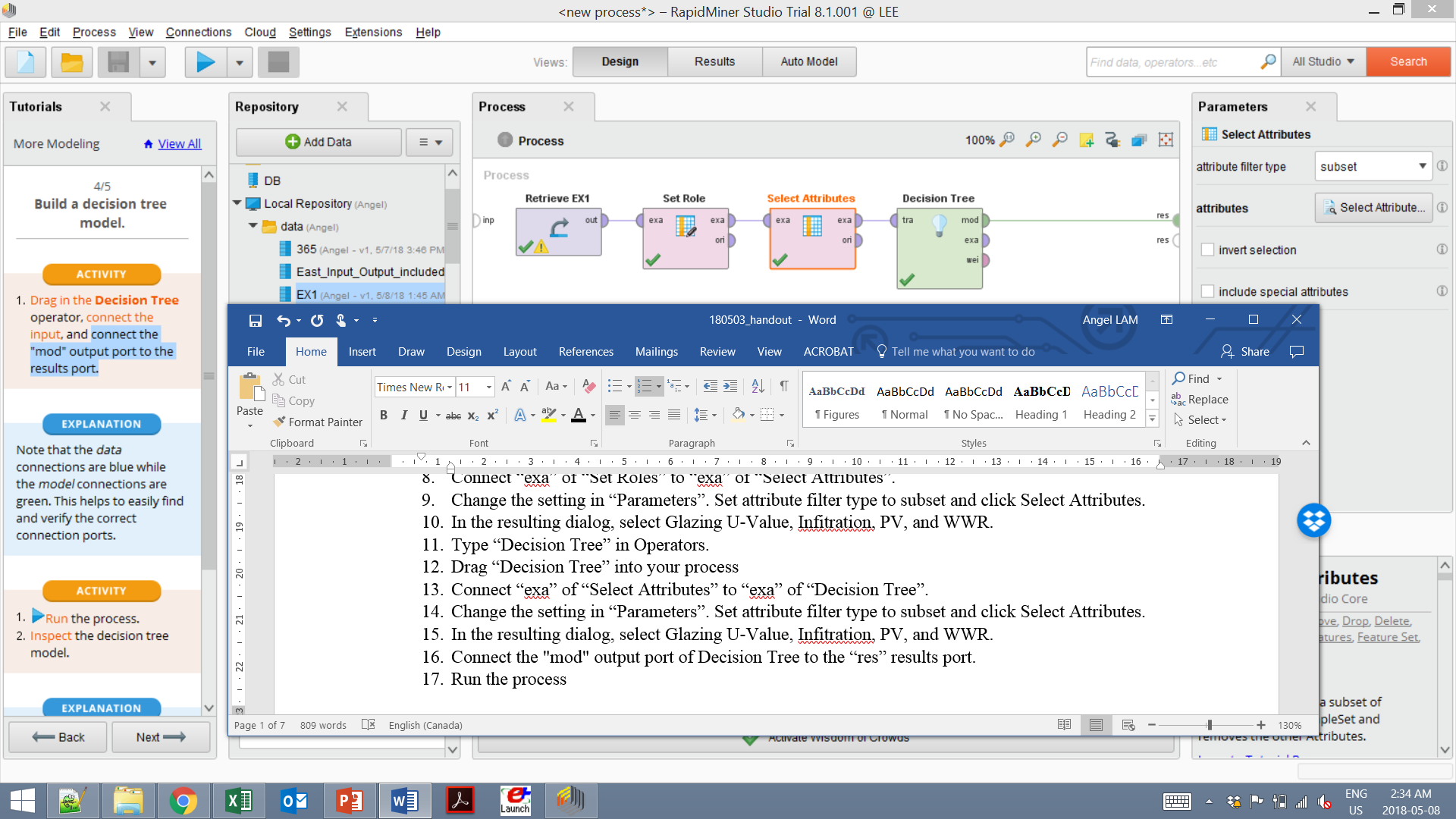
Result



What is **Predictive model**?

