

# **COMSATS UNIVERSITY ISLAMABAD**



## **SUBMITTED BY**

**NAME:** NOOR UL AIN

**REG. NO.:** FA19-BSE-168

**SECTION:** BSE-7C

## **SUBMITTED TO**

**TEACHER:** Ms. MAMUNA FATIMA

**SUBJECT:** DATA WAREHOUSING & DATA MINING

**LAB TASK**

# Table of Contents

K-means clustering:.....	3
classification through naïve bayes: .....	5
association rule mining:.....	7
k-mediod clustering:.....	9
classification through svm:.....	11
preprocessing on twitter covid dataset: .....	13

## K-MEANS CLUSTERING:

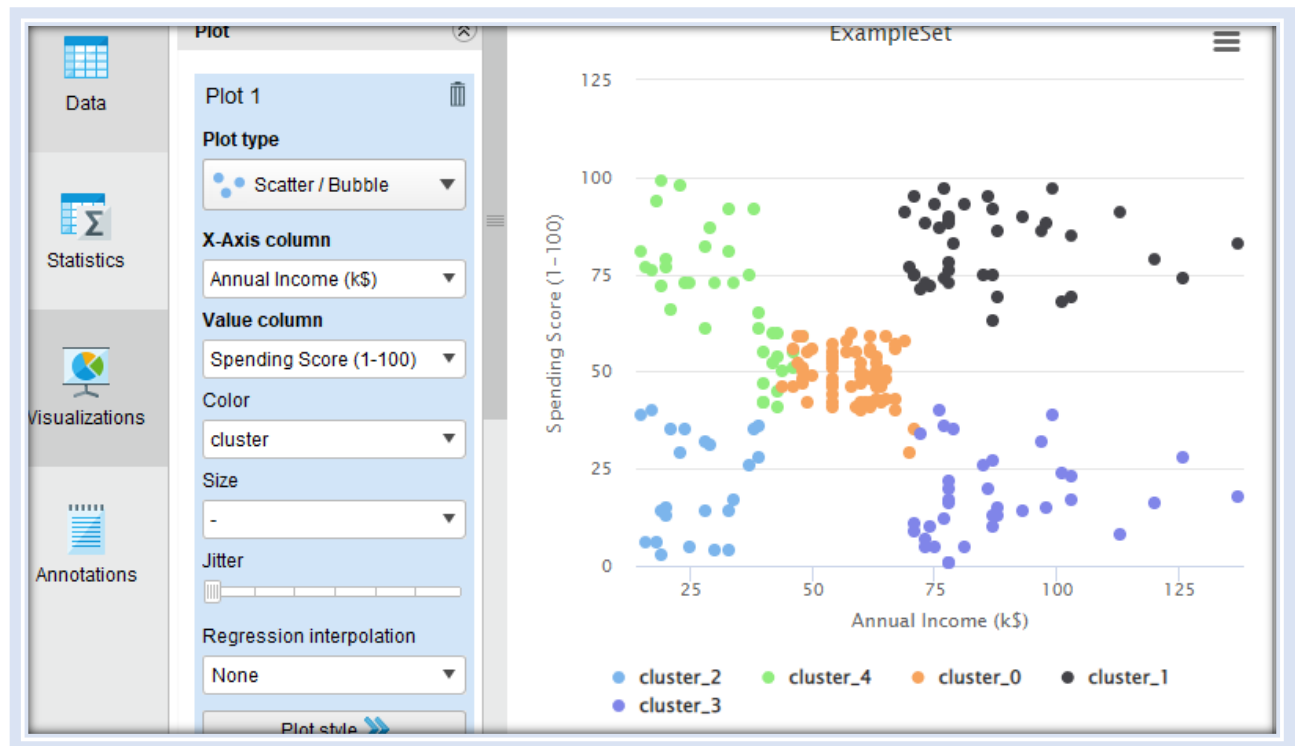
### DATASET BEFORE CLUSTERING:



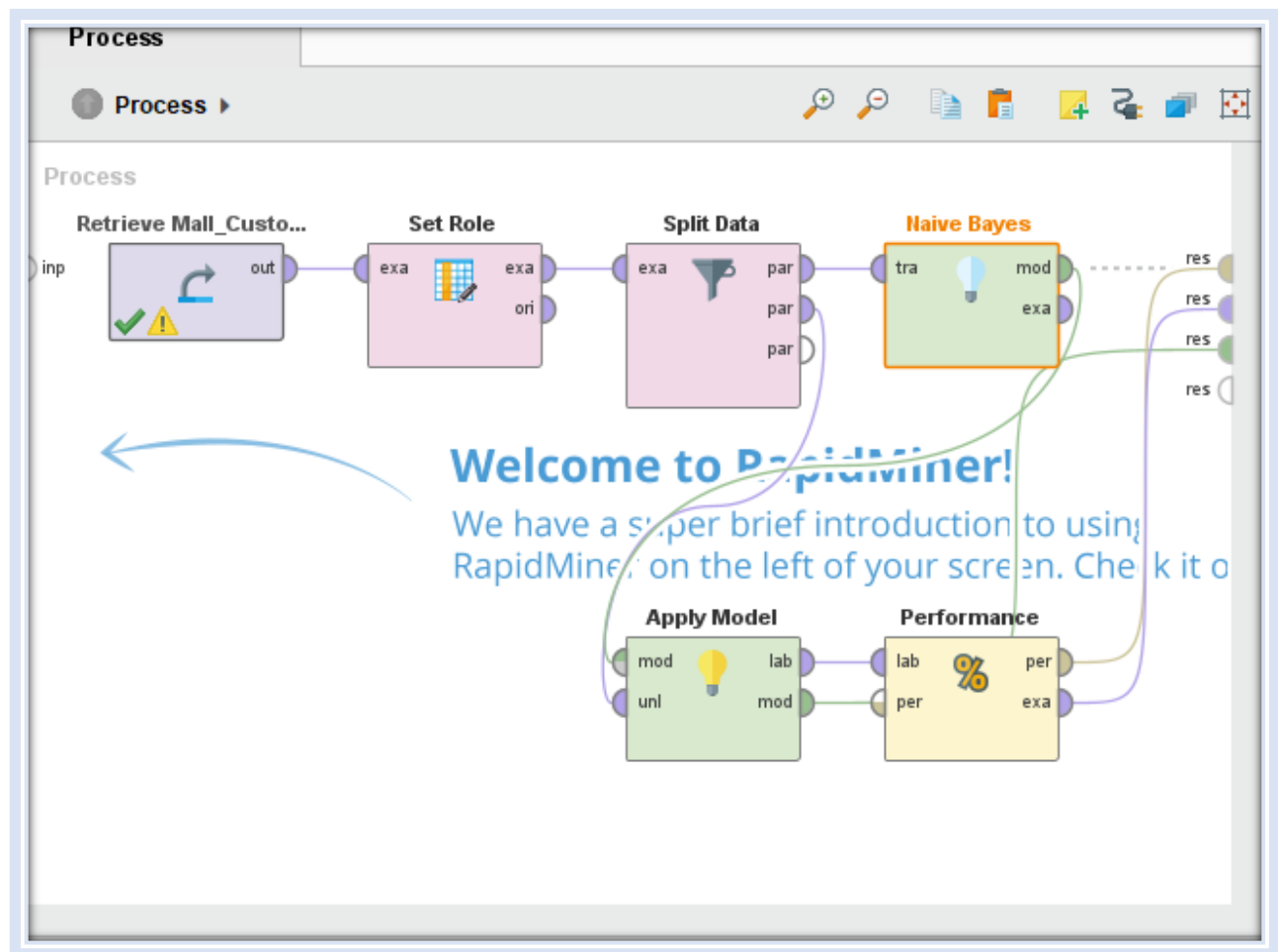
### Cluster Model

```
Cluster 0: 65 items
Cluster 1: 39 items
Cluster 2: 23 items
Cluster 3: 36 items
Cluster 4: 37 items
Total number of items: 200
```

## DATASET AFTER CLUSTERING:



## CLASSIFICATION THROUGH NAÏVE BAYES:



### SimpleDistribution

Distribution model for label attribute Gender

Class Male (0.443)

4 distributions

Class Female (0.557)

