+

SULIT

Faculty of Computing

UNIVERSITI PENYELIDIKAN

UNIVERSITI TEKNOLOGI MALAYSIA

SEMESTER II - 2020 / 2021

ALTERNATIVE ASSESSMENT (INDIVIDUAL)

SUBJECT CODE	:	SCSP2753
SUBJECT NAME	**:**	**DATA MINING**
DATE	**:**	**23 JUNE – 1 JULY 2021**
DURATION	:	7 DAYS
SUBMISSION DATE	:	1 JULY 2021

INSTRUCTIONS:

This alternative assessment is assessed individually.
This alternative assessment consists of **TWO** parts.
Read **ALL** questions carefully and please answer **ALL** questions.
Submit your answer (report) via e-learning and submission later than the due date is not accepted.
Any form of plagiarism is not allowed.

Name	NUR ALEEYA SYAKILA BINTI MUHAMAD SUBIAN
I/C No.	000726100810
Year / Course	2 SECP
Section	01
Lecturer Name	DR ROZILAWATI BINTI DOLLAH

Table Of Contents

1.0 PART 1

1.1 Introduction

There are a lot of tools out there that we can use to do Preprocessing and Data Mining Task. For this alternative assessment, I will be choosing RapidMiner as my tool. This is because RapidMiner has a lot of resources that I can refer to incase of I couldn't understand or misleading concept. This alternative assessment required me to do text preprocessing and few data mining tasks. To perform any data mining algorithm, preprocess the data is a must and it must be done before proceeding with other steps. In this alternative assessment, for question 1, I will be performing text preprocessing for appendix 1 given in the question sheet. After did some research about preprocessing, I found that every basic preprocessing has the same steps which are tokenize, filter stopwords, generate n gram, filter token and stemming. While for question 2, based on the appendix 2, the dataset that is given to me is dataset 29. Before performing the supervised and unsupervised learning algorithm, I preprocess my dataset beforehand.

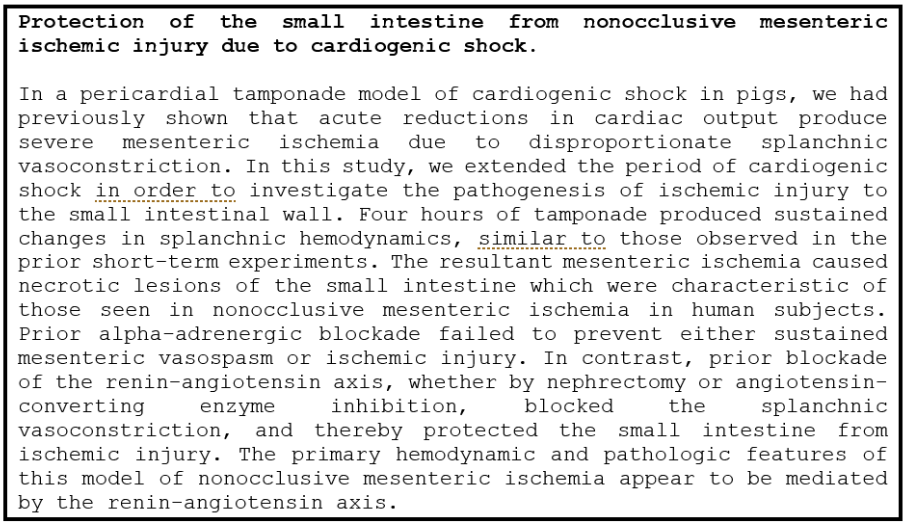
1.1.1 Text Preprocessing

Step 1 :

The data is in unstructured form where its format is variety such as in long text email or else. Therefore, I needed to convert it to the structured one by showing text only. To make it happened, I will be needed to do text preprocessing in RapidMiner. Firstly, I need to change the dataset from docx file into a xlsx file to make it easier for the RapidMiner to read the dataset.

Input :

APPENDIX 1

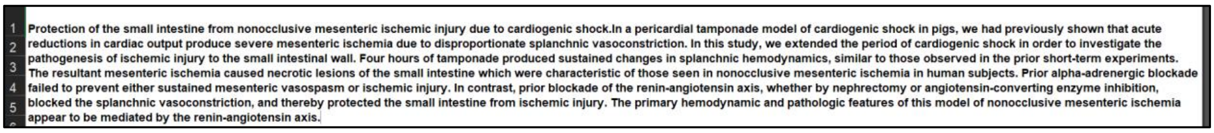


Protection of the small intestine from nonocclusive mesenteric ischemic injury due to cardiogenic shock.

In a pericardial tamponade model of cardiogenic shock in pigs, we had previously shown that acute reductions in cardiac output produce severe mesenteric ischemia due to disproportionate splanchnic vasoconstriction. In this study, we extended the period of cardiogenic shock in order to investigate the pathogenesis of ischemic injury to the small intestinal wall. Four hours of tamponade produced sustained changes in splanchnic hemodynamics, similar to those observed in the prior short-term experiments. The resultant mesenteric ischemia caused necrotic lesions of the small intestine which were characteristic of those seen in nonocclusive mesenteric ischemia in human subjects. Prior alpha-adrenergic blockade failed to prevent either sustained mesenteric vasospasm or ischemic injury. In contrast, prior blockade of the renin-angiotensin axis, whether by nephrectomy or angiotensin-converting enzyme inhibition, blocked the splanchnic vasoconstriction, and thereby protected the small intestine from ischemic injury. The primary hemodynamic and pathologic features of this model of nonocclusive mesenteric ischemia appear to be mediated by the renin-angiotensin axis.

Figure 1 : Raw dataset appendix 1 in docx

Output :



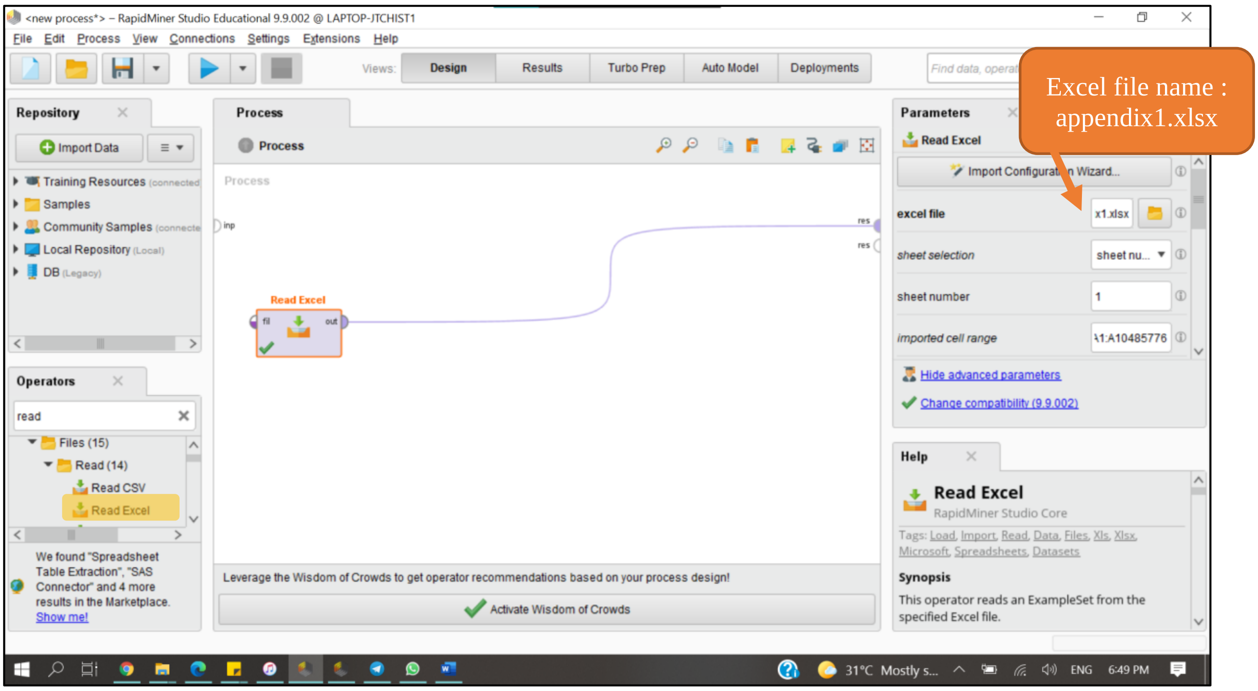
Protection of the small intestine from nonocclusive mesenteric ischemic injury due to cardiogenic shock.In a pericardial tamponade model of cardiogenic shock in pigs, we had previously shown that acute reductions in cardiac output produce severe mesenteric ischemia due to disproportionate splanchnic vasoconstriction. In this study, we extended the period of cardiogenic shock in order to investigate the pathogenesis of ischemic injury to the small intestinal wall. Four hours of tamponade produced sustained changes in splanchnic hemodynamics, similar to those observed in the prior short-term experiments. The resultant mesenteric ischemia caused necrotic lesions of the small intestine which were characteristic of those seen in nonocclusive mesenteric ischemia in human subjects. Prior alpha-adrenergic blockade failed to prevent either sustained mesenteric vasospasm or ischemic injury. In contrast, prior blockade of the renin-angiotensin axis, whether by nephrectomy or angiotensin-converting enzyme inhibition, blocked the splanchnic vasoconstriction, and thereby protected the small intestine from ischemic injury. The primary hemodynamic and pathologic features of this model of nonocclusive mesenteric ischemia appear to be mediated by the renin-angiotensin axis.

Figure 2 : Raw dataset appendix 1 in xlsx

Step 2 : Read Excel

After converting it to a data in xlsx file, the very first basic thing I need to do after opening the RapidMiner is to drag the ‘Read Excel’ from the Operators as the dataset is in excel file. Drag the Read Excel operator and read the dataset file.

Input :



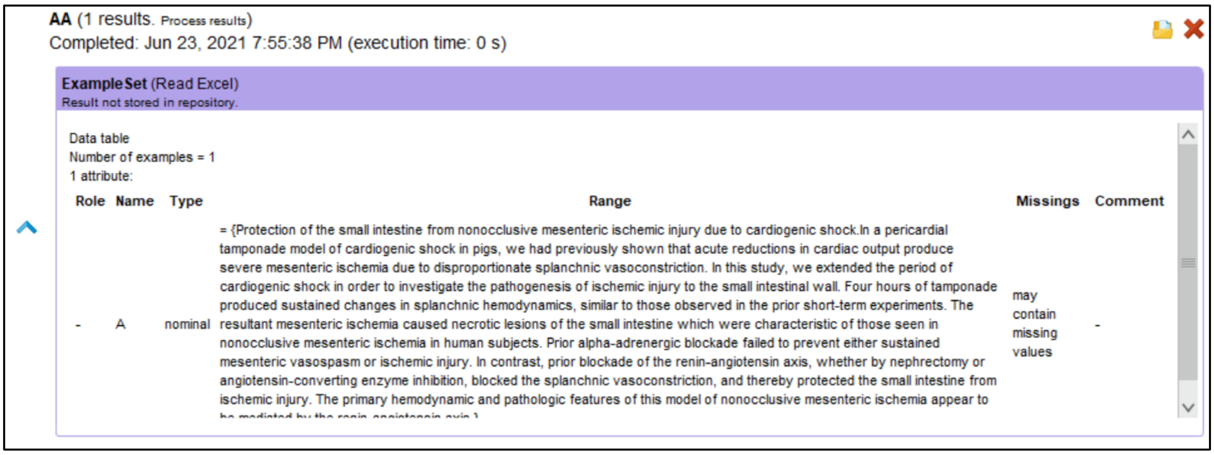
Excel file name : appendix1.xlsx

<new process*> – RapidMiner Studio Educational 9.9.002 @ LAPTOP-JTCHIST1
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
DB (Legacy)
Operators
read
Files (15)
Read (14)
Read CSV
Read Excel
We found "Spreadsheet Table Extraction", "SAS Connector" and 4 more results in the Marketplace. Show me!
Process
Read Excel
fil
out
inp
res
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Read Excel
Import Configuration Wizard...
Hide advanced parameters
Change compatibility (9.9.002)
Help
Read Excel
RapidMiner Studio Core
Tags: Load, Import, Read, Data, Files, Xls, Xlsx, Microsoft, Spreadsheets, Datasets
Synopsis
This operator reads an ExampleSet from the specified Excel file.
31°C Mostly s... ENG 6:49 PM

excel file: x1.xlsx
sheet selection: sheet nu...
sheet number: 1
imported cell range: A1:A10485776

Figure 3 : Choose Read Excel operator

Output :



AA (1 results. Process results)
Completed: Jun 23, 2021 7:55:38 PM (execution time: 0 s)

ExampleSet (Read Excel)
Result not stored in repository.

Data table
Number of examples = 1
1 attribute:

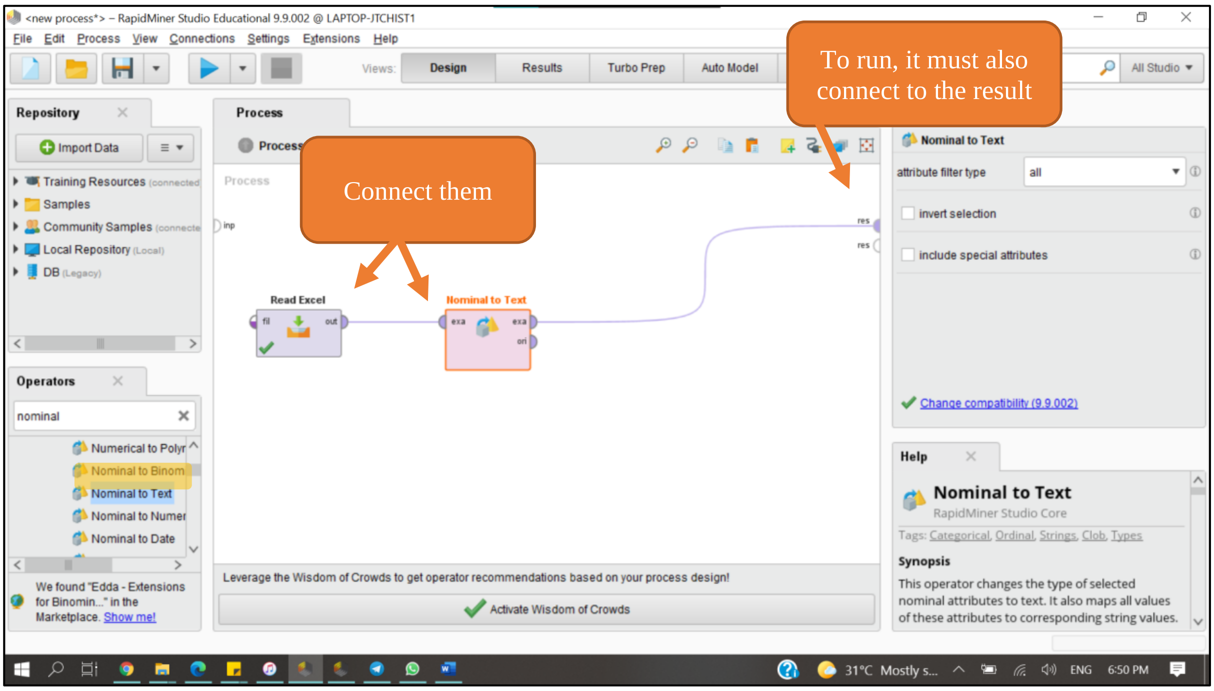
Role	Name	Type	Range	Missings	Comment
-	A	nominal	= {Protection of the small intestine from nonocclusive mesenteric ischemic injury due to cardiogenic shock.In a pericardial tamponade model of cardiogenic shock in pigs, we had previously shown that acute reductions in cardiac output produce severe mesenteric ischemia due to disproportionate splanchnic vasoconstriction. In this study, we extended the period of cardiogenic shock in order to investigate the pathogenesis of ischemic injury to the small intestinal wall. Four hours of tamponade produced sustained changes in splanchnic hemodynamics, similar to those observed in the prior short-term experiments. The resultant mesenteric ischemia caused necrotic lesions of the small intestine which were characteristic of those seen in nonocclusive mesenteric ischemia in human subjects. Prior alpha-adrenergic blockade failed to prevent either sustained mesenteric vasospasm or ischemic injury. In contrast, prior blockade of the renin-angiotensin axis, whether by nephrectomy or angiotensin-converting enzyme inhibition, blocked the splanchnic vasoconstriction, and thereby protected the small intestine from ischemic injury. The primary hemodynamic and pathologic features of this model of nonocclusive mesenteric ischemia appear to be mediated by the renin-angiotensin axis.}	may contain missing values	-

Figure 4 : The result after Read Excel

Step 3 : Nominal to Text

Then I needed to drag the ‘Nominal to Text’ operators. The Nominal to Text operator converts all nominal attributes to string attributes. Drag the ‘Nominal to Text’ operator. Connect the *output* of the ‘Read Excel’ operator to the *example* of the ‘Nominal to Text’ operator. Connect the output of ‘Read Excel’ operator to the example (input) of the ‘Nominal to Text’ operator

Input :



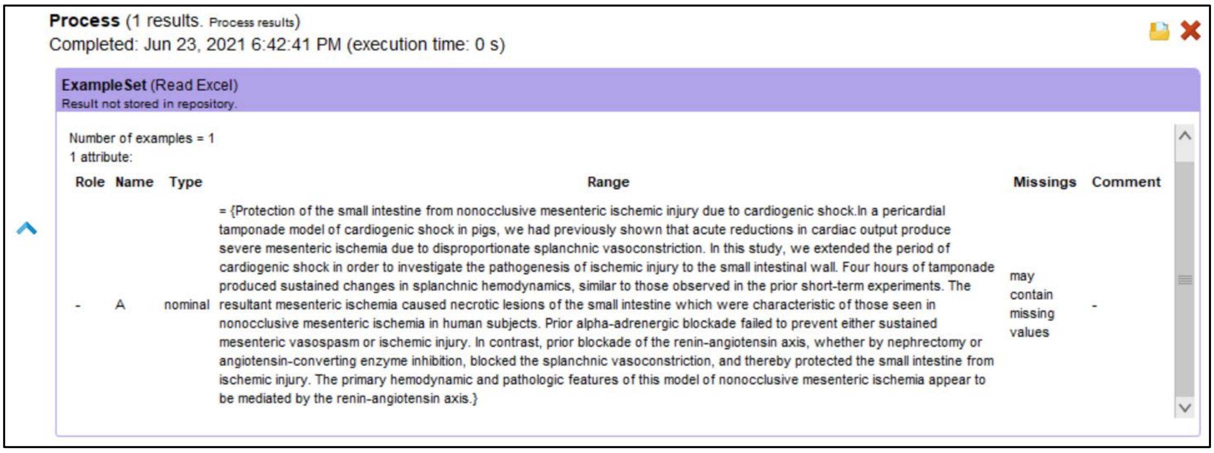
To run, it must also connect to the result

Connect them

<new process*> – RapidMiner Studio Educational 9.9.002 @ LAPTOP-JTCHIST1
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
All Studio
Repository
Import Data
Training Resources
Samples
Community Samples
Local Repository (Local)
DB (Legacy)
Process
Read Excel
Nominal to Text
Operators
nominal
Numerical to Polyn
Nominal to Binom
Nominal to Text
Nominal to Numer
Nominal to Date
We found "Edda - Extensions for Binomin..." in the Marketplace. Show me!
attribute filter type all
invert selection
include special attributes
Change compatibility (9.9.002)
Help
Nominal to Text
RapidMiner Studio Core
Tags: Categorical, Ordinal, Strings, Clob, Types
Synopsis
This operator changes the type of selected nominal attributes to text. It also maps all values of these attributes to corresponding string values.
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
31°C Mostly s... ENG 6:50 PM

Figure 5 : Choose the Nominal to Text operator

Output :



Process (1 results. Process results)
Completed: Jun 23, 2021 6:42:41 PM (execution time: 0 s)

ExampleSet (Read Excel)
Result not stored in repository.

Number of examples = 1
1 attribute:

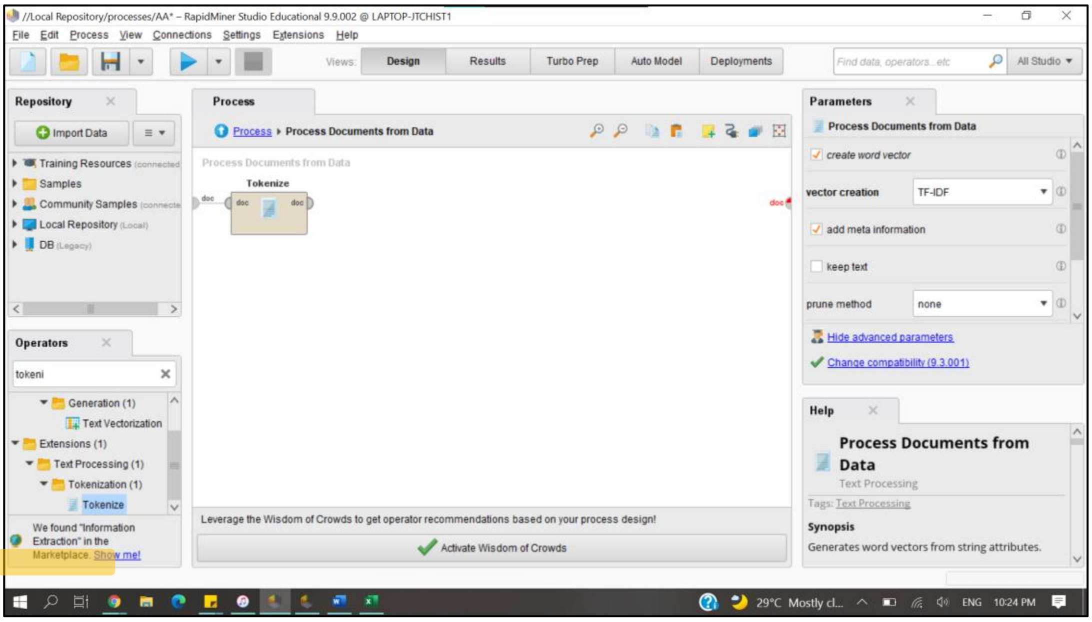
Role	Name	Type	Range	Missings	Comment
-	A	nominal	= {Protection of the small intestine from nonocclusive mesenteric ischemic injury due to cardiogenic shock.In a pericardial tamponade model of cardiogenic shock in pigs, we had previously shown that acute reductions in cardiac output produce severe mesenteric ischemia due to disproportionate splanchnic vasoconstriction. In this study, we extended the period of cardiogenic shock in order to investigate the pathogenesis of ischemic injury to the small intestinal wall. Four hours of tamponade produced sustained changes in splanchnic hemodynamics, similar to those observed in the prior short-term experiments. The resultant mesenteric ischemia caused necrotic lesions of the small intestine which were characteristic of those seen in nonocclusive mesenteric ischemia in human subjects. Prior alpha-adrenergic blockade failed to prevent either sustained mesenteric vasospasm or ischemic injury. In contrast, prior blockade of the renin-angiotensin axis, whether by nephrectomy or angiotensin-converting enzyme inhibition, blocked the splanchnic vasoconstriction, and thereby protected the small intestine from ischemic injury. The primary hemodynamic and pathologic features of this model of nonocclusive mesenteric ischemia appear to be mediated by the renin-angiotensin axis.}	may contain missing values	-

Figure 6 : The result after running the Nominal to Text operator

Step 4 : Tokenize

To continue the preprocessing, I needed to add ‘Process Documents from Data’ operator first. After that, proceed to the tokenize process. Tokenization is a pre-processing strategy what breaks a stream of text into words, expressions, images, or other significant components called tokens. After clicking the ‘Process Document from Data’, drag the ‘Tokenize’ operator to proceed.

