

Business Intelligence Technologies: Practical 1: Classification using Decision Trees

Weighting: 4% Due date: 10 Dec 2022

Using the heart.csv data provided, construct an initial ‘default’ decision tree for classifying whether a patient has less chance, or more chance of having a heart attack, and summarise the performance of the model.

Improve the performance by changing the decision tree parameters. Suggest reasons for why the performance has improved, and explain the results.

Document your processes, showing step by step screens and outputs, including an explanation of the final parameters you used, and what effect they had on your model.

Assessment Criteria:

Correct and fully documented processes: 2 marks

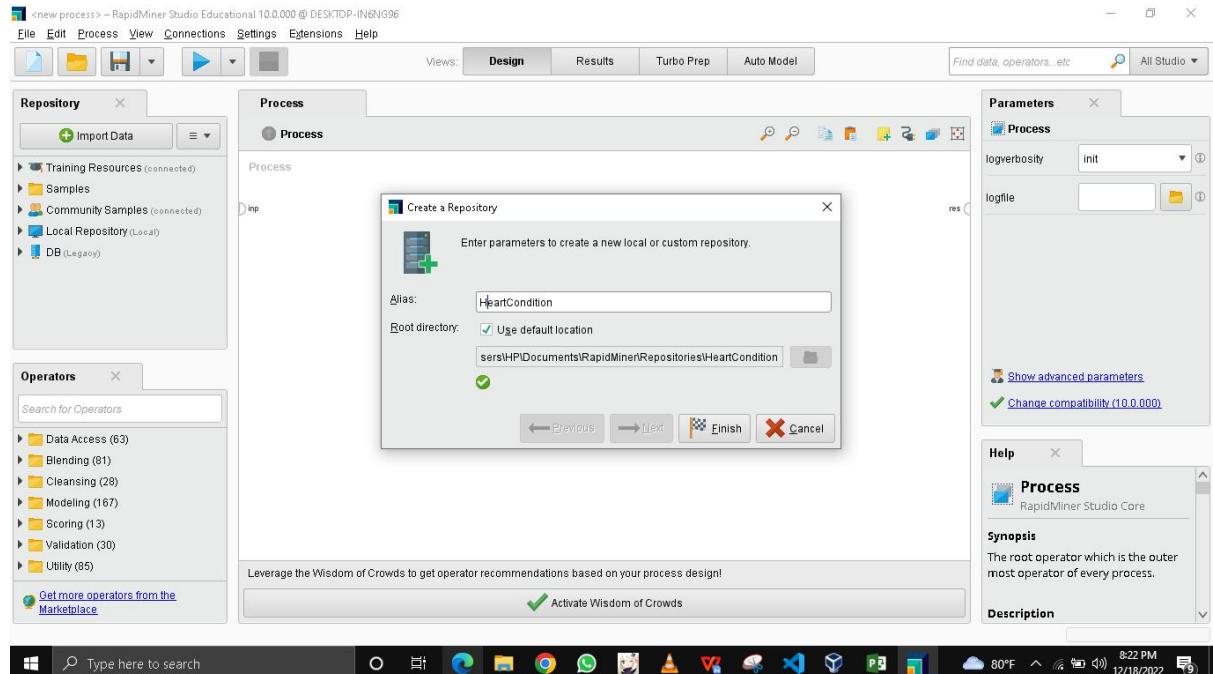
Explanation demonstrating sound understanding of principles of classification using Decision Trees : 2 marks

Description of the Heart dataset:

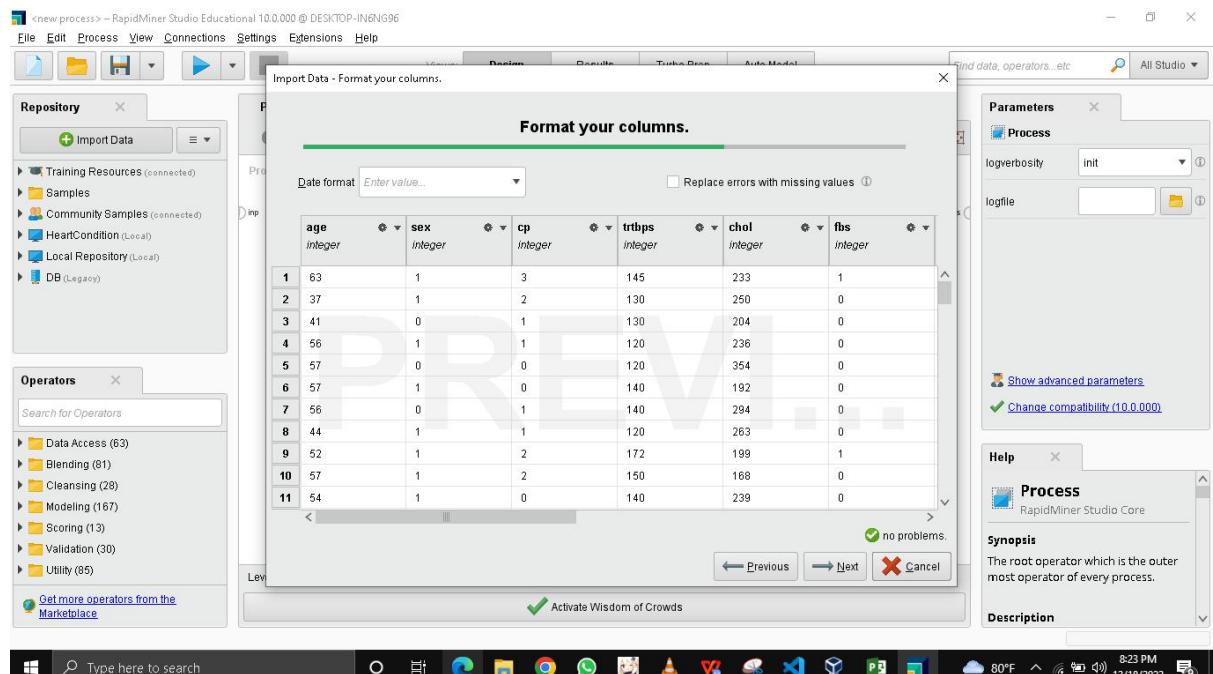
- Age : Age of the patient
- Sex : Sex of the patient
- exng: exercise induced angina (1 = yes; 0 = no)
- caa: number of major vessels coloured by fluroscopy (0-3)
- cp : Chest Pain type chest pain type
 - Value 1: typical angina
 - Value 2: atypical angina
 - Value 3: non-anginal pain
 - Value 4: asymptomatic
- trtbps : resting blood pressure (in mm Hg)
- chol : cholestorol in mg/dl fetched via BMI sensor
- fbs : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
- restecg : resting electrocardiographic results
 - Value 0: normal
 - Value 1: having ST-T wave abnormality
 - Value 2: showing probable or definite left ventricular hypertrophy
- thalachh : maximum heart rate achieved
- oldpeak: ST depression induced by exercise relative to rest (-2.6 to 6.2)
- slp: the slope of the peak exercise ST ssegment (0=upsloping, 1=flat, 2=Down)
- thall: 0=normal; 1=fixed defect; 2=reversable defect
- output (ie the label) : 0= less chance of heart attack 1= more chance of heart attack

Practical Documentation

1. Created a Repository labeled “HeartCondition”



2. Imported the heartData table into the HeartCondition Repository with the table's default attributes unchanged as shown below



S <new process> – RapidMiner Studio Educational 10.0.000 @ DESKTOP-IN6NG96

File Edit Process View Connections Settings Extensions Help

Import Data - Format your columns.

Format your columns.

Date format Enter value... Replace errors with missing values

	fbs	restecg	thalachh	exng	oldpeak	sip
	integer	integer	integer	integer	real	integer
1	1	0	150	0	2.300	0
2	0	1	187	0	3.500	0
3	0	0	172	0	1.400	2
4	0	1	178	0	0.800	2
5	0	1	163	1	0.600	2
6	0	1	148	0	0.400	1
7	0	0	153	0	1.300	1
8	0	1	173	0	0.000	2
9	1	1	162	0	0.500	2
10	0	1	174	0	1.600	2
11	0	1	160	0	1.200	2

no problems. Previous Next Cancel

Activate Wisdom of Crowds

Repository Operators

Parameters

Help

Process RapidMiner Studio Core

Synopsis The root operator which is the outer most operator of every process.

Description

Type here to search

8:23 PM 12/18/2022

S <new process> – RapidMiner Studio Educational 10.0.000 @ DESKTOP-IN6NG96

File Edit Process View Connections Settings Extensions Help

Import Data - Format your columns.

Format your columns.

Date format Enter value... Replace errors with missing values

	exng	oldpeak	sip	caa	thall	output
	integer	real	integer	integer	integer	integer
1	0	2.300	0	0	1	1
2	0	3.500	0	0	2	1
3	0	1.400	2	0	2	1
4	0	0.800	2	0	2	1
5	1	0.600	2	0	2	1
6	0	0.400	1	0	1	1
7	0	1.300	1	0	2	1
8	0	0.000	2	0	3	1
9	0	0.500	2	0	3	1
10	0	1.600	2	0	2	1
11	0	1.200	2	0	2	1

no problems. Previous Next Cancel

Activate Wisdom of Crowds

Repository Operators

Parameters

Help

Process RapidMiner Studio Core

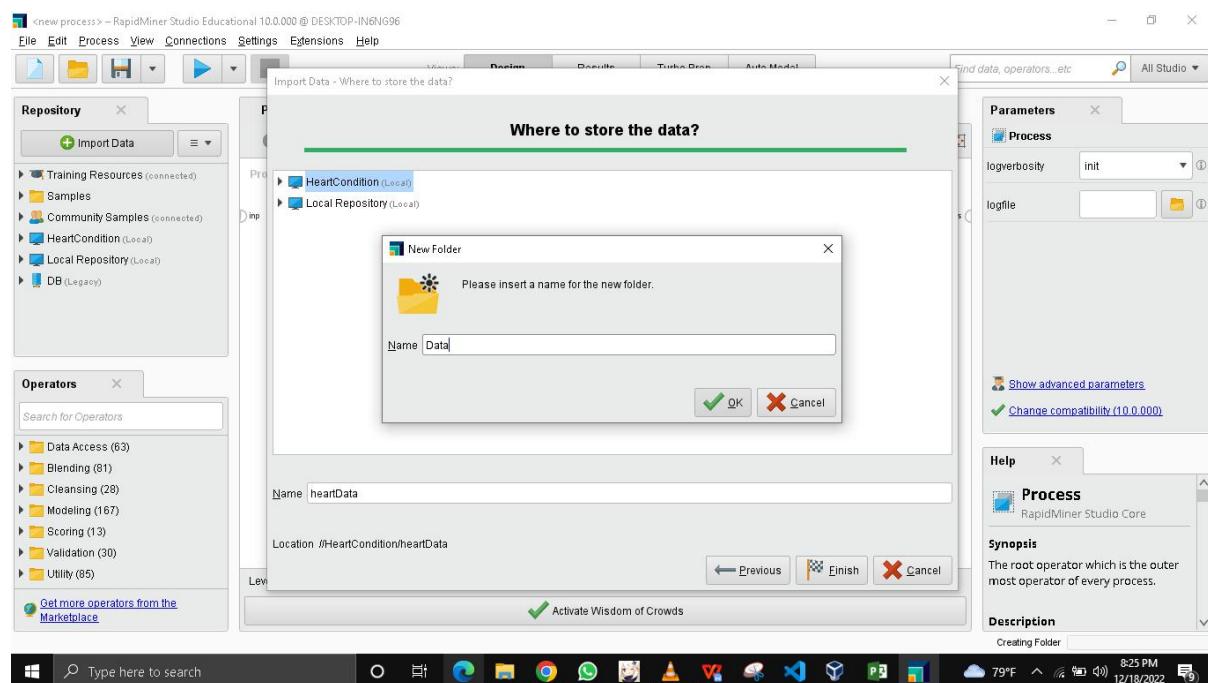
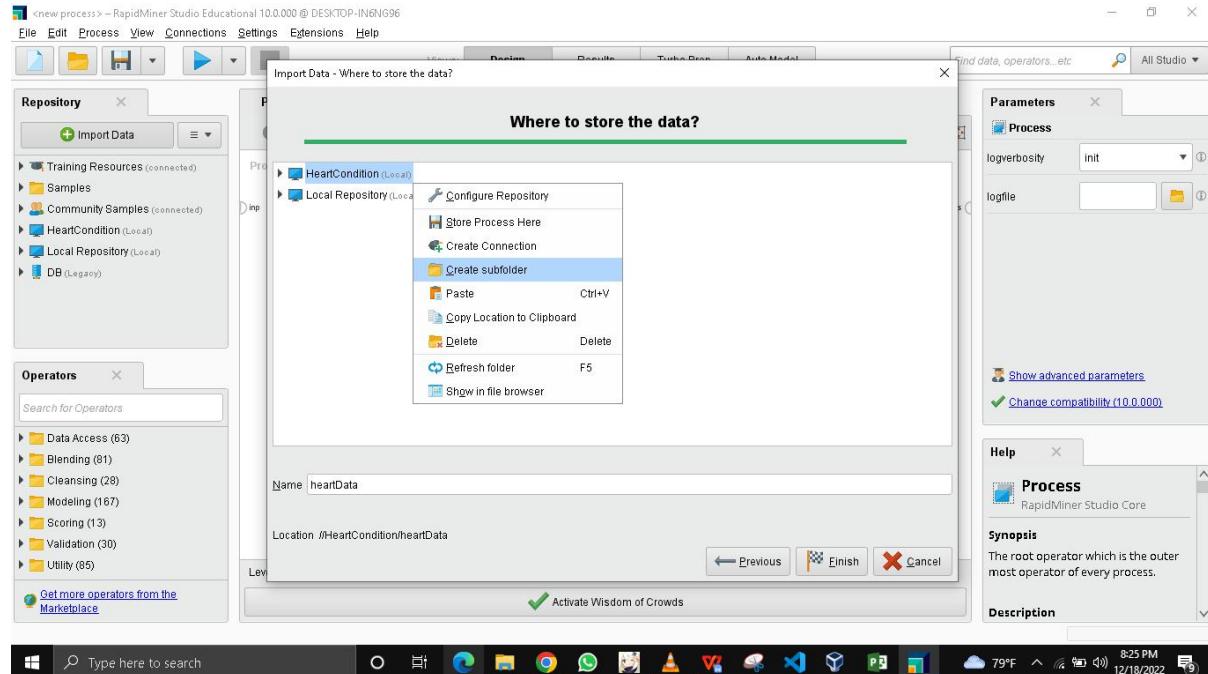
Synopsis The root operator which is the outer most operator of every process.

