DATA MINING LAB

Course Code: ISL74
Credit: **0:0:1:0**
Contact Hours: 14P

Note: The dataset considered should have at least 10 attributes and minimum of 50 records.

Course Content

Part – A (2 sessions)
Use of Rapidminer tool **Importing and Exporting data.** **Data Preprocessing** • Data Cleaning • Aggregation • Normalization • Sampling • Variable Selection **Modeling and evaluation** • Association mining using measures • Decision Tree Classification and evaluate them • Ensemble Method Classification and evaluate them • Clustering with evaluation
Part – B (8 sessions)
Perform the following tasks using R/Python programming 1. Read the dataset and perform data preprocessing on this dataset. 2. Visualize the data and datasets and identify anomalies and identify types of data preprocessing to be performed 3. Find useful patterns and associations using the Apriori approach. 4. Find useful patterns and associations using the Frequent Pattern Tree Approach. 5. Model Classifiers, evaluate performance and visualize the results. 6. Use Ensemble methods of classification to model a dataset, evaluate the performance and visualize the results. 7. Group the data in a dataset based on similarity by using partitioning methods. 8. Group the data in a dataset based on similarity by using Density Based methods.
Part-C(4 sessions)
Students will be assigned dataset based on an application. This dataset needs to be explored through suitable visualizations. Data in the data set is to be mined to produce essential

interpretations. The results obtained have to be evaluated for various metrics depending on the type of approach used.

Text Books:

1. Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining, Pearson Education, 2005.
2. Jiawei Han and MichelineKamber: Data Mining - Concepts and Techniques, 2nd Edition, Morgan Kaufmann Publisher, 2006.

Reference Books:

1. Arun K Pujari: Data Mining Techniques, 2nd Edition, Universities Press, 2009.
2. G. K. Gupta: Introduction to Data Mining with Case Studies, 3rd Edition, PHI, New Delhi, 2009.

Course Outcomes:

At the end of the course, students will be able to

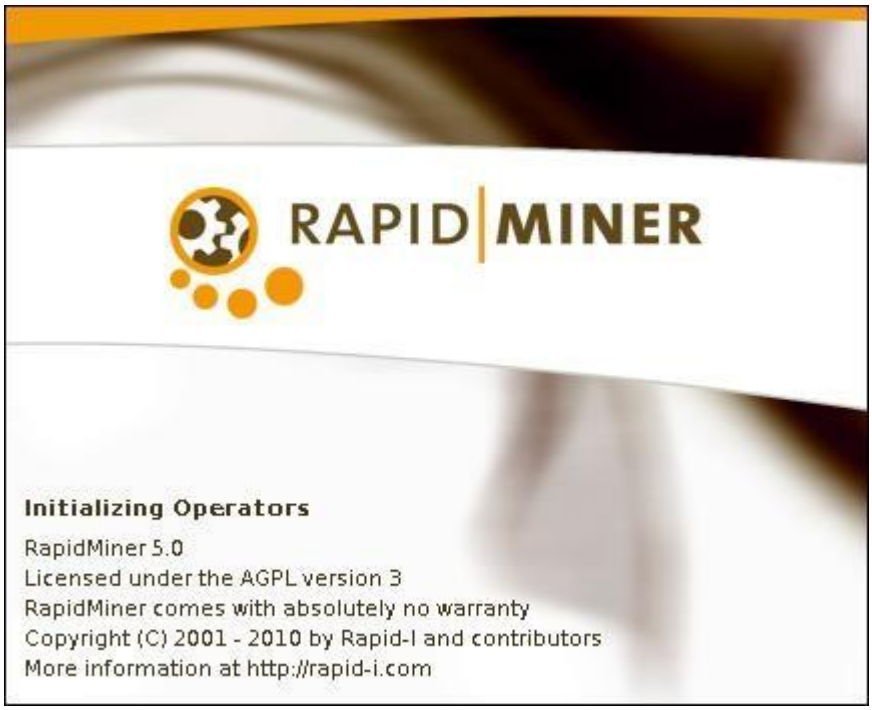
1.	Create, import and mine data using R-programming/Python (PO – 1,2,3,5,9) (PSO - 1,2)
2.	Import data, preprocess them and perform mining using data mining tools. (PO – 1,2,3,5,6,9,10,12) (PSO - 1,2)
3.	Design and develop preprocessing techniques , perform data mining tasks , analyze and evaluate the obtained result.(PO – 2,3,4,5,6,9,10,12) (PSO - 1,2,3)

Solution Manual

Introduction to Rapid Miner Tool:

Rapid Miner is a tool for experimenting with machine learning and data mining algorithms. An experiment is a set of operators that perform different tasks in the data input/output, data transformation, preprocessing, attribute selection, learning, and evaluation. The experiments can be described visually as a process.