4 distributions

## CONFUSION MATRIX FOR PERFORMANCE:

PerformanceVector (Performance) ExampleSet (//Local Repository/Mall\_Customers)

history SimpleDistribution (Naive Bayes) ExampleSet (Apply Model)

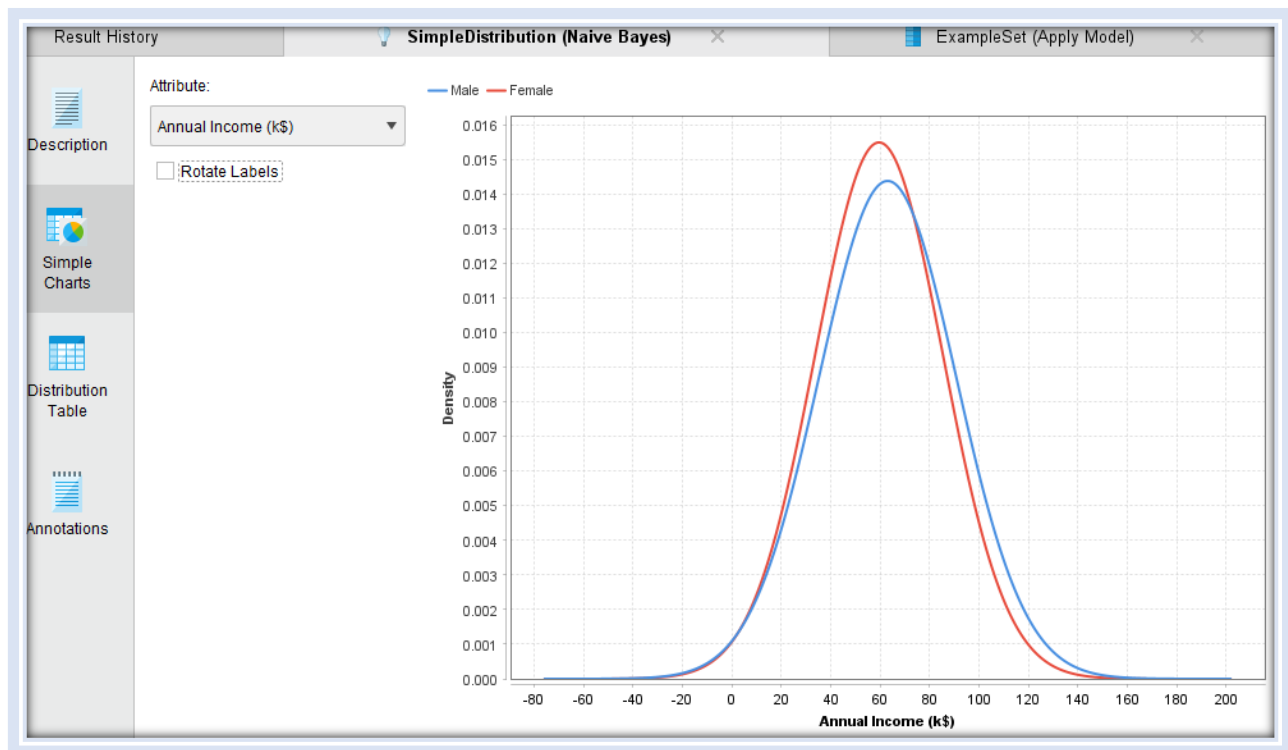
Criterion  
accuracy

Table View Plot View

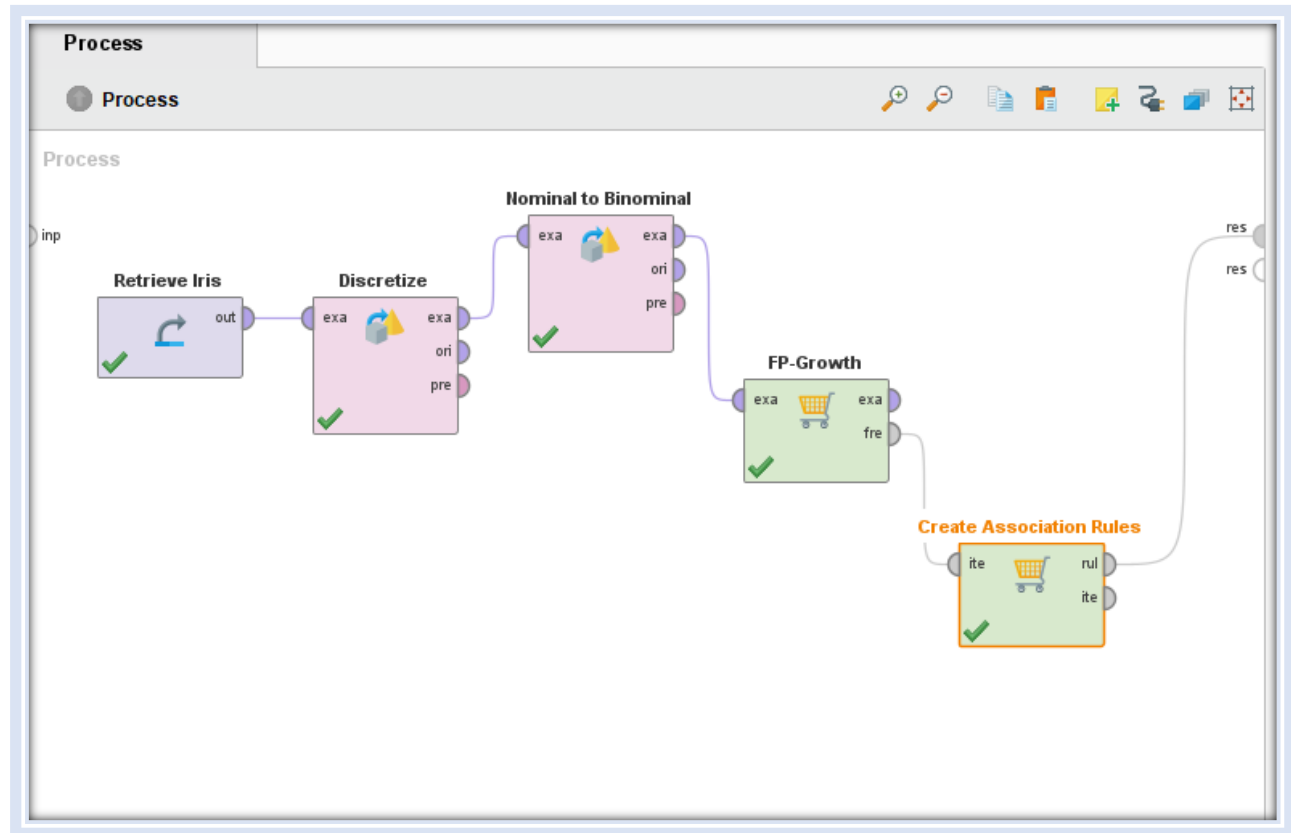
accuracy: 60.00%

	true Male	true Female	class precision
pred. Male	12	10	54.55%
pred. Female	14	24	63.16%
class recall	46.15%	70.59%	