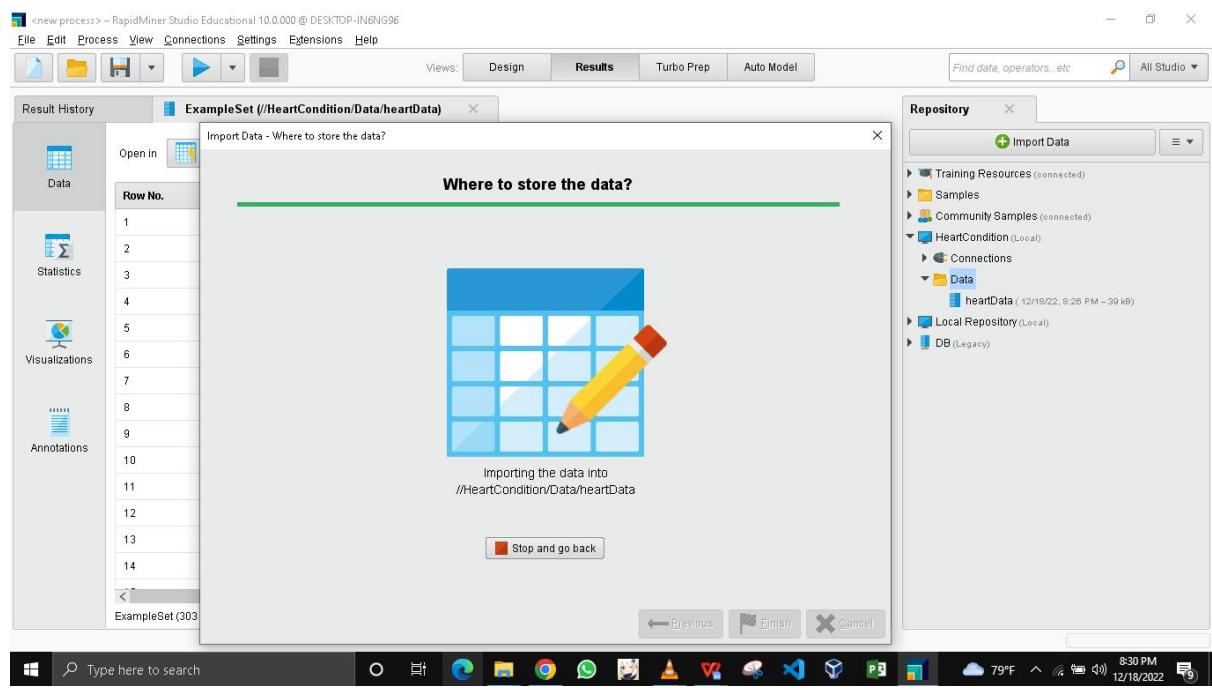
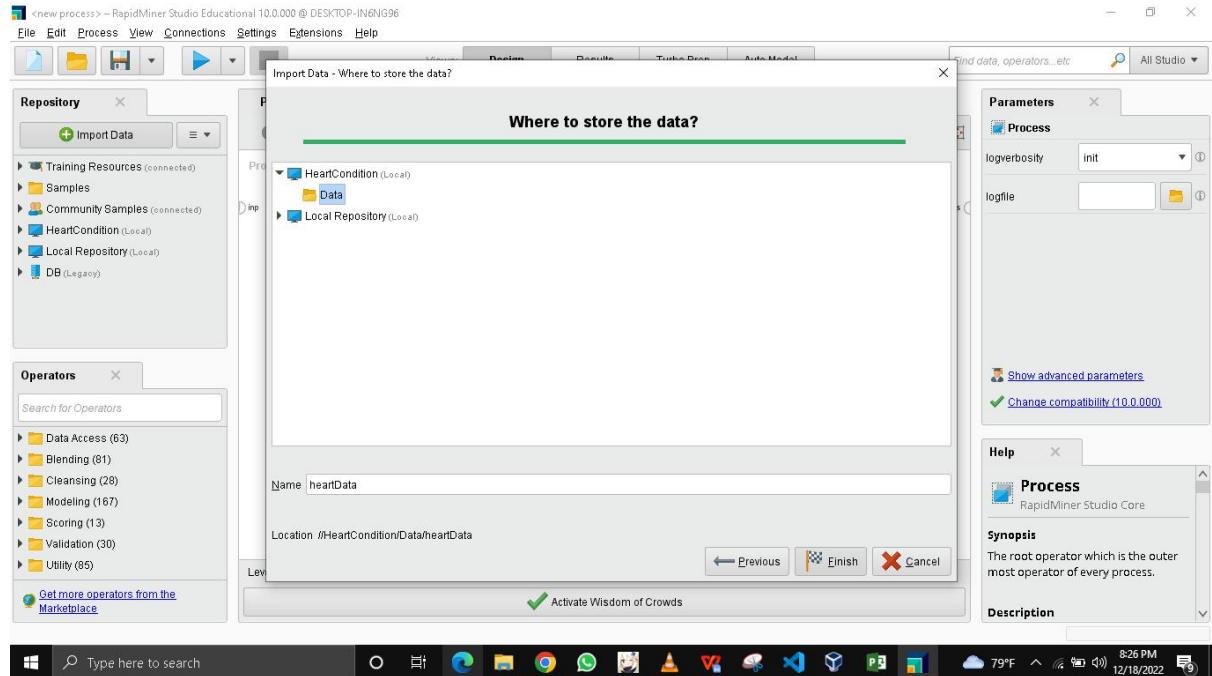
Description

Type here to search

8:24 PM 12/18/2022

3. Created a “Data” Folder within the Heart Condition Repository in which the heartData table was imported into





Below are the results of the completed import with the ranges of the values within the heartData table shown for further detail

The screenshot shows the RapidMiner Studio interface with the 'Results' tab selected. On the left, there's a sidebar with 'Result History' and three main sections: 'Data', 'Statistics', and 'Visualizations'. The 'Data' section contains a table of 303 examples with 14 attributes: Row No., age, sex, cp, trbps, chol, fbs, restecg, and thalachl. The 'Statistics' section shows summary statistics for each attribute. The 'Visualizations' section is currently empty. To the right, the 'Repository' pane is open, showing the 'HeartCondition' project with its connections and data. A detailed view of the 'heartData' table is shown, listing the number of examples (303), attributes (14), and a schema table with columns: Role, Name, Type, Range, and Missing. The 'fbs' attribute is highlighted.

This screenshot is identical to the one above, but it shows the 'heartData' table after attribute renaming. The renamed attributes are: fbs, restecg, thalachh, exng, oldpeak, and slp. The 'Statistics' section now includes these new attribute names in its summary table. The rest of the interface, including the repository and system status bar, remains the same.

Result History

ExampleSet (/HeartCondition/Data/heartData)

Views: Design Results Turbo Prep Auto Model

Find data, operators...etc All Studio

Data

Row No.	age	sex	cp	trbps	chol	fbs	restecg	thalachl
1	63	1	3	145	233	1	0	150
2	37	1	2	130	250	0	1	187
3	41	0	1	130	204	0	0	172
4	56	1	1	120	236	0	1	178
5	57	0	0	120	354	0	1	163
6	57	1	0	140	192	0	1	148
7	56	0	1	140	294	0	0	153
8	44	1	1	120	263	0	1	173
9	52	1	2	172	199	1	1	162
10	57	1	2	150	168	0	1	174
11	54	1	0	140	239	0	1	160
12	48	0	2	130	275	0	1	139
13	49	1	1	130	266	0	1	171
14	64	1	3	110	211	0	0	144

ExampleSet (303 examples, 0 special attributes, 14 regular attributes)

Repository

Import Data

Training Resources (connected)

Samples

Community Samples (connected)

HeartCondition (Local)

Connections

Data

heartData (12/18/22, 8:26 PM - 39 kB)

heartData

Data table

Number of examples = 303

14 attributes:

Role	Name	Type	Range	Missing
exng	# integer	= [0 - 1]	= 0	
oldpeak	# real	= [0 - 6.200]	= 0	
slp	# integer	= [0 - 2]	= 0	
caa	# integer	= [0 - 4]	= 0	
thall	# integer	= [0 - 3]	= 0	
output	# integer	= [0 - 1]	= 0	

Press 'F3' for focus.

Windows Taskbar: Type here to search, File Explorer, Edge, Google Chrome, WhatsApp, FileZilla, Vivaldi, Paint, File Manager, PowerShell, Task View, Cloud, 79°F, 8:31 PM, 12/18/2022

4. Retrieved the heartData table into the Process panel

Repository

Import Data

Training Resources (connected)

Samples

Community Samples (connected)

HeartCondition (Local)

Connections

Data

heartData (12/18/22, 8:26 PM)

Local Repository (Local)

DB (Legacy)

Process

Process

Retrieve heartData

Operators

Search for Operators

- Data Access (63)
- Blending (61)
- Cleansing (28)
- Modeling (167)
- Scoring (13)
- Validation (30)
- Utility (85)