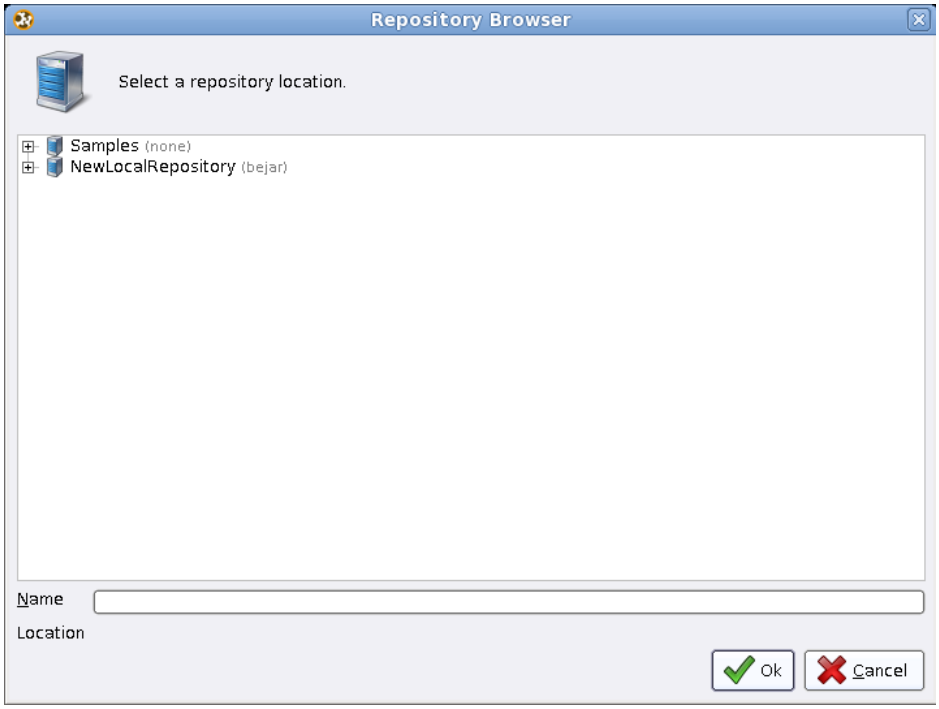
Rapid Miner - First contact



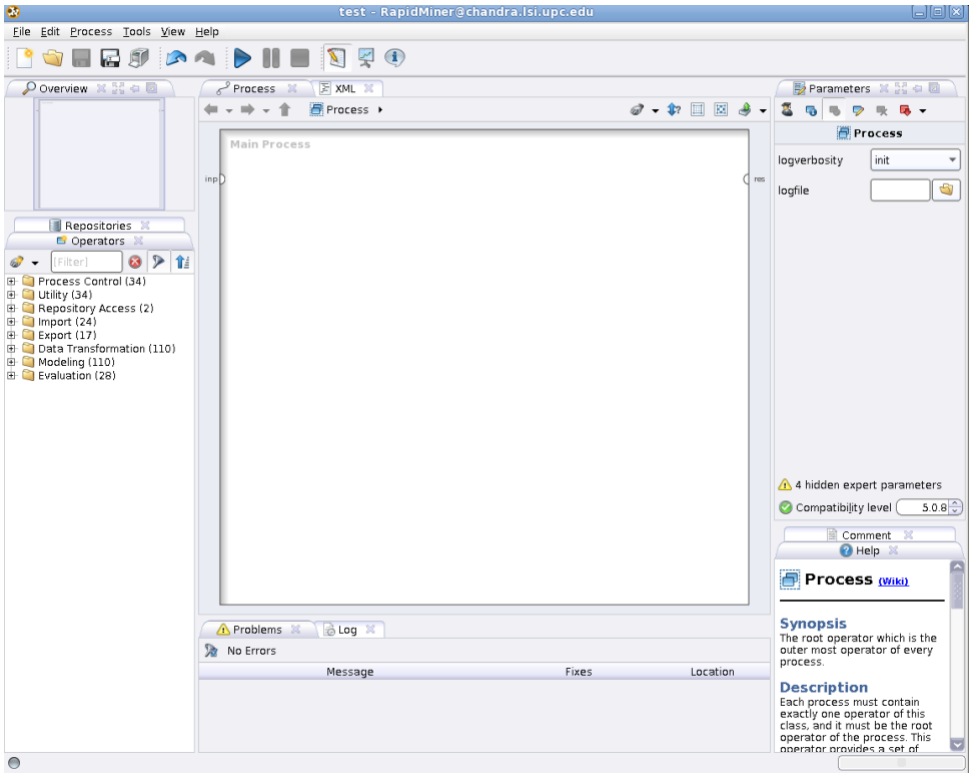
Rapid Miner - Setting a repository

Rapid Miner – Perspectives

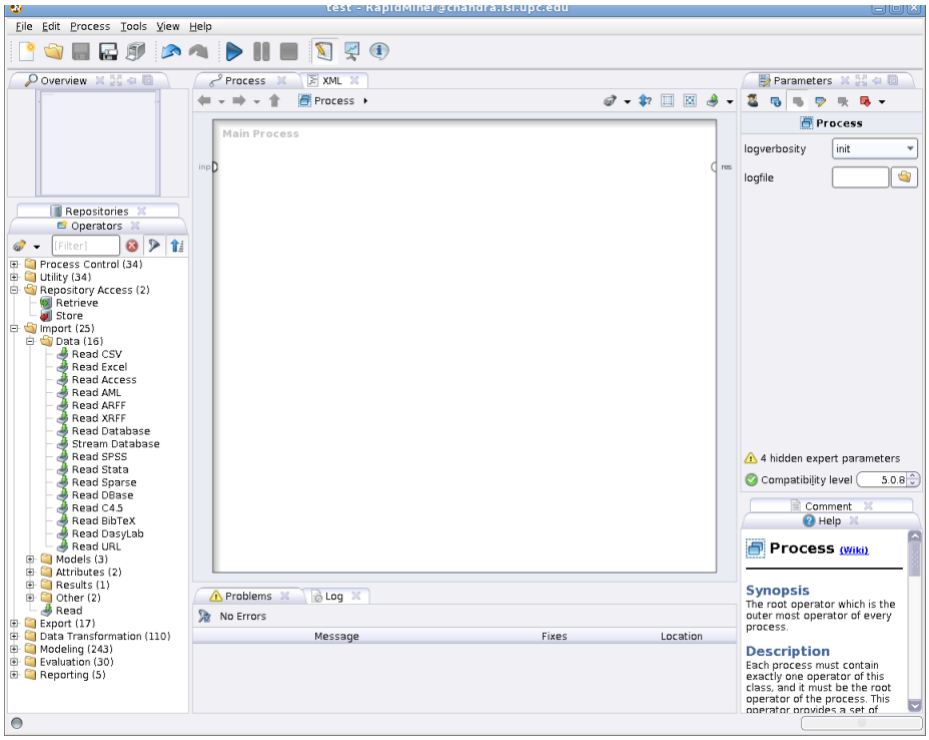
Rapid Miner - New project



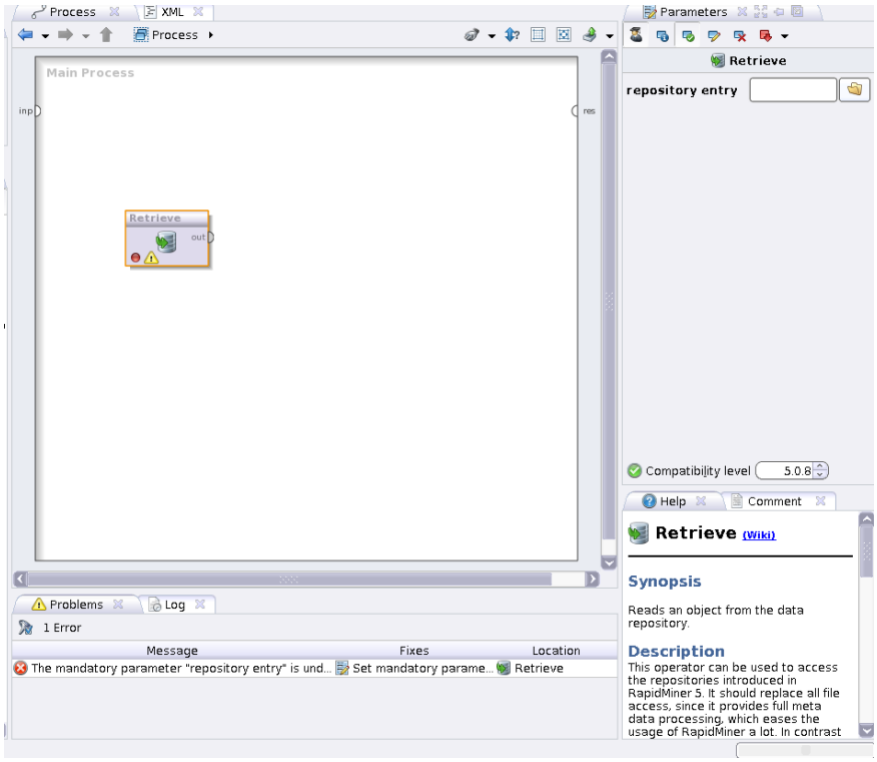
Rapid Miner - Process perspective



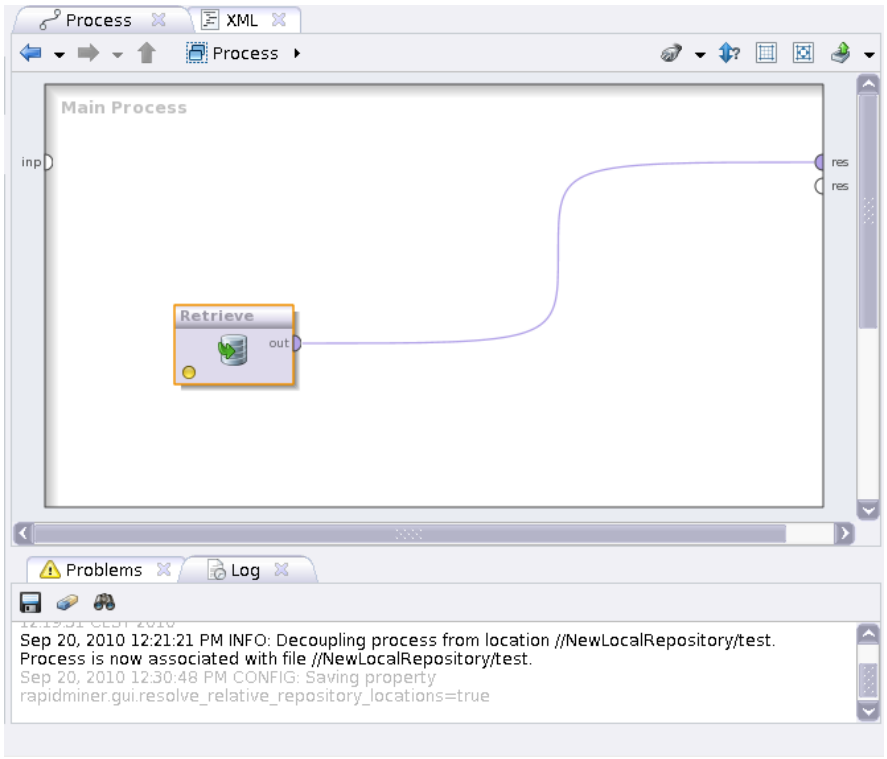
Rapid Miner – Operators



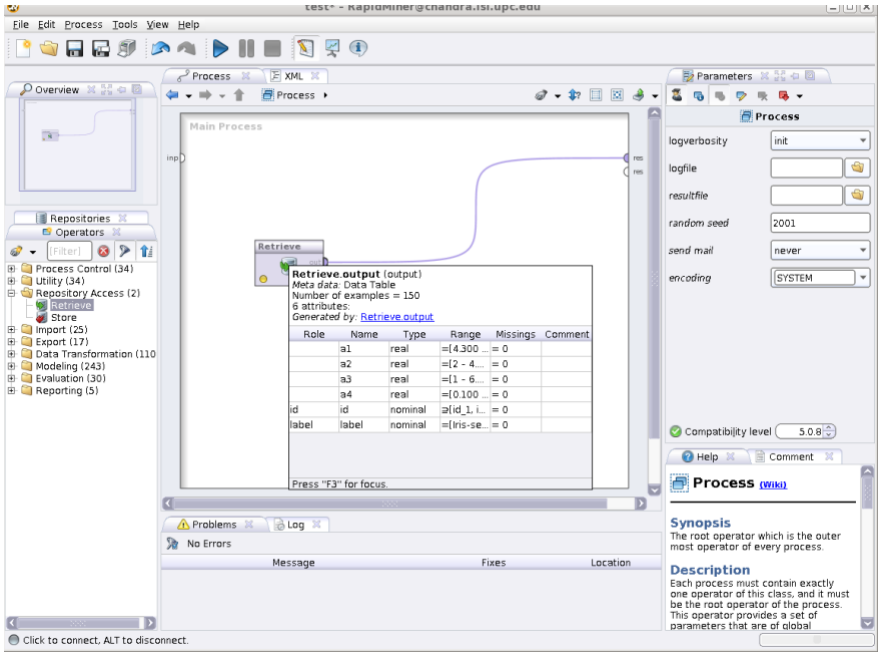
Rapid Miner - Adding operators



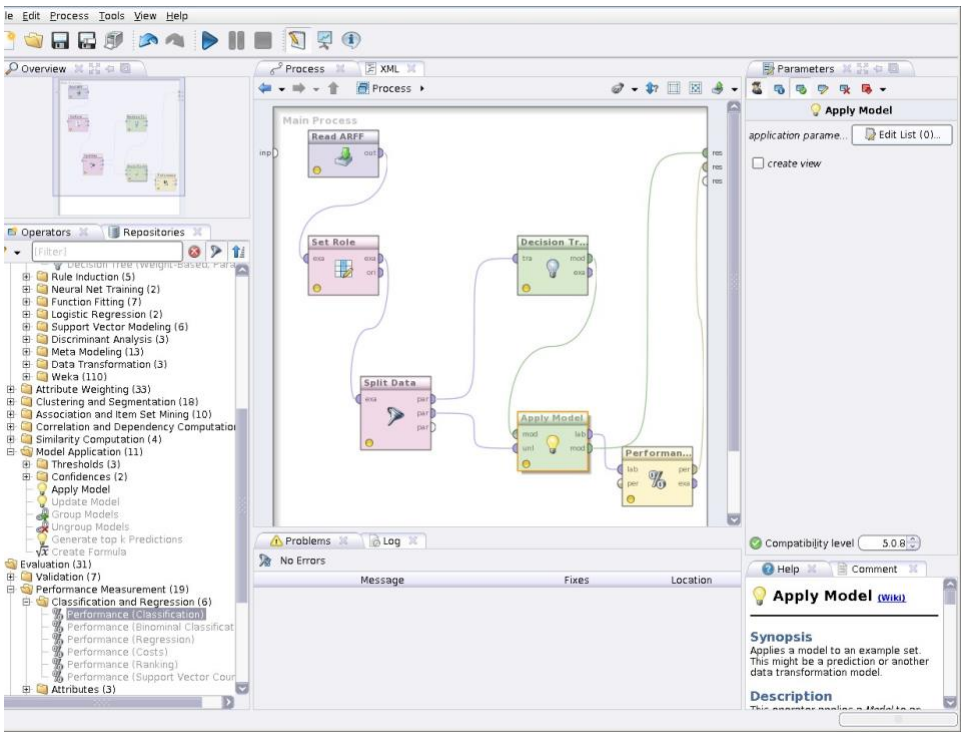
Rapid Miner - Connecting operators



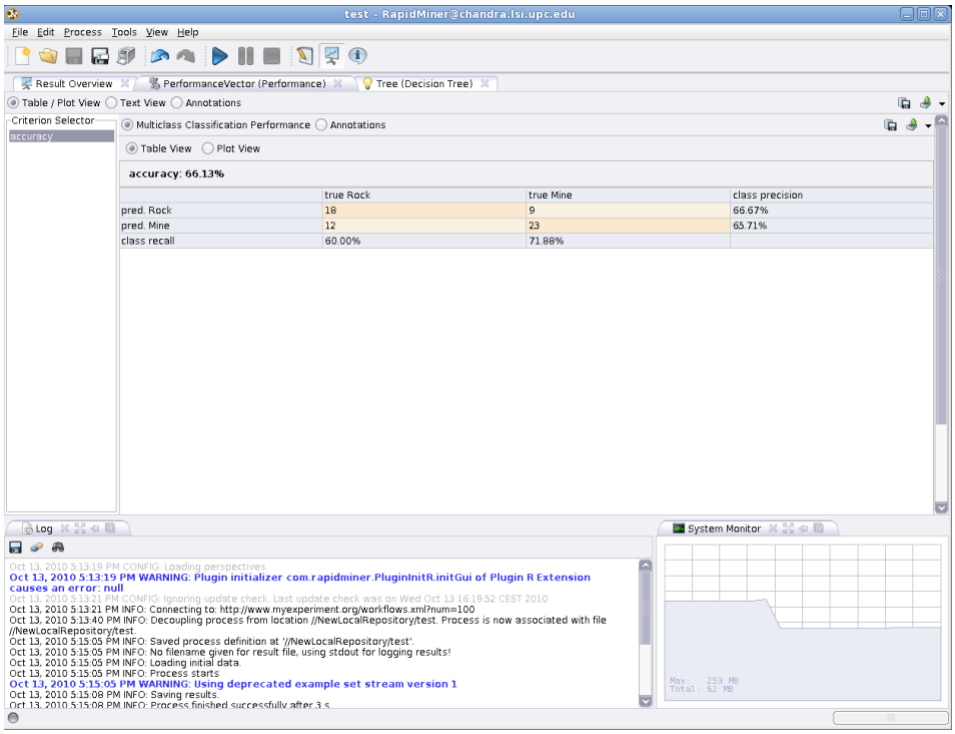
Rapid Miner – Metadata

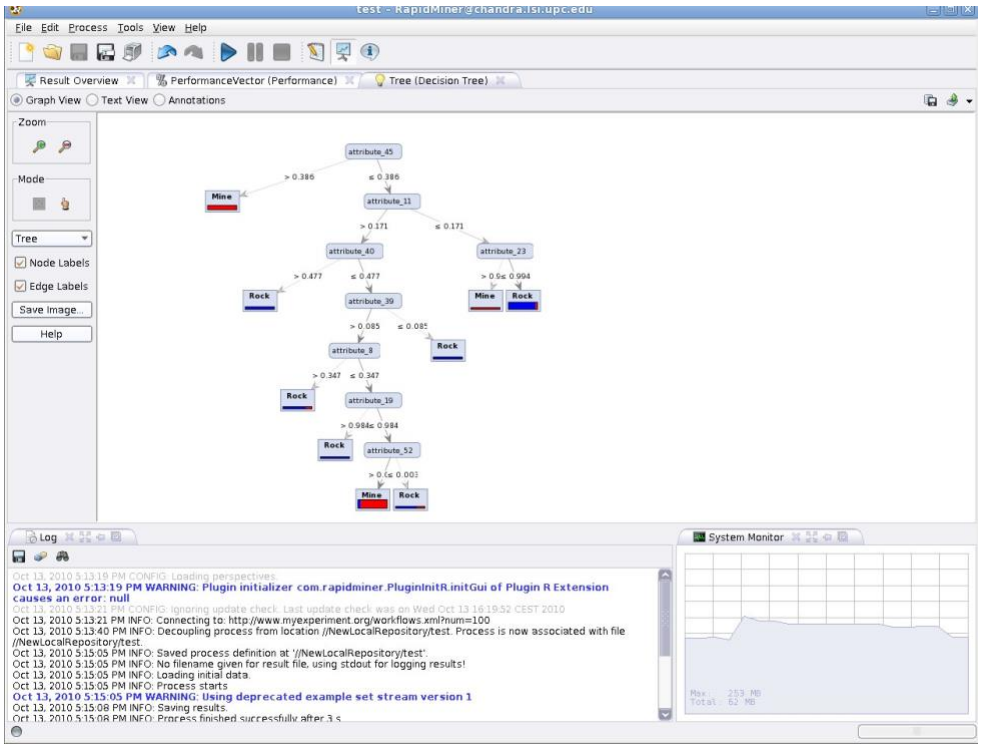


Rapid Miner - Complex Process



Rapid Miner - Results Perspective





STEP 1: DATA SET CREATION

Data set can be created in excel file or CSV format file. To create data set in excel follow the instructions:

Step 1: Get your data set ready in Excel. Your set can have as many rows and columns as you want, with or without a heading row.