Input :



Parameters: Process Documents from Data

☑ create word vector

vector creation: TF-IDF

☑ add meta information

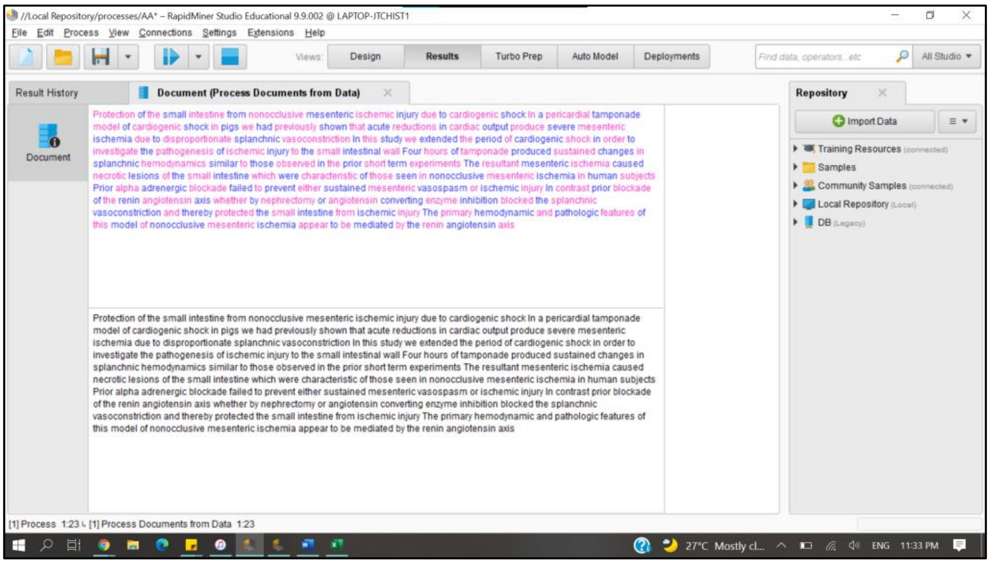
☐ keep text

prune method: none

//Local Repository/processes/AA* – RapidMiner Studio Educational 9.9.002 @ LAPTOP-JTCHIST1
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Repository
Import Data
Training Resources
Samples
Community Samples
Local Repository (Local)
DB (Legacy)
Operators
tokeni
Generation (1)
Text Vectorization
Extensions (1)
Text Processing (1)
Tokenization (1)
Tokenize
We found "Information Extraction" in the Marketplace. Show me!
Process
Process ▸ Process Documents from Data
Tokenize
doc
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Hide advanced parameters
Change compatibility (9.3.001)
Help
Process Documents from Data
Text Processing
Tags: Text Processing
Synopsis
Generates word vectors from string attributes.
29°C Mostly cl... ENG 10:24 PM

Figure 8 : Add ‘Tokenize’ operator

Output :



Document (Process Documents from Data)

Protection of the small intestine from nonocclusive mesenteric ischemic injury due to cardiogenic shock In a pericardial tamponade model of cardiogenic shock in pigs we had previously shown that acute reductions in cardiac output produce severe mesenteric ischemia due to disproportionate splanchnic vasoconstriction In this study we extended the period of cardiogenic shock in order to investigate the pathogenesis of ischemic injury to the small intestinal wall Four hours of tamponade produced sustained changes in splanchnic hemodynamics similar to those observed in the prior short term experiments The resultant mesenteric ischemia caused necrotic lesions of the small intestine which were characteristic of those seen in nonocclusive mesenteric ischemia in human subjects Prior alpha adrenergic blockade failed to prevent either sustained mesenteric vasospasm or ischemic injury In contrast prior blockade of the renin angiotensin axis whether by nephrectomy or angiotensin converting enzyme inhibition blocked the splanchnic vasoconstriction and thereby protected the small intestine from ischemic injury The primary hemodynamic and pathologic features of this model of nonocclusive mesenteric ischemia appear to be mediated by the renin angiotensin axis

Protection of the small intestine from nonocclusive mesenteric ischemic injury due to cardiogenic shock In a pericardial tamponade model of cardiogenic shock in pigs we had previously shown that acute reductions in cardiac output produce severe mesenteric ischemia due to disproportionate splanchnic vasoconstriction In this study we extended the period of cardiogenic shock in order to investigate the pathogenesis of ischemic injury to the small intestinal wall Four hours of tamponade produced sustained changes in splanchnic hemodynamics similar to those observed in the prior short term experiments The resultant mesenteric ischemia caused necrotic lesions of the small intestine which were characteristic of those seen in nonocclusive mesenteric ischemia in human subjects Prior alpha adrenergic blockade failed to prevent either sustained mesenteric vasospasm or ischemic injury In contrast prior blockade of the renin angiotensin axis whether by nephrectomy or angiotensin converting enzyme inhibition blocked the splanchnic vasoconstriction and thereby protected the small intestine from ischemic injury The primary hemodynamic and pathologic features of this model of nonocclusive mesenteric ischemia appear to be mediated by the renin angiotensin axis

[1] Process 1:23 ↳ [1] Process Documents from Data 1:23

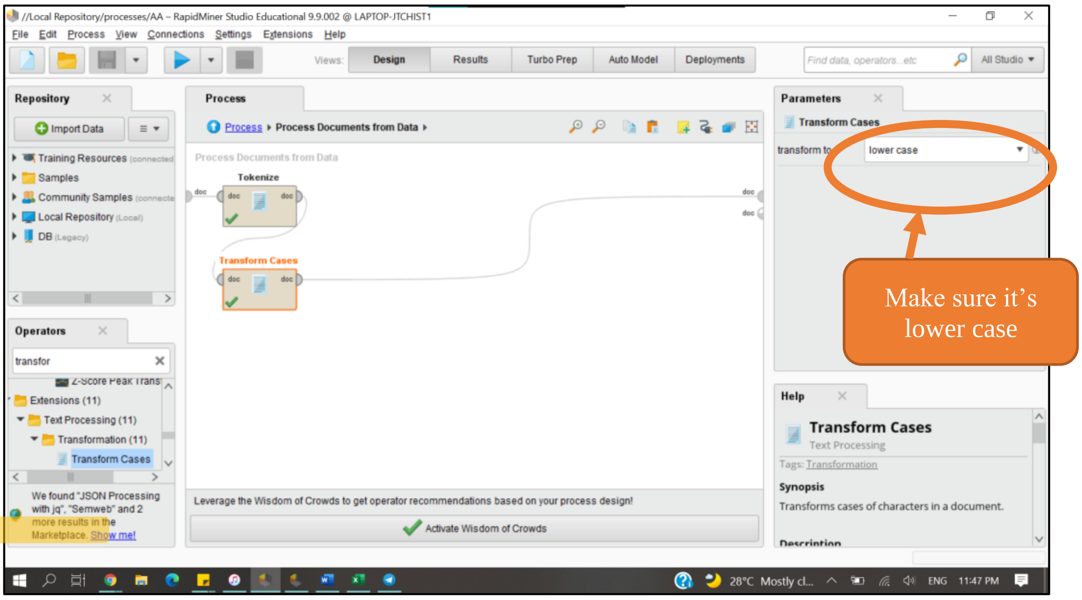
//Local Repository/processes/AA* – RapidMiner Studio Educational 9.9.002 @ LAPTOP-JTCHIST1
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Result History
Document
Repository
Import Data
Training Resources
Samples
Community Samples
Local Repository (Local)
DB (Legacy)
27°C Mostly cl... ENG 11:33 PM

Figure 9 : The result after the execution of the tokenize step

Step 5 : Lower Cases

Then I needed to transform the dataset into lower cases. This step can help in situations where the dataset isn't exceptionally huge and fundamentally assists with consistency of anticipated output. Drag the Transform Cases operator into the process and make sure it is set to transform to lower case. After the execution, a list of lower-case words has been created. There are total of 96 words produced in the list.

Input :



//Local Repository/processes/AA – RapidMiner Studio Educational 9.9.002 @ LAPTOP-JTCHIST1

Transform Cases

transform to: lower case

Make sure it's lower case

Transform Cases

Text Processing

Tags: Transformation

Synopsis

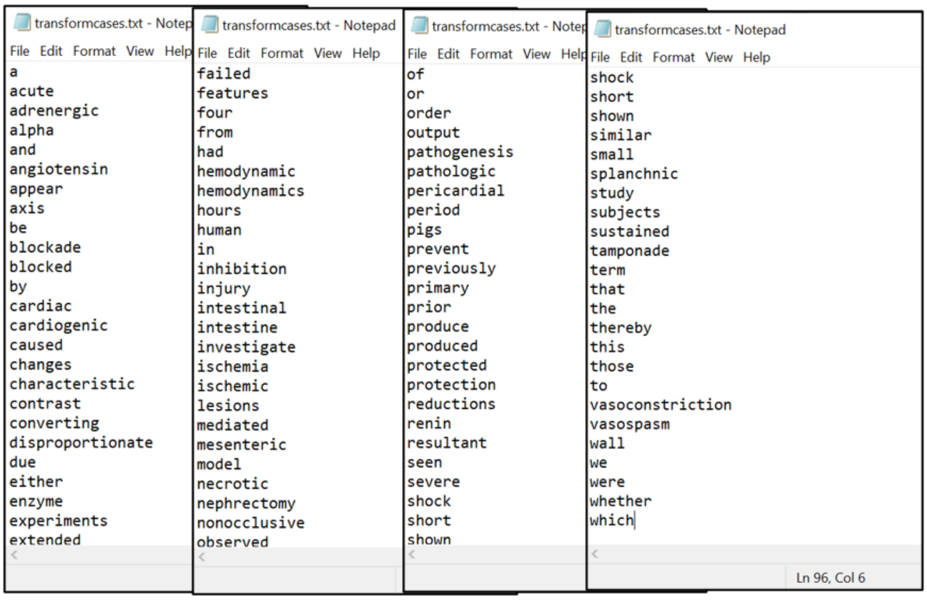
Transforms cases of characters in a document.

We found "JSON Processing with jq", "Semweb" and 2 more results in the Marketplace. Show me!

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

Figure 10 : Add 'Transform Cases' operator

Output :



```
a
acute
adrenergic
alpha
and
angiotensin
appear
axis
be
blockade
blocked
by
cardiac
cardiogenic
caused
changes
characteristic
contrast
converting
disproportionate
due
either
enzyme
experiments
extended
```

```
failed
features
four
from
had
hemodynamic
hemodynamics
hours
human
in
inhibition
injury
intestinal
intestine
investigate
ischemia
ischemic
lesions
mediated
mesenteric
model
necrotic
nephrectomy
nonocclusive
observed
```

```
of
or
order
output
pathogenesis
pathologic
pericardial
period
pigs
prevent
previously
primary
prior
produce
produced
protected
protection
reductions
renin
resultant
seen
severe
shock
short
shown
```

```
shock
short
shown
similar
small
splanchnic
study
subjects
sustained
tamponade
term
that
the
thereby
this
those
to
vasoconstriction
vasospasm
wall
we
were
whether
which
```

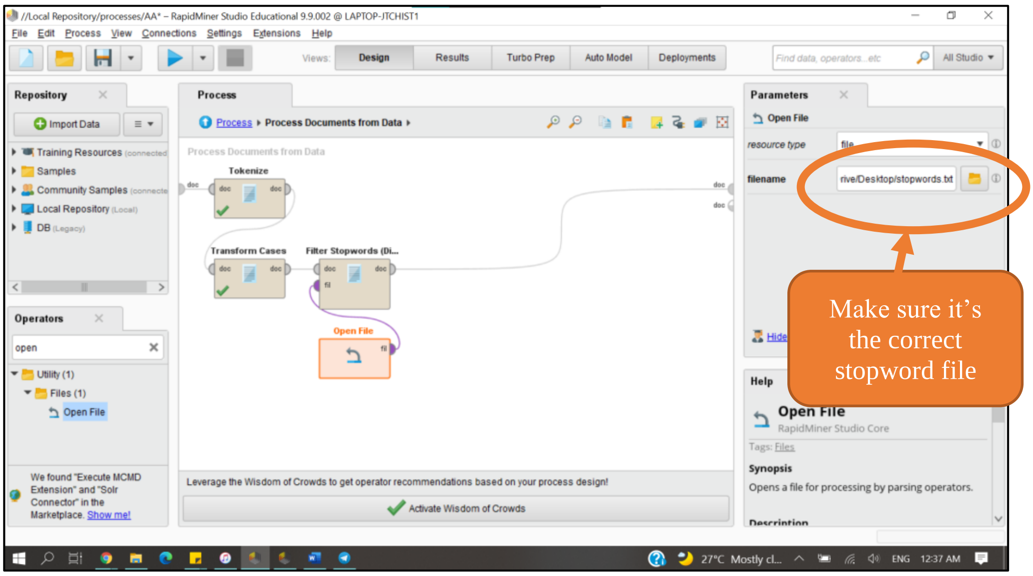
Ln 96, Col 6

Figure 11 : Result of transform cases process

Step 6 : Filter Stopwords

After transforming the dataset into lower cases, next step is filtering all stopwords. Stopwords are common words like “a”, “the”, “is”, “are” and etc. We removed those common words so that we can focus on important words. In this, I used the filter stopwords by dictionary and manually add a text file that contains a thousand of stopwords that has been listed. To do so, I needed to add the ‘Open File’ operator. After the execution, a list of words without stopwords has been generated and the total of words generated are 29 after going through the Filtering Stopwords process.

Input :


//Local Repository/processes/AA* – RapidMiner Studio Educational 9.9.002 @ LAPTOP-JTCHIST1
File Edit Process View Connections Settings Extensions Help
Design Results Turbo Prep Auto Model Deployments
Repository
Import Data
Training Resources
Samples
Community Samples
Local Repository
DB
Operators
open
Utility (1)
Files (1)
Open File
Process
Process Documents from Data
Tokenize
Transform Cases
Filter Stopwords (Di...
Open File
Parameters
Open File
Help
Open File
RapidMiner Studio Core
Activate Wisdom of Crowds

resource type: file
filename: rive/Desktop/stopwords.txt

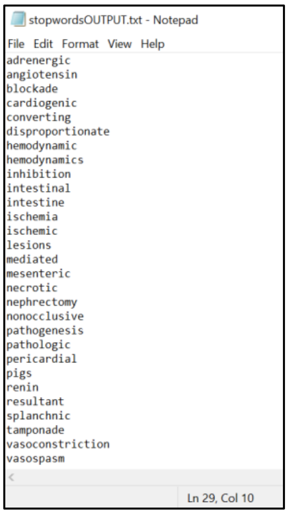
Make sure it’s the correct stopword file

Synopsis
Opens a file for processing by parsing operators.

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

Figure 12 : Add the ‘Filter Stopword by dictionary’ operator and the stopword file manually

Output :



```
adrenergic
angiotensin
blockade
cardiogenic
converting
disproportionate
hemodynamic
hemodynamics
inhibition
intestinal
intestine
ischemia
ischemic
lesions
mediated
mesenteric
necrotic
nephrectomy
nonocclusive
pathogenesis
pathologic
pericardial
pigs
renin
resultant
splanchnic
tamponade
vasoconstriction
vasospasm
```

Figure 12 : List of words without stopwords

Step 7 : Generate n grams (terms)

Next, we need to generate n-Grams. Clearly, language has a successive nature, thus the request in which words show up in the content matters a great deal. This operator permits us to comprehend the context of a sentence regardless of whether there are a few words missing. To do this, drag the 'Generate n-Grams' operator onto the process.

Input :

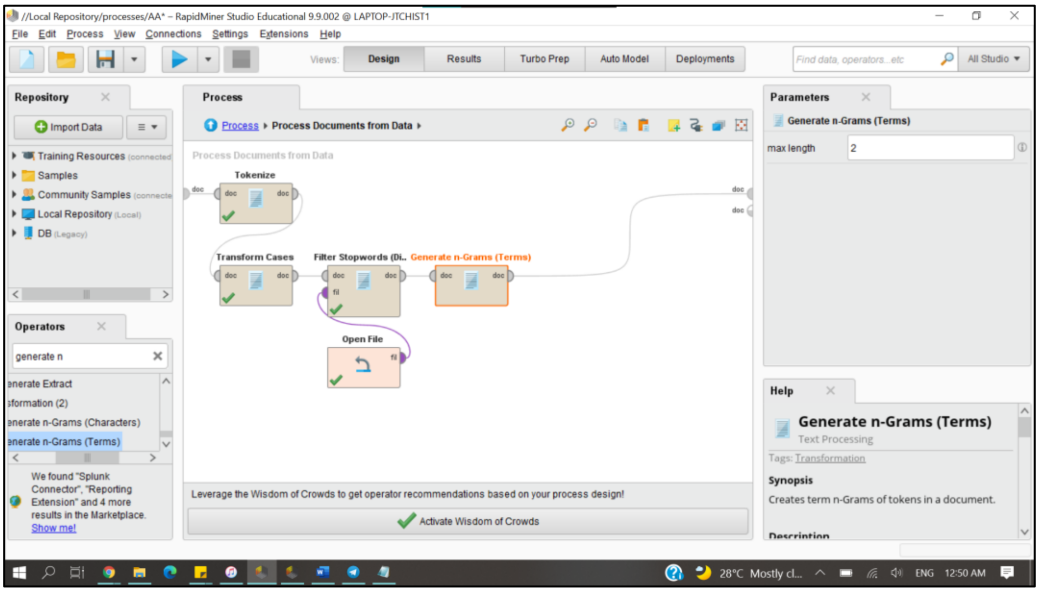
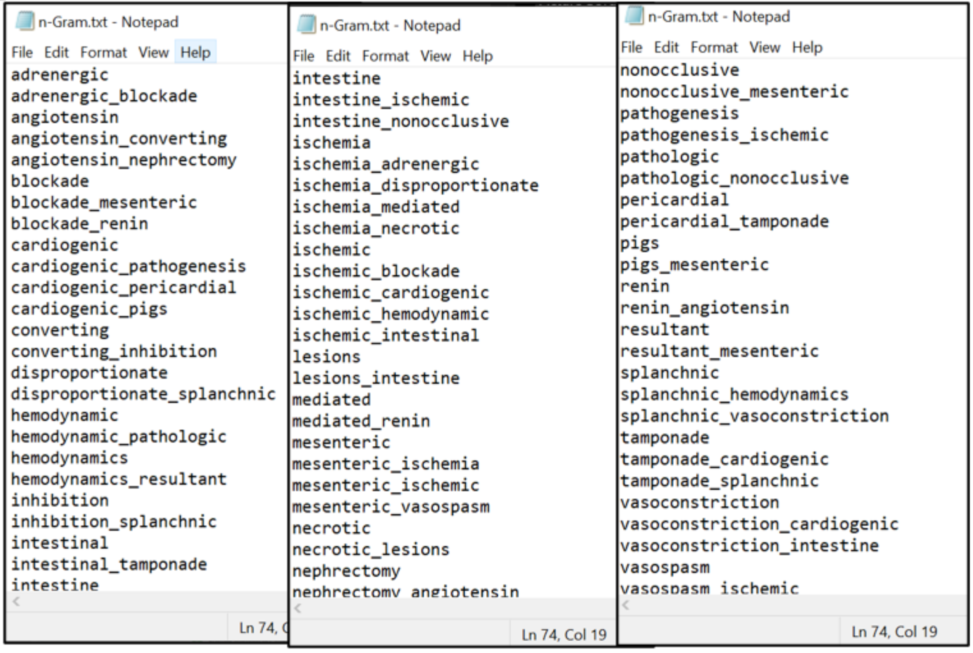

//Local Repository/processes/AA* – RapidMiner Studio Educational 9.9.002 @ LAPTOP-JTCHIST1
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Repository
Import Data
Training Resources
Samples
Community Samples
Local Repository (Local)
DB (Legacy)
Operators
generate n
enerate Extract
sformation (2)
enerate n-Grams (Characters)
enerate n-Grams (Terms)
We found "Splunk Connector", "Reporting Extension" and 4 more results in the Marketplace. Show me!
Process
Process ▸ Process Documents from Data ▸
Process Documents from Data
Tokenize
Transform Cases
Filter Stopwords (Di..
Generate n-Grams (Terms)
Open File
doc
fil
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Generate n-Grams (Terms)
max length 2
Help
Generate n-Grams (Terms)
Text Processing
Tags: Transformation
Synopsis
Creates term n-Grams of tokens in a document.
Description
28°C Mostly cl... ENG 12:50 AM