Get more operators from the Marketplace

Parameters

repository entry heartData

Parameters

Show advanced parameters

Help

Retrieve

RapidMiner Studio Core

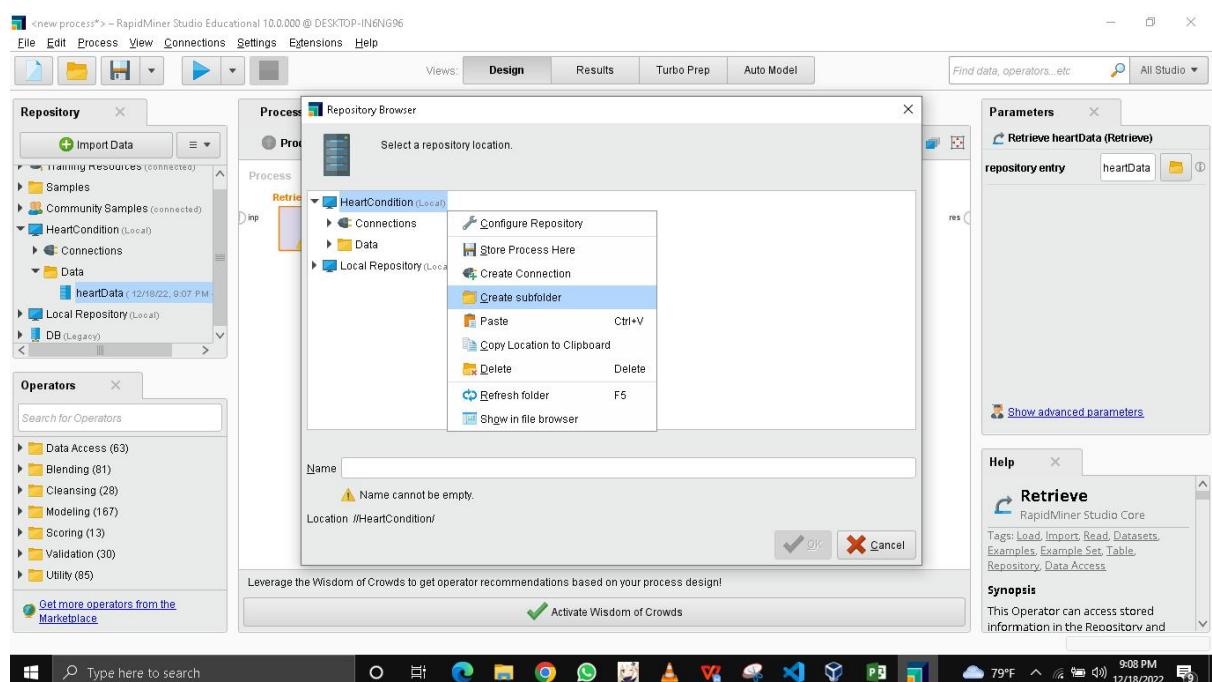
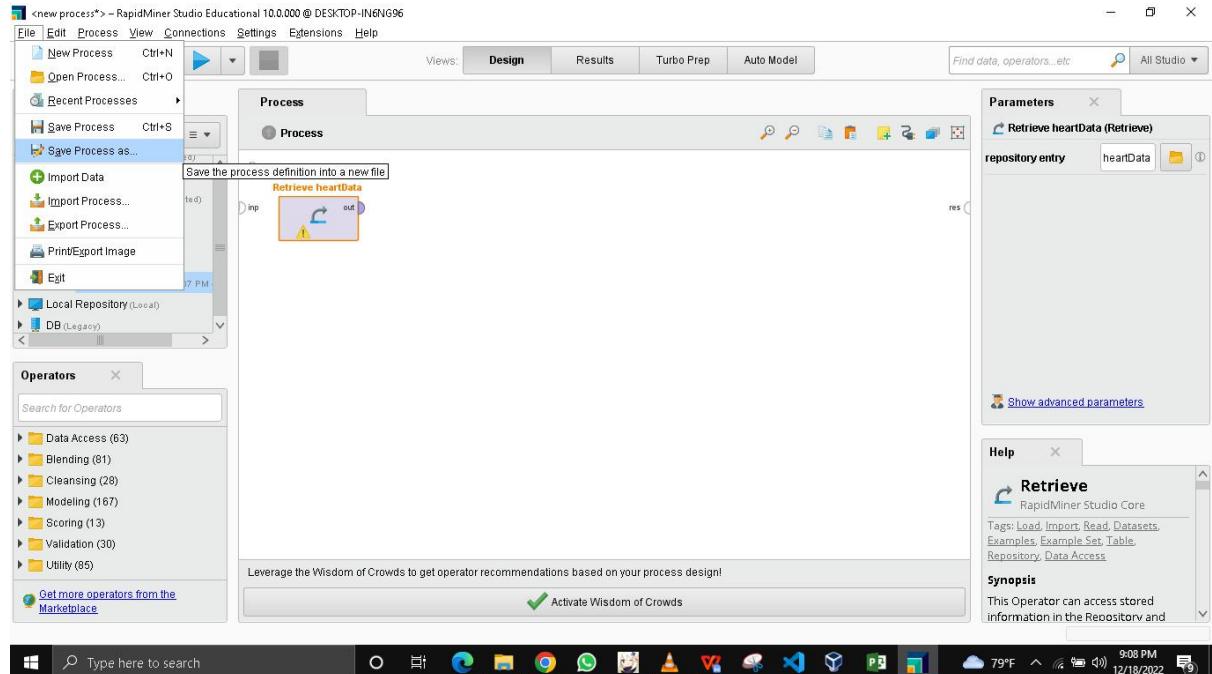
Tags: Load, Import, Read, Datasets, Examples, Example Set, Table, Repository, Data Access

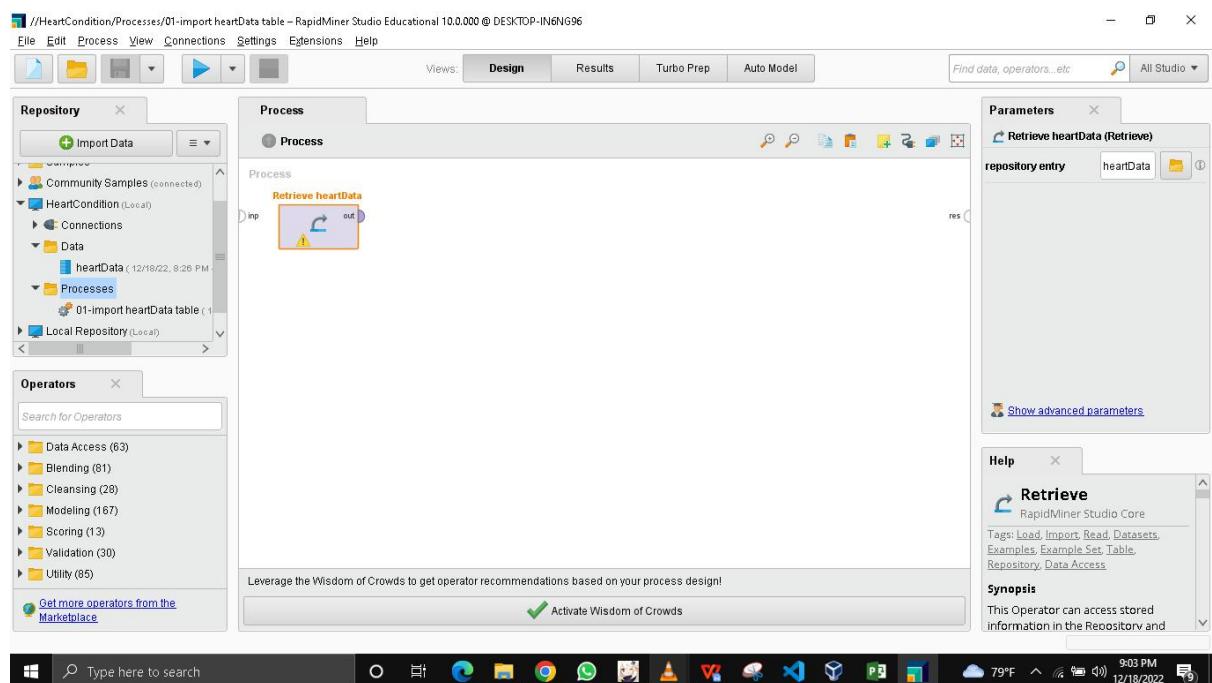
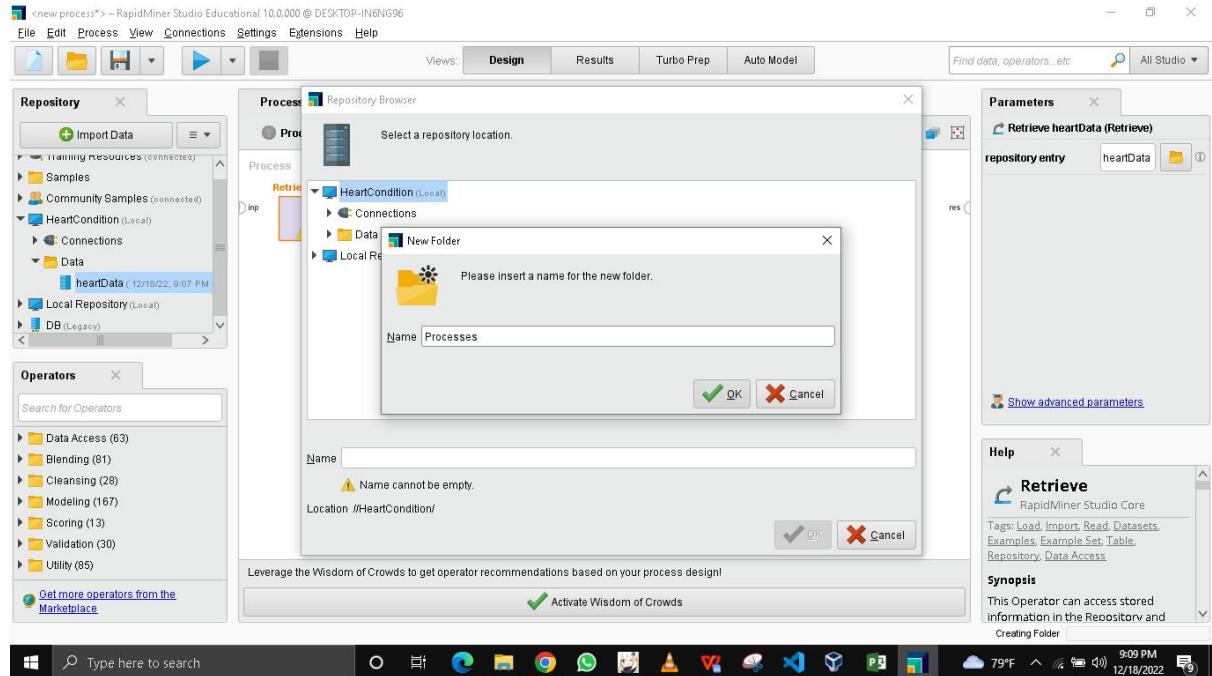
Synopsis

This Operator can access stored information in the Repository and

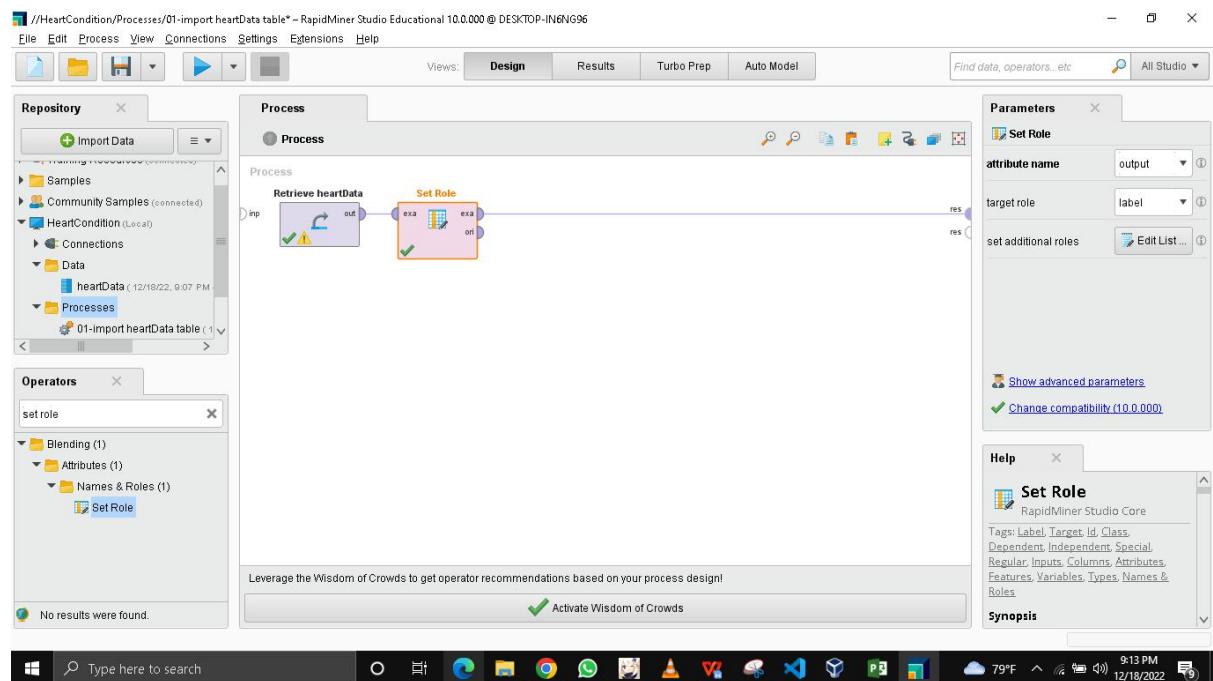
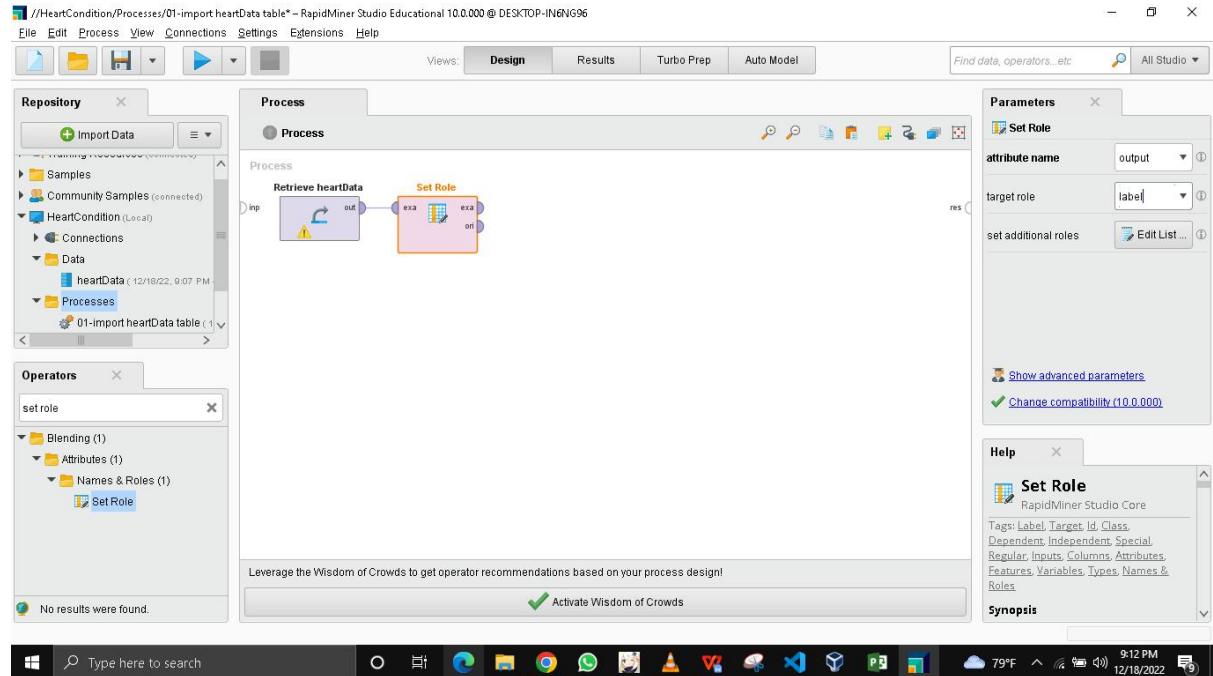
Windows Taskbar: Type here to search, File Explorer, Edge, Google Chrome, WhatsApp, FileZilla, Vivaldi, Paint, File Manager, PowerShell, Task View, Cloud, 79°F, 8:32 PM, 12/18/2022