Step 2: Insert two empty columns to the left of your data. To do this, you can right-click within the A column and click "insert" twice. Alternatively, you can click into the A column, select "Insert" from the menu bar, and click on "Columns."

Step 3: Type = RAND() into your first empty cell. This should be in the A column, under any heading row. The function will generate a random number between 0 and 1 in that cell.

Step 4: Copy the Rand() formula and paste it all the way down your A column. Make sure that every piece of data in your set now has a random number next to it.

Step 5: Highlight the whole column of random numbers. Copy these values.

Step 6: Paste the random values in the B column. Make sure you use "Paste Special" and select the "Values" option. This will tell Excel to copy the values but not the formulas. The RAND() formula recalculates a new random number each time you do anything else in the spreadsheet, but copying the values over to a new column will prevent them from changing in future.

Step 7: Sort your random values. Select the whole B column. Click on the icon for sorting (or go to "Data" in the menu bar and hit "Sort") and select "Ascending."

- Make sure you select "Expand the Selection" and then "Sort" so that the other columns are rearranged along with the B column.
- You can now delete your A and/or B columns. You won't need them again unless you want to sort the selection again.

Step 8: Select your random sample. You can pick however many rows or cells you want for your sample size. Just take the data you want, starting at the top of the list, to make up your sample. Because they were ordered according to random numbers, the sample will be a random selection from your data set, too.

STEP 2: PREPROCESSING

The preprocessing techniques that can be applied are as follows:

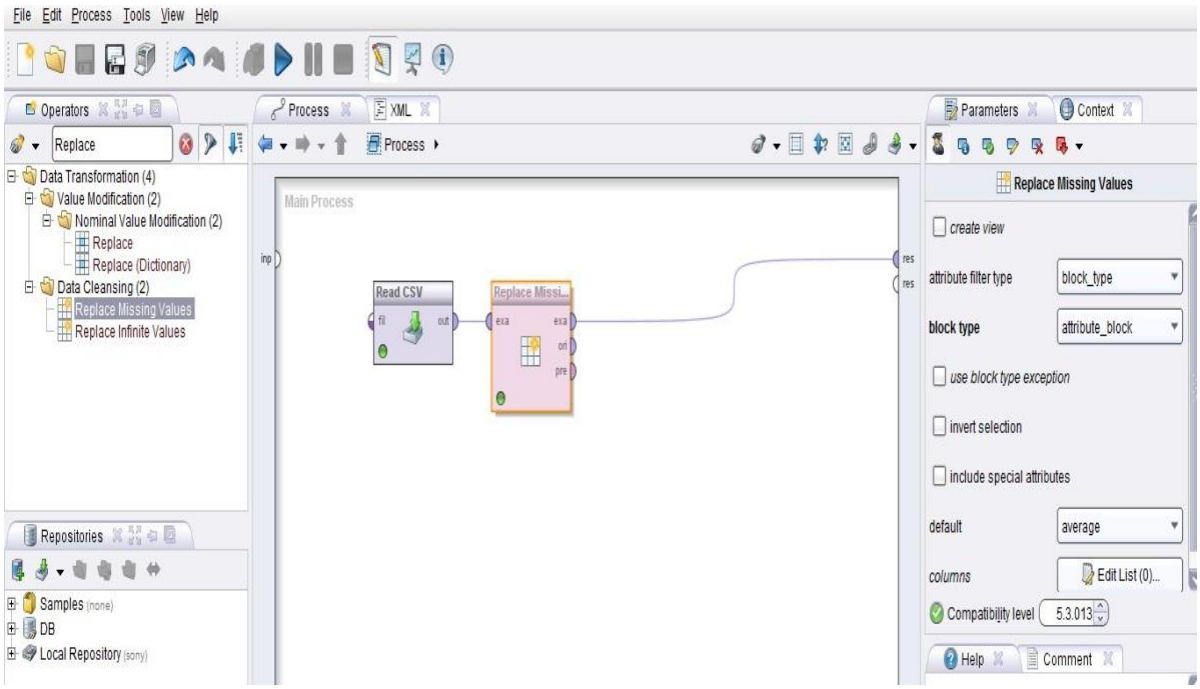
a. <u>Handling Missing Values</u>

<u>Steps</u>:-

1) Create Data set in excel then save as csv (comma delimited) format .