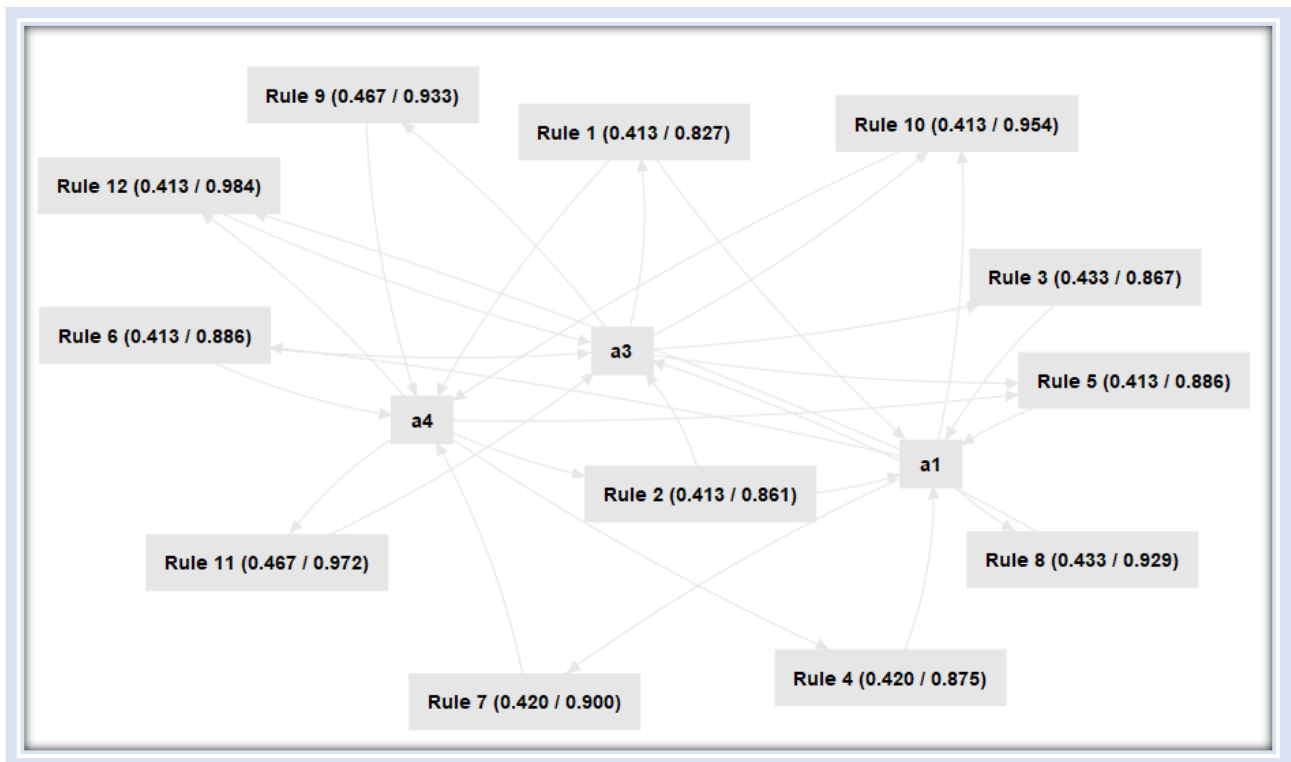
## VISUALIZATION:



## ASSOCIATION RULE MINING:



No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain
2	a4	a3, a1	0.413	0.861	0.955	-0.547
3	a3	a1	0.433	0.867	0.956	-0.567
4	a4	a1	0.420	0.875	0.959	-0.540
5	a3, a4	a1	0.413	0.886	0.964	-0.520
6	a1	a3, a4	0.413	0.886	0.964	-0.520
7	a1	a4	0.420	0.900	0.968	-0.513
8	a1	a3	0.433	0.929	0.977	-0.500
9	a3	a4	0.467	0.933	0.978	-0.533
10	a3, a1	a4	0.413	0.954	0.986	-0.453
11	a4	a3	0.467	0.972	0.991	-0.493
12	a4, a1	a3	0.413	0.984	0.995	-0.427



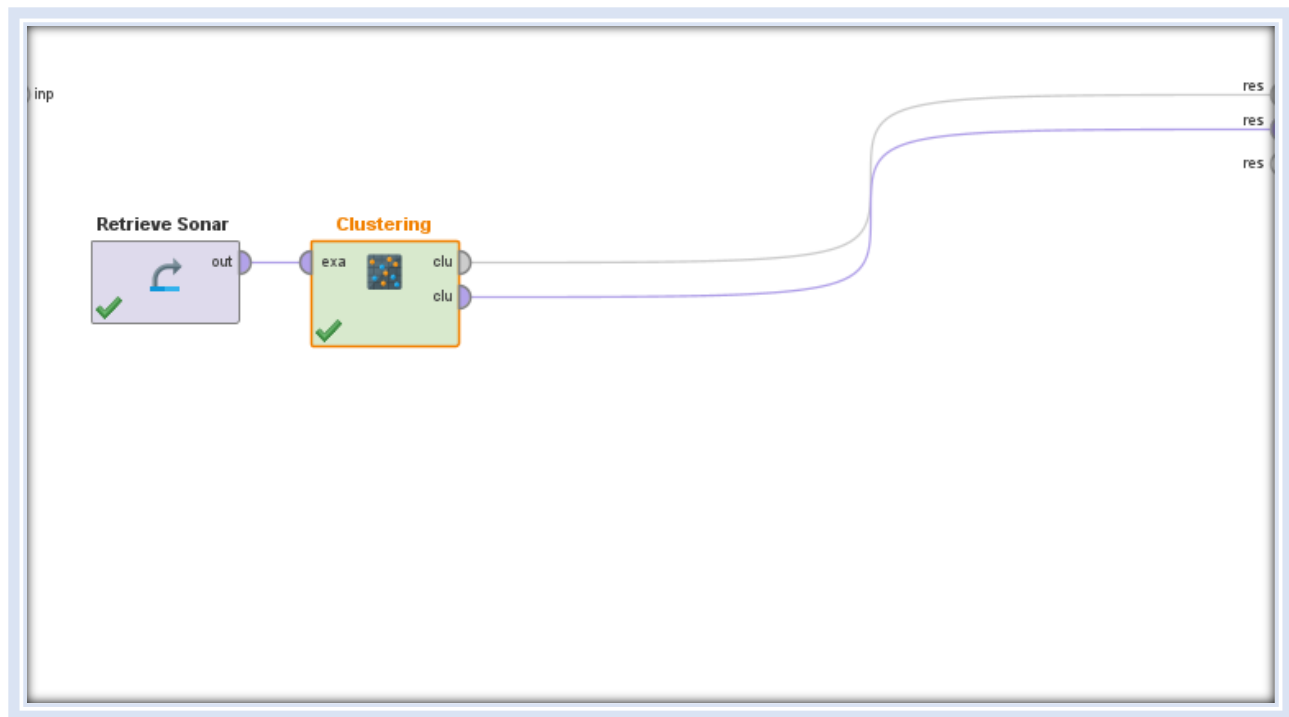
## AssociationRules

### Association Rules

```
[a3] --> [a4, a1] (confidence: 0.827)
[a4] --> [a3, a1] (confidence: 0.861)
[a3] --> [a1] (confidence: 0.867)
[a4] --> [a1] (confidence: 0.875)
[a3, a4] --> [a1] (confidence: 0.886)
[a1] --> [a3, a4] (confidence: 0.886)
[a1] --> [a4] (confidence: 0.900)
[a1] --> [a3] (confidence: 0.929)
[a3] --> [a4] (confidence: 0.933)
[a3, a1] --> [a4] (confidence: 0.954)
[a4] --> [a3] (confidence: 0.972)
[a4, a1] --> [a3] (confidence: 0.984)
```