5. Created a “Processes” Folder and saved the process as “01-import heartData table” in the “Processes” Folder





6. Defined the “Output” column within the heartData table as the Label using the “Set Role” operator



Shown below is a preview of the changes applied to the table after the change

Row No.	output	age	sex	cp	trtbps	chol	fbs	restecg
1	1	63	1	3	145	233	1	0
2	1	37	1	2	130	250	0	1
3	1	41	0	1	130	204	0	0
4	1	56	1	1	120	236	0	1
5	1	57	0	0	120	354	0	1
6	1	57	1	0	140	192	0	1
7	1	56	0	1	140	294	0	0
8	1	44	1	1	120	263	0	1
9	1	52	1	2	172	199	1	1
10	1	57	1	2	150	168	0	1
11	1	54	1	0	140	238	0	1
12	1	48	0	2	130	275	0	1
13	1	49	1	1	130	266	0	1
14	1	64	1	3	110	211	0	0

7. Filtered the heartData Table to only show fields under the “Output” column containing defined values using the “Filter Examples” operator

The process flow in the Design view:

```

graph LR
    A[Retrieve heartData] -- out --> B[Set Role]
    B -- exa --> C[Filter Examples]
    
```

The 'Filter Examples' operator parameters:

- filters: invert filter

The 'Operators' panel shows:

- filter example
- Blending (2)
- Examples (2)
- Filter (2) (highlighted)
- Filter Example Range

The 'Help' panel for 'Filter Examples' states:

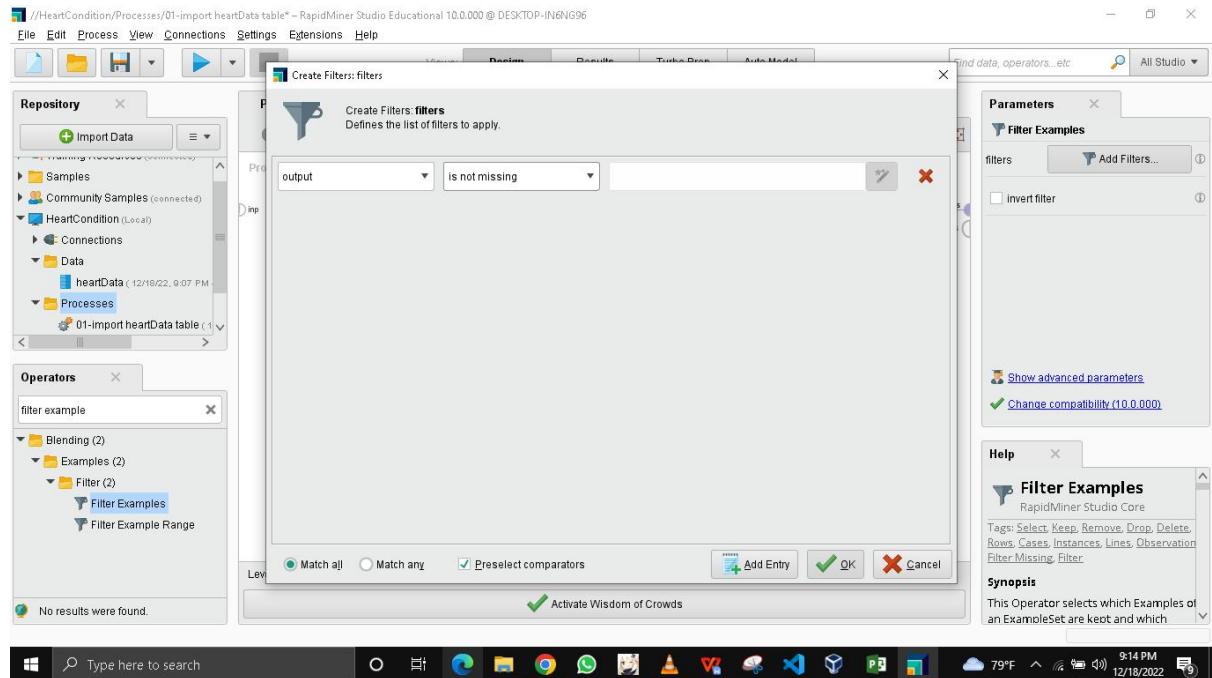
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Activate Wisdom of Crowds

Filter Examples
RapidMiner Studio Core

Tags: Select, Keep, Remove, Drop, Delete, Rows, Cases, Instances, Lines, Observation, Filter Missing, Filter

Synopsis
This Operator selects which Examples of an ExampleSet are kept and which

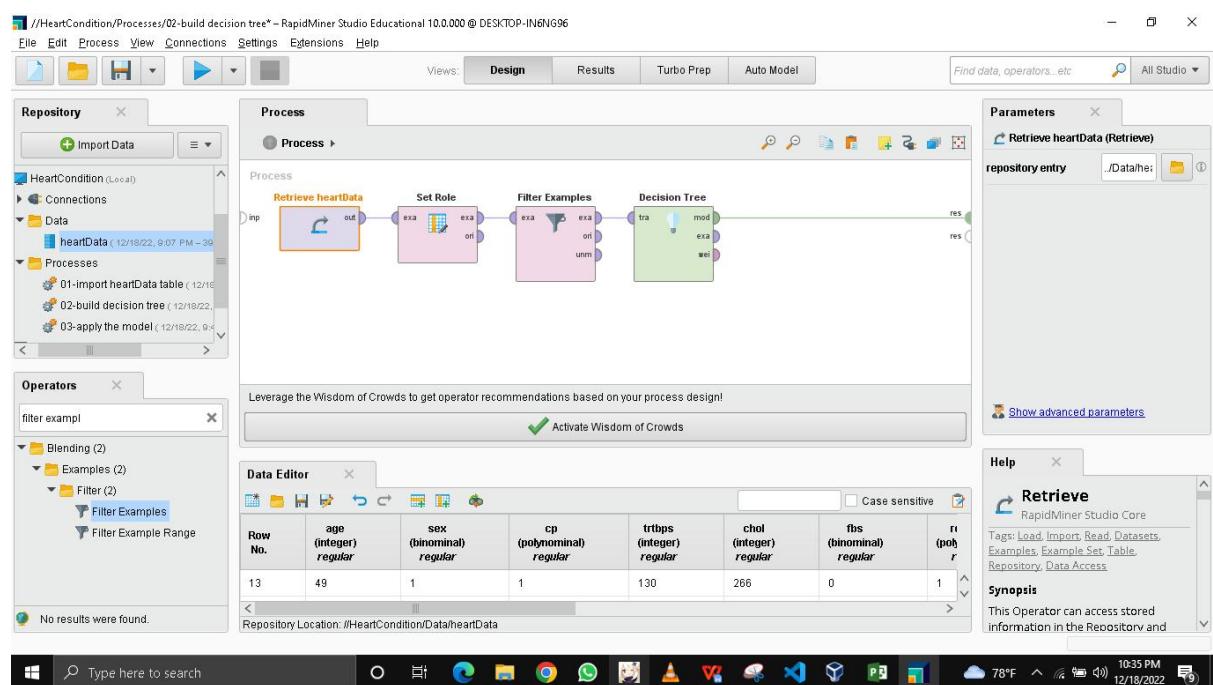
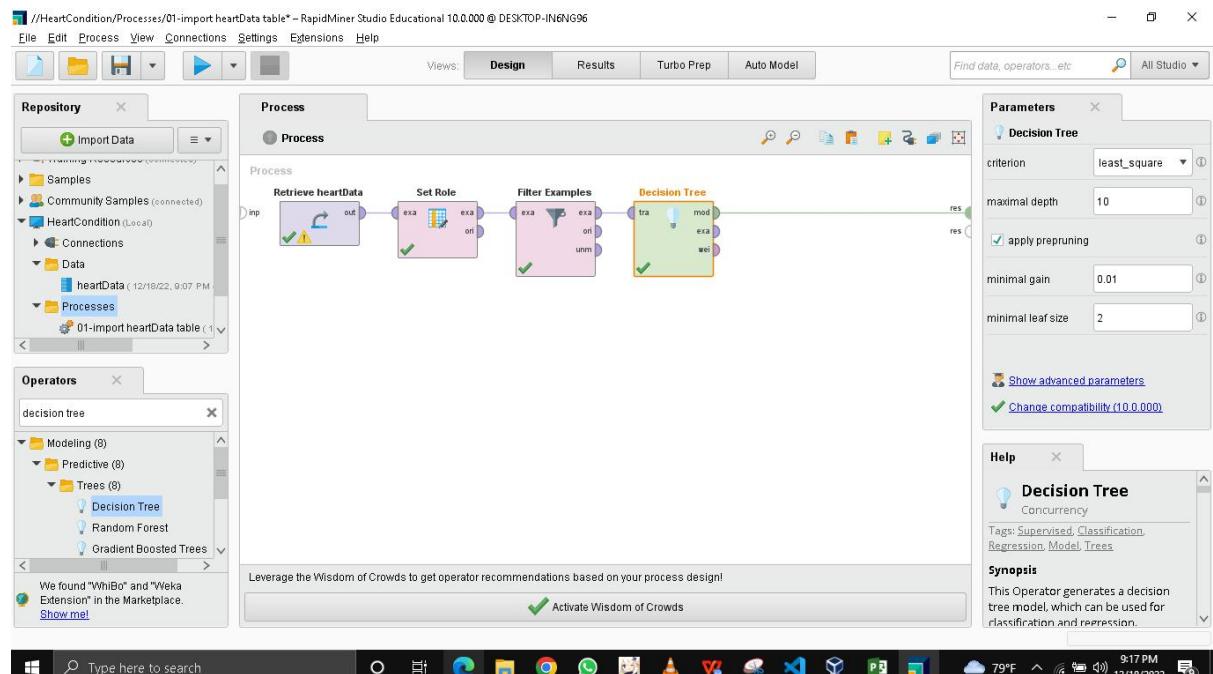