a1	a2	a3
123	2	TRUE
22	3	FALSE
24	4	TRUE
323	5	
23		
54		
654		

2) Click Import and drag Read CSV to workspace.
3) Go to Import Configuration Wizard
   -Import the csv file
   -Press Next, make File encoding as System
   -Make Column Separation as Comma ","
   - Press Next and Finish.
4) Click Data Transformation -> Data Cleansing -> Replace Missing Values. Drag it to Workspace.
5) Click on Replace Missing Values
   -Make attribute filter type as "block_type" and default as "average".
6) Connect the connections.
7) Play and get the Result as Data View.



Result :

Row No.	a1	a2	a3
1	123	2	TRUE
2	22	3	FALSE
3	24	4	TRUE
4	323	5	TRUE
5	23	4	TRUE
6	54	4	TRUE
7	654	4	TRUE

b. Redundancy Techniques

Steps:-

8) Create Data set in excel then save as csv (comma delimited) format .

a1	a2	a3
123	2	TRUE
22	3	FALSE
24	4	TRUE
323	5	TRUE
23	5	TRUE
54	2	FALSE
654	3	TRUE

9) Click Import and drag Read CSV to workspace.

10)Go to Import Configuration Wizard

-Import the csv file

-Press Next, make File encoding as System

-Make Column Separation as Comma ","

- Press Next and Finish.

11)Click Data Transformation -> Filtering -> Remove Duplicates. Drag it to Workspace.

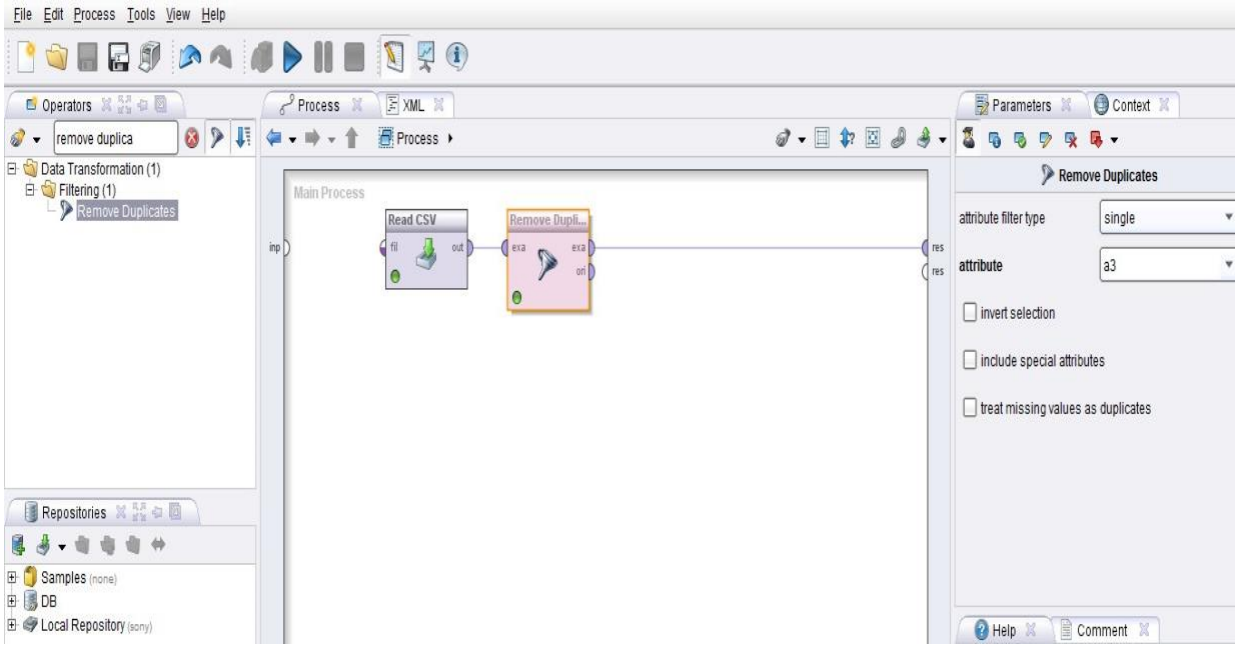
12)Click on Remove Duplicates

-Select attribute filter type as "single"

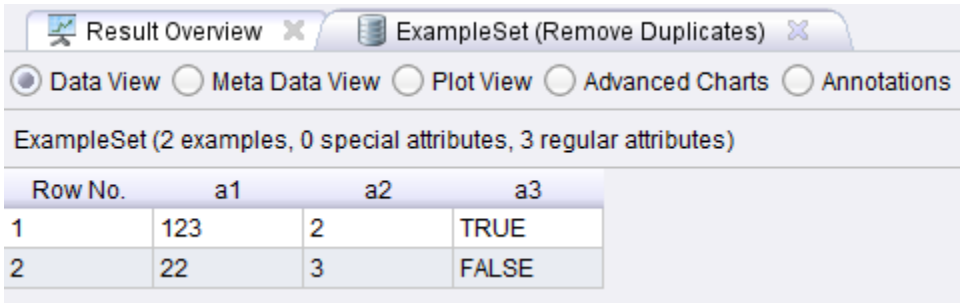
-Select attribute as "a3"

13)Connect the connections.

14)Play and get the Result as Data View.