Figure 13 : Add the 'Generate n-Gram' operator

Output :


n-Gram.txt - Notepad

```
adrenergic
adrenergic_blockade
angiotensin
angiotensin_converting
angiotensin_nephrectomy
blockade
blockade_mesenteric
blockade_renin
cardiogenic
cardiogenic_pathogenesis
cardiogenic_pericardial
cardiogenic_pigs
converting
converting_inhibition
disproportionate
disproportionate_splanchnic
hemodynamic
hemodynamic_pathologic
hemodynamics
hemodynamics_resultant
inhibition
inhibition_splanchnic
intestinal
intestinal_tamponade
intestine
```

n-Gram.txt - Notepad

```
intestine
intestine_ischemic
intestine_nonocclusive
ischemia
ischemia_adrenergic
ischemia_disproportionate
ischemia_mediated
ischemia_necrotic
ischemic
ischemic_blockade
ischemic_cardiogenic
ischemic_hemodynamic
ischemic_intestinal
lesions
lesions_intestine
mediated
mediated_renin
mesenteric
mesenteric_ischemia
mesenteric_ischemic
mesenteric_vasospasm
necrotic
necrotic_lesions
nephrectomy
nephrectomy_angiotensin
```

n-Gram.txt - Notepad

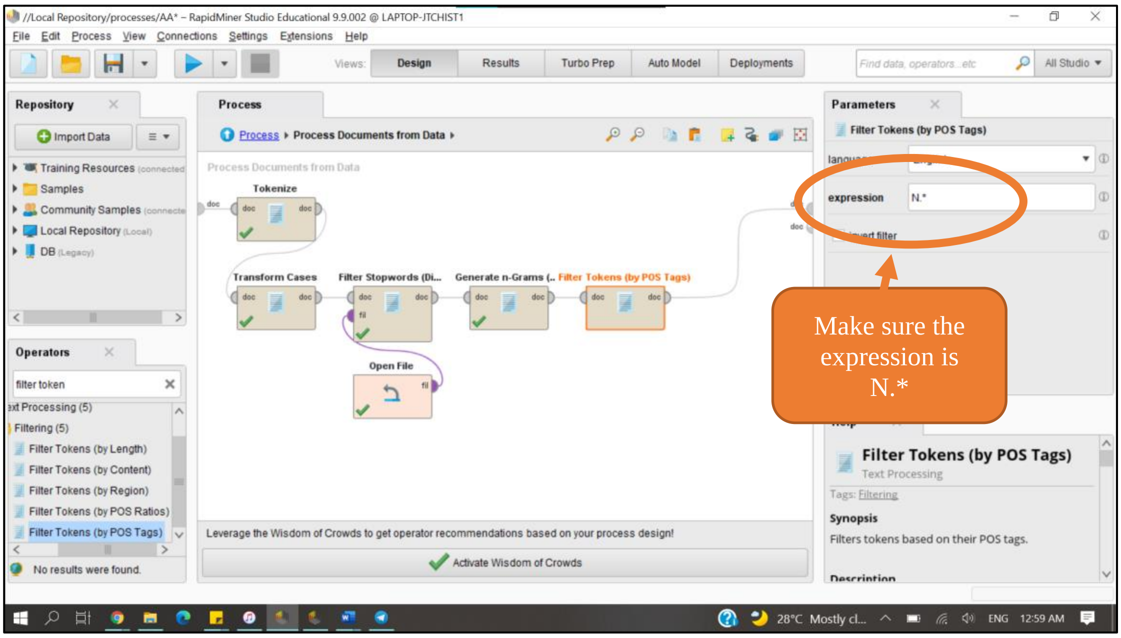
```
nonocclusive
nonocclusive_mesenteric
pathogenesis
pathogenesis_ischemic
pathologic
pathologic_nonocclusive
pericardial
pericardial_tamponade
pigs
pigs_mesenteric
renin
renin_angiotensin
resultant
resultant_mesenteric
splanchnic
splanchnic_hemodynamics
splanchnic_vasoconstriction
tamponade
tamponade_cardiogenic
tamponade_splanchnic
vasoconstriction
vasoconstriction_cardiogenic
vasoconstriction_intestine
vasospasm
vasospasm_ischemic
```

Figure 14 : The result after going through the Generate n-Gram process

Step 8 : Filter Tokens by POS Tags

The next step is Filter Tokens by POS Tags. A POS tag (or part-of-speech tag) is a special label assigned to each token (word) in a text corpus to indicate the part of speech and often also other grammatical categories such such as tense, number (plural/singular), case etc. To complete this, drag the ‘Filter Tokens by POS Tags’ operator and set the expression to N.*

Input :



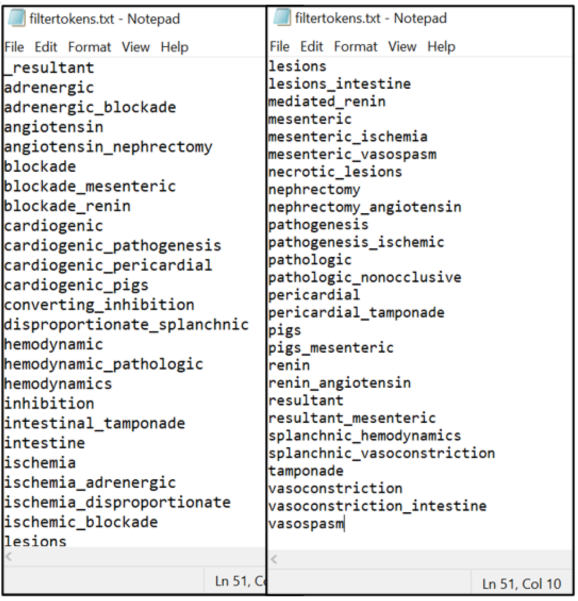
//Local Repository/processes/AA* – RapidMiner Studio Educational 9.9.002 @ LAPTOP-JTCHIST1
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Repository
Import Data
Training Resources
Samples
Community Samples
Local Repository (Local)
DB (Legacy)
Operators
filter token
Filtering (5)
Filter Tokens (by Length)
Filter Tokens (by Content)
Filter Tokens (by Region)
Filter Tokens (by POS Ratios)
Filter Tokens (by POS Tags)
No results were found.
Process
Process ▸ Process Documents from Data ▸
Tokenize
Transform Cases
Filter Stopwords (Di...
Generate n-Grams (...
Filter Tokens (by POS Tags)
Open File
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Parameters
Filter Tokens (by POS Tags)
Filter Tokens (by POS Tags)
Text Processing
Tags: Filtering
Synopsis
Filters tokens based on their POS tags.
Description
28°C Mostly cl... ENG 12:59 AM

expression: N.*

Make sure the expression is N.*

Figure 15 : Add the ‘Filter Tokens by POS Tags’ operator

Output :



filtertokens.txt - Notepad
File Edit Format View Help
Ln 51, C
filtertokens.txt - Notepad
File Edit Format View Help
Ln 51, Col 10

```
_resultant
adrenergic
adrenergic_blockade
angiotensin
angiotensin_nephrectomy
blockade
blockade_mesenteric
blockade_renin
cardiogenic
cardiogenic_pathogenesis
cardiogenic_pericardial
cardiogenic_pigs
converting_inhibition
disproportionate_splanchnic
hemodynamic
hemodynamic_pathologic
hemodynamics
inhibition
intestinal_tamponade
intestine
ischemia
ischemia_adrenergic
ischemia_disproportionate
ischemic_blockade
lesions
```

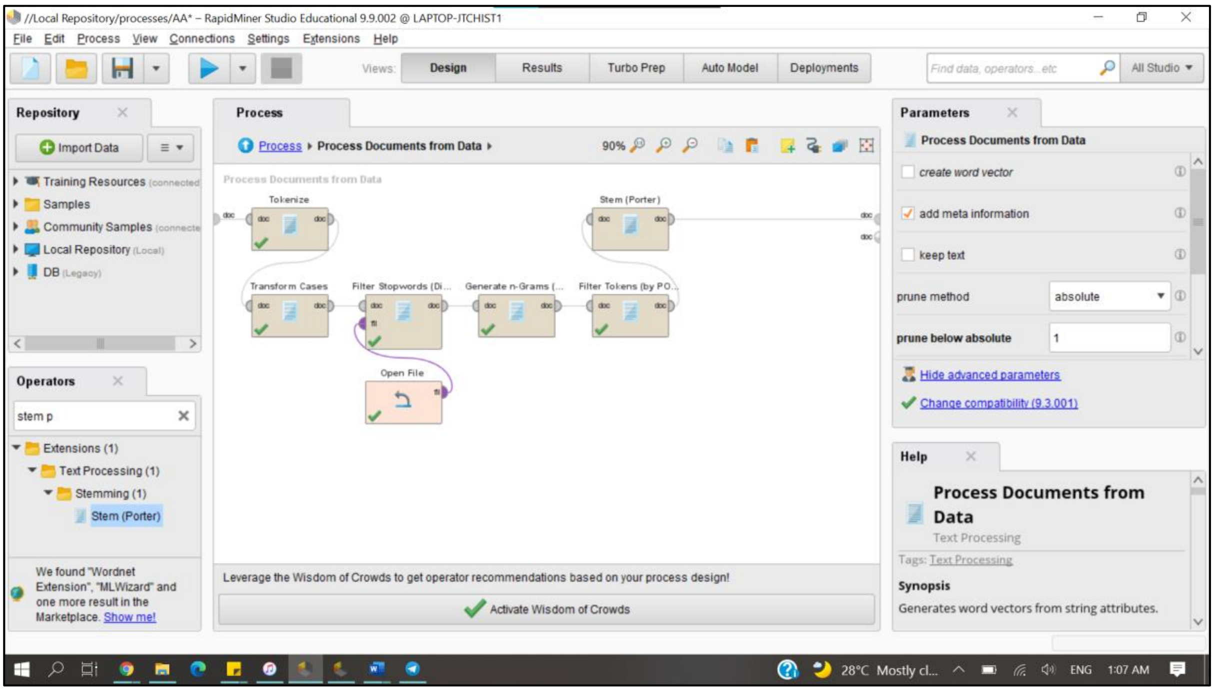
```
lesions
lesions_intestine
mediated_renin
mesenteric
mesenteric_ischemia
mesenteric_vasospasm
necrotic_lesions
nephrectomy
nephrectomy_angiotensin
pathogenesis
pathogenesis_ischemic
pathologic
pathologic_nonocclusive
pericardial
pericardial_tamponade
pigs
pigs_mesenteric
renin
renin_angiotensin
resultant
resultant_mesenteric
splanchnic_hemodynamics
splanchnic_vasoconstriction
tamponade
vasoconstriction
vasoconstriction_intestine
vasospasm
```

Figure 16 : The list of word generated after filter token process has been executed

Step 9 : Stem (Porter)

Then, I needed to do the stemming step. The Porter stemming algorithm (or 'Porter stemmer') is a process for removing the commoner morphological and in flexional endings from words in English. To do this, simply drag the 'Stem(Porter)' operator.

Input :



Process Documents from Data

☐ create word vector

☑ add meta information

☐ keep text

prune method: absolute

prune below absolute: 1

Hide advanced parameters

Change compatibility (9.3.001)

We found "Wordnet Extension", "MLWizard" and one more result in the Marketplace. Show me!

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

Activate Wisdom of Crowds

Process Documents from Data

Text Processing

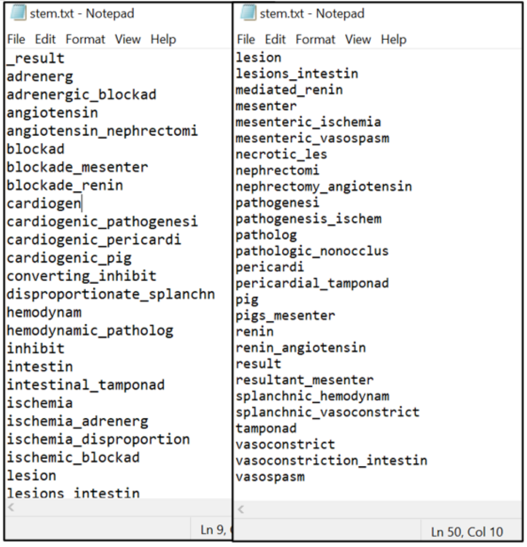
Tags: Text Processing

Synopsis

Generates word vectors from string attributes.

Figure 17 : Add the 'Stem (Porter)' operator into the process

Output :



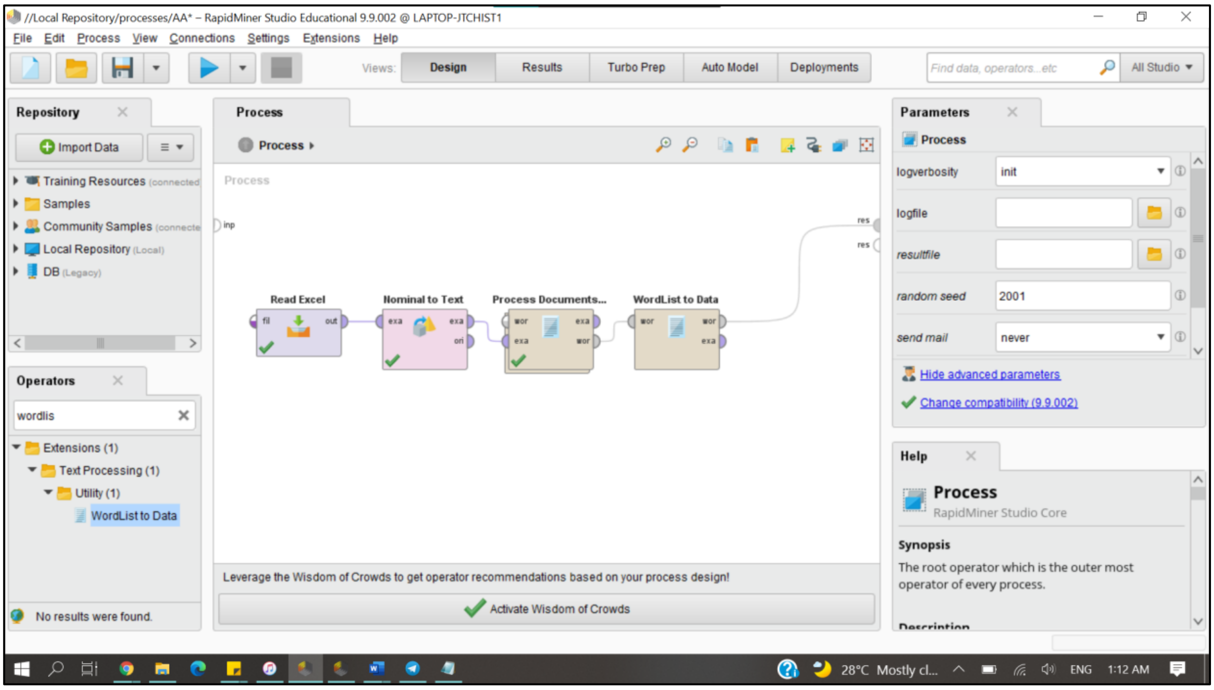
```
_result
adrenerg
adrenergic_blockad
angiotensin
angiotensin_nephrectomi
blockad
blockade_mesenter
blockade_renin
cardiogen
cardiogenic_pathogenesi
cardiogenic_pericardi
cardiogenic_pig
converting_inhibit
disproportionate_splanchn
hemodynam
hemodynamic_patholog
inhibit
intestin
intestinal_tamponad
ischemia
ischemia_adrenerg
ischemia_disproportion
ischemic_blockad
lesion
lesions_intestin
```

```
lesion
lesions_intestin
mediated_renin
mesenter
mesenteric_ischemia
mesenteric_vasospasm
necrotic_les
nephrectomi
nephrectomy_angiotensin
pathogenesi
pathogenesis_ischem
patholog
pathologic_nonocclus
pericardi
pericardial_tamponad
pig
pigs_mesenter
renin
renin_angiotensin
result
resultant_mesenter
splanchnic_hemodynam
splanchnic_vasoconstrict
tamponad
vasoconstrict
vasoconstriction_intestin
vasospasm
```

Figure 18 : A list of words generated after the execution of Stem (Porter) process

Step 10 : Wordlist to Data

Input :


//Local Repository/processes/AA* – RapidMiner Studio Educational 9.9.002 @ LAPTOP-JTCHIST1
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
DB (Legacy)
Operators
Extensions (1)
Text Processing (1)
Utility (1)
WordList to Data
Process
Read Excel
Nominal to Text
Process Documents...
WordList to Data
Parameters
Help
Activate Wisdom of Crowds
Hide advanced parameters
Change compatibility (9.9.002)
28°C Mostly cl...
ENG
1:12 AM

Operators search: wordlis

No results were found.

Parameters – Process:

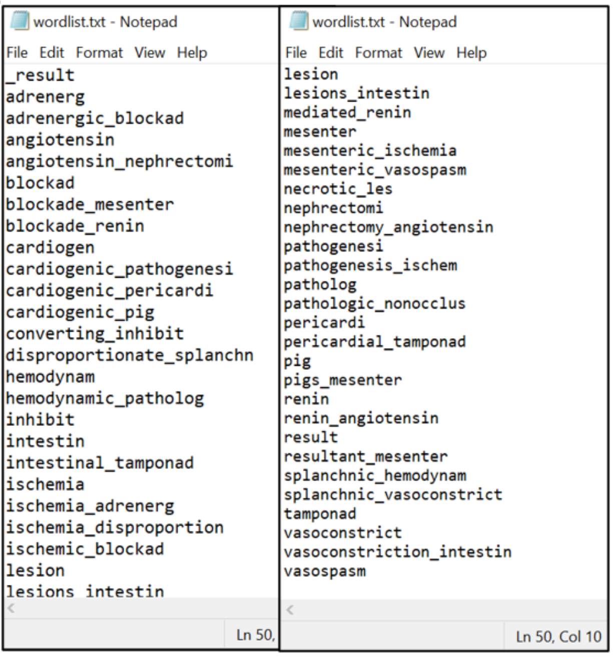
logverbosity: init
logfile:
resultfile:
random seed: 2001
send mail: never

Help – Process
RapidMiner Studio Core
Synopsis
The root operator which is the outer most operator of every process.
Description

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

Figure 19 : Add the 'Wordlist to Data' operator

Output :



```
_result
adrenerg
adrenergic_blockad
angiotensin
angiotensin_nephrectomi
blockad
blockade_mesenter
blockade_renin
cardiogen
cardiogenic_pathogenesi
cardiogenic_pericardi
cardiogenic_pig
converting_inhibit
disproportionate_splanchn
hemodynam
hemodynamic_patholog
inhibit
intestin
intestinal_tamponad
ischemia
ischemia_adrenerg
ischemia_disproportion
ischemic_blockad
lesion
lesions_intestin
```

```
lesion
lesions_intestin
mediated_renin
mesenter
mesenteric_ischemia
mesenteric_vasospasm
necrotic_les
nephrectomi
nephrectomy_angiotensin
pathogenesi
pathogenesis_ischem
patholog
pathologic_nonocclus
pericardi
pericardial_tamponad
pig
pigs_mesenter
renin
renin_angiotensin
result
resultant_mesenter
splanchnic_hemodynam
splanchnic_vasoconstrict
tamponad
vasoconstrict
vasoconstriction_intestin
vasospasm
```

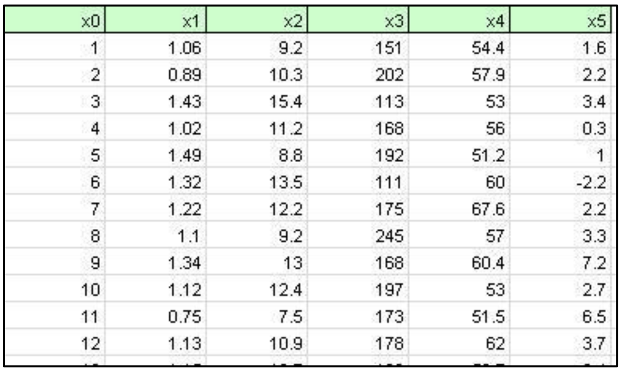
Figure 20 : The final result

1.2 Data Mining Task

1.2.1 Supervised Learning

a)

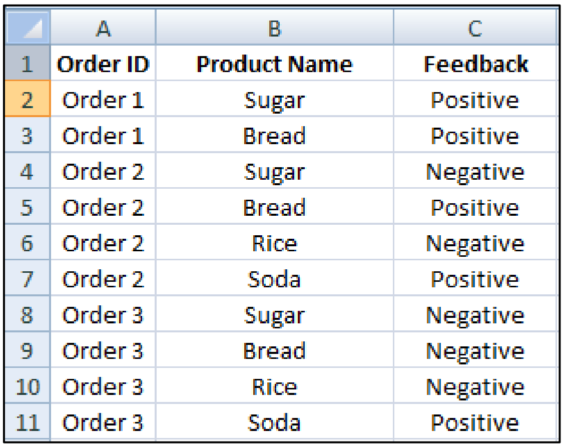
The one supervised learning algorithm that I can apply for appendix 2 is Classification. There are 2 common algorithms for the supervised which are Regression and Classification. Numerical data is made of numbers and it has two more categories that falls under it which are continuous and discrete. To use the Regression, we need the data in continuous target variable. Figure below shows the example of continuous data.



x0	x1	x2	x3	x4	x5
1	1.06	9.2	151	54.4	1.6
2	0.89	10.3	202	57.9	2.2
3	1.43	15.4	113	53	3.4
4	1.02	11.2	168	56	0.3
5	1.49	8.8	192	51.2	1
6	1.32	13.5	111	60	-2.2
7	1.22	12.2	175	67.6	2.2
8	1.1	9.2	245	57	3.3
9	1.34	13	168	60.4	7.2
10	1.12	12.4	197	53	2.7
11	0.75	7.5	173	51.5	6.5
12	1.13	10.9	178	62	3.7

Figure 21 : Continuous data

While on the other hand, we also have Categorical data. Categorical data is made of words. Under it. It has two more categories which are ordinal and nominal. Classification algorithm is used to interpret the nominal data. Just like the appendix 2, it its made of words. Therefore, the most suitable supervised learning algorithm is Classification. Figure below shows an example of nominal data.