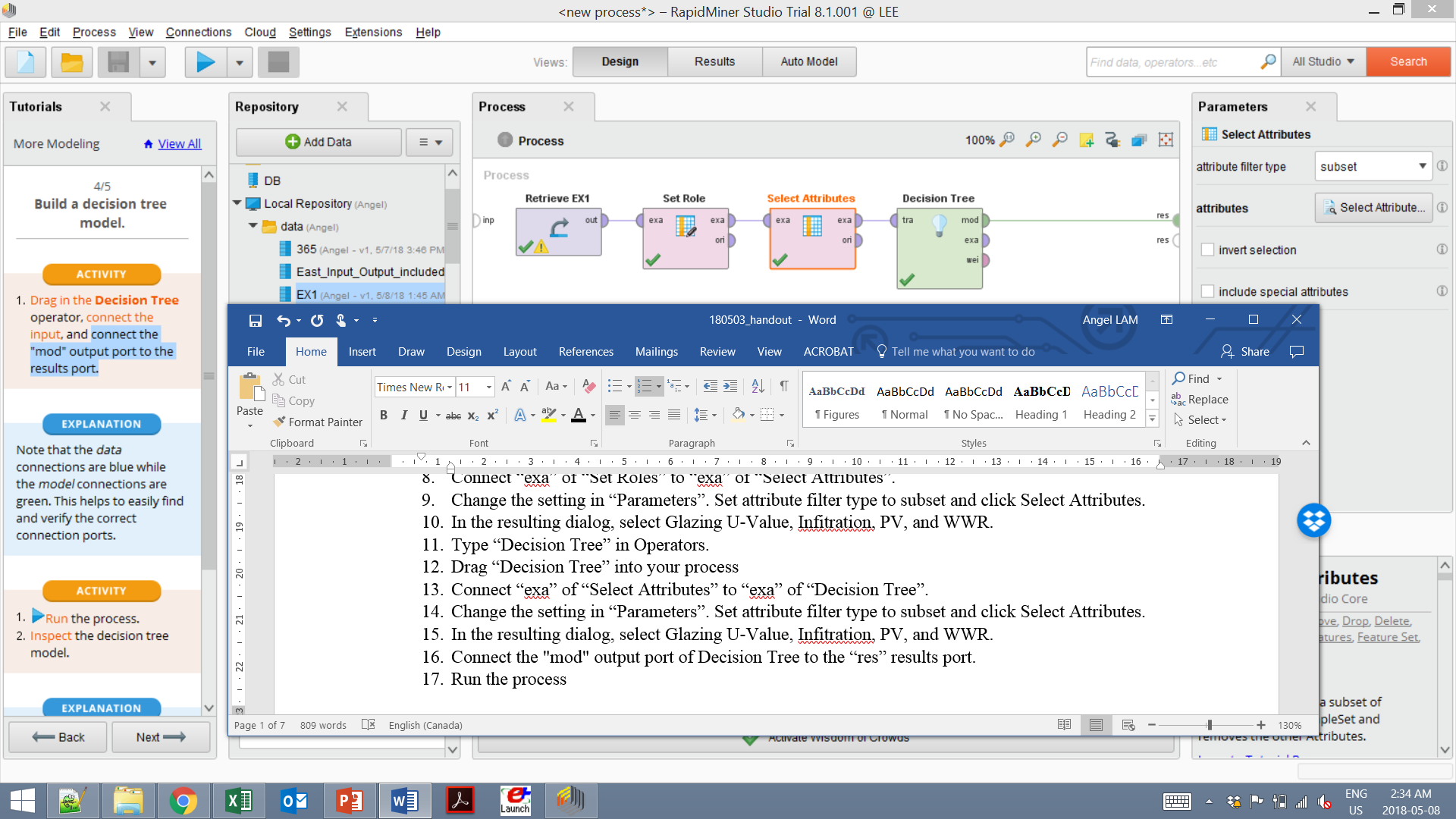
Predictive model is a set of machine learning techniques which search for patterns in big data sets and use those patterns to create predictions for new situations. Those predictions can be categorical (this is called classification learning) or numerical (regression learning).