Result :


Result Overview | ExampleSet (Remove Duplicates)

Data View | Meta Data View | Plot View | Advanced Charts | Annotations

ExampleSet (2 examples, 0 special attributes, 3 regular attributes)

Row No.	a1	a2	a3
1	123	2	TRUE
2	22	3	FALSE

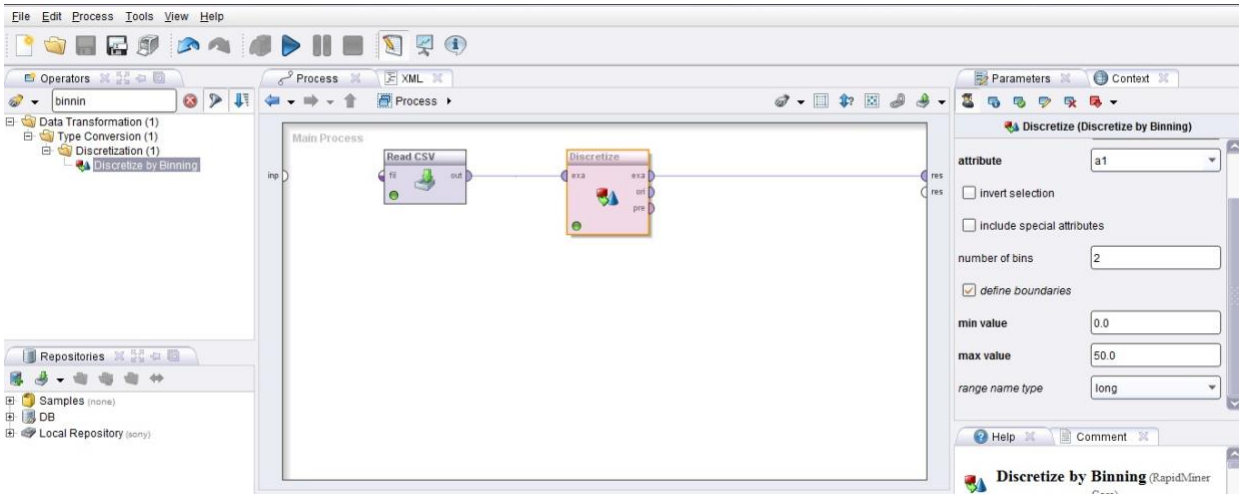
c. **Binning:**

Steps :-

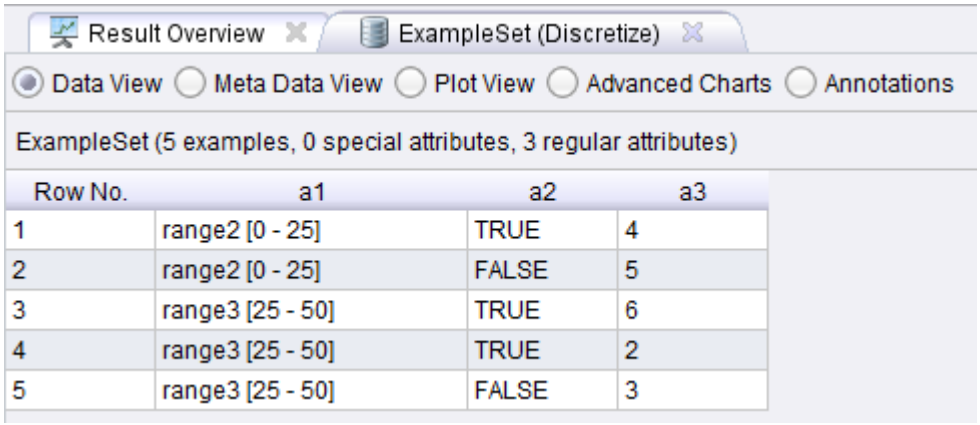
1) Create Data set in excel then save as csv (comma delimited) format .

a1	a2	a3
10	TRUE	4
20	FALSE	5
30	TRUE	6
40	TRUE	2
50	FALSE	3

2) Click Import and drag Read CSV to workspace.
3) Go to Import Configuration Wizard
   -Import the csv file
   -Press Next, make File encoding as System
   -Make Column Separation as Comma ","
   - Press Next and Finish.
4) Click Data Transformation -> Type Conversion -> Discretization -> Discretize by Binning. Drag it to workspace.
5) Click Discretize by Binning
   -Make attribute filter type as "single"
   -Make attribute as "a1"
   -Make number of bins as 2
   -Make min value as 0
   -Make max value as 50
6) Connect the connections.
7) Play and get the Result as Data View.



Result :



Row No.	a1	a2	a3
1	range2 [0 - 25]	TRUE	4
2	range2 [0 - 25]	FALSE	5
3	range3 [25 - 50]	TRUE	6
4	range3 [25 - 50]	TRUE	2
5	range3 [25 - 50]	FALSE	3

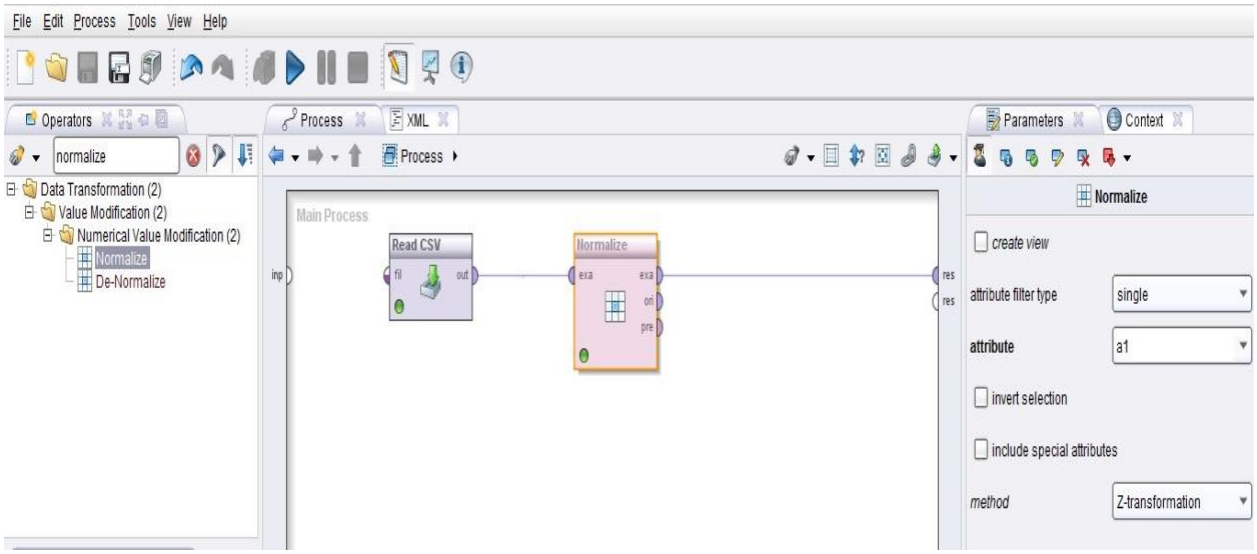
d. Normalization Techniques

Steps:-

1) Create Data set in excel then save as csv (comma delimited) format .

tid	a1	a2	a3
1	-0.598	2	TRUE
2	-0.598	3	FALSE
3	1.195	4	TRUE
4	1.195	5	TRUE
5	0	5	TRUE
6	-1.195	2	FALSE
7	0	3	TRUE

2) Click Import and drag Read CSV to workspace.
3) Go to Import Configuration Wizard
   -Import the csv file
   -Press Next, make File encoding as System
   -Make Column Separation as Comma ","
   -Make tid as "label"
   - Press Next and Finish.
4) Click Data Transformation -> Value Modification -> Numerical Value Modification -> Normalize. Drag it to Workspace.
5) Click on Normalize
   -Select attribute filter type as "single"
   -Select attribute as "a1"
   -Select method as "Z-Transformation"
6) Connect the connections.
7) Play and get the Result as Data View.