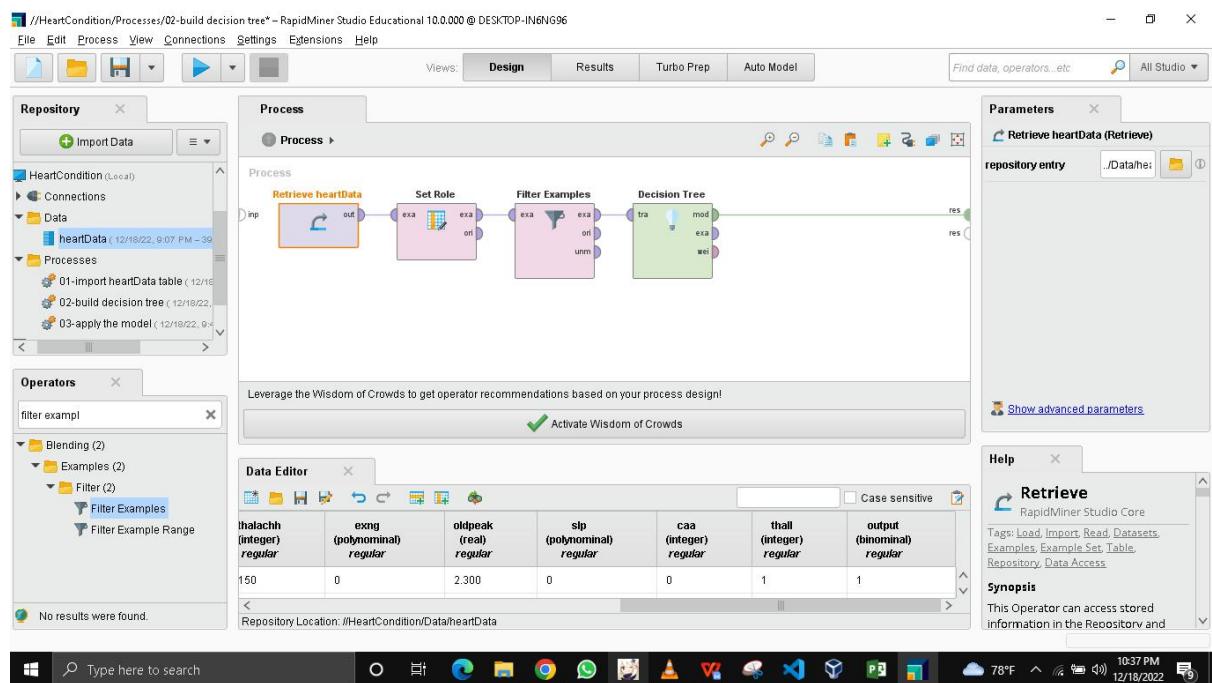
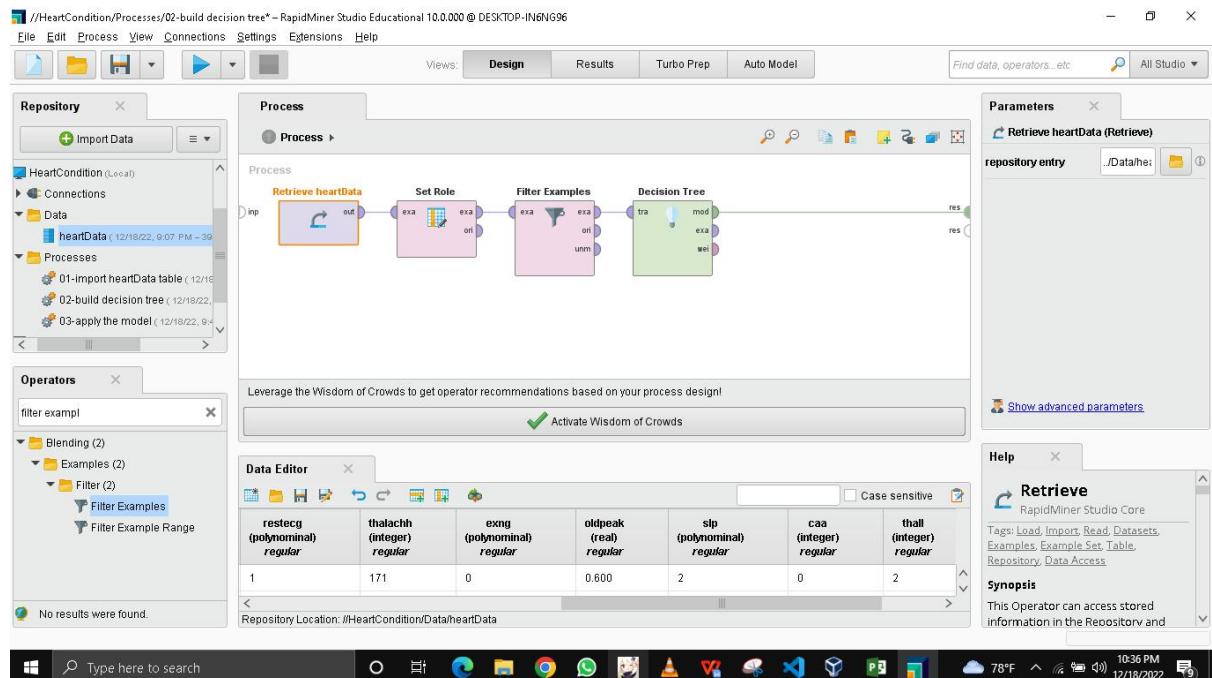
The screenshot shows the RapidMiner Studio interface with the 'ExampleSet (Filter Examples)' view open. The view displays a table of 303 examples from the 'heartData' table. The columns are: Row No., output, age, sex, cp, trtbps, chol, fbs, and restecg. The 'Repository' panel on the right shows the project structure, including 'HeartCondition' and '01-import heartData table'. The Windows taskbar at the bottom shows various application icons.

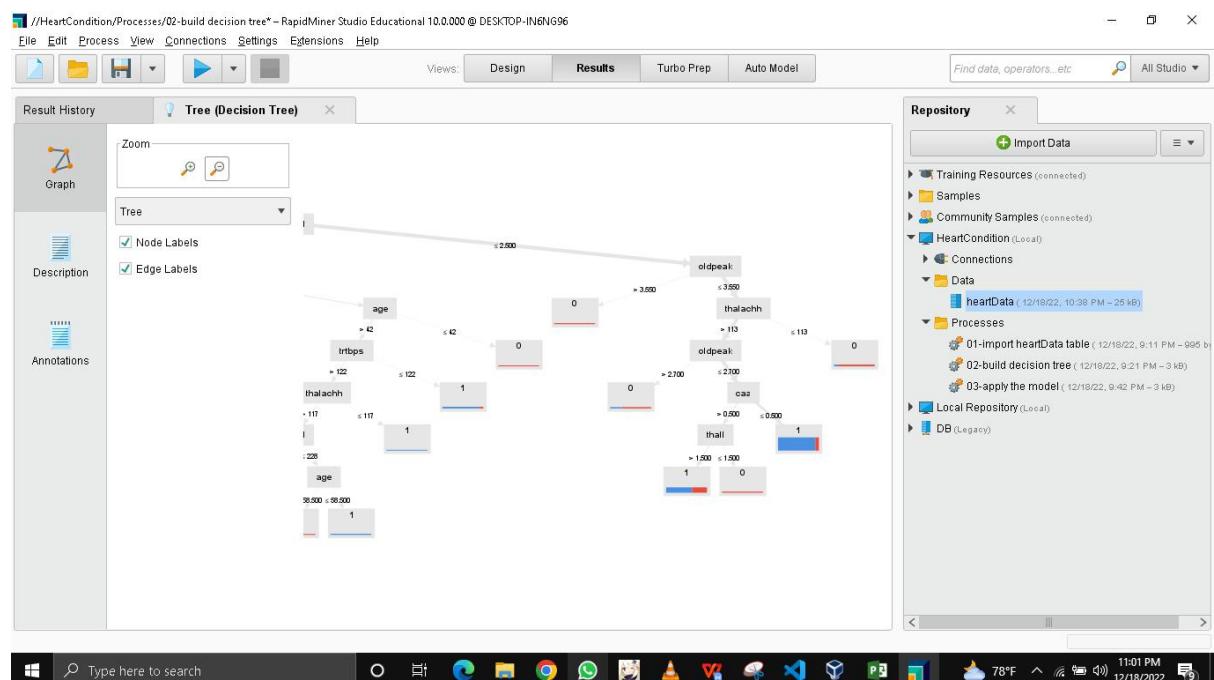
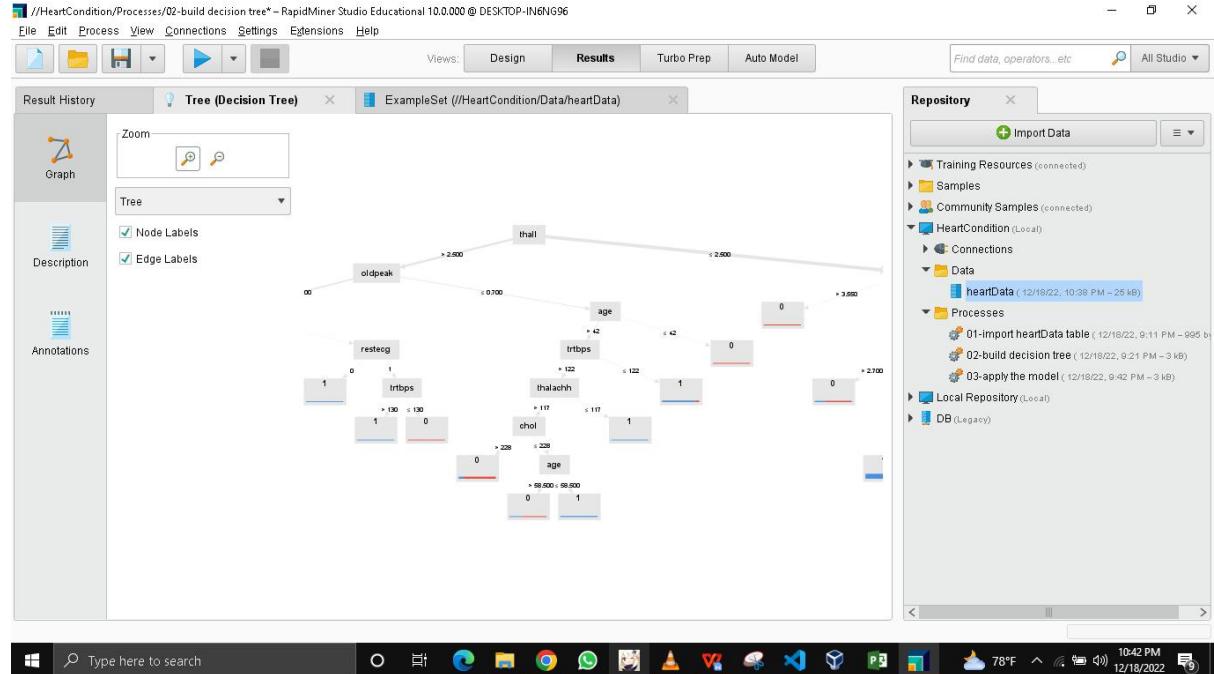
Row No.	output	age	sex	cp	trtbps	chol	fbs	restecg
1	1	63	1	3	145	233	1	0
2	1	37	1	2	130	250	0	1
3	1	41	0	1	130	204	0	0
4	1	56	1	1	120	236	0	1
5	1	57	0	0	120	354	0	1
6	1	57	1	0	140	192	0	1
7	1	56	0	1	140	294	0	0
8	1	44	1	1	120	263	0	1
9	1	52	1	2	172	199	1	1
10	1	57	1	2	150	168	0	1
11	1	54	1	0	140	239	0	1
12	1	48	0	2	130	275	0	1
13	1	49	1	1	130	266	0	1
14	1	64	1	3	110	211	0	0

No changes take place as all the fields under the output column consist of defined values

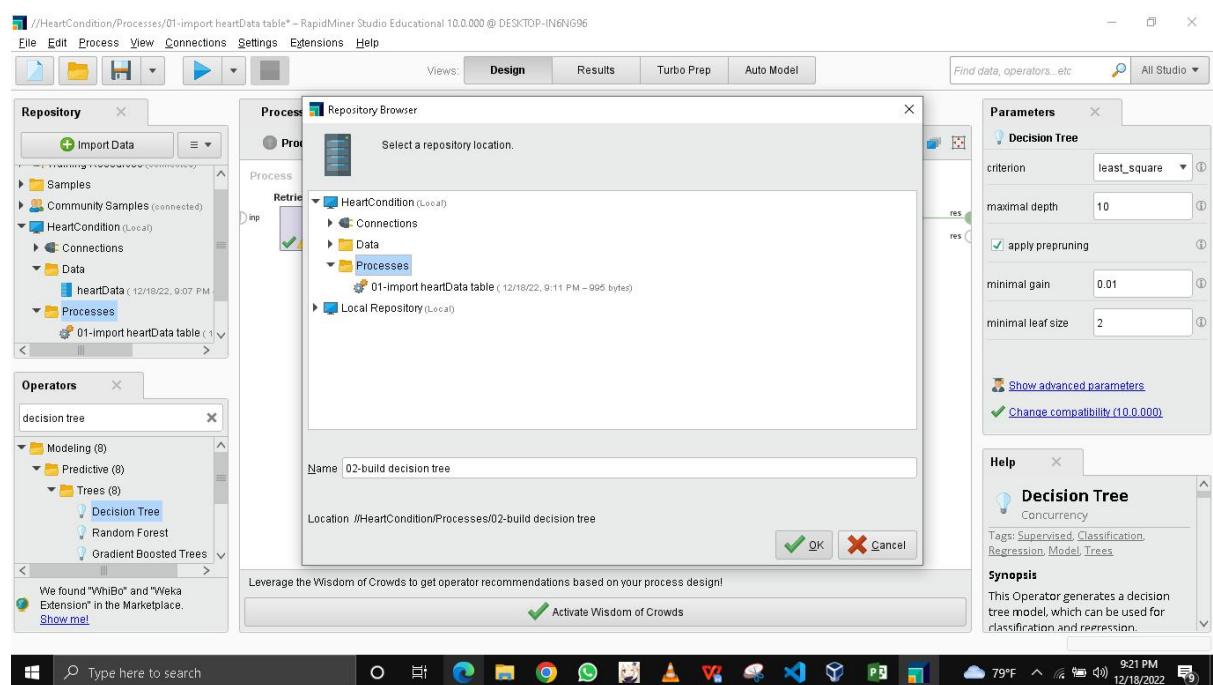
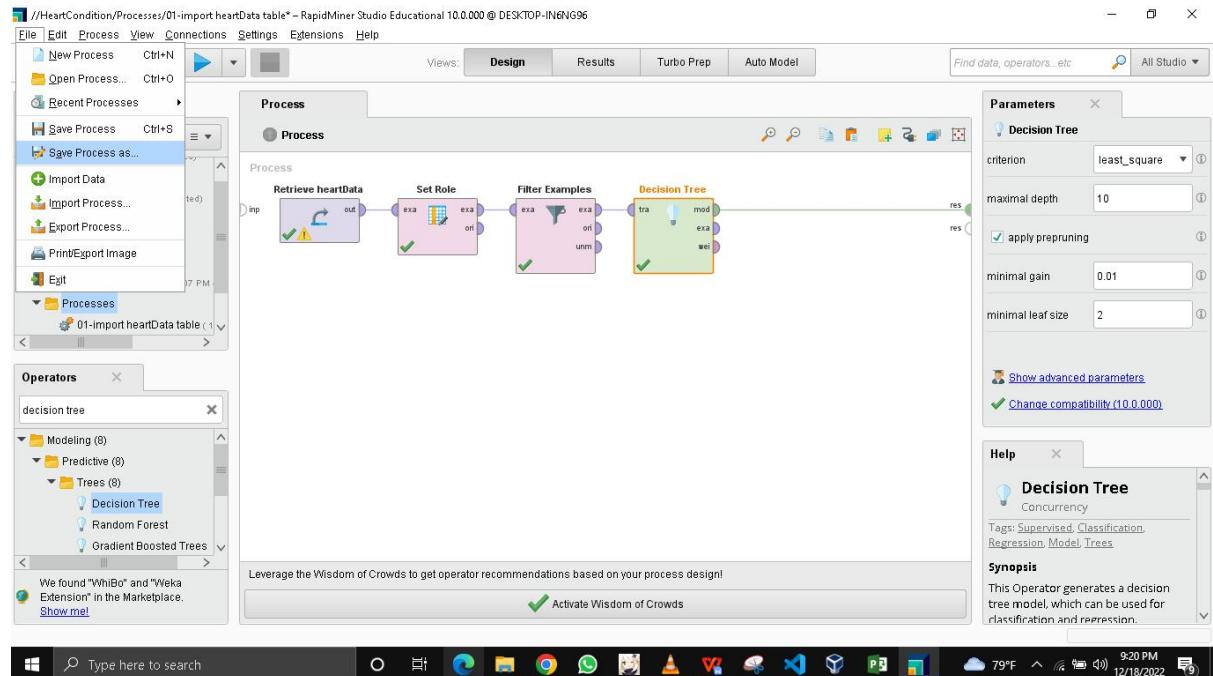
8. Created a Decision Tree of the heartData table using the “Decision Tree” operator after having changed the table attributes to better suited attributes