	A	B	C
1	Order ID	Product Name	Feedback
2	Order 1	Sugar	Positive
3	Order 1	Bread	Positive
4	Order 2	Sugar	Negative
5	Order 2	Bread	Positive
6	Order 2	Rice	Negative
7	Order 2	Soda	Positive
8	Order 3	Sugar	Negative
9	Order 3	Bread	Negative
10	Order 3	Rice	Negative
11	Order 3	Soda	Positive

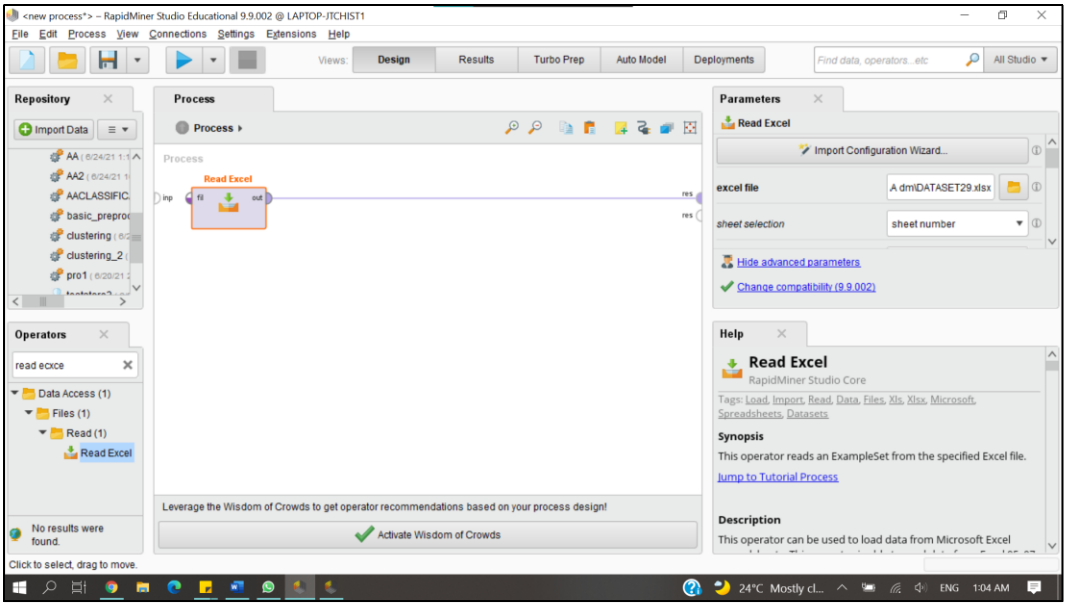
Figure 22 : Nominal data

b)

Step 1 : Read Excel

To prepare the dataset for training, I read my dataset 29 that I already converted into an excel file by using the Read Excel operator.

Input :


<new process*> – RapidMiner Studio Educational 9.9.002 @ LAPTOP-JTCHIST1
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
AA
AA2
AACLASSIFIC
basic_preproc
clustering
clustering_2
pro1
Process
Read Excel
inp
fil
out
res
res
Operators
read ecxce
Data Access (1)
Files (1)
Read (1)
Read Excel
No results were found.
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Click to select, drag to move.
24°C Mostly cl... ENG 1:04 AM

Parameters

Read Excel

Import Configuration Wizard...

excel file: A dm\DATASET29.xlsx

sheet selection: sheet number

Hide advanced parameters

Change compatibility (9.9.002)

Help

Read Excel

RapidMiner Studio Core

Tags: Load, Import, Read, Data, Files, Xls, Xlsx, Microsoft, Spreadsheets, Datasets

Synopsis

This operator reads an ExampleSet from the specified Excel file.

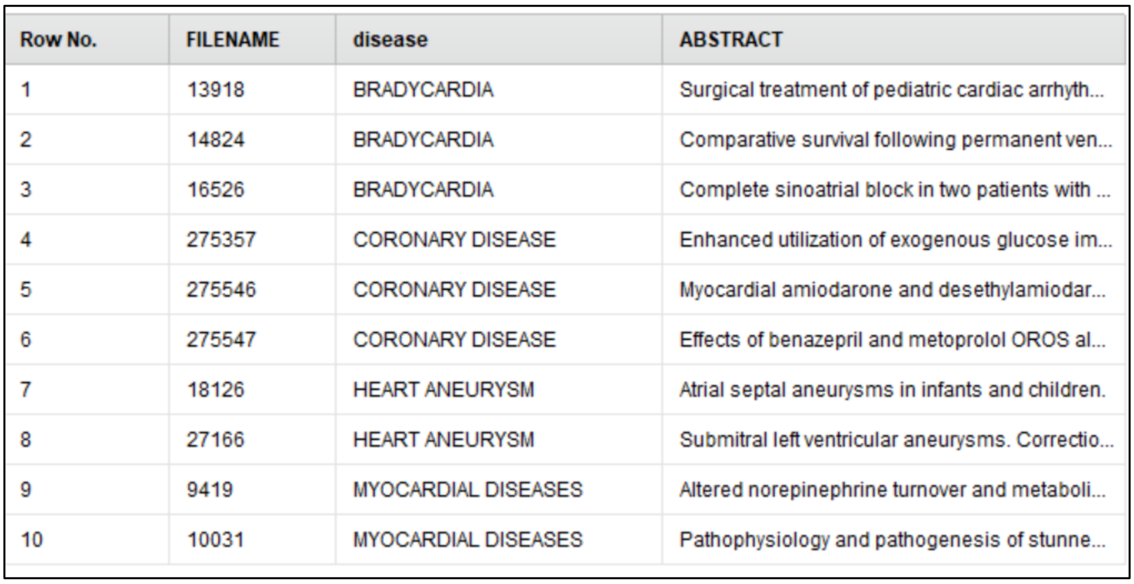
Jump to Tutorial Process

Description

This operator can be used to load data from Microsoft Excel

Figure 23 : Read the data from excel

Output :



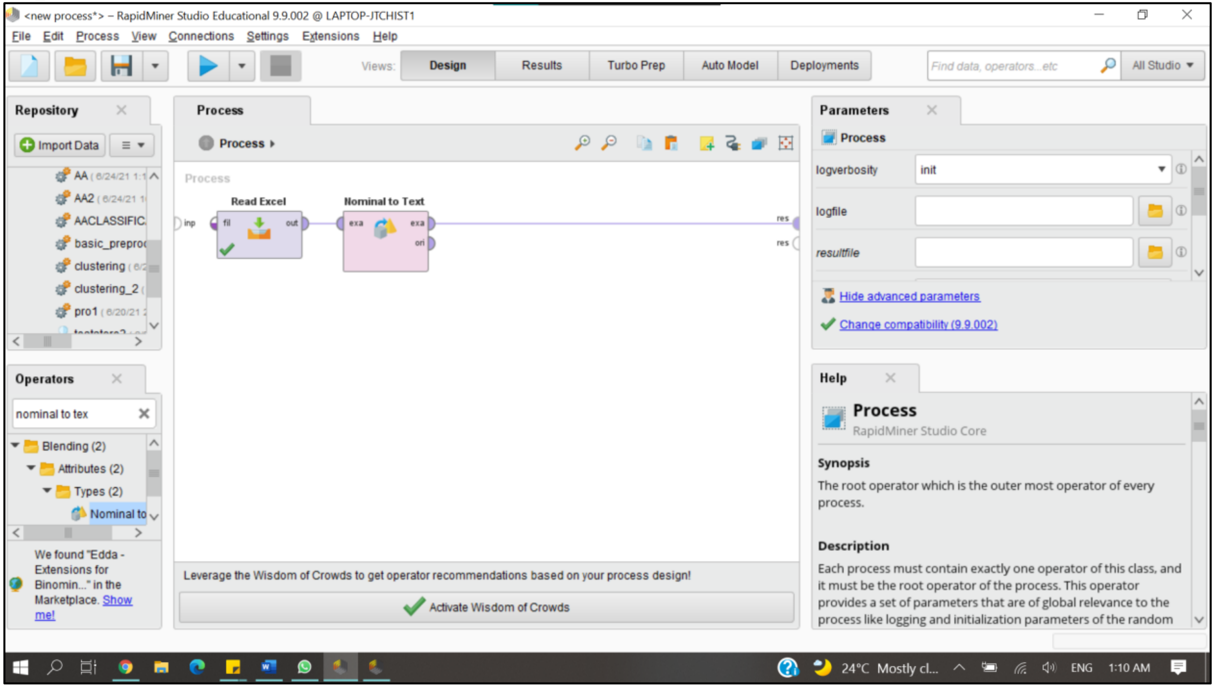
Row No.	FILENAME	disease	ABSTRACT
1	13918	BRADYCARDIA	Surgical treatment of pediatric cardiac arrhyth...
2	14824	BRADYCARDIA	Comparative survival following permanent ven...
3	16526	BRADYCARDIA	Complete sinoatrial block in two patients with ...
4	275357	CORONARY DISEASE	Enhanced utilization of exogenous glucose im...
5	275546	CORONARY DISEASE	Myocardial amiodarone and desethylamiodar...
6	275547	CORONARY DISEASE	Effects of benazepril and metoprolol OROS al...
7	18126	HEART ANEURYSM	Atrial septal aneurysms in infants and children.
8	27166	HEART ANEURYSM	Submitral left ventricular aneurysms. Correctio...
9	9419	MYOCARDIAL DISEASES	Altered norepinephrine turnover and metaboli...
10	10031	MYOCARDIAL DISEASES	Pathophysiology and pathogenesis of stunne...

Figure 24 : Data that has been read using read excel

Step 2 : Nominal To Text

Next, I added the Nominal to Text operator. The Nominal to Text operator converts all nominal attributes to string attributes.

Input :



We found "Edda - Extensions for Binomin..." in the Marketplace. Show me!

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

Process

logverbosity: init

logfile:

resultfile:

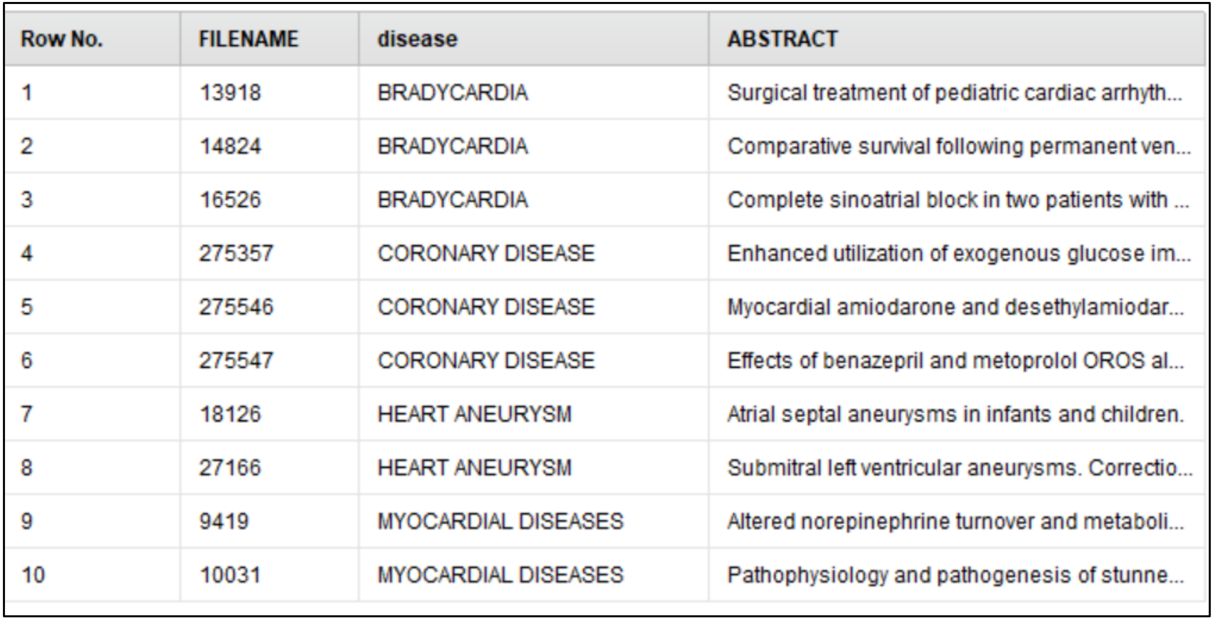
Process
RapidMiner Studio Core

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

Description
Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of global relevance to the process like logging and initialization parameters of the random

Figure 25 : Add the Nominal To Text operator

Output :



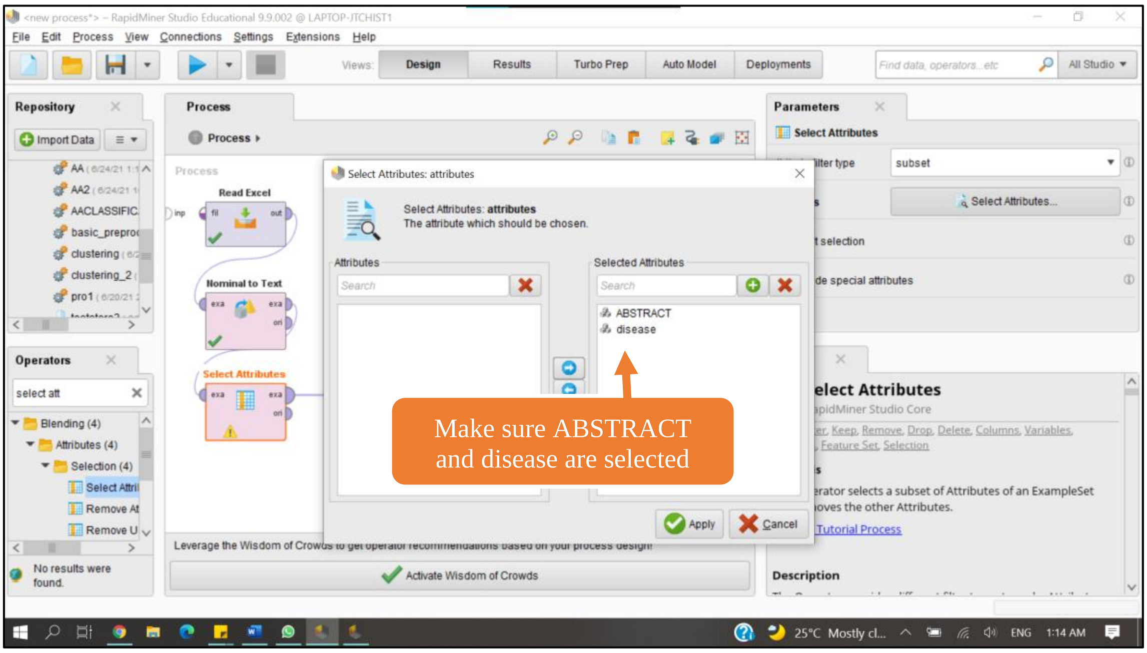
Row No.	FILENAME	disease	ABSTRACT
1	13918	BRADYCARDIA	Surgical treatment of pediatric cardiac arrhyth...
2	14824	BRADYCARDIA	Comparative survival following permanent ven...
3	16526	BRADYCARDIA	Complete sinoatrial block in two patients with ...
4	275357	CORONARY DISEASE	Enhanced utilization of exogenous glucose im...
5	275546	CORONARY DISEASE	Myocardial amiodarone and desethylamiodar...
6	275547	CORONARY DISEASE	Effects of benazepril and metoprolol OROS al...
7	18126	HEART ANEURYSM	Atrial septal aneurysms in infants and children.
8	27166	HEART ANEURYSM	Submitral left ventricular aneurysms. Correctio...
9	9419	MYOCARDIAL DISEASES	Altered norepinephrine turnover and metaboli...
10	10031	MYOCARDIAL DISEASES	Pathophysiology and pathogenesis of stunne...

Figure 26 : The data after has changed to text

Step 3 : Select Attribute

Next, I will select the attributes. From here I selected ABSTRACT and disease as the chosen attributes.

Input :



Select Attributes: attributes

Select Attributes: **attributes**
The attribute which should be chosen.

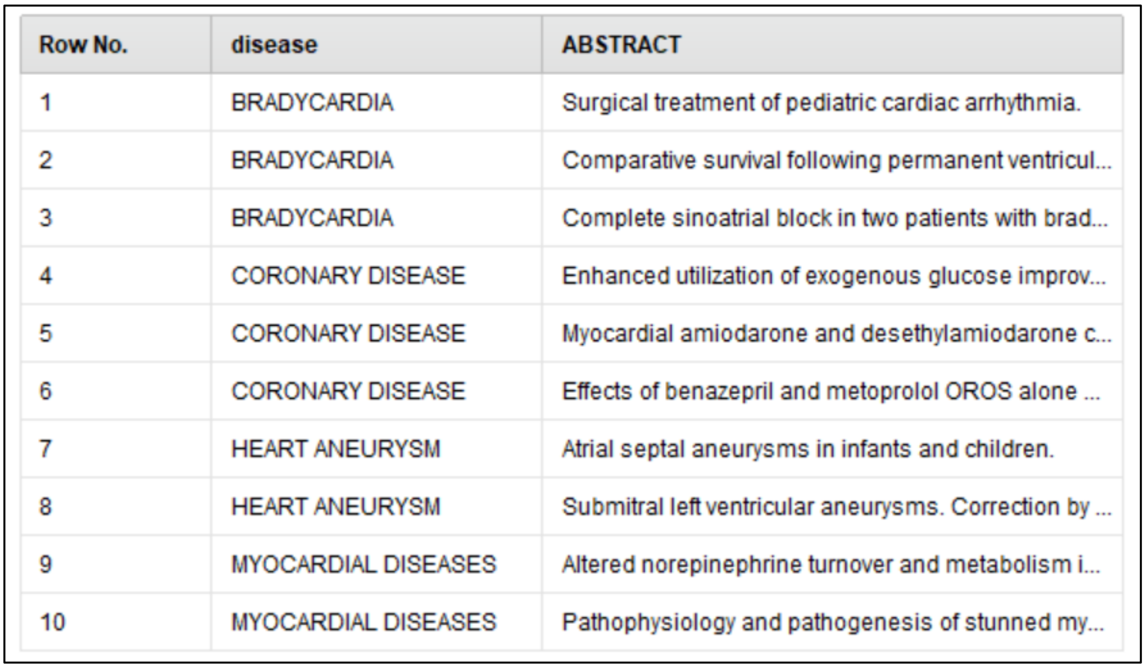
Selected Attributes
ABSTRACT
disease

Make sure ABSTRACT and disease are selected

<new process*> – RapidMiner Studio Educational 9.9.002 @ LAPTOP-JTCHIST1
File Edit Process View Connections Settings Extensions Help
Design
Results
Turbo Prep
Auto Model
Deployments
Repository
Import Data
Operators
select att
Process
Read Excel
Nominal to Text
Select Attributes
Attributes
Search
Apply
Cancel
Parameters
Select Attributes
subset
Select Attributes...
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!
Activate Wisdom of Crowds
Description

Figure 27 : Select Attribute

Output :



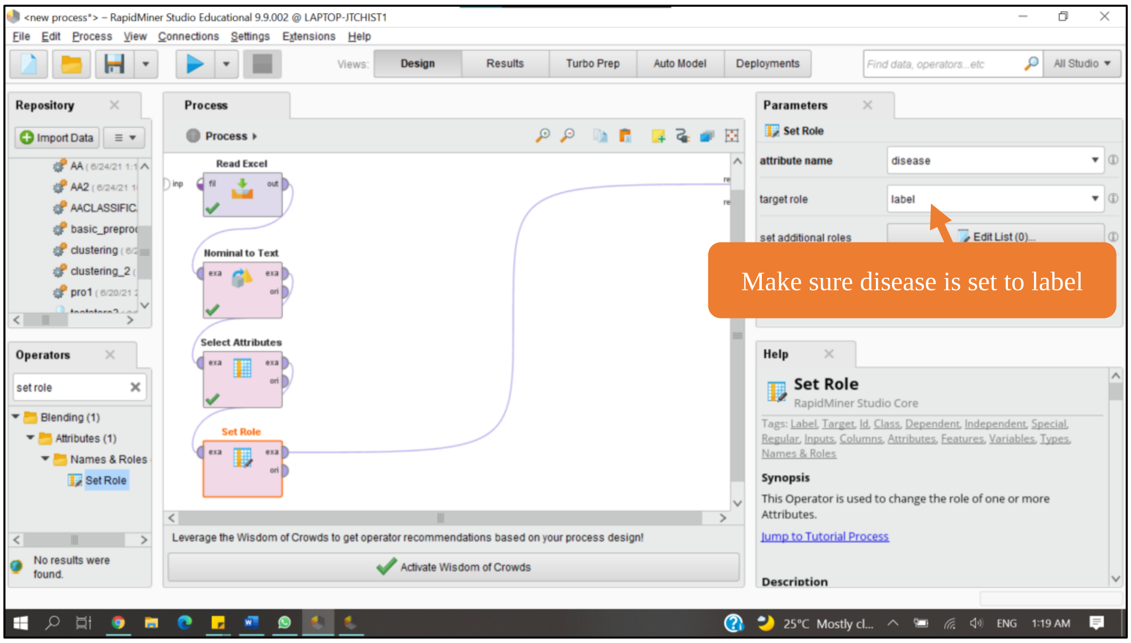
Row No.	disease	ABSTRACT
1	BRADYCARDIA	Surgical treatment of pediatric cardiac arrhythmia.
2	BRADYCARDIA	Comparative survival following permanent ventricul...
3	BRADYCARDIA	Complete sinoatrial block in two patients with brad...
4	CORONARY DISEASE	Enhanced utilization of exogenous glucose improv...
5	CORONARY DISEASE	Myocardial amiodarone and desethylamiodarone c...
6	CORONARY DISEASE	Effects of benazepril and metoprolol OROS alone ...
7	HEART ANEURYSM	Atrial septal aneurysms in infants and children.
8	HEART ANEURYSM	Submitral left ventricular aneurysms. Correction by ...
9	MYOCARDIAL DISEASES	Altered norepinephrine turnover and metabolism i...
10	MYOCARDIAL DISEASES	Pathophysiology and pathogenesis of stunned my...

Figure 28 : The result after choosing specific attributes

Step 4 : Set Role

Then, I set the role for disease as label. Label is a special role. An Attribute with the id role acts as an identifier for the Examples. It should be unique for all Examples.

Input :



<new process*> – RapidMiner Studio Educational 9.9.002 @ LAPTOP-JTCHIST1
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Repository
Import Data
Process
Read Excel
Nominal to Text
Select Attributes
Set Role
Operators
set role
Blending (1)
Attributes (1)
Names & Roles
Set Role
Parameters
Set Role

attribute name: disease

target role: label

set additional roles: Edit List (0)...

Make sure disease is set to label

Set Role
RapidMiner Studio Core

Tags: Label, Target, Id, Class, Dependent, Independent, Special, Regular, Inputs, Columns, Attributes, Features, Variables, Types, Names & Roles

Synopsis

This Operator is used to change the role of one or more Attributes.