**Exercise 3 – Data cleansing – Remove outliers**

1. Drag the EX2 into your process.
2. Type “Select Attributes” in Operators.
3. Drag “Select Attributes” into your process
4. Connect “out” of “EX2” to “exa” of “Select Attributes”.
5. Change the setting in “Parameters”. Set attribute filter type to subset and click Select Attributes.
6. In the resulting dialog, select Glazing U-Value, Net-Zero, PV, SHGC, and WWR.
7. Type “Detect Outlier (Distances)” in Operators.
8. Drag “Detect Outlier (Distances)” into your process
9. Connect “exa” of “Set Roles” to “exa” of “Select Attributes”.
10. Type “Filter Examples” in Operators.
11. Drag “Filter Examples” into your process
12. Connect “exa” of “Select Attributes” to “exa” of “Filter Examples””.
13. In its Parameters, add a new filter with *Outlier*, *equals*, and *false* as values
14. Run  the process

Data cleansing is to identify unusual cases and remove them from the data set. In some situations, the outliers are simply the result of an incorrect measurement and should be removed from the data set.

**Exercise 4 – To build rule model**

1. Drag the EX3 into your process.
2. Type “Set Role” in Operators.
3. Drag “Set Role” into your process
4. Connect “out” of “retrieve EX3” to “exa” of “Set Roles”.
5. Change the setting in “Parameters”. Select “Net Zero” in “attribute name” and “target role” into “label”
6. Type “Select Attributes” in Operators.
7. Drag “Select Attributes” into your process
8. Connect “exa” of “Set Roles” to “exa” of “Select Attributes”.
9. Change the setting in “Parameters”. Set attribute filter type to subset and click Select Attributes.
10. In the resulting dialog, select Glazing U-Value, Net-Zero, PV, and WWR.
11. Type “Rule Induction” in Operators.
12. Drag “Rule Induction” into your process
13. Connect “mod” output port of “Rule Inudction” to “res” results port.
14. Run  the process

Result

if WWR > 0.345 then No (219 / 3)

if PV > 0.146 then Yes (0 / 43)

if WWR > 0.162 and PV ≤ 0.136 then No (32 / 0)

if WWR ≤ 0.131 then Yes (0 / 5)

if Glazing U-Value > 1.857 then No (8 / 0)

if Glazing U-Value > 1.158 and WWR ≤ 0.302 then Yes (0 / 7)

else No (1 / 0)

correct: 315 out of 318 training examples.

**Interpretation**

if WWR > 0.345 then No (219 / 3)