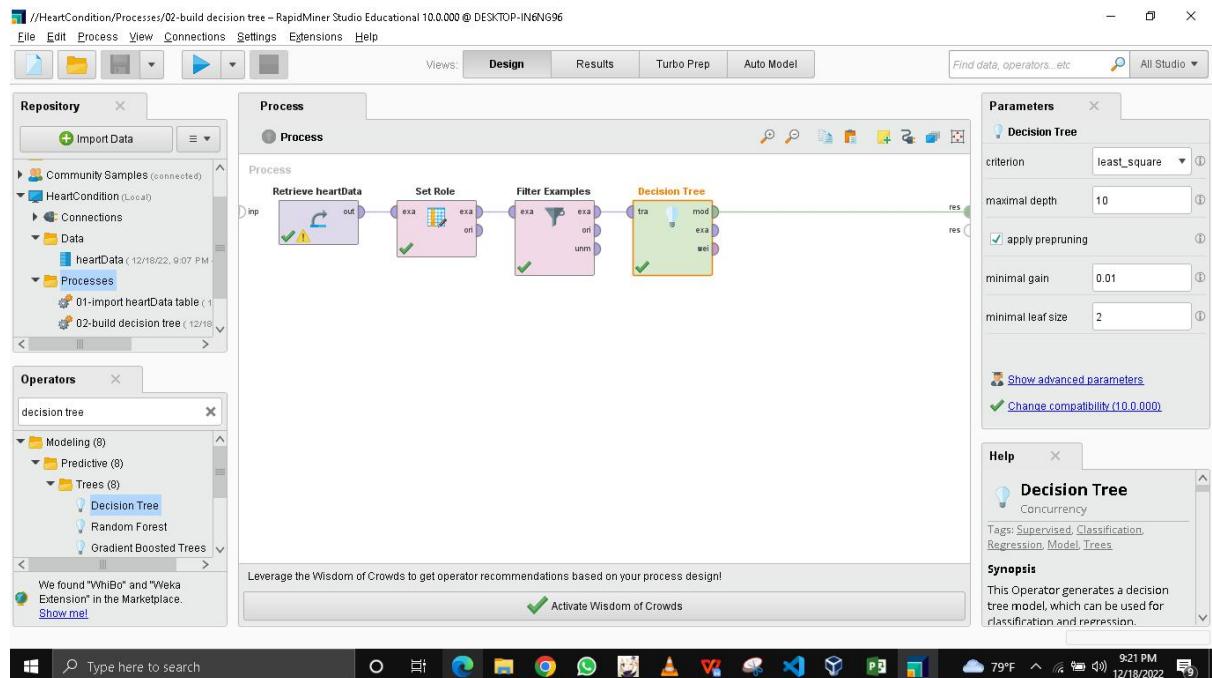




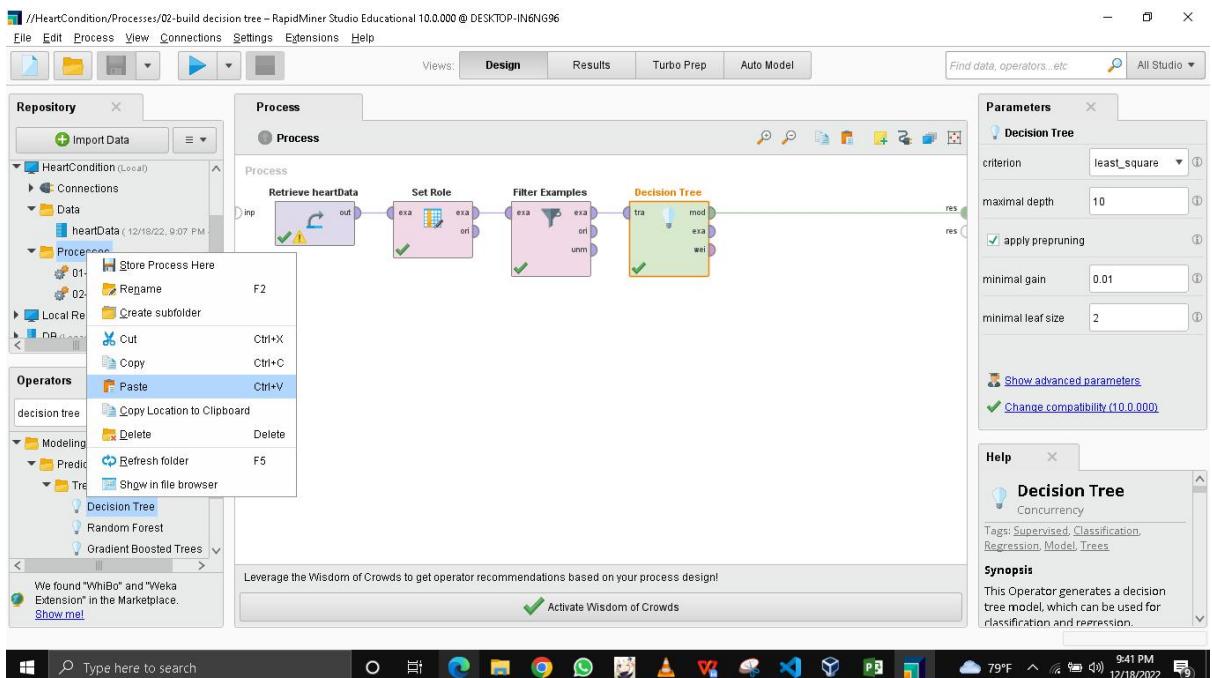


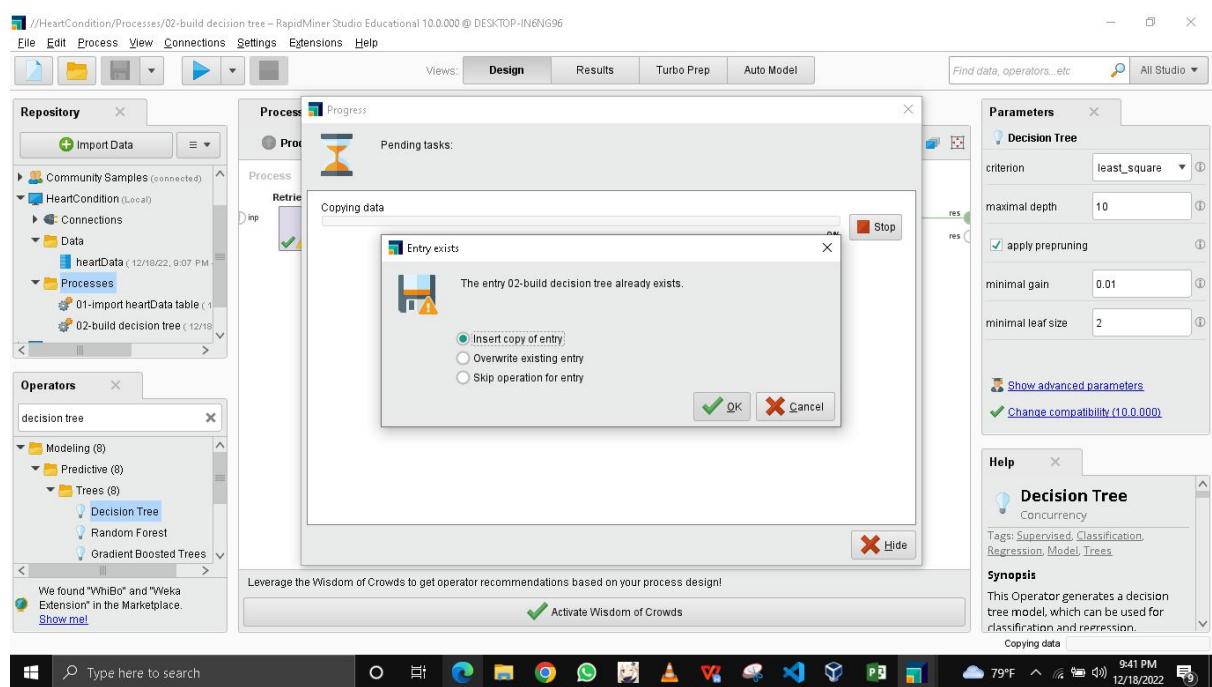
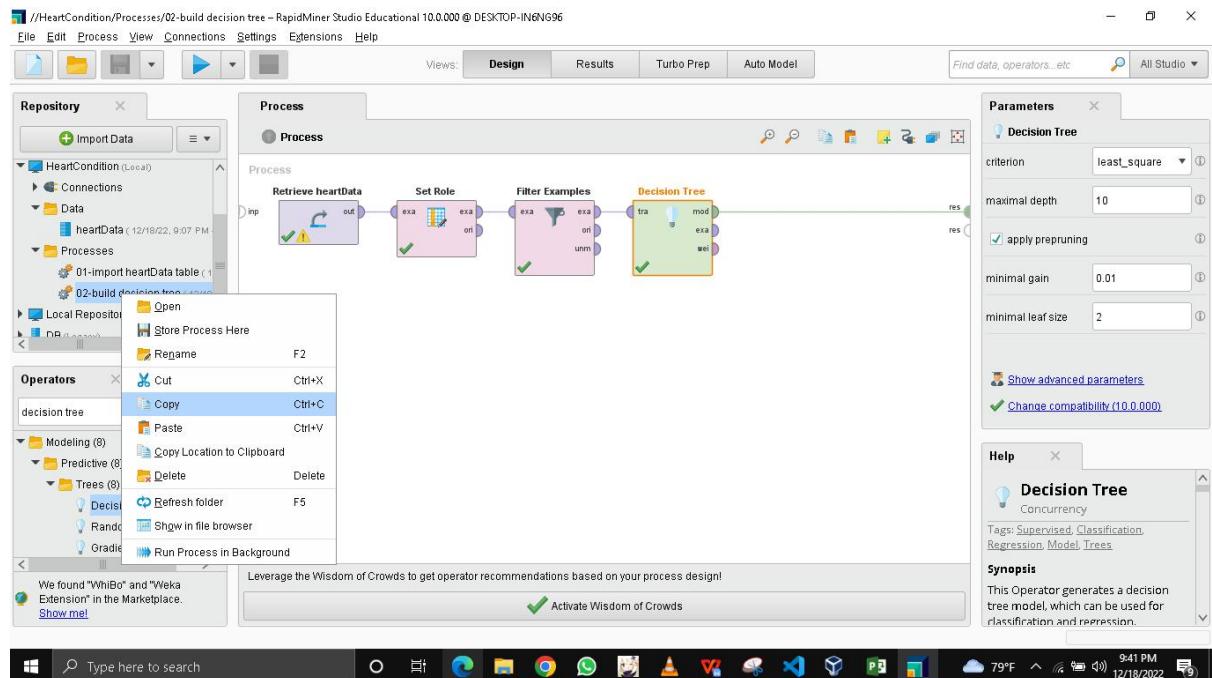
9. Saved the process as “02-build decision tree” in the “Processes” Folder