Jump to Tutorial Process

Description

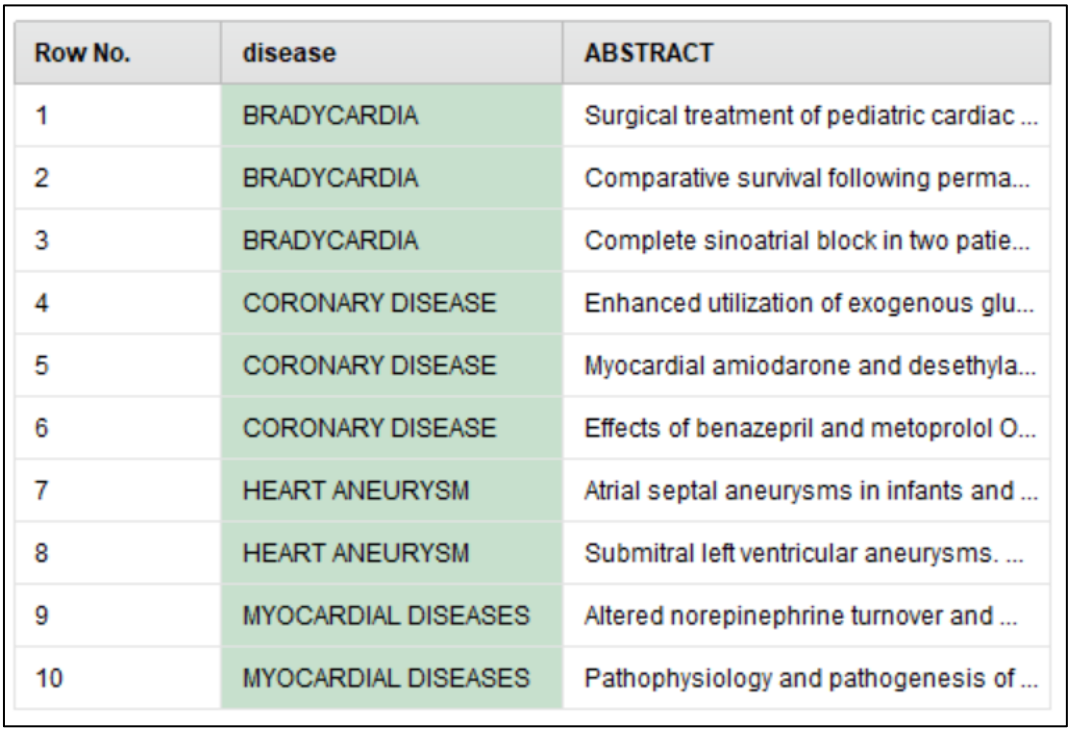
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Activate Wisdom of Crowds

No results were found.

Figure 29 : Set the role label for Disease

Output :



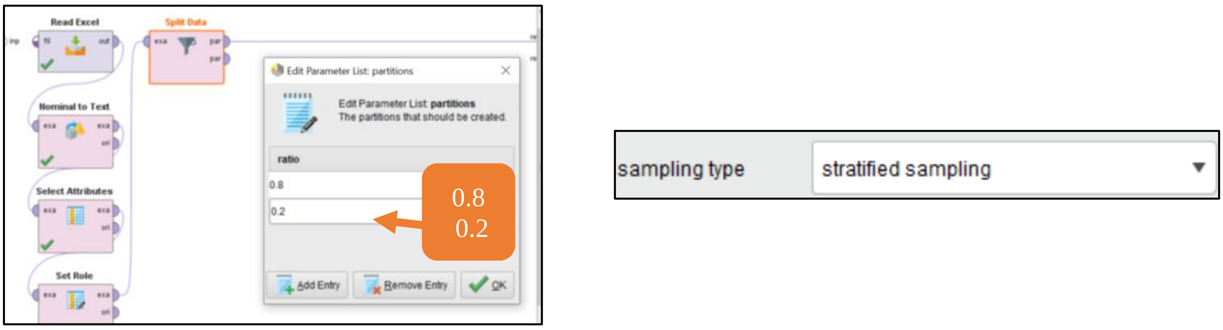
Row No.	disease	ABSTRACT
1	BRADYCARDIA	Surgical treatment of pediatric cardiac ...
2	BRADYCARDIA	Comparative survival following perma...
3	BRADYCARDIA	Complete sinoatrial block in two patie...
4	CORONARY DISEASE	Enhanced utilization of exogenous glu...
5	CORONARY DISEASE	Myocardial amiodarone and desethyla...
6	CORONARY DISEASE	Effects of benazepril and metoprolol O...
7	HEART ANEURYSM	Atrial septal aneurysms in infants and ...
8	HEART ANEURYSM	Submitral left ventricular aneurysms. ...
9	MYOCARDIAL DISEASES	Altered norepinephrine turnover and ...
10	MYOCARDIAL DISEASES	Pathophysiology and pathogenesis of ...

Figure 30 : The disease after has been set role as label

Step 5 : Split Data

Continue, I split data into training and testing dataset by 0.8 (80%) and 0.2 (20%) respectively. Splitting data into training and testing sets is a significant piece of assessing data mining models. By utilizing comparable information for training and testing, we can limit the impacts of information errors and better comprehend the qualities of the model. [2] After the completion I keep the testing data into a file data named teststore1.

Input :



Edit Parameter List: partitions

Edit Parameter List: **partitions**
The partitions that should be created.

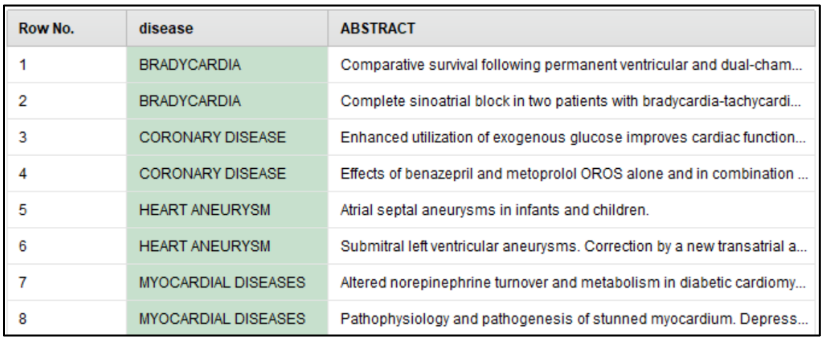
ratio
0.8
0.2

0.8
0.2

sampling type: stratified sampling

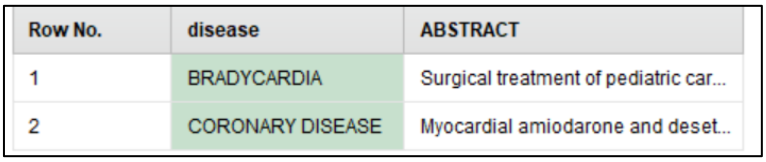
Figure 31 : Split Data into 0.8 and 0.2 and change the sampling type

Output :



Row No.	disease	ABSTRACT
1	BRADYCARDIA	Comparative survival following permanent ventricular and dual-cham...
2	BRADYCARDIA	Complete sinoatrial block in two patients with bradycardia-tachycardi...
3	CORONARY DISEASE	Enhanced utilization of exogenous glucose improves cardiac function...
4	CORONARY DISEASE	Effects of benazepril and metoprolol OROS alone and in combination ...
5	HEART ANEURYSM	Atrial septal aneurysms in infants and children.
6	HEART ANEURYSM	Submitral left ventricular aneurysms. Correction by a new transatrial a...
7	MYOCARDIAL DISEASES	Altered norepinephrine turnover and metabolism in diabetic cardiomy...
8	MYOCARDIAL DISEASES	Pathophysiology and pathogenesis of stunned myocardium. Depress...

Figure 32 : The training dataset (80%)



Row No.	disease	ABSTRACT
1	BRADYCARDIA	Surgical treatment of pediatric car...
2	CORONARY DISEASE	Myocardial amiodarone and deset...

Figure 33 : The testing dataset (20%)

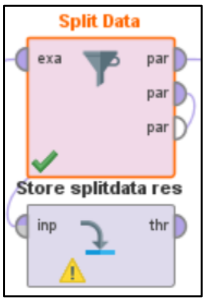


Figure 34 : Store the result into a data file

Step 6 : Preprocessing

Before starting any process, I should do data cleaning to my dataset. Therefore, I added Process Document from Data operator and do all preprocessing steps inside the operator.

Input :

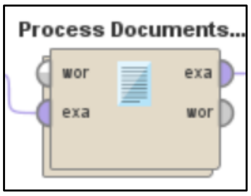


Figure 35 : Add Process Documents

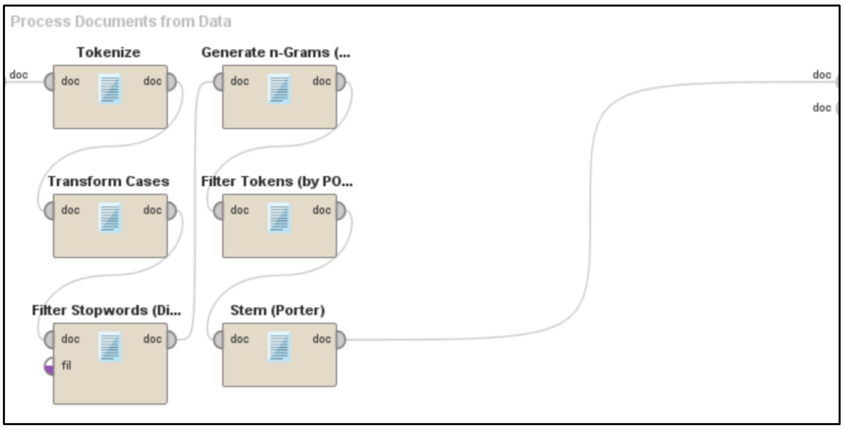
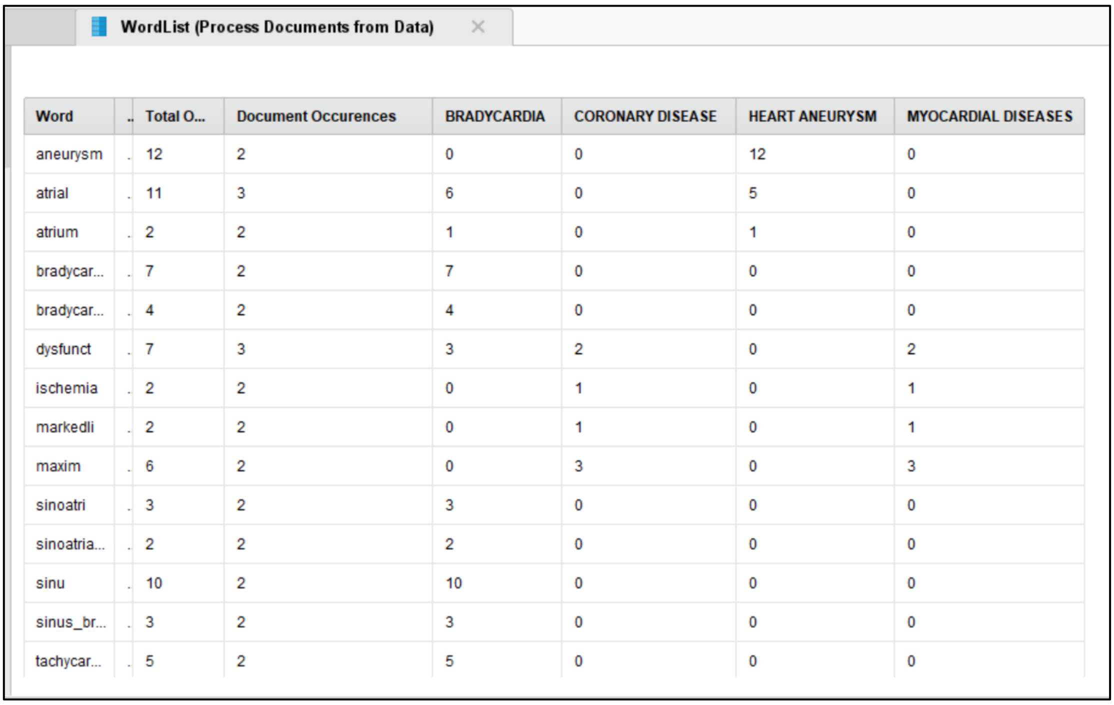


Figure 36 : Perform Preprocessing steps

Output :


WordList (Process Documents from Data)

Word	..	Total O...	Document Occurences	BRADYCARDIA	CORONARY DISEASE	HEART ANEURYSM	MYOCARDIAL DISEASES
aneurysm	.	12	2	0	0	12	0
atrial	.	11	3	6	0	5	0
atrium	.	2	2	1	0	1	0
bradycar...	.	7	2	7	0	0	0
bradycar...	.	4	2	4	0	0	0
dysfunct	.	7	3	3	2	0	2
ischemia	.	2	2	0	1	0	1
markedli	.	2	2	0	1	0	1
maxim	.	6	2	0	3	0	3
sinoatri	.	3	2	3	0	0	0
sinoatria...	.	2	2	2	0	0	0
sinu	.	10	2	10	0	0	0
sinus_br...	.	3	2	3	0	0	0
tachycar...	.	5	2	5	0	0	0

Figure 37 : Data after going through preprocessing

Step 7 : Cross Validation

Next, I added Cross Validation operator. Cross-validation is fundamentally utilized in applied machine learning to assess the ability of a machine learning model on unseen data. That is, to utilize a restricted example to assess how the model is relied upon to act overall when used to make expectations on data not utilized during the preparation of the model. [3]

Input :

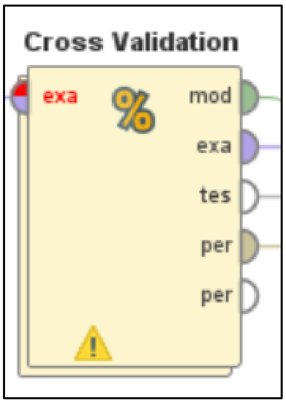


Figure 38 : Add Cross Validation operator

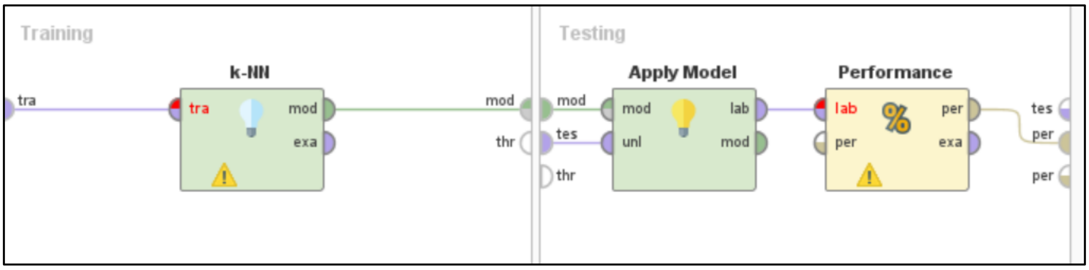
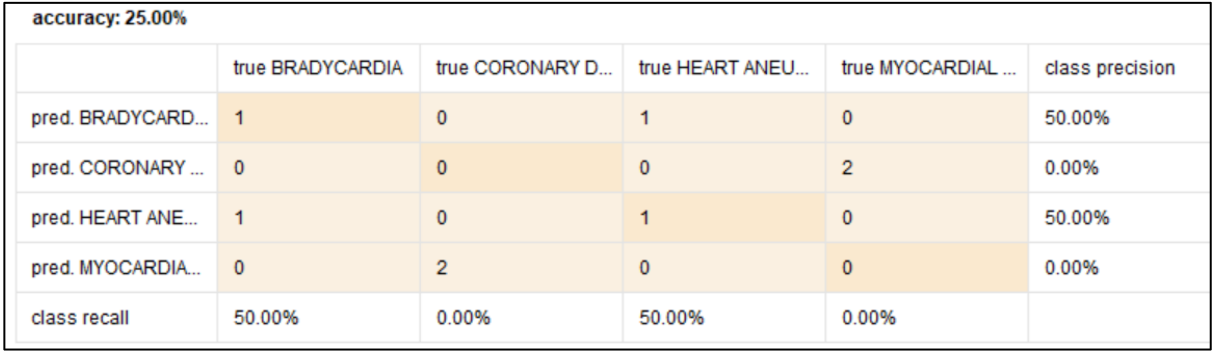


Figure 39 : Inside Cross Validation

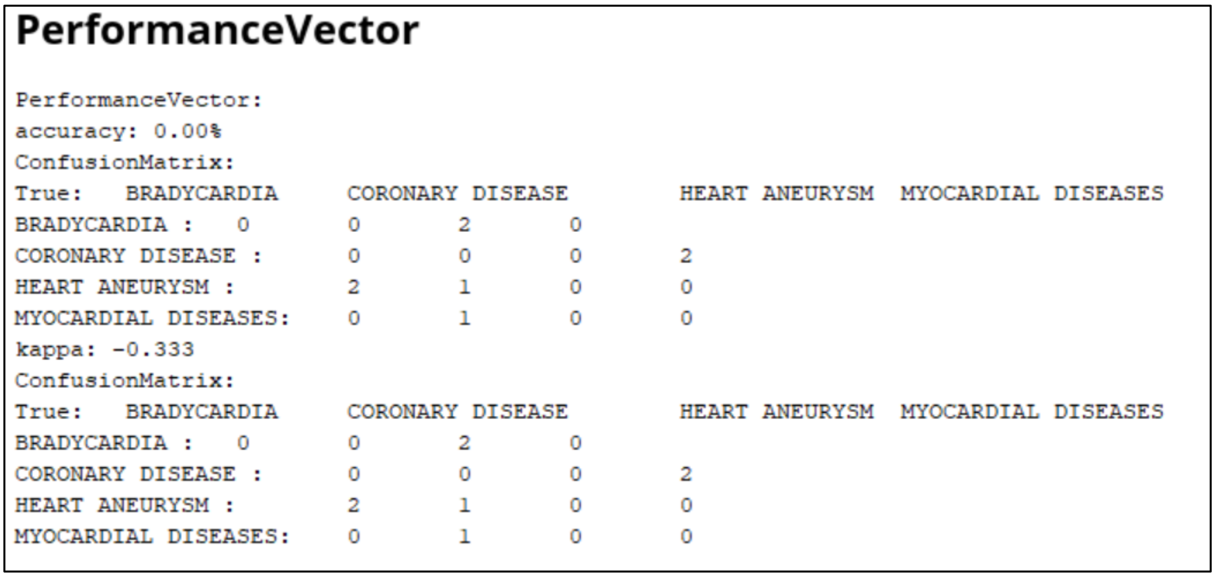
Output :



accuracy: 25.00%

	true BRADYCARDIA	true CORONARY D...	true HEART ANEU...	true MYOCARDIAL ...	class precision
pred. BRADYCARD...	1	0	1	0	50.00%
pred. CORONARY ...	0	0	0	2	0.00%
pred. HEART ANE...	1	0	1	0	50.00%
pred. MYOCARDIA...	0	2	0	0	0.00%
class recall	50.00%	0.00%	50.00%	0.00%	

Figure 40 : The accuracy



PerformanceVector

```
PerformanceVector:
accuracy: 0.00%
ConfusionMatrix:
True:   BRADYCARDIA     CORONARY DISEASE        HEART ANEURYSM  MYOCARDIAL DISEASES
BRADYCARDIA :   0       0       2       0
CORONARY DISEASE :      0       0       0       2
HEART ANEURYSM :        2       1       0       0
MYOCARDIAL DISEASES:    0       1       0       0
kappa: -0.333
ConfusionMatrix:
True:   BRADYCARDIA     CORONARY DISEASE        HEART ANEURYSM  MYOCARDIAL DISEASES
BRADYCARDIA :   0       0       2       0
CORONARY DISEASE :      0       0       0       2
HEART ANEURYSM :        2       1       0       0
MYOCARDIAL DISEASES:    0       1       0       0
```

Figure 41 : Performance Vector

Testing

Step 1 : Get the data and test

To start the testing, I fetched my data and testing dataset from the file that I have stored before during the process Training.

Input :

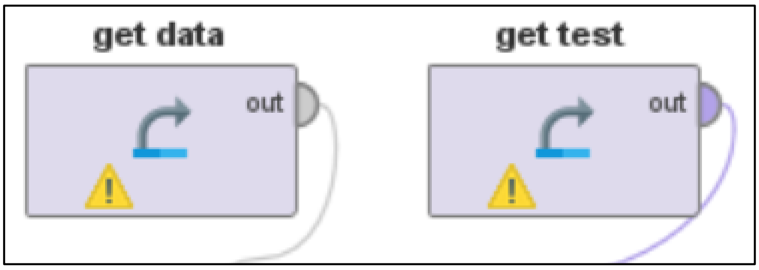
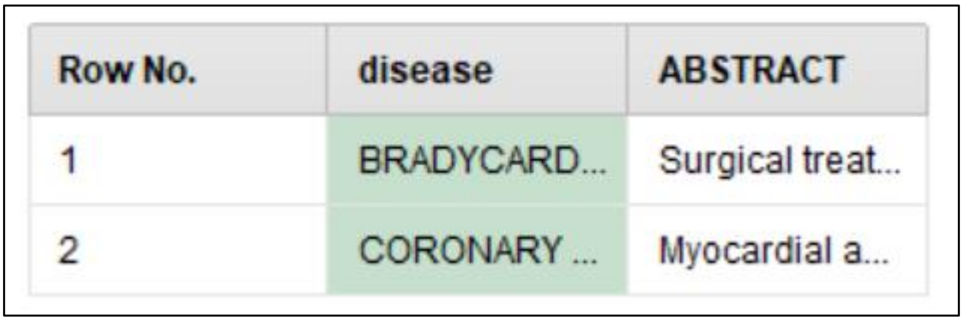


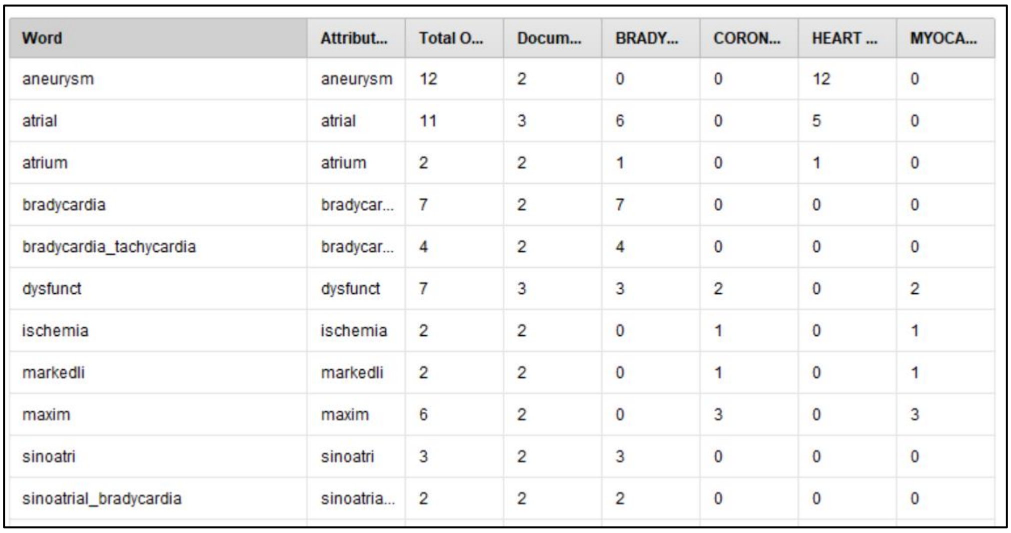
Figure 42 : Get Data and Test

Output :



Row No.	disease	ABSTRACT
1	BRADYCARD...	Surgical treat...
2	CORONARY ...	Myocardial a...