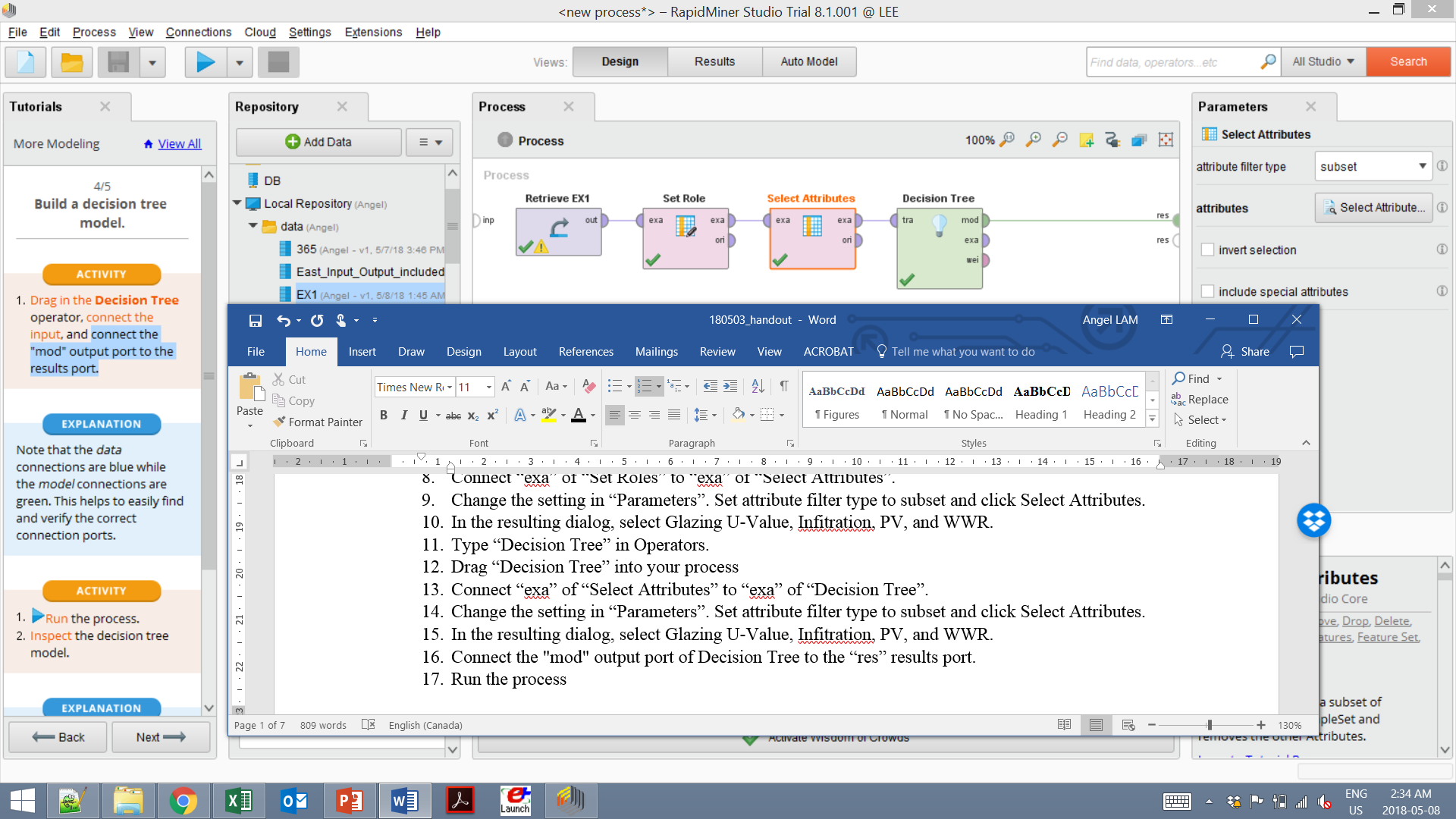
For WWR > 0.345, 219 cases are not net-zero designs, 3 cases are net-zero designs.

**Exercise 5 - To cluster**

In database EX1, the scattered plot between Glazing U-value and SHGC



In this exercise, you will learn how to separate the scattered plot into 10 clusters.

1. Drag the EX1 into your process.
2. Type “Select Attributes” in Operators.
3. Drag “Select Attributes” into your process
4. Connect “exa” of “Set Roles” to “exa” of “Select Attributes”.
5. Change the setting in “Parameters”. Set attribute filter type to subset and click Select Attributes.
6. In the resulting dialog, select Glazing U-Value and SHGC.
7. Drag “Clustering” into your process.
8. Change the setting in “Parameters”. Set k from 2 to 10
9. Connect “clu” output port of “Clustering” to “res” results port.
10. Run  the process

Result

Cluster 0: 2379 items

Cluster 1: 2537 items

Cluster 2: 2369 items

Cluster 3: 2618 items

Cluster 4: 2442 items

Cluster 5: 2459 items

Cluster 6: 2412 items

Cluster 7: 2524 items

Cluster 8: 2483 items

Cluster 9: 2353 items

Total number of items: 24576