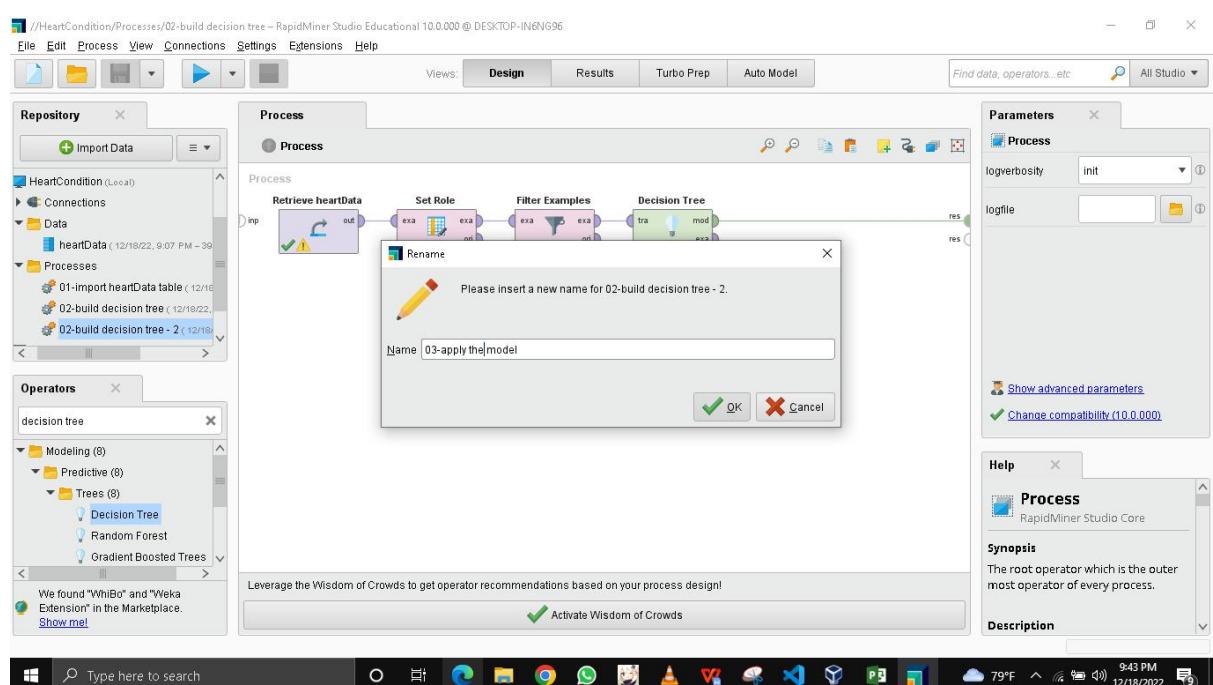
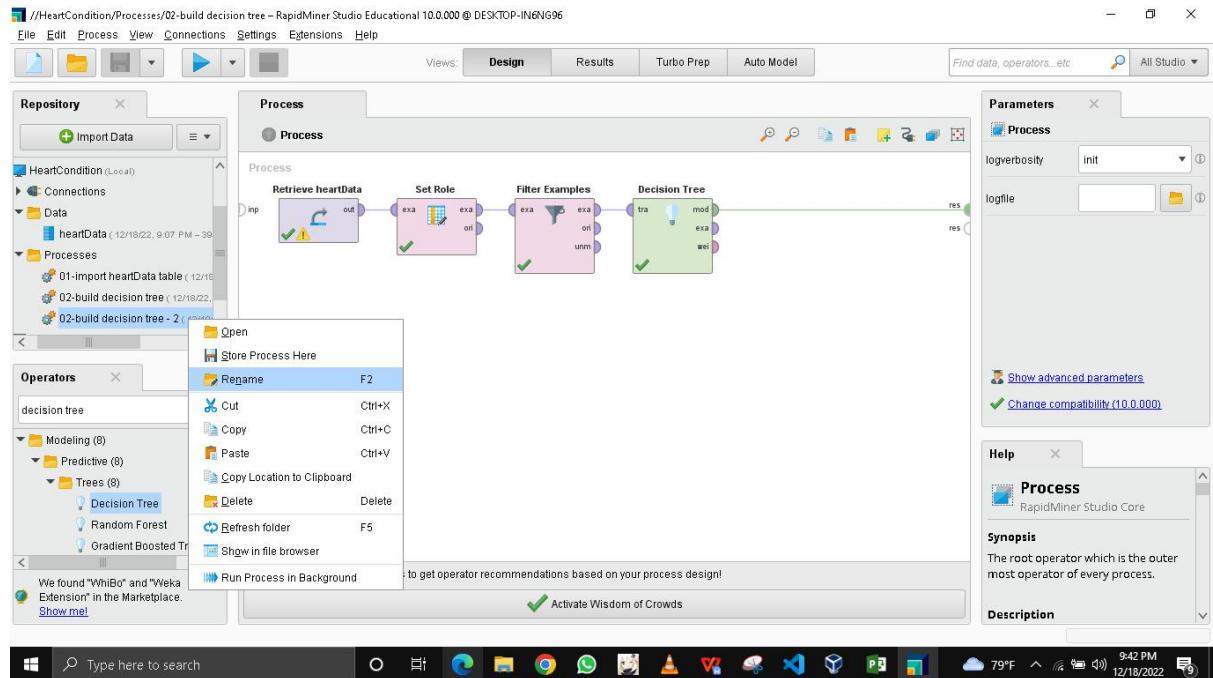


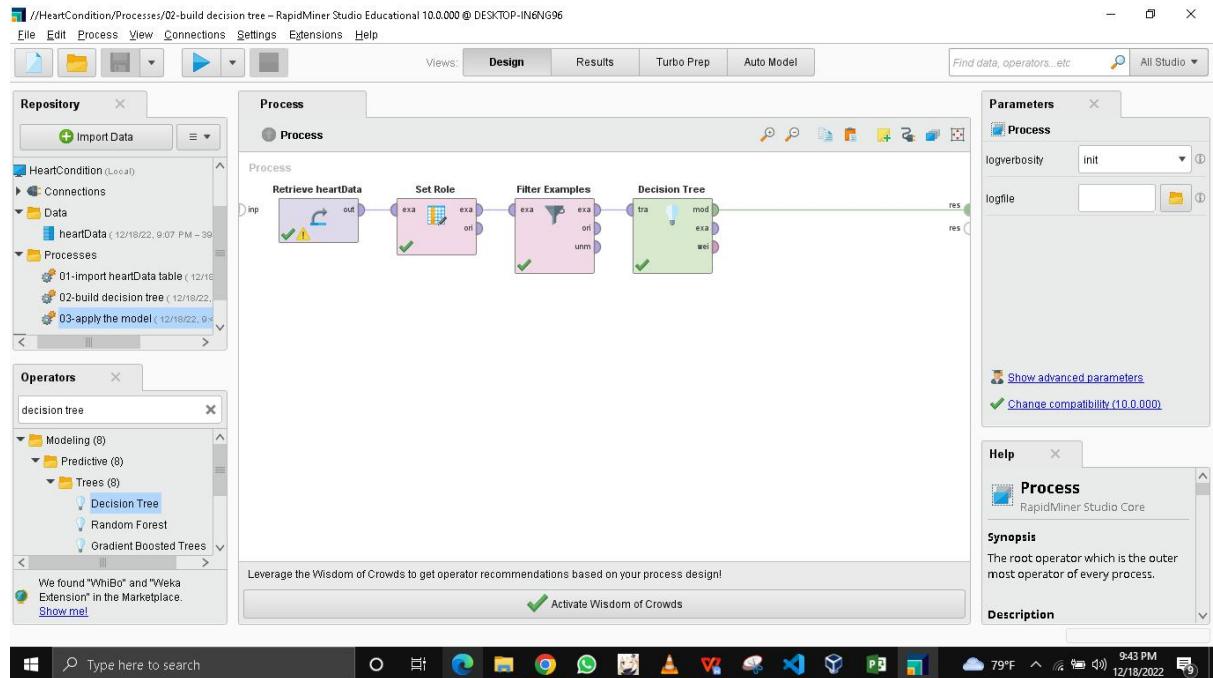


10. Created “03-apply the model” process using a copy of the “02- build decision tree” process

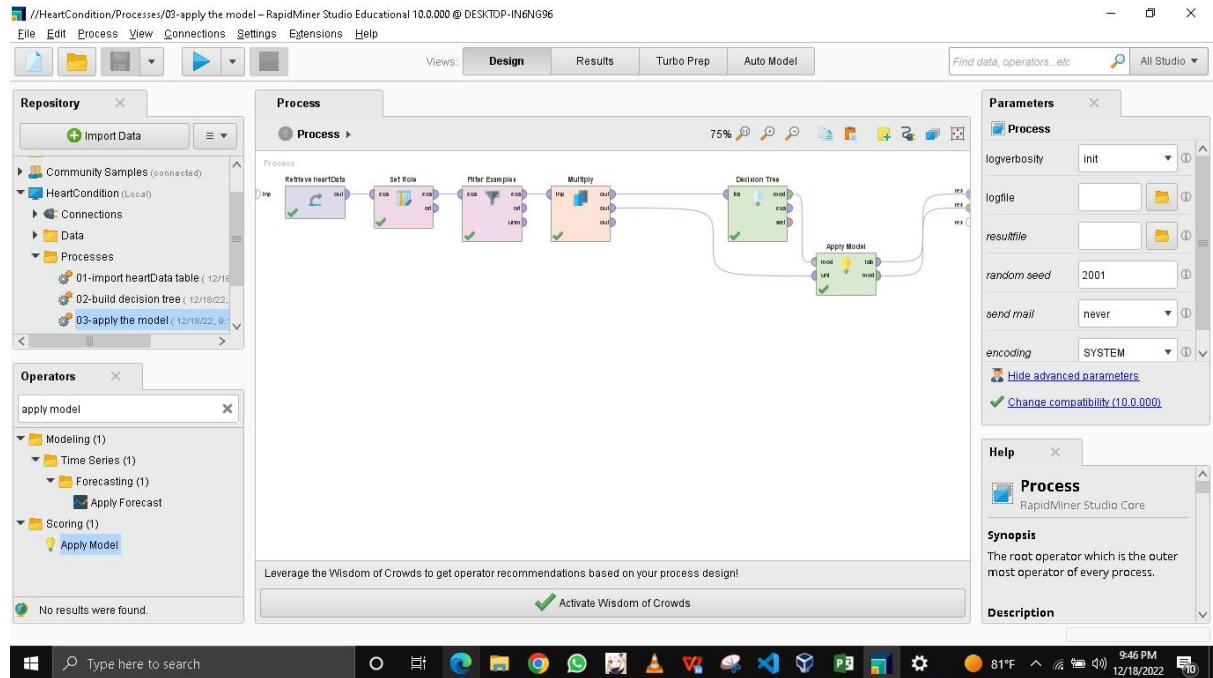








11. Used the “Multiply” and “Apply Model” operators to create a copy of the filtered data, and then get a prediction and the likeliness of the predicted results of the output respectively



//HeartCondition/Processes/03-apply the model – RapidMiner Studio Educational 10.0.000 @ DESKTOP-IN6NG96

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Find data, operators...etc All Studio

Result History Tree (Decision Tree) ExampleSet (Apply Model)

Data Statistics Visualizations Annotations

Row No. ↑	output	prediction(0...)	confidence(1)	confidence(0)	age	sex	cp	trtbs
1	1	1	0.914	0.086	63	1	3	145
2	1	0	0.286	0.714	37	1	2	130
3	1	1	0.914	0.086	41	0	1	130
4	1	1	0.914	0.086	56	1	1	120
5	1	1	0.914	0.086	57	0	0	120
6	1	1	0.914	0.086	57	1	0	140
7	1	1	0.914	0.086	56	0	1	140
8	1	1	0.900	0.100	44	1	1	120
9	1	1	1	0	52	1	2	172
10	1	1	0.914	0.086	57	1	2	150
11	1	1	0.914	0.086	54	1	0	140
12	1	1	0.914	0.086	48	0	2	130
13	1	1	0.914	0.086	49	1	1	130
14	1	1	0.914	0.086	64	1	3	110

ExampleSet (303 examples, 4 special attributes, 13 regular attributes)

Repository Import Data Training Resources (connected) Samples Community Samples (connected) HeartCondition (Local) Connections Data Processes Local Repository (Local) DB (Legacy)

//HeartCondition/Processes/03-apply the model – RapidMiner Studio Educational 10.0.000 @ DESKTOP-IN6NG96

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Find data, operators...etc All Studio

Result History Tree (Decision Tree) ExampleSet (Apply Model)