Figure 43 : Result for get Test



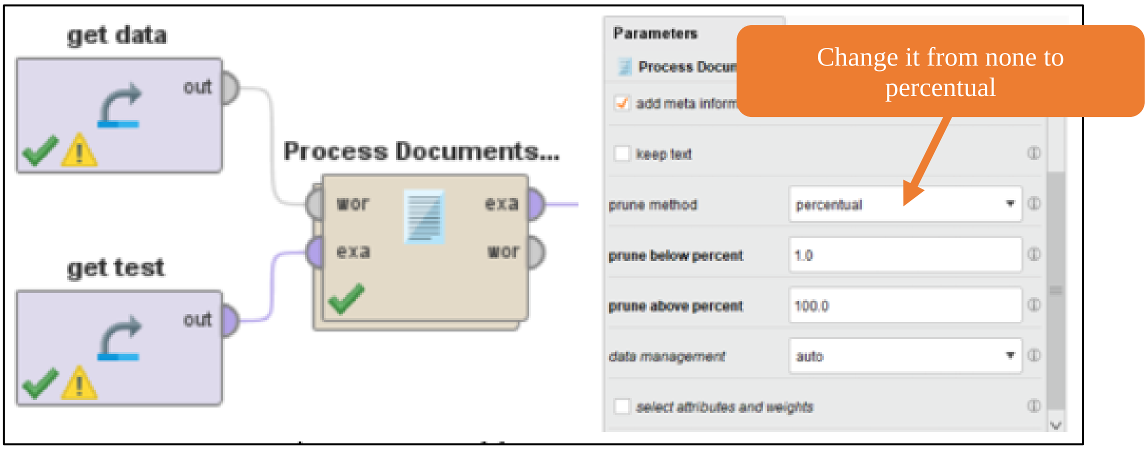
Word	Attribut...	Total O...	Docum...	BRADY...	CORON...	HEART ...	MYOCA...
aneurysm	aneurysm	12	2	0	0	12	0
atrial	atrial	11	3	6	0	5	0
atrium	atrium	2	2	1	0	1	0
bradycardia	bradycar...	7	2	7	0	0	0
bradycardia_tachycardia	bradycar...	4	2	4	0	0	0
dysfunct	dysfunct	7	3	3	2	0	2
ischemia	ischemia	2	2	0	1	0	1
markedli	markedli	2	2	0	1	0	1
maxim	maxim	6	2	0	3	0	3
sinoatri	sinoatri	3	2	3	0	0	0
sinoatrial_bradycardia	sinoatria...	2	2	2	0	0	0

Figure 44 : Result for get Data

Step 2 : Preprocess

To double the secure, I preprocessed both data again and the prune method is percentual.

Input :



Parameters

Process Docum

☑ add meta inform

□ keep text

prune method: percentual

prune below percent: 1.0

prune above percent: 100.0

data management: auto

□ select attributes and weights

Change it from none to percentual

Figure 45 : Add Process Document

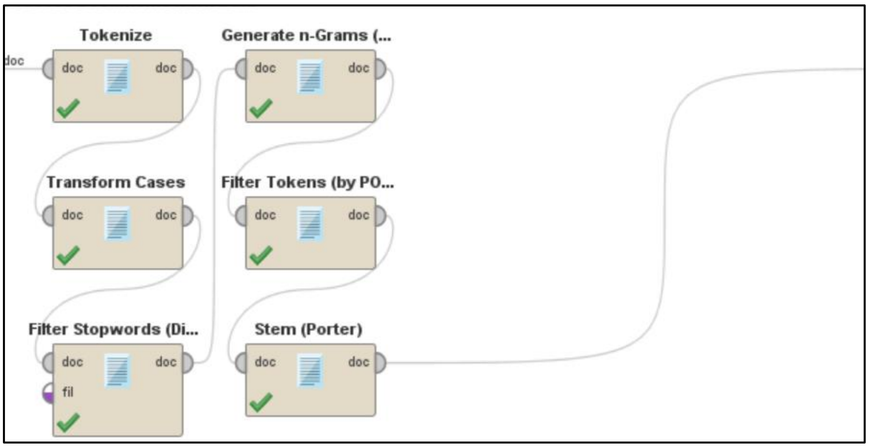
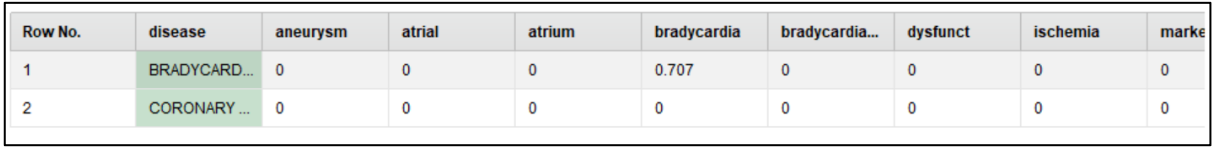


Figure 46 : Inside of the process document

Output :



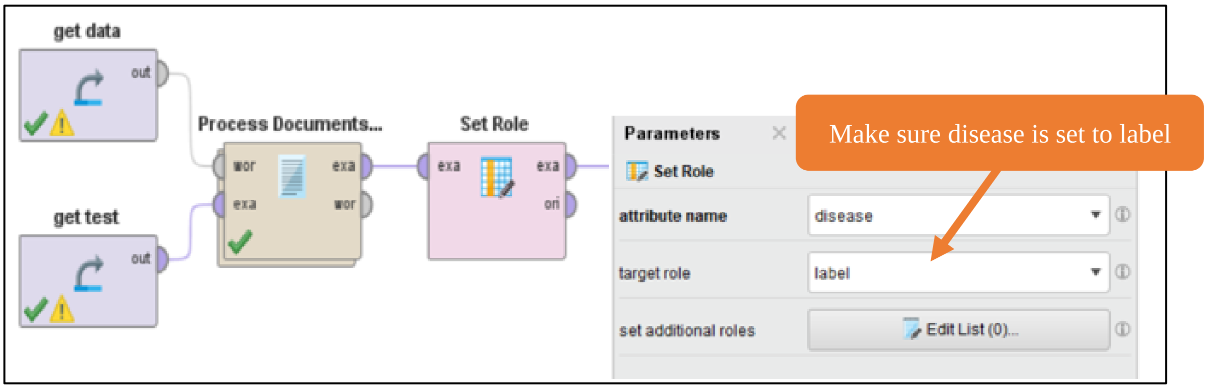
Row No.	disease	aneurysm	atrial	atrium	bradycardia	bradycardia...	dysfunct	ischemia	marke
1	BRADYCARD...	0	0	0	0.707	0	0	0	0
2	CORONARY ...	0	0	0	0	0	0	0	0

Figure 47 : The result after going through preprocess

Step 3 : Set Role

Next, I set role for disease as the label. This step is important. Setting an attribute to label is a very important role. An Attribute with the id role goes about as an identifier for the Examples. It should be unique for all Examples.

Input :



Parameters

Set Role

attribute name: disease

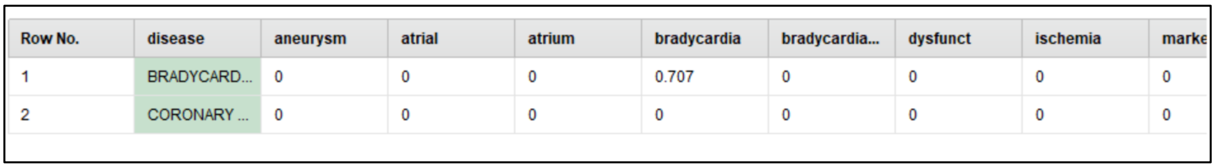
target role: label

set additional roles: Edit List (0)...

Make sure disease is set to label

Figure 48 : Set Role for attribute disease

Output :



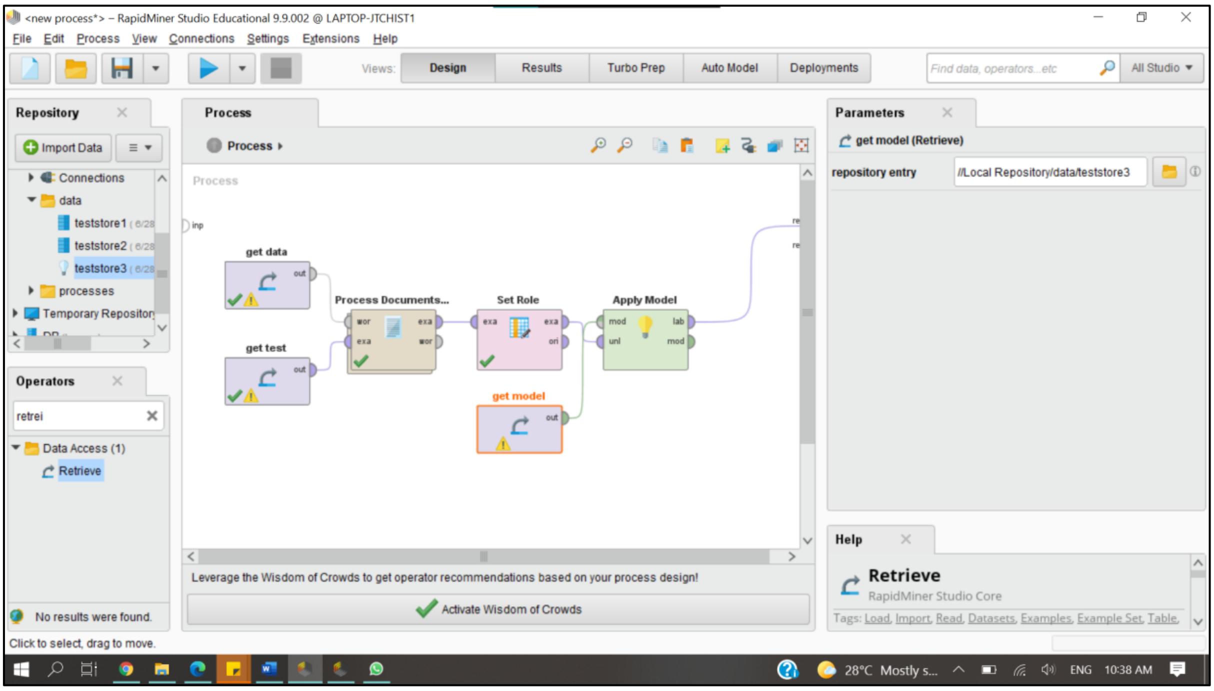
Row No.	disease	aneurysm	atrial	atrium	bradycardia	bradycardia...	dysfunct	ischemia	marke
1	BRADYCARD...	0	0	0	0.707	0	0	0	0
2	CORONARY ...	0	0	0	0	0	0	0	0

Figure 49 : The result after set role

Step 4 : Apply Model

Lastly, I will apply the model. I added the Apply Model operator. As I add the operator, it will ask for the model. Therefore, I imported the model that has been generated during the Training process that I kept in a file by using the Retrieve operator.

Input :



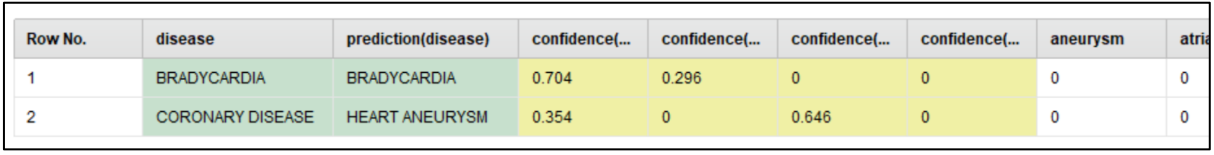
get model (Retrieve)

repository entry //Local Repository/data/teststore3

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

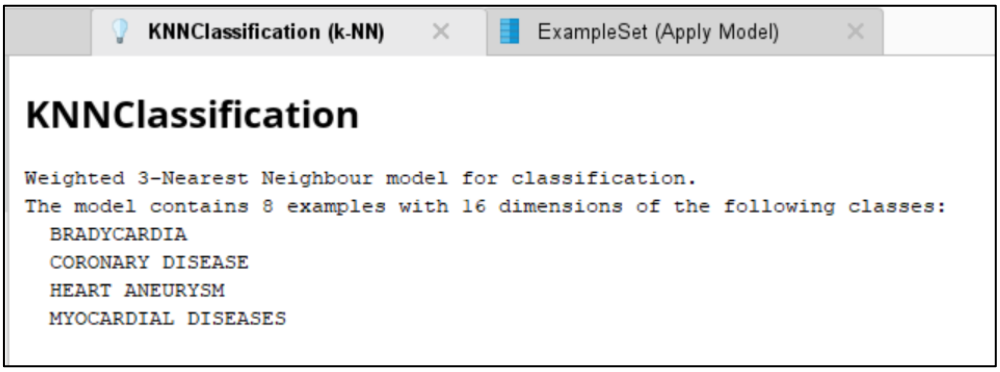
Figure 50 : Add Apply Model operator and get Model operator

Output :



Row No.	disease	prediction(disease)	confidence(...	confidence(...	confidence(...	confidence(...	aneurysm	atria
1	BRADYCARDIA	BRADYCARDIA	0.704	0.296	0	0	0	0
2	CORONARY DISEASE	HEART ANEURYSM	0.354	0	0.646	0	0	0

Figure 51 : The final result of the classification

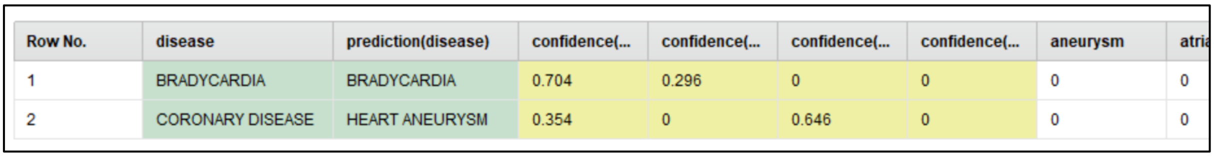


KNNClassification

```
Weighted 3-Nearest Neighbour model for classification.
The model contains 8 examples with 16 dimensions of the following classes:
  BRADYCARDIA
  CORONARY DISEASE
  HEART ANEURYSM
  MYOCARDIAL DISEASES
```

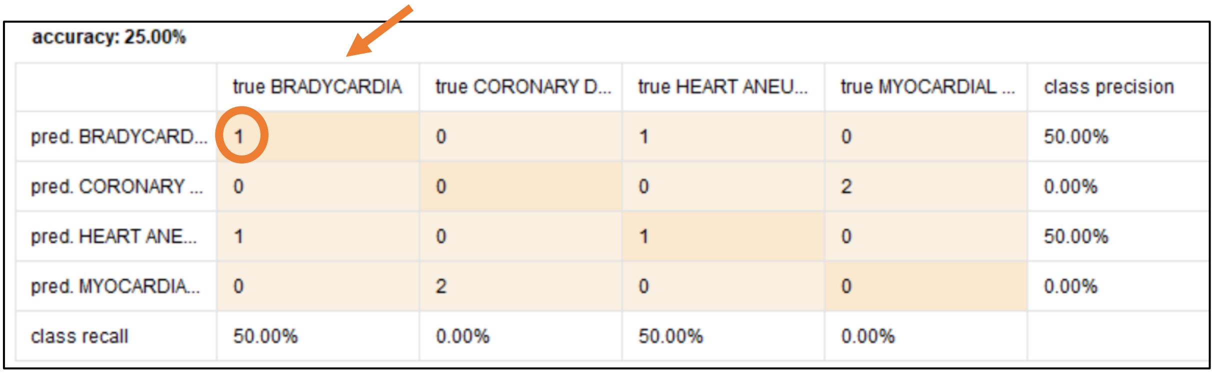
Figure 52 : KNNClassification

c) Performance Measures



Row No.	disease	prediction(disease)	confidence(...	confidence(...	confidence(...	confidence(...	aneurysm	atri
1	BRADYCARDIA	BRADYCARDIA	0.704	0.296	0	0	0	0
2	CORONARY DISEASE	HEART ANEURYSM	0.354	0	0.646	0	0	0

Figure 53 : The result for Testing dataset using KNN algorithm



accuracy: 25.00%

	true BRADYCARDIA	true CORONARY D...	true HEART ANEU...	true MYOCARDIAL ...	class precision
pred. BRADYCARD...	1	0	1	0	50.00%
pred. CORONARY ...	0	0	0	2	0.00%
pred. HEART ANE...	1	0	1	0	50.00%
pred. MYOCARDIA...	0	2	0	0	0.00%
class recall	50.00%	0.00%	50.00%	0.00%	

Figure 54 : The result for Training dataset using KNN algorithm

Based on figure 54, the total item in the table could total up to 8, which is equal to the training dataset. It has been shown that the accuracy of the training dataset is 25%. To understand the training result better, we should look at BRADYCARDIA. It predicted BRADYCARDIA to be true is only 1 while the another one it predicted it to be HEART ANEURYSM. This could cause the precision to be 50% while recall is also 50%. Even though the accuracy is low, we can't say that the whole classification process using KNN is totally wrong. The accuracy is actually depends on number of data used, the less data been used will cause the result to be a little inaccurate and insufficient.

1.2.2 Unsupervised Learning Algorithm

Unsupervised learning is likewise an exceptionally normal kind of machine learning. It contrasts from regulated learning in that the data that has no label (unlabeled). One unsupervised learning algorithm that I can apply for given appendix 2 is Clustering. There are many algorithms that falls under the unsupervised learning algorithm such as clustering, association and dimensionality reduction. Clustering is gathering a set of items in such a way that objects in a same cluster are more comparable than to those objects having a place with different cluster. While, association rules are tied in with discovering relationship among things inside huge business data sets. Clustering algorithms are for the most part utilized when we need to make the clusters dependent on the characteristic of the data focuses. For both algorithms, the target variable is not available as the data that is going to use is the unlabeled.

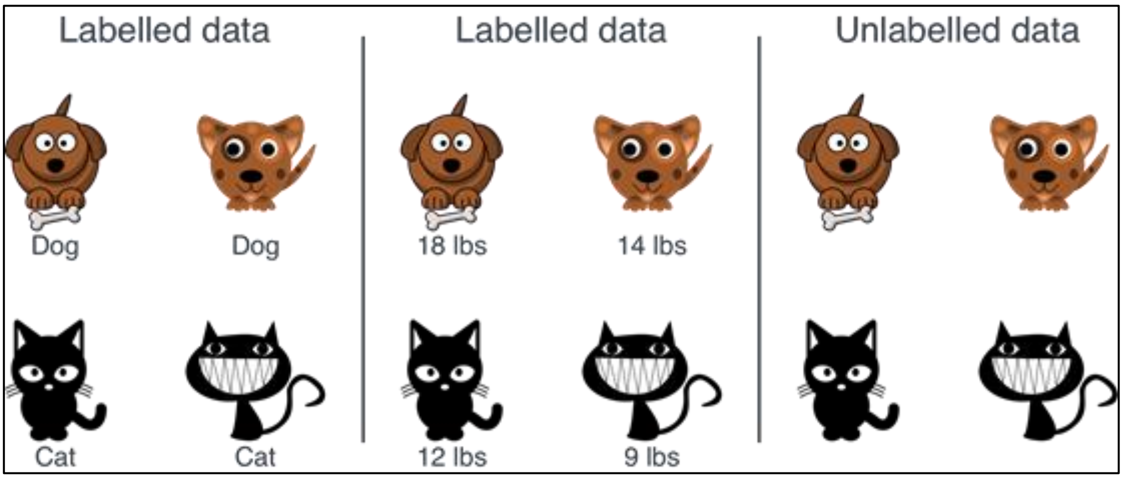


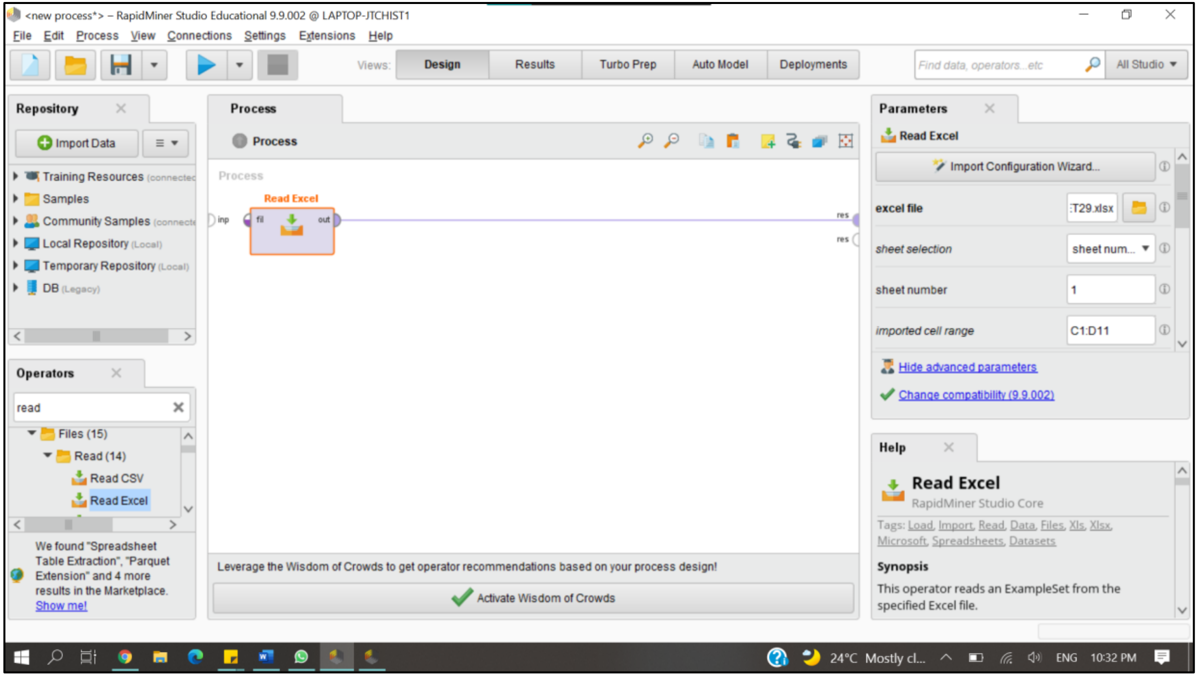
Figure 55: Visualization of data [4]

e)

Step 1 : Read Excel

The very first step is to read the dataset. For this question, I needed to refer appendix 2 and my dataset is dataset 29. Therefore, I changed the format from .docx to .xlsx so that the tool can read it easily by using the 'Read Excel' operator.