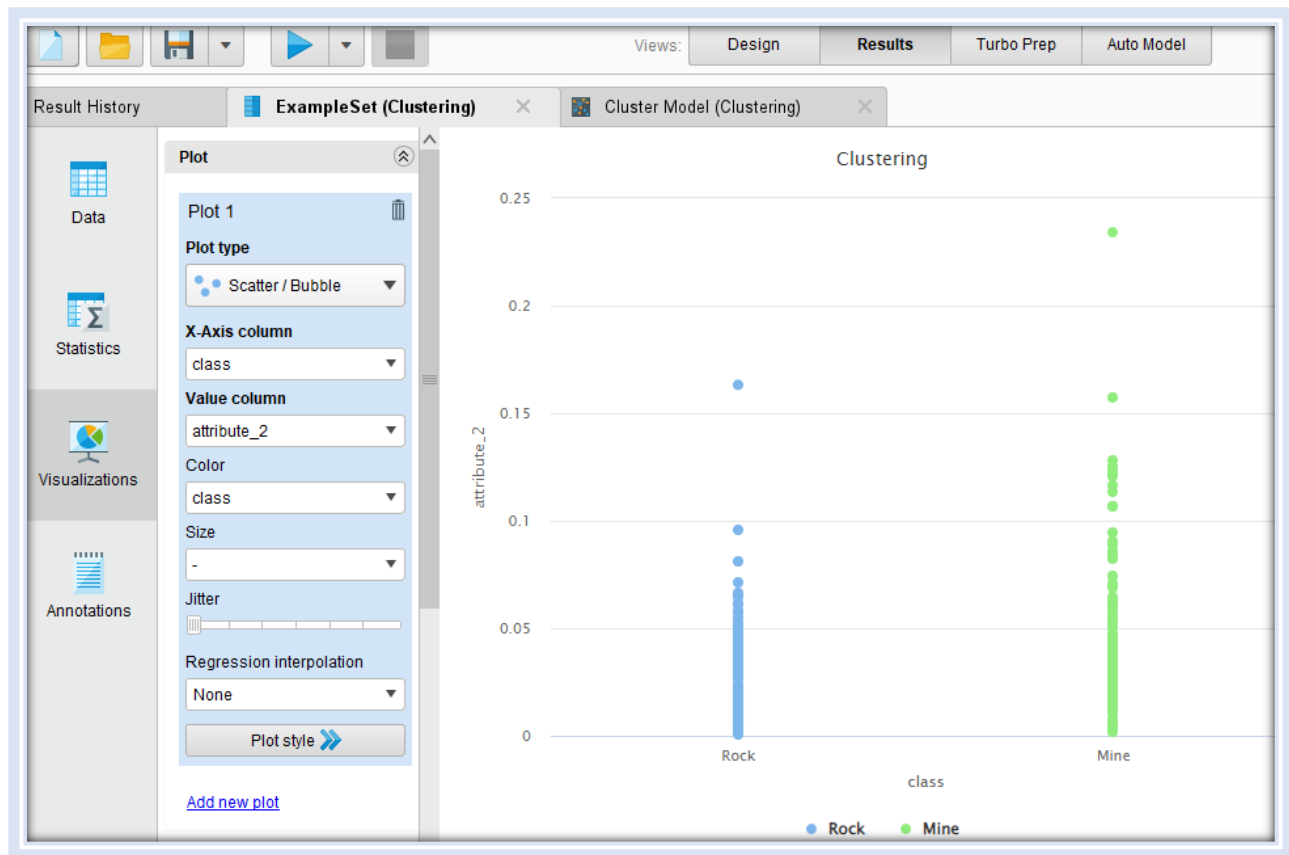
## K-MEDIOD CLUSTERING:



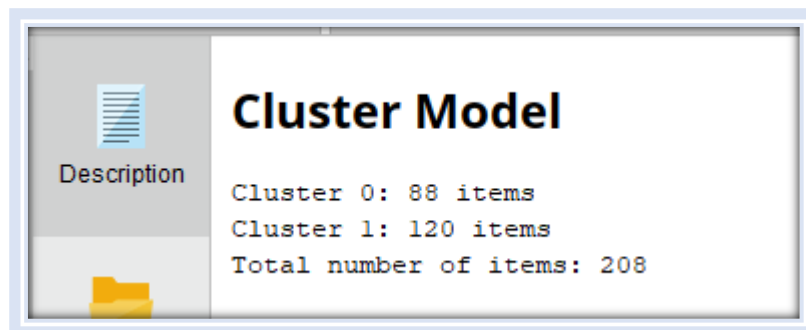
Open in [Turbo Prep](#) [Auto Model](#) Filter (208 / 208 examples): [all](#)

Row No.	id	class	cluster	attribute_1	attribute_2	attribute_3	attribute_4	attribute_5
1	1	Rock	cluster_1	0.020	0.037	0.043	0.021	0.095
2	2	Rock	cluster_0	0.045	0.052	0.084	0.069	0.118
3	3	Rock	cluster_0	0.026	0.058	0.110	0.108	0.097
4	4	Rock	cluster_1	0.010	0.017	0.062	0.021	0.021
5	5	Rock	cluster_0	0.076	0.067	0.048	0.039	0.059
6	6	Rock	cluster_0	0.029	0.045	0.028	0.017	0.038
7	7	Rock	cluster_1	0.032	0.096	0.132	0.141	0.167
8	8	Rock	cluster_0	0.052	0.055	0.084	0.032	0.116
9	9	Rock	cluster_0	0.022	0.037	0.048	0.048	0.065
10	10	Rock	cluster_1	0.016	0.017	0.035	0.007	0.019
11	11	Rock	cluster_1	0.004	0.006	0.015	0.034	0.031
12	12	Rock	cluster_0	0.012	0.031	0.017	0.031	0.036
13	13	Rock	cluster_0	0.008	0.009	0.005	0.025	0.034
14	14	Rock	cluster_1	0.009	0.006	0.025	0.049	0.120

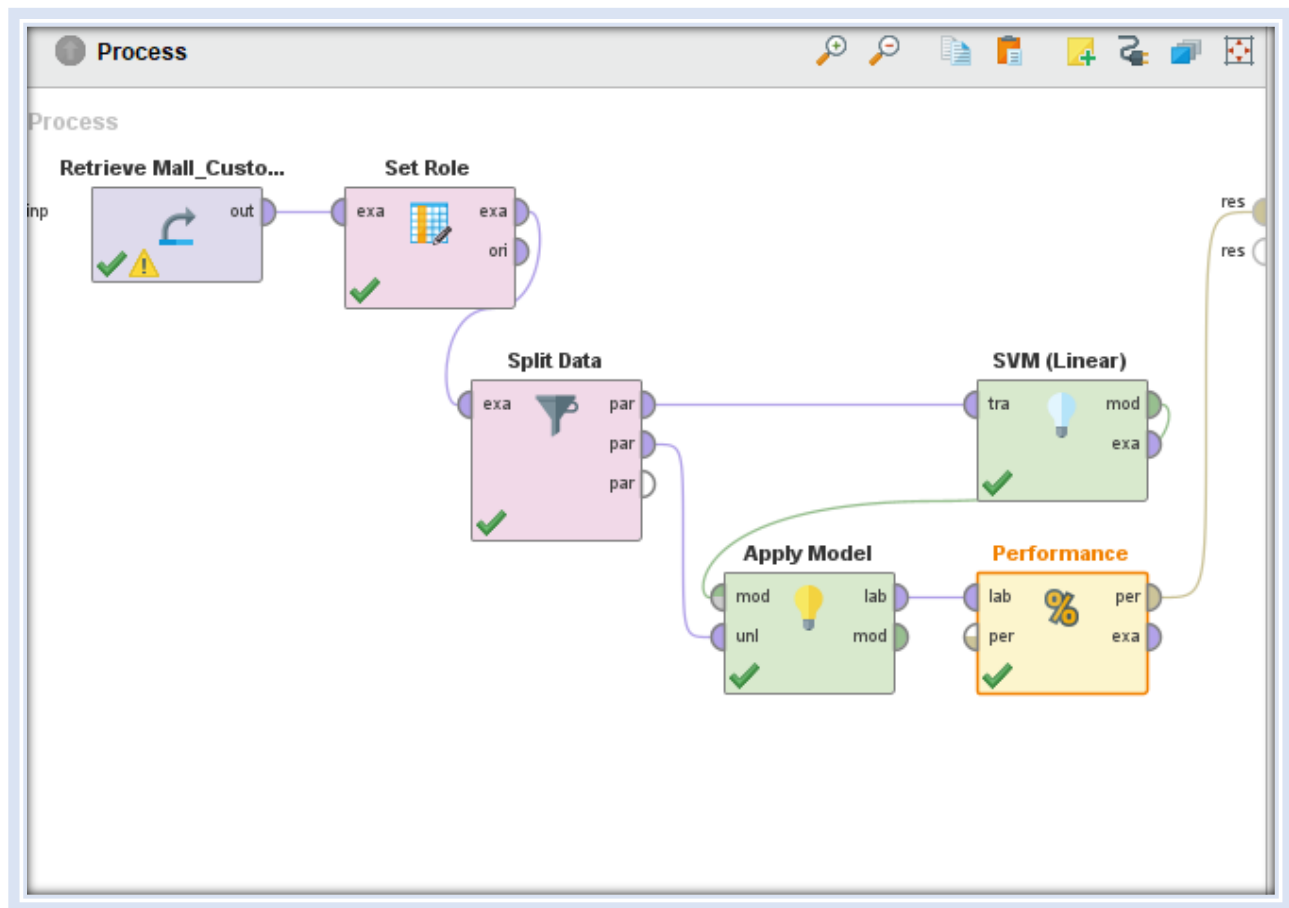
## Visualization:



## Description:



## CLASSIFICATION THROUGH SVM:



## CONFUSION MATRIX:

PerformanceVector (Performance)

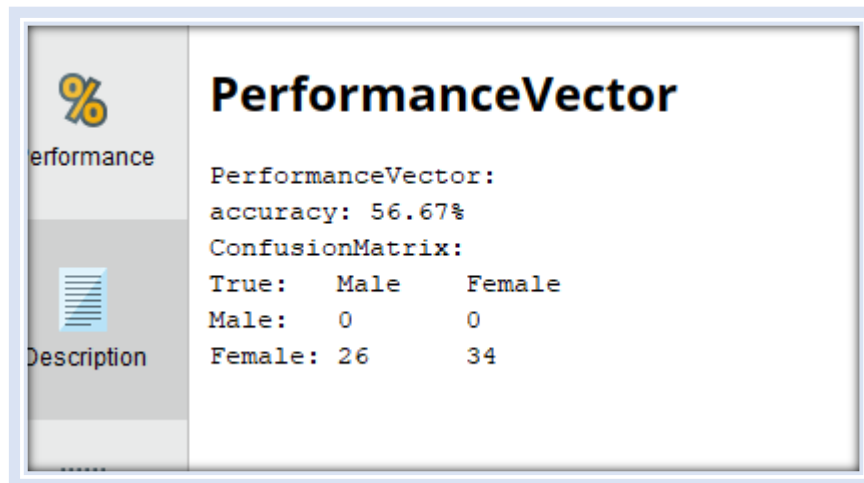
Criterion: accuracy

Table View Plot View

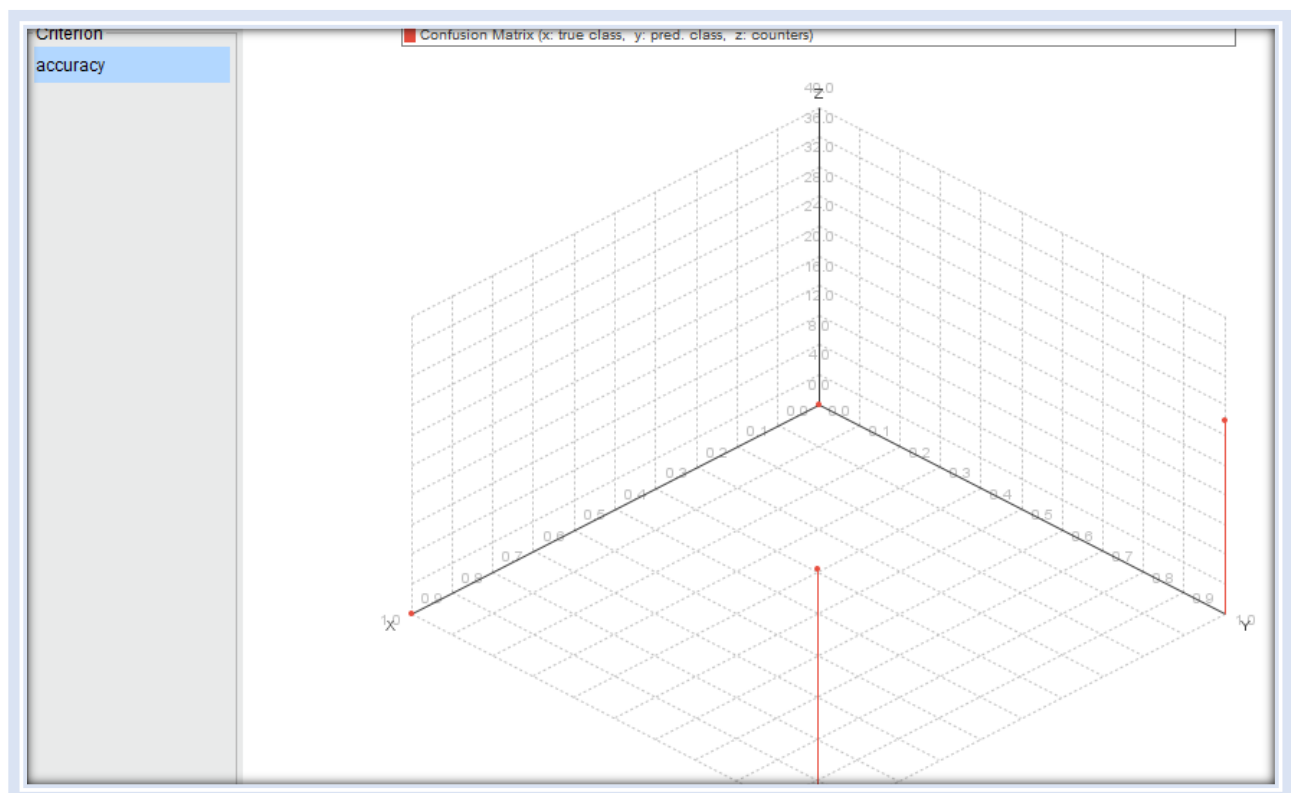
accuracy: 56.67%

	true Male	true Female	class precision
pred. Male	0	0	0.00%
pred. Female	26	34	56.67%
class recall	0.00%	100.00%	

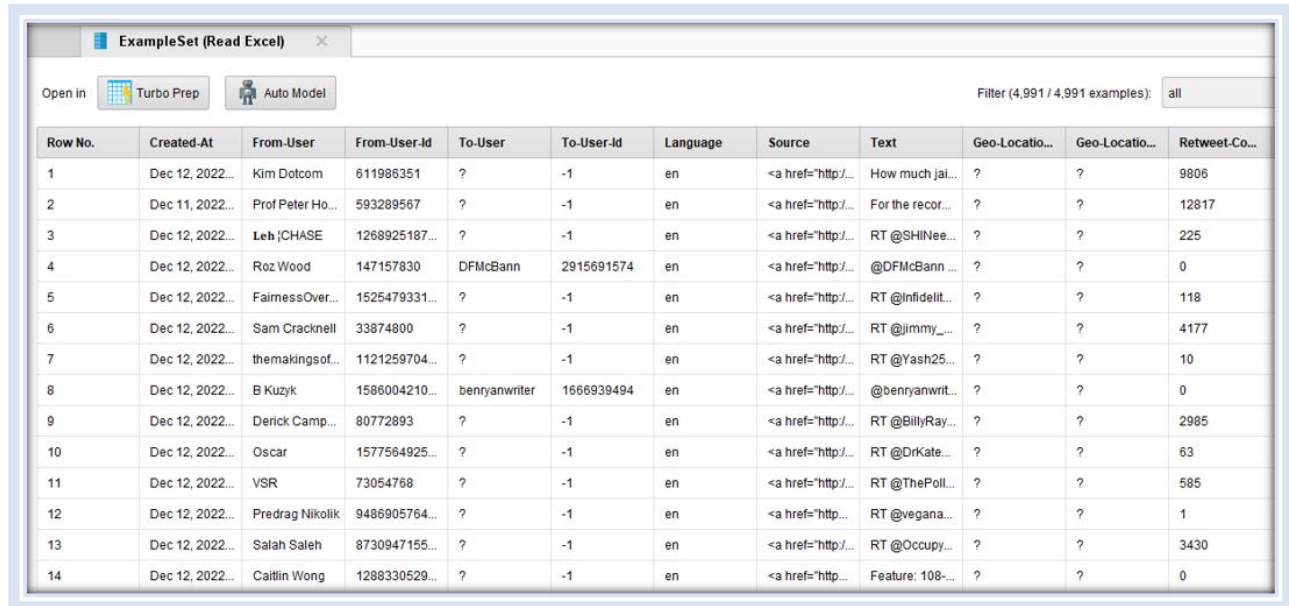
## DESCRIPTION:



## VISUALIZATION:

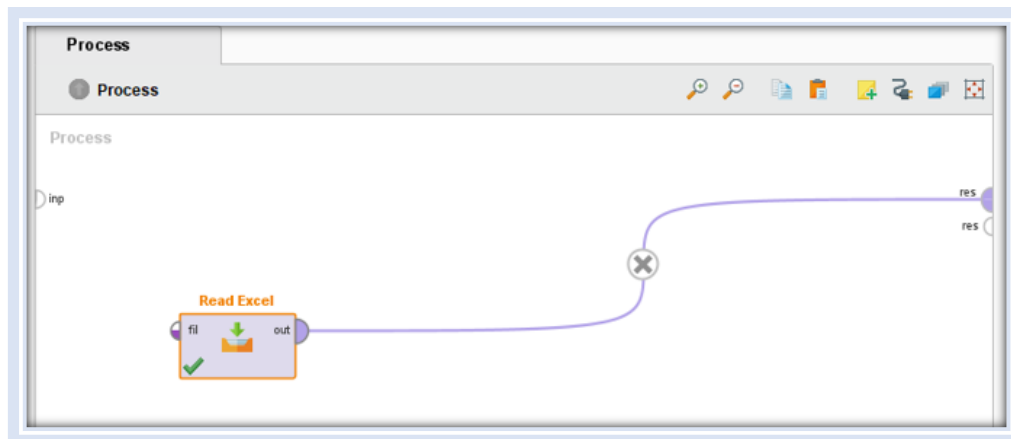


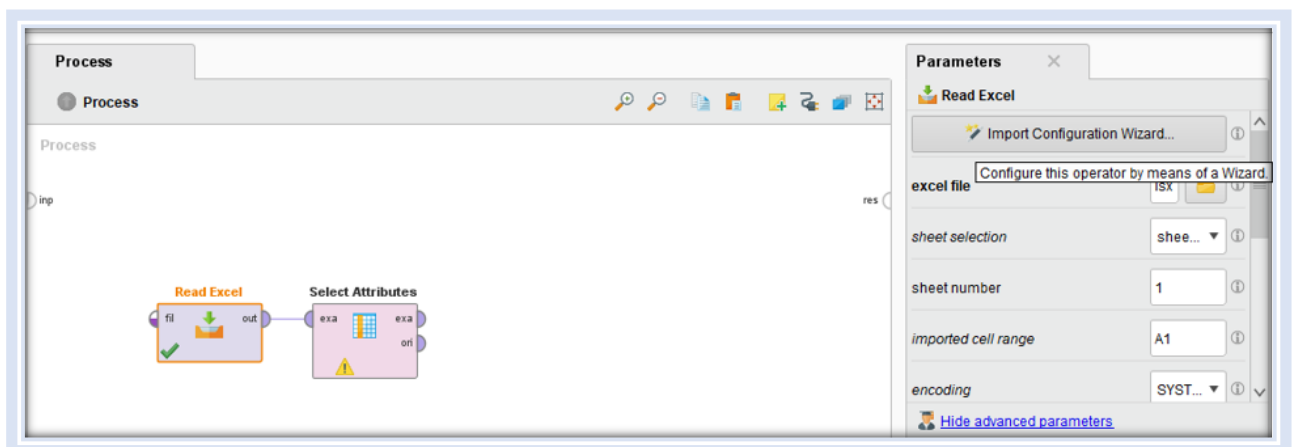
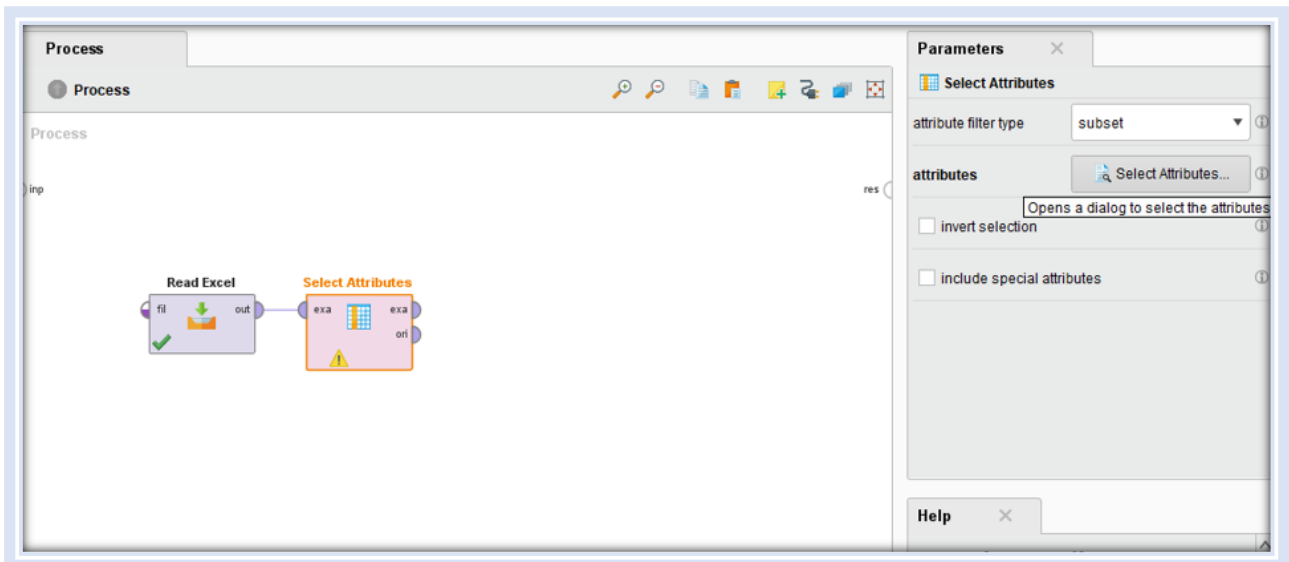
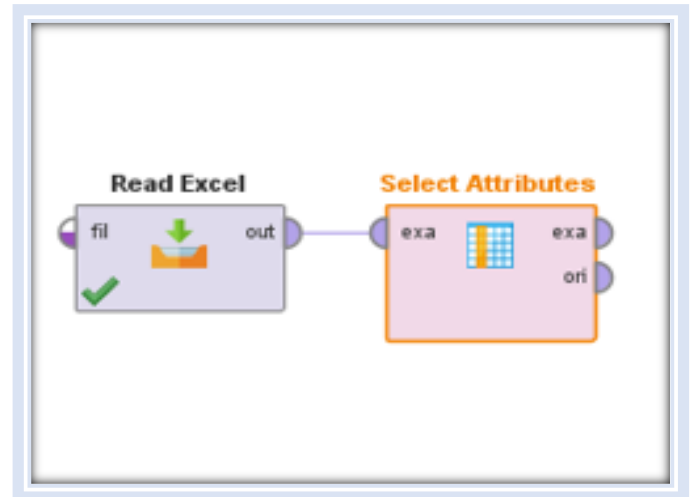
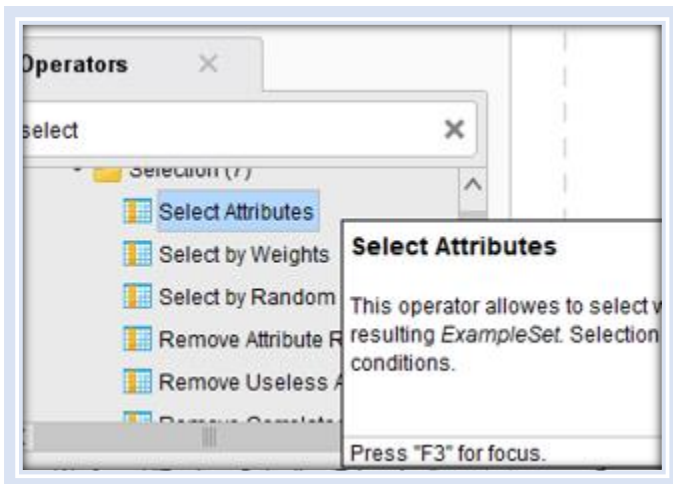
## PREPROCESSING ON TWITTER COVID DATASET:



Row No.	Created-At	From-User	From-User-Id	To-User	To-User-Id	Language	Source	Text	Geo-Location	Geo-Location	Retweet-Co...
1	Dec 12, 2022...	Kim Dotcom	611986351	?	-1	en	<a href="http/...	How much jai...	?	?	9806
2	Dec 11, 2022...	Prof Peter Ho...	593289567	?	-1	en	<a href="http/...	For the recor...	?	?	12817
3	Dec 12, 2022...	Leh JCHASE	1268925187...	?	-1	en	<a href="http/...	RT @SHINee...	?	?	225
4	Dec 12, 2022...	Roz Wood	147157830	DFMcBann	2915691574	en	<a href="http/...	@DFMcBann ...	?	?	0
5	Dec 12, 2022...	FairnessOver...	1525479331...	?	-1	en	<a href="http/...	RT @Infidelit...	?	?	118
6	Dec 12, 2022...	Sam Cracknell	33874800	?	-1	en	<a href="http/...	RT @jimmy_...	?	?	4177
7	Dec 12, 2022...	themakingsof...	1121259704...	?	-1	en	<a href="http/...	RT @Yash25...	?	?	10
8	Dec 12, 2022...	B Kuzyk	1586004210...	benryanwriter	1666939494	en	<a href="http/...	@benryanwrit...	?	?	0
9	Dec 12, 2022...	Derick Camp...	80772893	?	-1	en	<a href="http/...	RT @BillyRay...	?	?	2985
10	Dec 12, 2022...	Oscar	1577564925...	?	-1	en	<a href="http/...	RT @OrKate...	?	?	63
11	Dec 12, 2022...	VSR	73054768	?	-1	en	<a href="http/...	RT @ThePoll...	?	?	585
12	Dec 12, 2022...	Predrag Nikolic	9486905764...	?	-1	en	<a href="http/...	RT @vegana...	?	?	1
13	Dec 12, 2022...	Salah Saleh	8730947155...	?	-1	en	<a href="http/...	RT @Occupy...	?	?	3430
14	Dec 12, 2022...	Caitlin Wong	1288330529...	?	-1	en	<a href="http/...	Feature: 108-...	?	?	0

In text attribute where tweets are there, we firstly remove the http links and then tags @





Import Data - Select the cells to import.

### Select the cells to import.

Sheet: Search Twitter Cell range: A:L Select All ☒ Define header row: 1

	A	B	C	D	E	F	G	H	I	J
1	Created...	From-Us...	From-Us...	To-User	To-User-Id	Language	Source	Text	Geo-Loc...	Geo-L
2	Dec 12, ...	Kim Dotc...	6119863...		-1	en	<a href="...	How mu...		
3	Dec 11, ...	Prof Pete...	5932895...		-1	en	<a href="...	For the r...		
4	Dec 12, ...	Leh !CH...	1268925...		-1	en	<a href="...	RT @SH...		
5	Dec 12, ...	Roz Wood	1471578...	DFMcBa...	2915691...	en	<a href="...	@DFMc...		
6	Dec 12, ...	Fairness...	1525479...		-1	en	<a href="http://twitter.com/download/iphone" rel="no...	for iPhone</a>		
7	Dec 12, ...	Sam Cra...	33874800		-1	en				
8	Dec 12, ...	themaki...	1121259...		-1	en	<a href="...	RT @Ya...		
9	Dec 12, ...	B Kuzyk	1586004...	benryan...	1666939...	en	<a href="...	@benrya...		
10	Dec 12, ...	Derick C...	80772893		-1	en	<a href="...	RT @Bill...		
11	Dec 12, ...	Oscar	1577564...		-1	en	<a href="...	RT @Dr...		
12	Dec 12, ...	VSR	73054768		-1	en	<a href="...	RT @Th...		
13	Dec 12, ...	Predrag ...	9486905...		-1	en	<a href="...	RT @veg...		

Previous Next Cancel

Import Data - Format your columns.

### Format your columns.

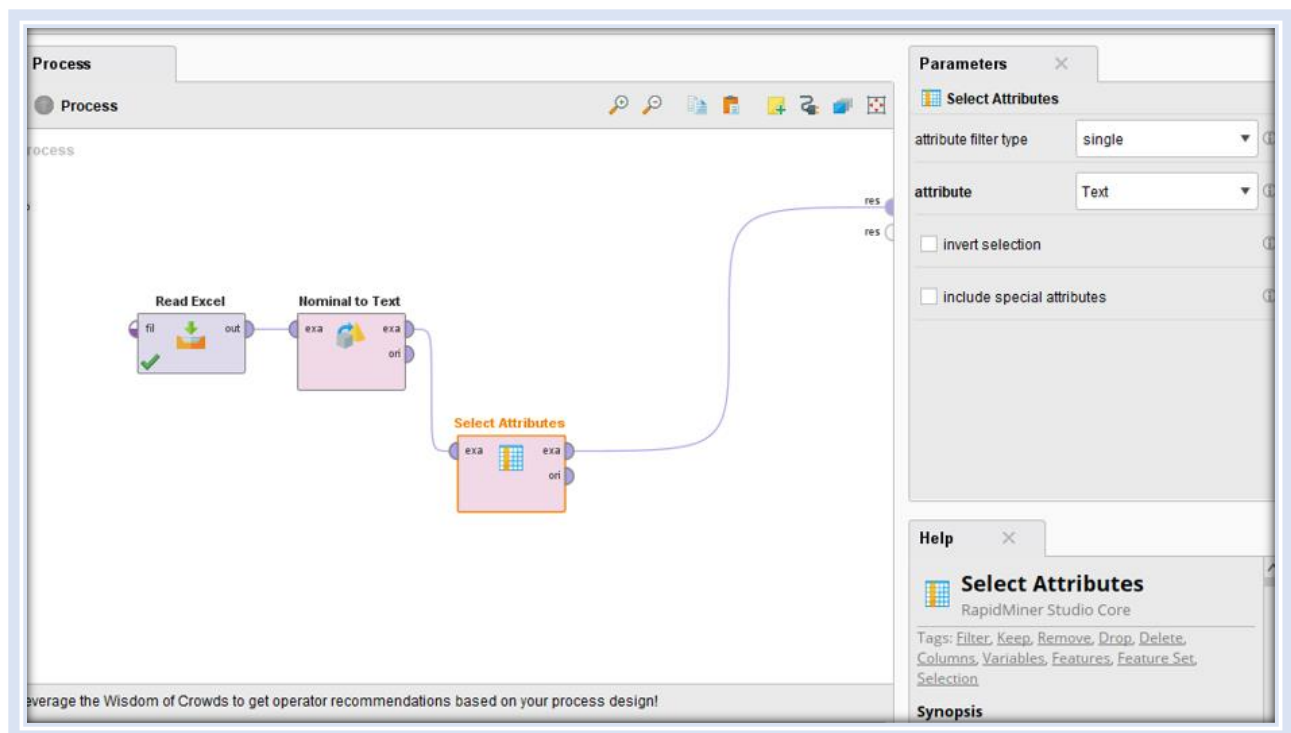
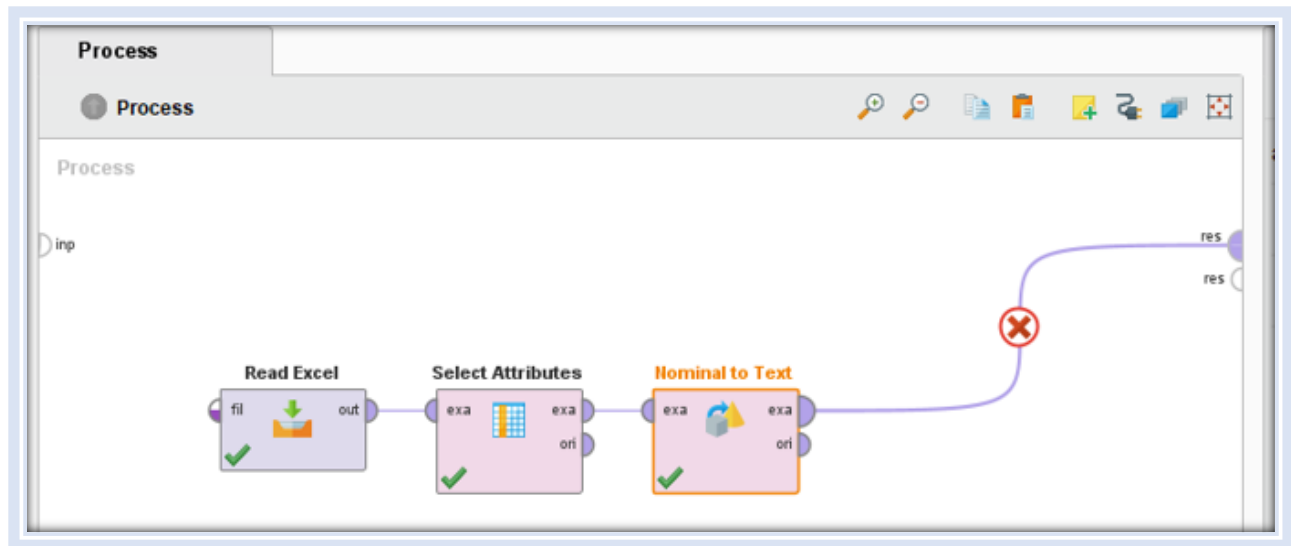
☐ Replace errors with missing values ⓘ