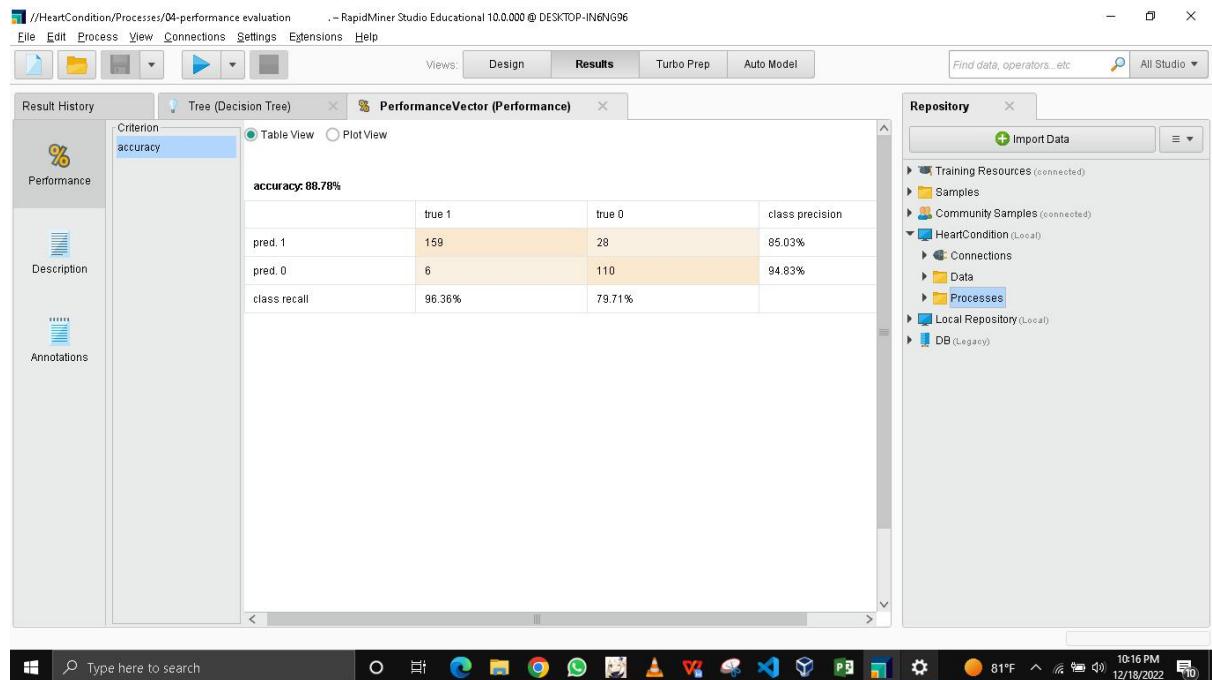
Data Statistics Visualizations Annotations

fbs	restecg	thalachh	exmg	oldpeak	sip	caa ↓	thall
0	1	169	0	0	2	4	2
0	1	144	0	0.400	1	4	3
0	1	173	0	0	2	4	2
0	1	173	0	0	2	4	2
1	0	143	1	0.100	1	4	3
0	1	146	0	1.800	1	3	3
1	1	147	0	0.100	2	3	3
1	0	173	0	0	2	3	2
0	0	108	1	1.500	1	3	2
0	0	114	0	1	1	3	3
0	0	131	1	2.200	1	3	3
0	0	145	0	6.200	0	3	3
0	1	139	0	2	1	3	3
1	0	132	1	1.800	2	3	3

ExampleSet (303 examples, 4 special attributes, 13 regular attributes)

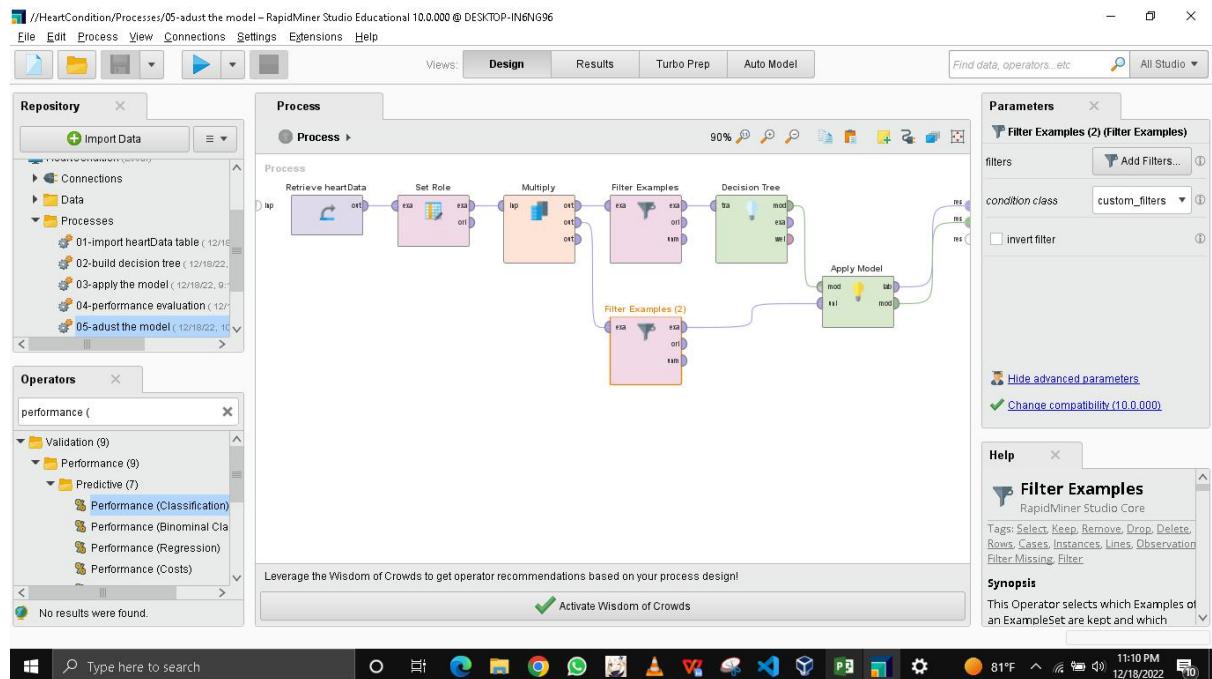
Repository Import Data Training Resources (connected) Samples Community Samples (connected) HeartCondition (Local) Connections Data Processes Local Repository (Local) DB (Legacy)

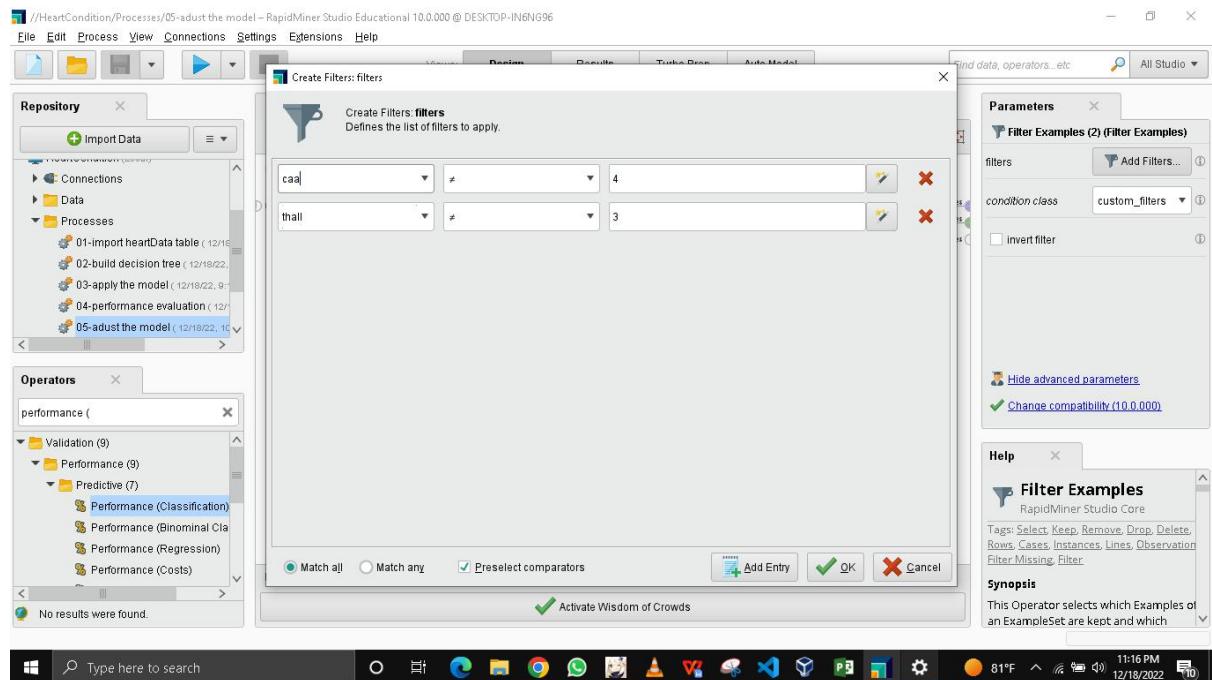
11. Added the “Performance (Classification)” operator to the model to evaluate the model’s performance in terms of overall prediction accuracy and saved the process as “04-performance evaluation”



As indicated by the screenshot above the overall prediction accuracy of the current model is at 88.78% however this is not indicative of realistic results

12. Created an “05-adjust the model” process with the “Multiply” operator repurposed and re-positioned in order to be able to feed the same set of data into differently separate filters with the parameters of filtration for the bottom filter as shown in the following screen captures below





With the reason being to eliminate and prevent undefined anomalies from interfering with the accuracy of the results as the defined range for the category "caa" as shown below is only in between (0-3) with the value 4 not included in the predefined range

- caa: number of major vessels coloured by fluroscopy (0-3)

Whilst the category labeled "thall" consists only of definitions for the numbers 0-2 without a definition existing for the undefined integer 3 as shown below

- thall: 0=normal; 1=fixed defect; 2=reversible defect

With these anomalies taken into consideration the results of the post filtered data are as follows with the number of examples having decreased from 303 to 183 in the ExampleSet

Result History Tree (Decision Tree) ExampleSet (Apply Model)

Data Statistics Visualizations Annotations

Open in: Turbo Prep Auto Model

Filter (183 / 183 examples): all

Row No.	output	prediction(0...)	confidence(1)	confidence(0)	age	sex	cp	trtbs
1	1	1	0.914	0.086	63	1	3	145
2	1	0	0.286	0.714	37	1	2	130
3	1	1	0.914	0.086	41	0	1	130
4	1	1	0.914	0.086	56	1	1	120
5	1	1	0.914	0.086	57	0	0	120
6	1	1	0.914	0.086	57	1	0	140
7	1	1	0.914	0.086	56	0	1	140
8	1	1	0.914	0.086	57	1	2	150
9	1	1	0.914	0.086	54	1	0	140
10	1	1	0.914	0.086	48	0	2	130
11	1	1	0.914	0.086	49	1	1	130
12	1	1	0.914	0.086	64	1	3	110
13	1	1	0.914	0.086	58	0	3	150
14	1	1	0.914	0.086	50	0	2	120

ExampleSet (183 examples, 4 special attributes, 13 regular attributes)

Result History Tree (Decision Tree) ExampleSet (Apply Model)

Data Statistics Visualizations Annotations

Open in: Turbo Prep Auto Model

Filter (183 / 183 examples): all

fbs	restecg	thalachh	exng	oldpeak	slp	caa ↓	thall
1	0	173	0	0	2	3	2
0	0	108	1	1.500	1	3	2
0	0	109	0	2.400	1	3	2
0	0	162	1	0	2	3	2
0	1	120	1	0	1	3	1
1	1	106	0	1.900	1	3	2
0	0	126	0	0.800	2	3	2
0	2	140	0	4.400	0	3	1
0	1	151	0	1.800	2	2	2
0	1	162	0	0.400	2	2	2
1	1	175	0	0	2	2	2
0	0	178	0	0.800	2	2	2
0	1	179	0	0	2	2	2
0	1	122	0	2	1	2	2

ExampleSet (183 examples, 4 special attributes, 13 regular attributes)

13. Tested the adjusted model using the “Performance (Classification)” operator and saved the process as “06-performance re-evaluation” with the results indicating an accuracy of 84.15% as shown below

//HeartCondition/Processes/06-performance re-evaluation - RapidMiner Studio Educational 10.0.000 @ DESKTOP-IN6NG96

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model

Find data, operators...etc All Studio

Result History Tree (Decision Tree) PerformanceVector (Performance)

Criterion accuracy

accuracy: 84.15%

	true 1	true 0	class precision
pred. 1	131	26	83.44%
pred. 0	3	23	88.46%
class recall	97.76%	46.94%	

Repository

- Import Data
- Training Resources (connected)
- Samples
- Community Samples (connected)
- HeartCondition (Local)
 - Connections
 - Data
 - Processes
- Local Repository (Local)
- DB (Legacy)

Type here to search

11:29 PM 81°F 12/18/2022