Input :



<new process*> – RapidMiner Studio Educational 9.9.002 @ LAPTOP-JTCHIST1
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Temporary Repository (Local)
DB (Legacy)
Operators
read
Files (15)
Read (14)
Read CSV
Read Excel
Process
Read Excel
inp
fil
out
res
res
Parameters
Import Configuration Wizard...
Hide advanced parameters
Change compatibility (9.9.002)
Help
Activate Wisdom of Crowds
24°C Mostly cl... ENG 10:32 PM

We found "Spreadsheet Table Extraction", "Parquet Extension" and 4 more results in the Marketplace. Show me!

Read Excel

excel file: :T29.xlsx

sheet selection: sheet num...

sheet number: 1

imported cell range: C1:D11

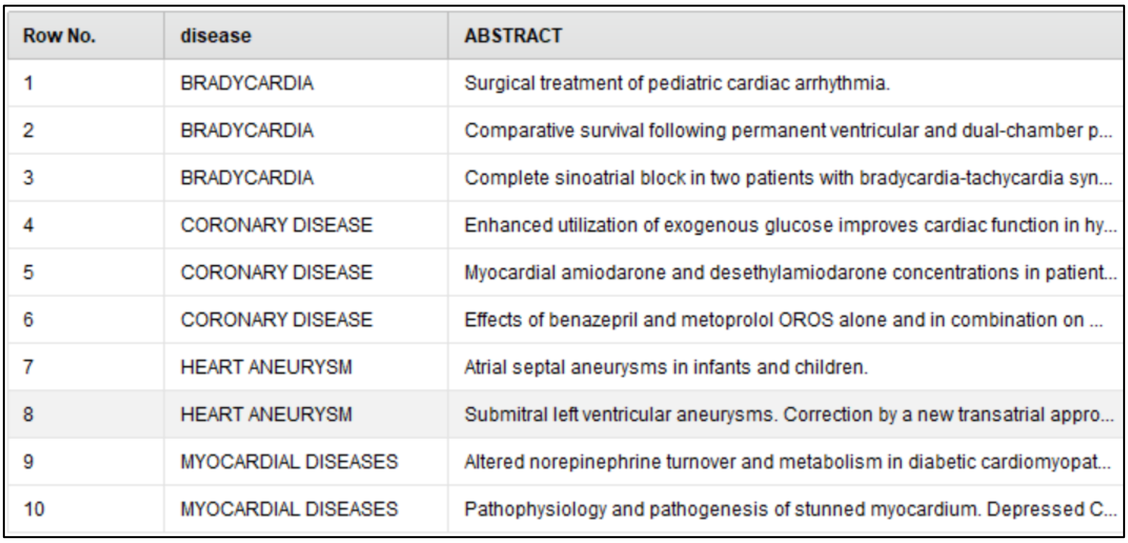
Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Read Excel
RapidMiner Studio Core
Tags: Load, Import, Read, Data, Files, Xls, Xlsx, Microsoft, Spreadsheets, Datasets

Synopsis
This operator reads an ExampleSet from the specified Excel file.

Figure 56 : Read excel operator

Output :



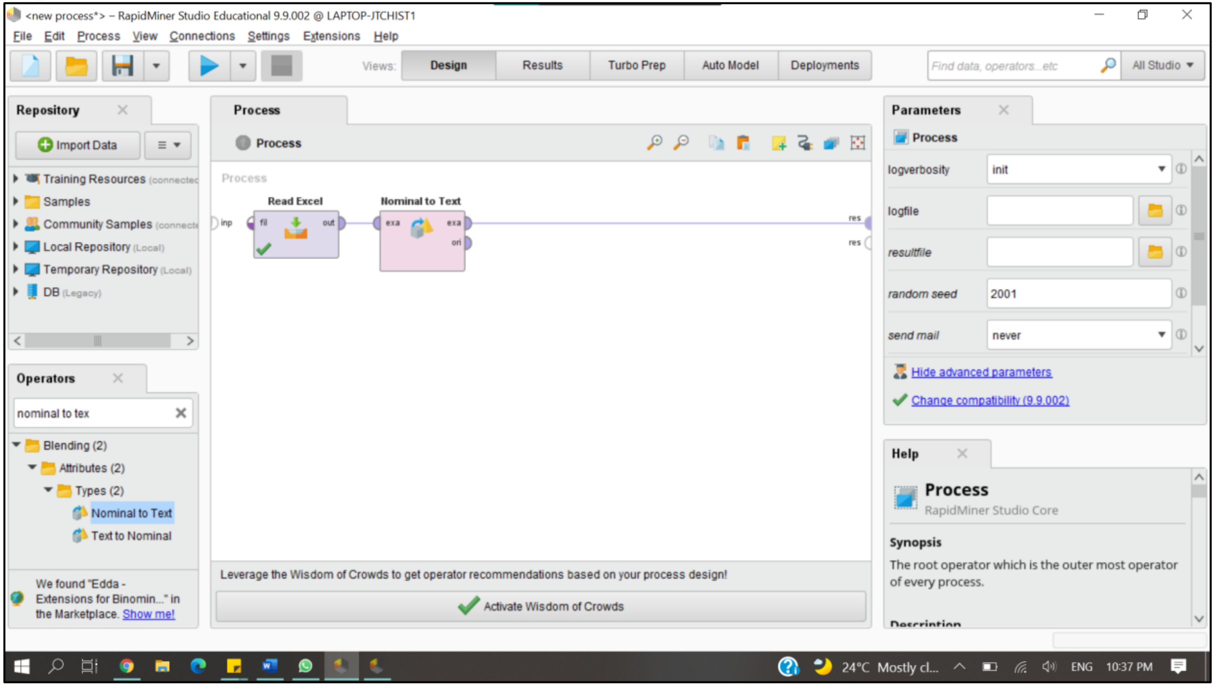
Row No.	disease	ABSTRACT
1	BRADYCARDIA	Surgical treatment of pediatric cardiac arrhythmia.
2	BRADYCARDIA	Comparative survival following permanent ventricular and dual-chamber p...
3	BRADYCARDIA	Complete sinoatrial block in two patients with bradycardia-tachycardia syn...
4	CORONARY DISEASE	Enhanced utilization of exogenous glucose improves cardiac function in hy...
5	CORONARY DISEASE	Myocardial amiodarone and desethylamiodarone concentrations in patient...
6	CORONARY DISEASE	Effects of benazepril and metoprolol OROS alone and in combination on ...
7	HEART ANEURYSM	Atrial septal aneurysms in infants and children.
8	HEART ANEURYSM	Submitral left ventricular aneurysms. Correction by a new transatrial appro...
9	MYOCARDIAL DISEASES	Altered norepinephrine turnover and metabolism in diabetic cardiomyopat...
10	MYOCARDIAL DISEASES	Pathophysiology and pathogenesis of stunned myocardium. Depressed C...

Figure 57 : The data that has been read by the tool

Step 2 : Nominal to Text

Next, I added the Nominal to Text operator. The Nominal to Text operator converts all nominal attributes to string attributes.

Input :


File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Temporary Repository (Local)
DB (Legacy)
Operators
Blending (2)
Attributes (2)
Types (2)
Nominal to Text
Text to Nominal
Process
Read Excel
Nominal to Text
inp
fil
out
exa
exa
ori
res
res
Activate Wisdom of Crowds
Parameters
Help
24°C Mostly cl...
ENG
10:37 PM

<new process*> – RapidMiner Studio Educational 9.9.002 @ LAPTOP-JTCHIST1

Operators search: nominal to tex

We found "Edda - Extensions for Binomin..." in the Marketplace. Show me!

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

Parameters – Process

logverbosity: init

logfile:

resultfile:

random seed: 2001

send mail: never

Hide advanced parameters

Change compatibility (9.9.002)

Help – Process

RapidMiner Studio Core

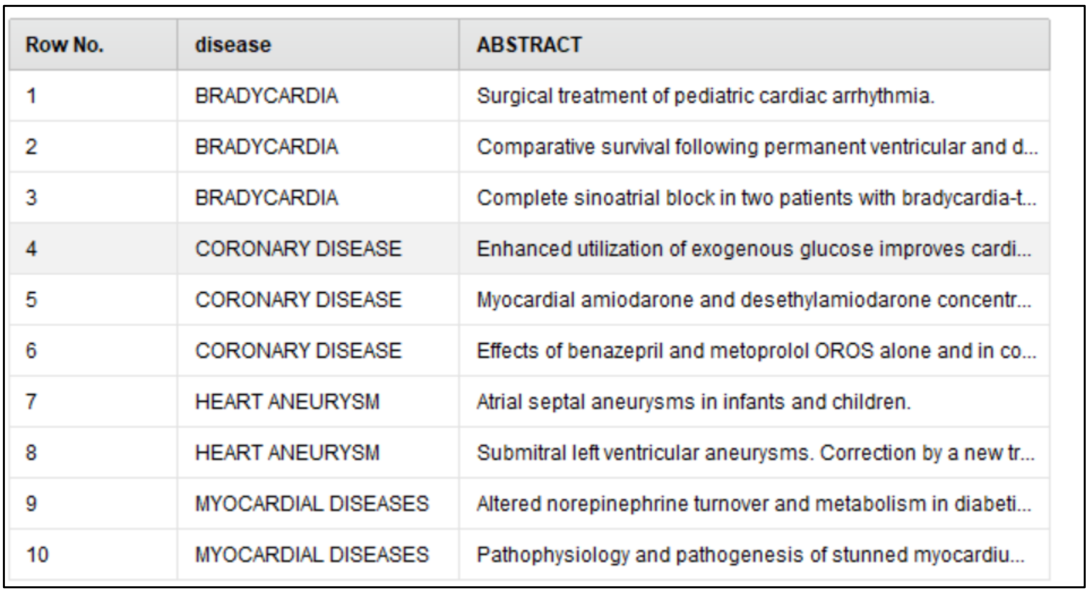
Synopsis

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

Description

Figure 58 : Nominal to Text operator

Output :



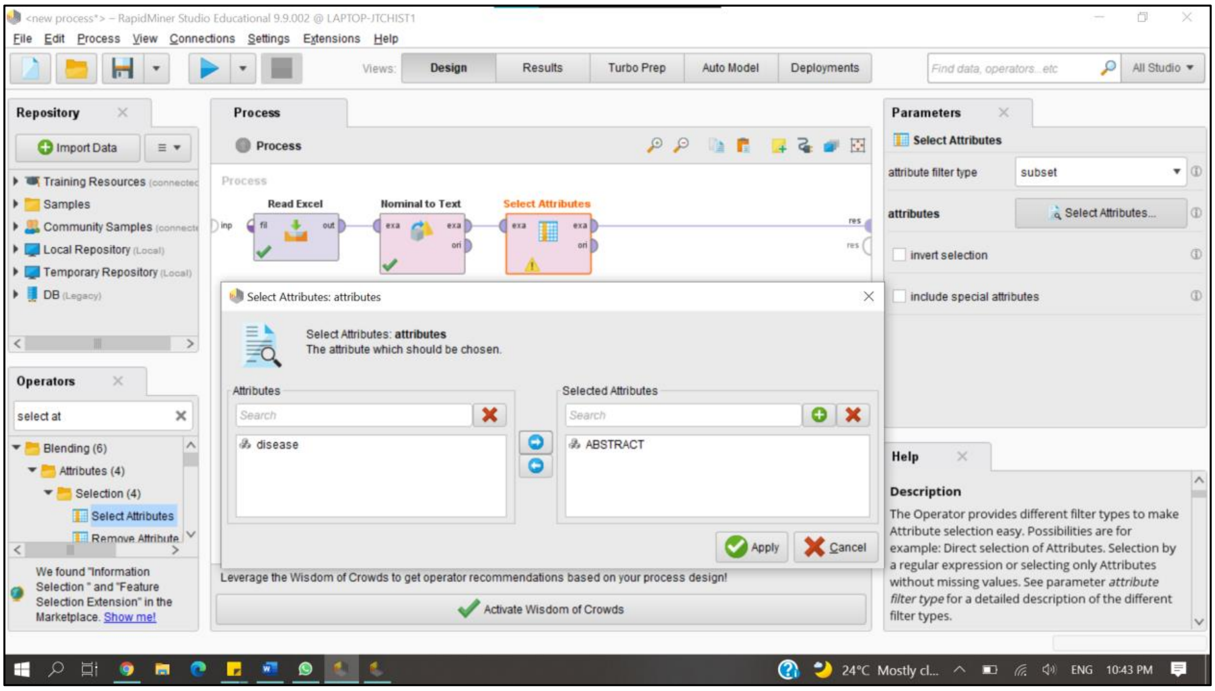
Row No.	disease	ABSTRACT
1	BRADYCARDIA	Surgical treatment of pediatric cardiac arrhythmia.
2	BRADYCARDIA	Comparative survival following permanent ventricular and d...
3	BRADYCARDIA	Complete sinoatrial block in two patients with bradycardia-t...
4	CORONARY DISEASE	Enhanced utilization of exogenous glucose improves cardi...
5	CORONARY DISEASE	Myocardial amiodarone and desethylamiodarone concentr...
6	CORONARY DISEASE	Effects of benazepril and metoprolol OROS alone and in co...
7	HEART ANEURYSM	Atrial septal aneurysms in infants and children.
8	HEART ANEURYSM	Submitral left ventricular aneurysms. Correction by a new tr...
9	MYOCARDIAL DISEASES	Altered norepinephrine turnover and metabolism in diabeti...
10	MYOCARDIAL DISEASES	Pathophysiology and pathogenesis of stunned myocardiu...

Figure 59 : The data has been changed from nominal to text

Step 3 : Select attribute

In this step, I needed to select the needed atrtribute only. Therefore I chose ABSTRACT as I am about to process the long text data.

Input :



<new process*> – RapidMiner Studio Educational 9.9.002 @ LAPTOP-JTCHIST1
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Temporary Repository (Local)
DB (Legacy)
Operators
select at
Blending (6)
Attributes (4)
Selection (4)
Select Attributes
Remove Attribute
Process
Read Excel
Nominal to Text
Select Attributes
Parameters
Select Attributes
Help

We found "Information Selection" and "Feature Selection Extension" in the Marketplace. Show me!

Select Attributes: attributes

Select Attributes: **attributes**
The attribute which should be chosen.

Attributes: disease

Selected Attributes: ABSTRACT

Apply Cancel

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

Activate Wisdom of Crowds

attribute filter type: subset

attributes: Select Attributes...

□ invert selection

□ include special attributes

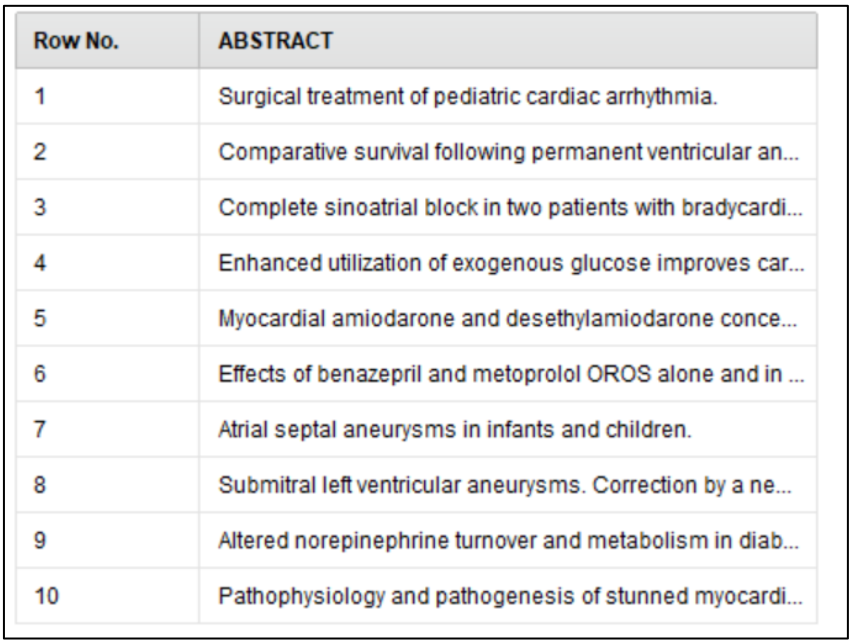
Description

The Operator provides different filter types to make Attribute selection easy. Possibilities are for example: Direct selection of Attributes. Selection by a regular expression or selecting only Attributes without missing values. See parameter *attribute filter type* for a detailed description of the different filter types.

24°C Mostly cl... ENG 10:43 PM

Figure 60 : Choosing the Abstract as attribute

Output :



Row No.	ABSTRACT
1	Surgical treatment of pediatric cardiac arrhythmia.
2	Comparative survival following permanent ventricular an...
3	Complete sinoatrial block in two patients with bradycardi...
4	Enhanced utilization of exogenous glucose improves car...
5	Myocardial amiodarone and desethylamiodarone conce...
6	Effects of benazepril and metoprolol OROS alone and in ...
7	Atrial septal aneurysms in infants and children.
8	Submitral left ventricular aneurysms. Correction by a ne...
9	Altered norepinephrine turnover and metabolism in diab...
10	Pathophysiology and pathogenesis of stunned myocardi...

Figure 61 : The latest data after choosing attribute abstract

Step 4 : Preprocessing

To do preprocessing, I needed to add the Process Documents from Data operator first. And then proceed the basic preprocessing in the Process Documents. Almost all process, preprocessing is the most basic and needed to be done first before we get deeper into other processes.

Input :

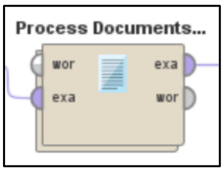


Figure 62 : Add Process Documents operator

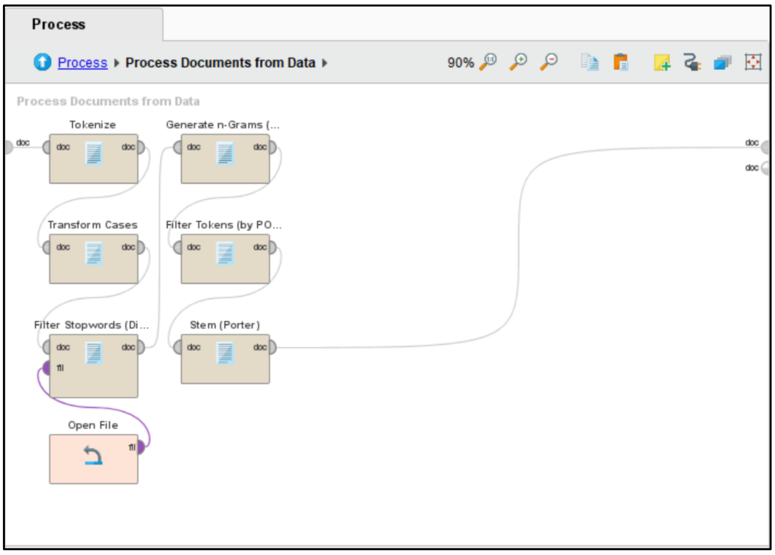
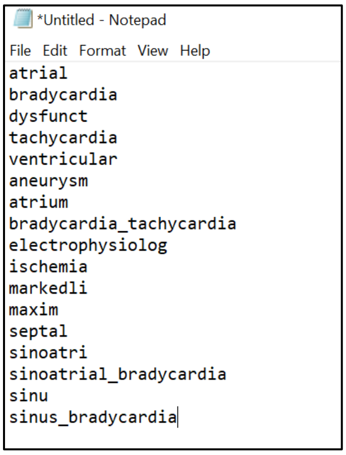


Figure 63 : All preprocessing basic step

Output :



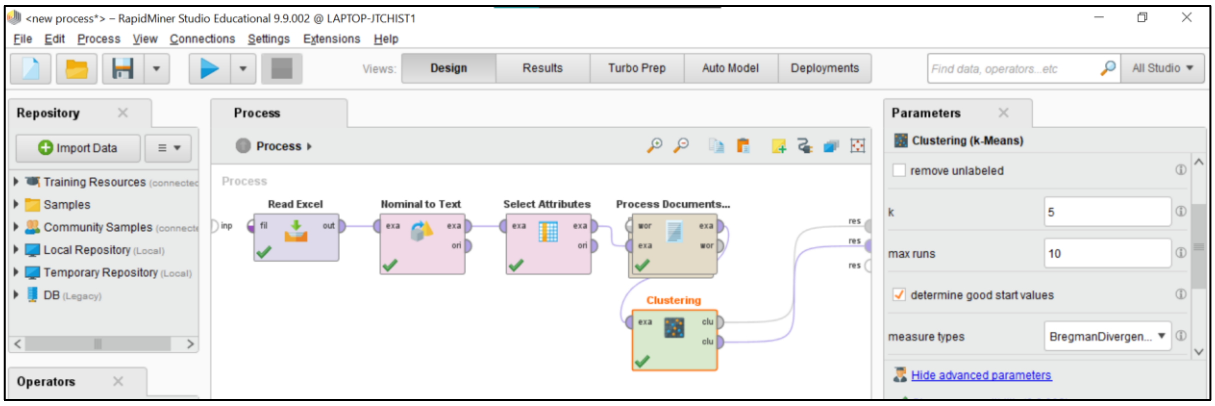
```
atrial
bradycardia
dysfunct
tachycardia
ventricular
aneurysm
atrium
bradycardia_tachycardia
electrophysiolog
ischemia
markedli
maxim
septal
sinoatri
sinoatrial_bradycardia
sinu
sinus_bradycardia
```

Figure 64 : The list of word generated after going through the preprocessing

Step 5 : K Means

To apply the clustering, I added the K-Means Clustering operator. This Operator performs clustering utilizing the k-means algorithm. Clustering assembles examples which are almost likely the same with one another.