Result :

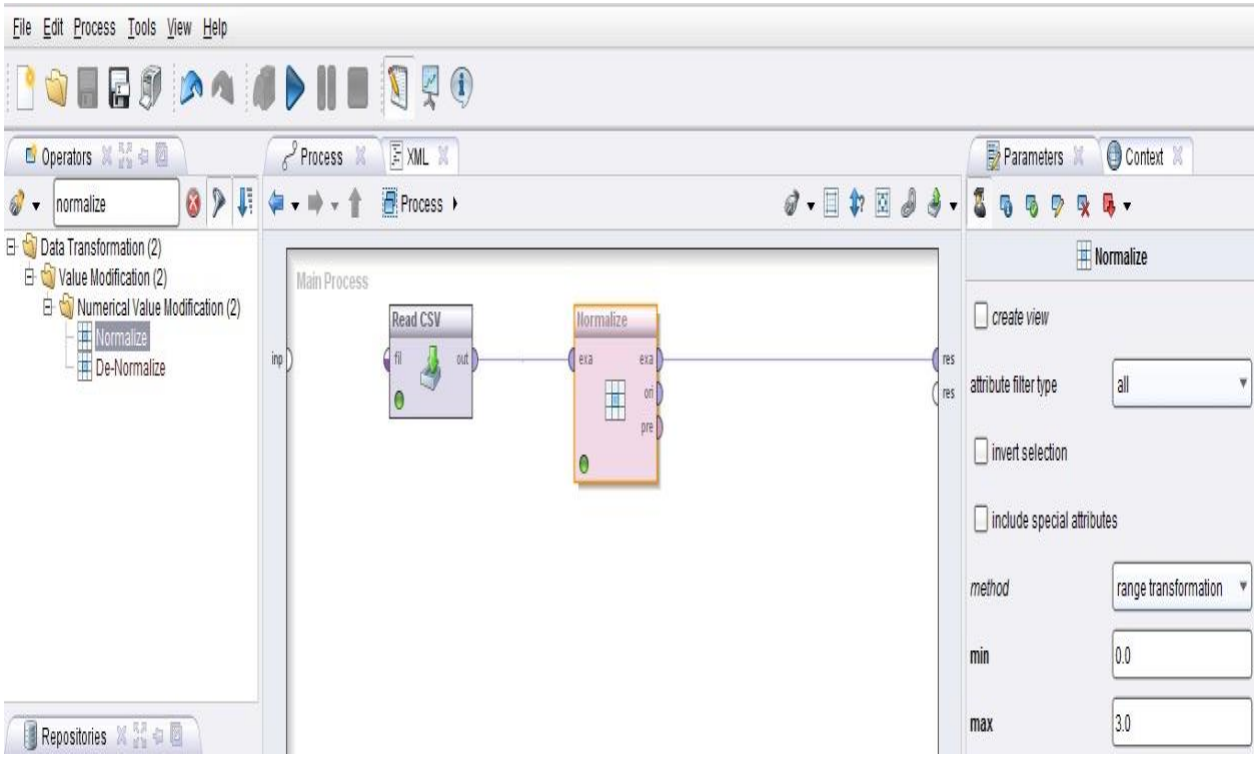
Result Overview | ExampleSet (Normalize)

Data View | Meta Data View | Plot View | Advanced Charts | Annotations

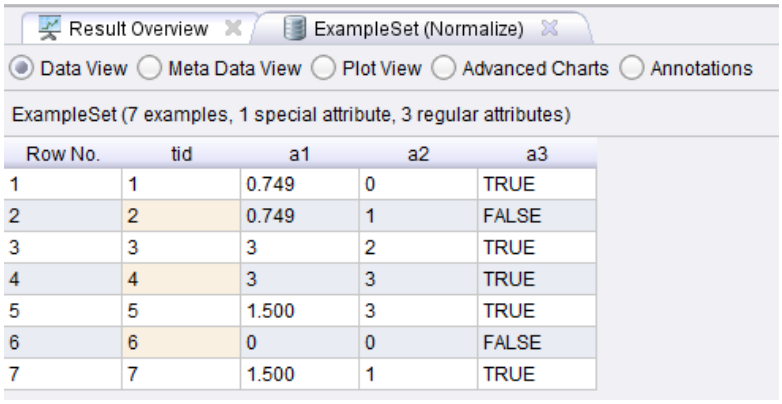
ExampleSet (7 examples, 1 special attribute, 3 regular attributes)

Row No.	tid	a1	a2	a3
1	1	-0.655	2	TRUE
2	2	-0.655	3	FALSE
3	3	1.309	4	TRUE
4	4	1.309	5	TRUE
5	5	0.000	5	TRUE
6	6	-1.309	2	FALSE
7	7	0.000	3	TRUE

8) Again, Click on Normalize
   -Select attribute filter type as “all”
   -Select method as “Range Transformation”
   -Select min as 0.0
   -Select max as 3.0
9) Connect the connections.
10) Play and get the Result as Data View.



Result :



Row No.	tid	a1	a2	a3
1	1	0.749	0	TRUE
2	2	0.749	1	FALSE
3	3	3	2	TRUE
4	4	3	3	TRUE
5	5	1.500	3	TRUE
6	6	0	0	FALSE
7	7	1.500	1	TRUE

Step 3:

Perform the following on the preprocessed dataset:

FP Growth Algorithm

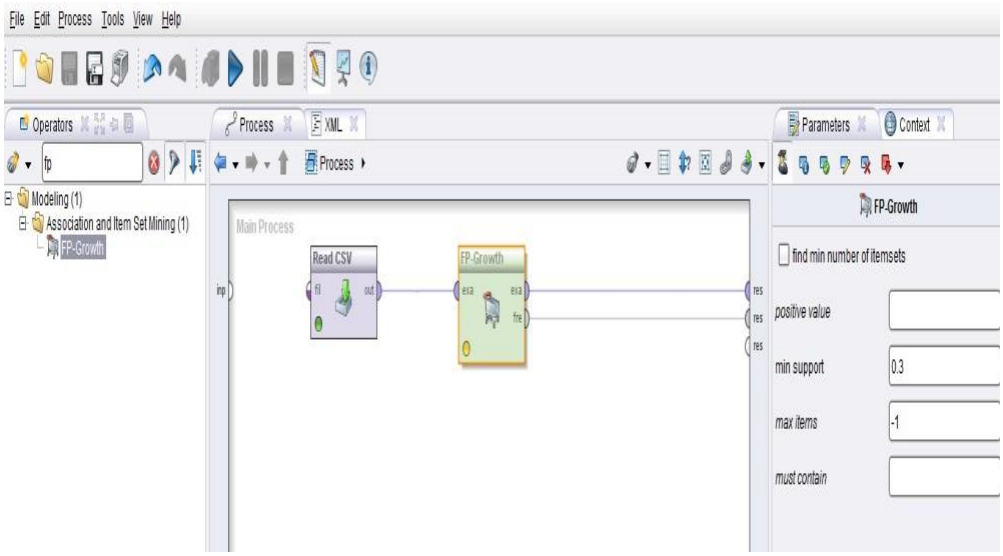
Steps :-

1) Create Data set in excel then save as csv (comma delimited) format .