	Created-At	From-User	From-User-Id	To-User	To-User-Id	Language
	date_time	polynomial	real	polynomial	real	polynomial
1	Dec 12, 2022, 3:...	Kim Dotcom	611986351.000	?	-1.000	en
2	Dec 11, 2022, 5:...	Prof Peter Hotez ...	593289567.000	?	-1.000	en
3	Dec 12, 2022, 2:...	Leh !CHASE	1268925187409...	?	-1.000	en
4	Dec 12, 2022, 2:...	Roz Wood	147157830.000	DFMcBann	2915691574.000	en
5	Dec 12, 2022, 2:...	FairnessOverEq...	1525479331383...	?	-1.000	en
6	Dec 12, 2022, 2:...	Sam Cracknell	33874800.000	?	-1.000	en
7	Dec 12, 2022, 2:...	themakingsof/...	1121259704720...	?	-1.000	en
8	Dec 12, 2022, 2:...	B Kuzyk	1586004210134...	benryanwriter	1666939494.000	en
9	Dec 12, 2022, 2:...	Derick Campbell ...	80772893.000	?	-1.000	en
10	Dec 12, 2022, 2:...	Oscar	1577564925093...	?	-1.000	en
11	Dec 12, 2022, 2:...	VSR	73054768.000	?	-1.000	en
12	Dec 12, 2022, 2:...	Predrag Nikolic	9486905764041...	?	-1.000	en

☒ no problems.

Previous Finish Cancel

☒ Activate Wisdom of Crowds Finish This spe





Row No.	Text
1	How much jail time does one get for killing millions of people? We need Nuremberg trials for the entire US Covid-19 Mafia. <a href="https://t.co/C62xyHSmrJ">https://t.co/C62xyHSmrJ</a>
2	For the record: Dr. Fauci has done nothing wrong, except serve our nation. In the meantime, Mr. Musk should know that 200,000 Americans needlessly lost their lives from Covid du...
3	RT @SHINeebble: [221212] Onew
4	@DFMcBann Not going for the winter Covid booster. I don't fancy the hospital again. I will take my chances
5	RT @Infideliter2022: Wifey's cousin was a bit concerned about the increase in Covid cases.
6	RT @jimmy_dore: For the record: Dr Fauci lied about masks, lied about herd immunity, about natural immunity, about vaccines stopping transmis...
7	RT @Yash25571056: "Medical staff across England have been put on alert for signs of the deadly camel flu as World Cup fans return from Qata...
8	@benryanwriter @JohnBrennan SNOWFLAKE Fauci is the cause of covid! Funding GAIN OF FUNCTION RESEARCH & LIKELY INVOLVED IN THE RELEASE FROM THE WUHAN...
9	RT @BillyRay5229: He was the one person who could beat Trump. And did.
10	RT @DrKate_Miller: I read the other day that there are 4000 people unaware that they have cancer because they've stayed away from the doct...
11	RT @ThePollLady: Rana Ayyub becomes 1st Indian to receive U.S. highest Press Freedom award.
12	RT @vegananne: @elonmusk @DrJBhattacharya Please shine the sunlight on those w/covid #vaccineinjuries & the apparent serious trials shortco...
13	RT @OccupyDemocrats: BREAKING: In his most dangerous Tweet yet, MAGA Twitter CEO Elon Musk calls for the Justice Department to "prosecute"...
14	Feature: 108-year-old Chinese sound after recovery from COVID-19
15	It's Still Worth Fighting Anti-Covid Vaccine Misinformation <a href="https://t.co/biDZZfOpP">https://t.co/biDZZfOpP</a>

ExampleSet (4,991 examples, 0 special attributes, 1 regular attribute)

### Process

Process

```

graph LR
    Inp((inp)) --> ReadExcel[Read Excel]
    ReadExcel --> NominalToText[Nominal to Text]
    NominalToText --> SelectAttributes[Select Attributes]
    SelectAttributes --> Replace[Replace]
    Replace --> Out((out))
  
```

### Parameters

**Replace**

attribute filter type: all

☐ invert selection

☐ include special attributes

replace what:

replace by:

✓ [Change compatibility \(10.0.000\)](#)

### Help

**Replace**  
RapidMiner Studio Core

Tags: [Map](#), [Change](#), [Regex](#), [Regular expressions](#), [Values](#)

**Synopsis**

This operator replaces parts of the values of selected nominal attributes matching a specified regular expression by a specified

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

✓ [Activate Wisdom of Crowds](#)

http links have been removed Successfully:

Row No.	Text
1	How much jail time does one get for killing millions of people? We need Nuremberg trials for the entire US Covid-19 Mafia.
2	For the record: Dr. Fauci has done nothing wrong, except serve our nation. In the meantime, Mr. Musk should know that 200,000 Americans need to protect their lives from Covid due
3	RT @SHINeebble: [221212] Onew
4	@DFMcBann Not going for the winter Covid booster. I don't fancy the hospital again. I will take my chances
5	RT @Infideliter2022: Wifey's cousin was a bit concerned about the increase in Covid cases.
6	RT @jimmy_dore: For the record: Dr Fauci lied about masks, lied about herd immunity, about natural immunity, about vaccines stopping transmis...
7	RT @Yash25571056: "Medical staff across England have been put on alert for signs of the deadly camel flu as World Cup fans return from Qata...
8	@benryanwriter @JohnBrennan SNOWFLAKE Fauci is the cause of covid! Funding GAIN OF FUNCTION RESEARCH & LIKELY INVOLVED IN THE RELEASE FROM THE WUHAN L...
9	RT @BillyRay5229: He was the one person who could beat Trump. And did.
10	RT @DrKate_Miller: I read the other day that there are 4000 people unaware that they have cancer because they've stayed away from the doc...
11	RT @ThePollLady: Rana Ayyub becomes 1st Indian to receive U.S. highest Press Freedom award.
12	RT @vegananne: @elonmusk @DrJBhattacharya Please shine the sunlight on those w/covid #vaccineinjuries & the apparent serious trials shortco...
13	RT @OccupyDemocrats: BREAKING: In his most dangerous Tweet yet, MAGA Twitter CEO Elon Musk calls for the Justice Department to "prosecute"...
14	Feature: 108-year-old Chinese sound after recovery from COVID-19
15	It's Still Worth Fighting Anti-Covid Vaccine Misinformation

ExampleSet (4,991 examples, 0 special attributes, 1 regular attribute)

Transform text to lower case:

The screenshot displays the Orange3 data mining software interface. The main workspace shows a workflow titled "Process Documents from Data" with two connected operators: "Tokenize" and "Transform Cases". A "Drag here" label is present in the center of the workspace. The left sidebar contains the "Repository" and "Operators" panels. The "Repository" panel shows a list of data sources under "Community Samples (connected)" and "Local Repository (Local)". The "Operators" panel shows a search for "transform cas" and a list of operators including "Transform Cases". The "Transform Cases" operator is highlighted, and a tooltip is displayed over it, stating: "Transform Cases: Transforms cases of characters in a document. Press 'F3' for focus." The right sidebar contains the "Parameters" and "Help" panels. The "Parameters" panel shows the "Transform Cases" operator with a "transform to" dropdown set to "lower case". The "Help" panel shows the "Transform Cases" operator with a "Text Processing" tag and a "Synopsis" section stating: "Transforms cases of characters in a document."