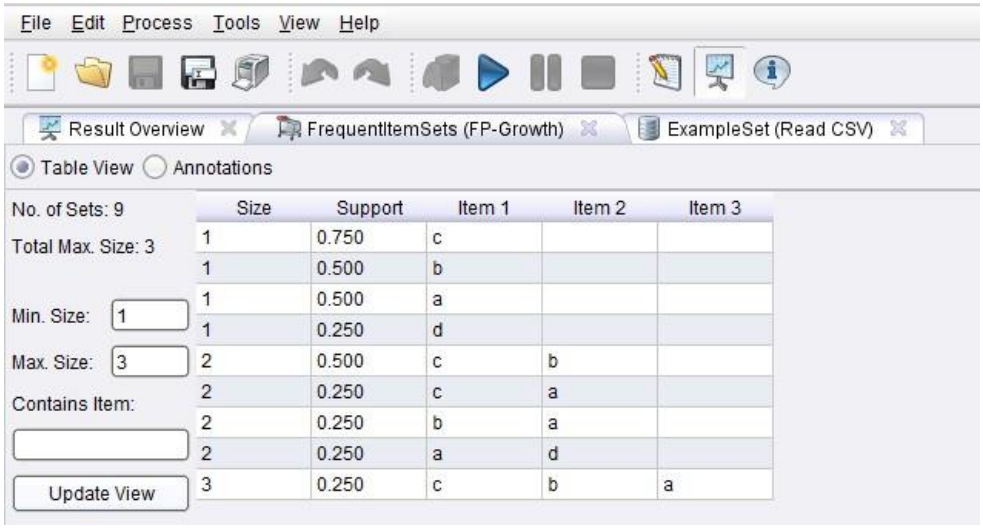
tid	a	b	c	d
1	0	0	1	0
2	1	1	1	0
3	0	1	1	0
4	1	0	0	1

11)Click Import and drag Read CSV to workspace.
12)Go to Import Configuration Wizard
-Import the csv file
-Press Next, make File encoding as System
-Make Column Separation as Comma ","
-Make tid as "label" and other attributes as "binomial".
- Press Next and Finish.

2) Click Modeling -> Association and Item Set Mining -> FP Growth. Drag it to workspace.
3) Connect the connections.
4) Click FP Growth
-Make min support as 0.2
-Make max items as -1
5) Play and get the Result as Table View.



Result :-



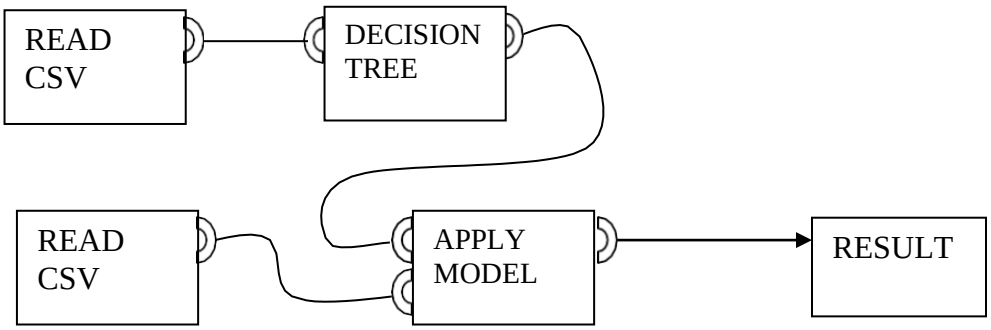
Size	Support	Item 1	Item 2	Item 3
1	0.750	c		
1	0.500	b		
1	0.500	a		
1	0.250	d		
2	0.500	c	b	
2	0.250	c	a	
2	0.250	b	a	
2	0.250	a	d	
3	0.250	c	b	a

Decision Tree:

STEPS:

- Import the CSV file for the training set and test set.
- Set the delimiter as comma [,].
- Drag and drop “Decision tree”.
- Drag and drop “Apply model”.
- Make connections as shown in figure.
- Run the process

SCHEMATIC DIAGRAM:



INPUT:

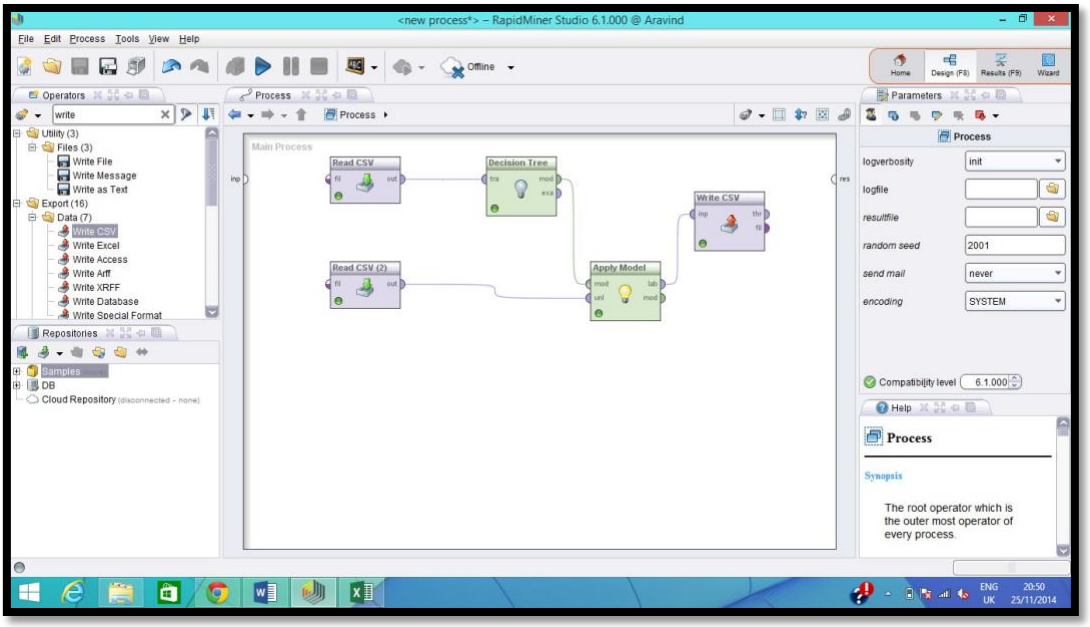
Training set:

Name	Body temperature	Gives birth	Four legged	Hibernates	Class label
Human	Warm	yes	no	no	yes
Pigeon	Warm	no	no	no	no
Elephant	Warm	yes	yes	no	yes
Leopard shark	Cold	yes	no	no	no
Turtle	Cold	no	yes	no	no
Penguin	Cold	no	no	no	no
Eel	Cold	no	no	no	no
Dolphin	warm	yes	no	no	yes
Spiny anteater	warm	no	yes	yes	yes
Gila monster	Cold	no	yes	yes	no

Test set:

Name	Body temperature	Gives birth	Four legged	Hibernates	Class label
Salamander	cold	no	yes	yes	
Guppy	cold	yes	no	no	
Eagle	warm	no	no	no	
Poorwill	warm	no	no	yes	
Platypus	warm	no	yes	yes	

SCREENSHOT:



OUTPUT:

Name	Body temperature	Gives birth	Four legged	Hibernates	Confidence (yes)	Confidence (no)	Prediction (Class label)
Salamander	cold	no	yes	yes	0	1	no
Guppy	cold	yes	no	no	0	1	no
Eagle	warm	no	no	no	0.5	0.5	no
Poorwill	warm	no	no	yes	0.5	0.5	no
Platypus	warm	no	yes	yes	0.5	0.5	no