Input :



<new process*> – RapidMiner Studio Educational 9.9.002 @ LAPTOP-JTCHIST1
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Temporary Repository (Local)
DB (Legacy)
Operators
Process
Read Excel
Nominal to Text
Select Attributes
Process Documents...
Clustering
Parameters
Clustering (k-Means)
Hide advanced parameters

□ remove unlabeled

k: 5

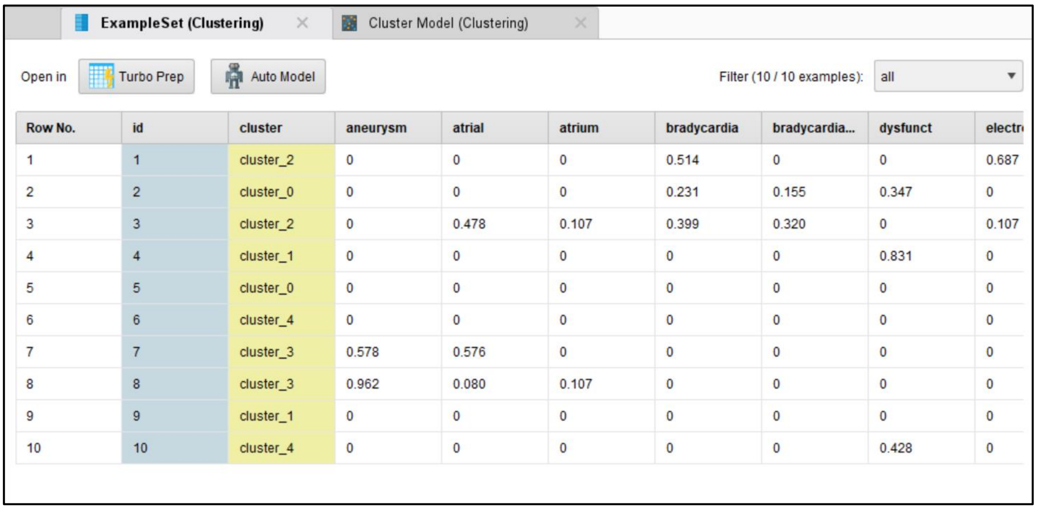
max runs: 10

☑ determine good start values

measure types: BregmanDivergen...

Figure 65 : Add the K-Means operator

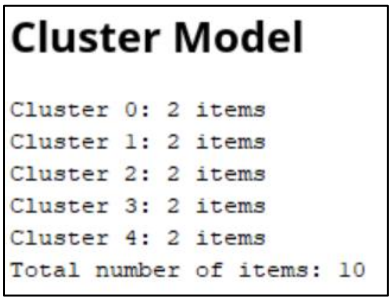
Output :



Filter (10 / 10 examples): all

Row No.	id	cluster	aneurysm	atrial	atrium	bradycardia	bradycardia...	dysfunct	electr
1	1	cluster_2	0	0	0	0.514	0	0	0.687
2	2	cluster_0	0	0	0	0.231	0.155	0.347	0
3	3	cluster_2	0	0.478	0.107	0.399	0.320	0	0.107
4	4	cluster_1	0	0	0	0	0	0.831	0
5	5	cluster_0	0	0	0	0	0	0	0
6	6	cluster_4	0	0	0	0	0	0	0
7	7	cluster_3	0.578	0.576	0	0	0	0	0
8	8	cluster_3	0.962	0.080	0.107	0	0	0	0
9	9	cluster_1	0	0	0	0	0	0	0
10	10	cluster_4	0	0	0	0	0	0.428	0

Figure 66 : Example Set



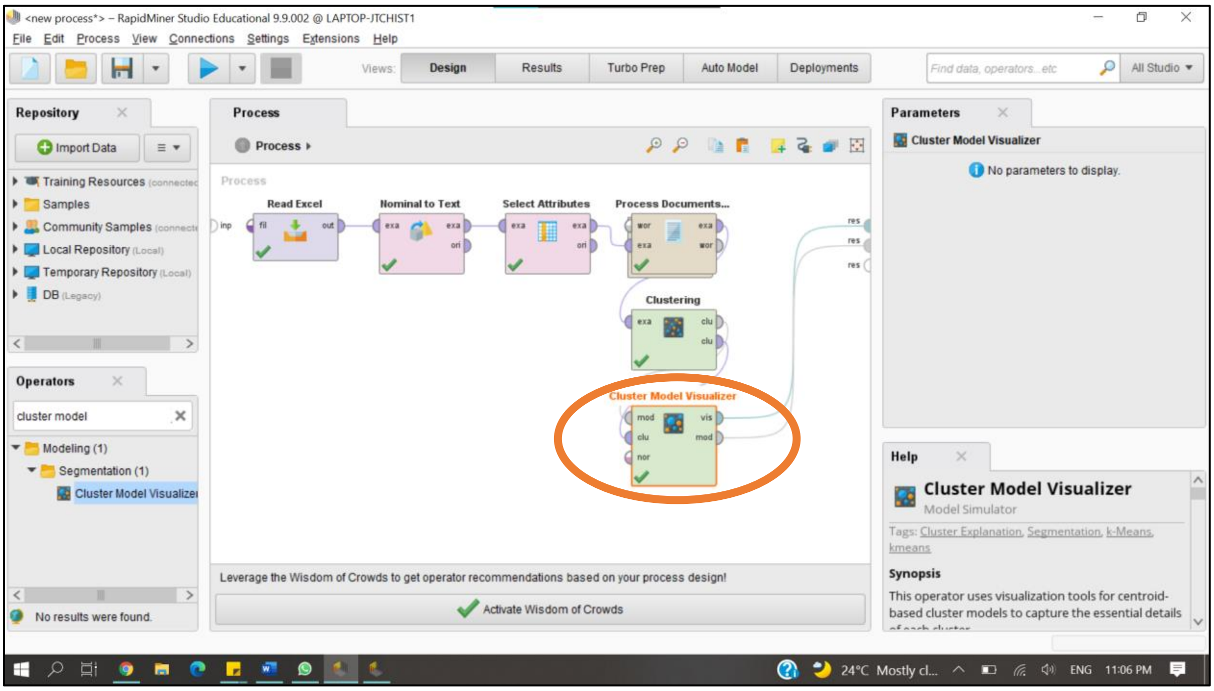
```
Cluster Model

Cluster 0: 2 items
Cluster 1: 2 items
Cluster 2: 2 items
Cluster 3: 2 items
Cluster 4: 2 items
Total number of items: 10
```

Figure 67 : Cluster Model

Step 6 : Cluster Model Visualizer

To visualize and understand the cluster model better, I added Cluster Model Visualizer. There will be no parameter needed.

Input :


<new process*> – RapidMiner Studio Educational 9.9.002 @ LAPTOP-JTCHIST1
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import Data
Training Resources (connected)
Samples
Community Samples (connected)
Local Repository (Local)
Temporary Repository (Local)
DB (Legacy)
Operators
cluster model
Modeling (1)
Segmentation (1)
Cluster Model Visualizer
Process
Read Excel
Nominal to Text
Select Attributes
Process Documents...
Clustering
Cluster Model Visualizer
Parameters
Cluster Model Visualizer
Help
Cluster Model Visualizer
Model Simulator
Tags: Cluster Explanation, Segmentation, k-Means, kmeans
Synopsis
Activate Wisdom of Crowds
24°C Mostly cl...
ENG
11:06 PM

No parameters to display.

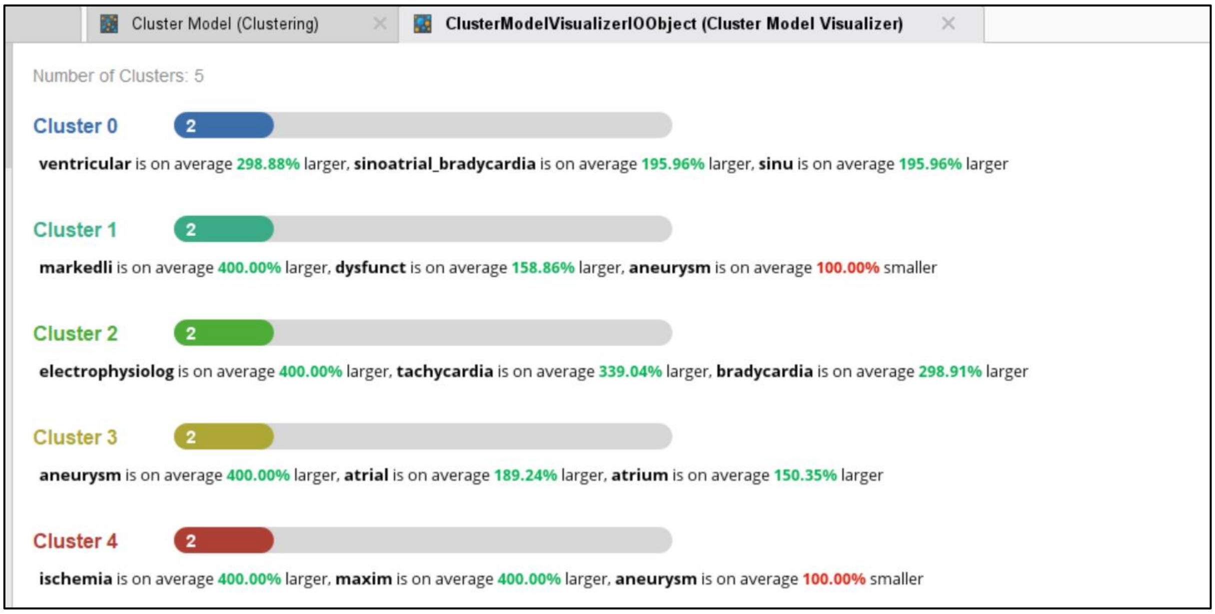
No results were found.

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

This operator uses visualization tools for centroid-based cluster models to capture the essential details of each cluster

Figure 68 : Add Cluster Model Visualizer

Output :


Cluster Model (Clustering)
ClusterModelVisualizerIOObject (Cluster Model Visualizer)

Number of Clusters: 5

Cluster 0 2

ventricular is on average 298.88% larger, **sinoatrial_bradycardia** is on average 195.96% larger, **sinu** is on average 195.96% larger

Cluster 1 2

markedli is on average 400.00% larger, **dysfunct** is on average 158.86% larger, **aneurysm** is on average 100.00% smaller

Cluster 2 2

electrophysiolog is on average 400.00% larger, **tachycardia** is on average 339.04% larger, **bradycardia** is on average 298.91% larger

Cluster 3 2

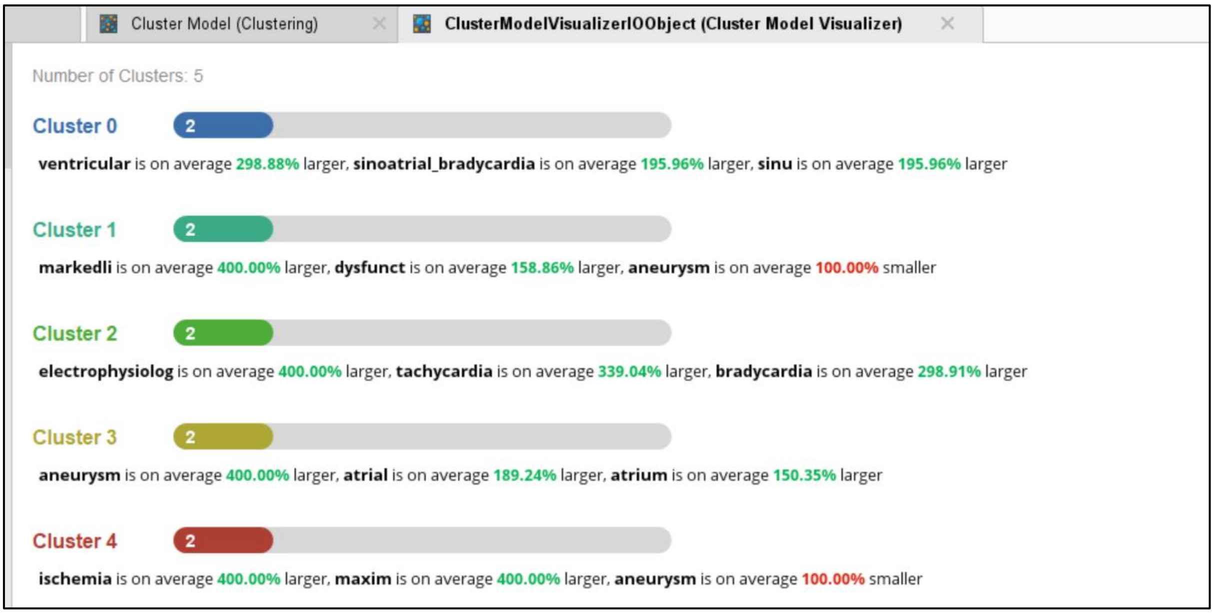
aneurysm is on average 400.00% larger, **atrial** is on average 189.24% larger, **atrium** is on average 150.35% larger

Cluster 4 2

ischemia is on average 400.00% larger, **maxim** is on average 400.00% larger, **aneurysm** is on average 100.00% smaller

Figure 69 : Output Cluster Model Visualizer

f) Performance Measures


Cluster Model (Clustering)
ClusterModelVisualizerIOObject (Cluster Model Visualizer)

Number of Clusters: 5

Cluster 0 — 2

ventricular is on average 298.88% larger, **sinoatrial_bradycardia** is on average 195.96% larger, **sinu** is on average 195.96% larger

Cluster 1 — 2

markedli is on average 400.00% larger, **dysfunct** is on average 158.86% larger, **aneurysm** is on average 100.00% smaller

Cluster 2 — 2

electrophysiolog is on average 400.00% larger, **tachycardia** is on average 339.04% larger, **bradycardia** is on average 298.91% larger

Cluster 3 — 2

aneurysm is on average 400.00% larger, **atrial** is on average 189.24% larger, **atrium** is on average 150.35% larger

Cluster 4 — 2

ischemia is on average 400.00% larger, **maxim** is on average 400.00% larger, **aneurysm** is on average 100.00% smaller

Figure 70 : Output Cluster Model Visualizer

Based on the figure above, it shows the final result for the Clustering process and it is visualized using the Cluster Model Operator. The total number of clusters is 5 (cluster 0 – cluster 4) since I set the value of K = 5 during the process of K-Means. To understand this better, from the figure, in cluster 3, aneurysm, atrial and atrium are been clustered together since they own almost the same characteristic. But we can see that the word aneurysm is 400% larger than atrial (189.24%) and atrium (150.35%). This simply means that aneurysm is the most popular word in the cluster 3.

2.0 PART 2

(a) Throughout completing this alternative assessment, I found that I learnt a lot of new things specifically in Data Mining. Before doing this alternative assessment, I thought it's very hard for me to understand any topic in this course. Since this alternative assessment is important as part of the courses, therefore I knew I will need to do some research any information regarding this project. Luckily, before starting the alternative assessment, me and my group members for project has started doing some project works especially in the text preprocessing step. Therefore, it wasn't awkward for me to start doing this alternative assessment by myself. I could say that I really enjoyed the experience of doing data mining task. I enjoyed doing the research on how to do data mining. The thing that I found frustrating about is that every knowledge that I applied during completing the alternative assessment, I could say that almost everything is wrong and misleading. That's when I realized that my efforts were kind of been wasted just like that. But nevertheless, it brings me to a new experience whereby I learnt from my mistakes.

(b) For me, the most difficult part in this alternative assessment is the part 1 question 2 where we needed to do data mining task. I found it's hard because I may can do the data mining task, but I don't really understand the result of it. That's when I knew I needed to do a lot more research than before. I gathered a lot more information about each supervised and unsupervised learning algorithm. The thing that I remembered to be effective is I started this alternative assessment earlier. As soon as Dr. Rozilawati gave it to us, I opened it and read everything through so that I could measure my strength of knowledge based on the difficulty of the questions. By the time I received this alternative assessment also, I knew that I could easily done the part 1 question 1. But the question 2, I seemed a bit taken aback since by that time my knowledge about doing data mining task is a bit low. The most ineffective I ever did during the completion of the alternative assessment is I didn't allocate the time wisely to complete the alternative assessment. After done with question 1 for part 1, I left the question 2 unanswered for a while because I thought that I needed to do more research about it. So, I did other courses' assignments first. Few days later, then I just realized that I haven't start anything for question 2, so I was panic and tried to do everything and completed the question 2, two days before the submission.

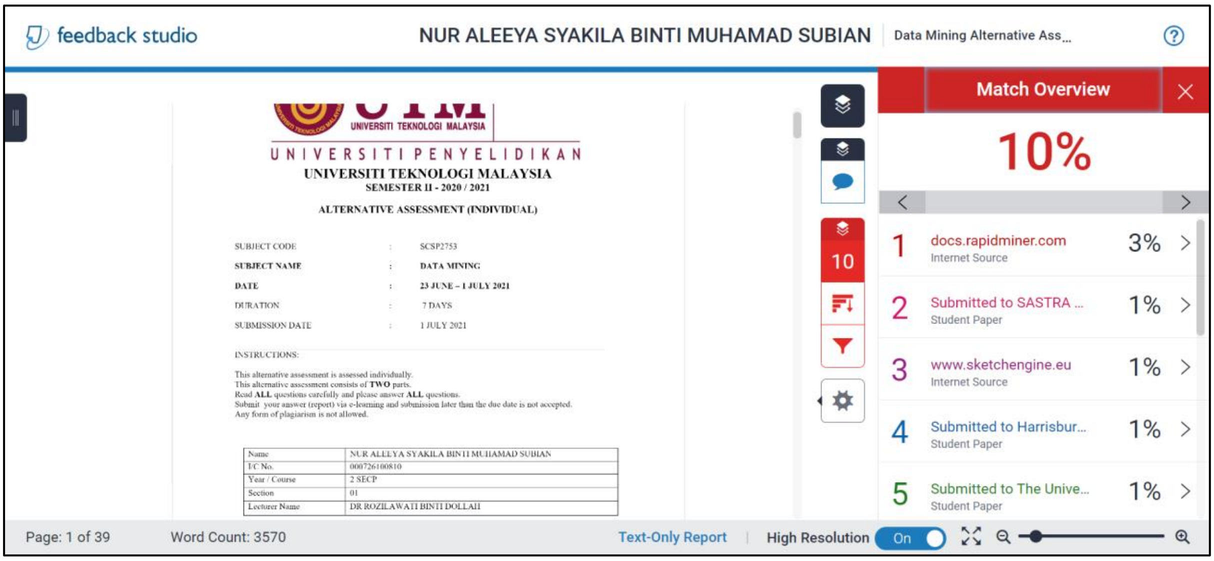
(c) There are few things that I think I needed to improve in this course. Firstly, in order for me to improve myself in this course, I should explore more other data mining tools. In the short time given, I chose RapidMiner for my tool for this alternative assessment because it has many resources for me to at least learn from. But I think after this I might want to explore tools such as Weka, Power BI etc. Secondly, I should expose myself with handling different type of dataset with a big amount of data. Since in the world, we have a lots of data type. Therefore, by having the experience to handle different types of data, it can enhance my knowledge about data mining, and I can also get the benefits from it. In my opinion, this Data Mining course can be applied in my future job. Since I'm major in Data Engineer, my scope of job definitely will be in this field too, which needed to handle big data. This course really helped to expose me to the real-world job especially in data.

(d) In a nutshell, in order for me to meet the learning needs, I should do a lot more findings about data mining task and engage myself handling big data more. To be success specifically in this course, I should try and error all data mining algorithm. Learn from mistakes and change to be better. Throughout this course, I found that I had developed new skills which are preprocess data from raw and process them through data mining task. Before entering this course, I don't even know how to clean data and I thought that I can't do it. After finishing this course, I realized that cleaning data wasn't really that hard if I practiced a lot. After all the practice and learning from our lecturer, Dr. Rozilawati, my skills increased

Reference

1. *(2) What is the difference between clustering and association rule mining? - Quora*. (2018). Quora.com. https://www.quora.com/What-is-the-difference-between-clustering-and-association-rule-mining
2. Minewiskan. (2018, May 8). *Training and Testing Data Sets*. Microsoft.com. https://docs.microsoft.com/en-us/analysis-services/data-mining/training-and-testing-data-sets?view=asallproducts-allversions
3. https://www.facebook.com/MachineLearningMastery. (2018, May 22). *A Gentle Introduction to k-fold Cross-Validation*. Machine Learning Mastery. https://machinelearningmastery.com/k-fold-cross-validation/
4. *2.1 What is the difference between labelled and unlabelled data? · Grokking Machine Learning MEAP V14*. (2021). Manning.com. https://livebook.manning.com/book/grokking-machine-learning/2-1-what-is-the-difference-between-labelled-and-unlabelled-data-/v-4/40

APPENDIX



NUR ALEEYA SYAKILA BINTI MUHAMAD SUBIAN | Data Mining Alternative Ass...

UNIVERSITI TEKNOLOGI MALAYSIA

UNIVERSITI PENYELIDIKAN

UNIVERSITI TEKNOLOGI MALAYSIA

SEMESTER II - 2020 / 2021

ALTERNATIVE ASSESSMENT (INDIVIDUAL)

SUBJECT CODE	:	SCSP2753
SUBJECT NAME	**:**	**DATA MINING**
DATE	**:**	**23 JUNE – 1 JULY 2021**
DURATION	:	7 DAYS
SUBMISSION DATE	:	1 JULY 2021

INSTRUCTIONS:

This alternative assessment is assessed individually.
This alternative assessment consists of **TWO** parts.
Read **ALL** questions carefully and please answer **ALL** questions.
Submit your answer (report) via e-learning and submission later than the due date is not accepted.
Any form of plagiarism is not allowed.

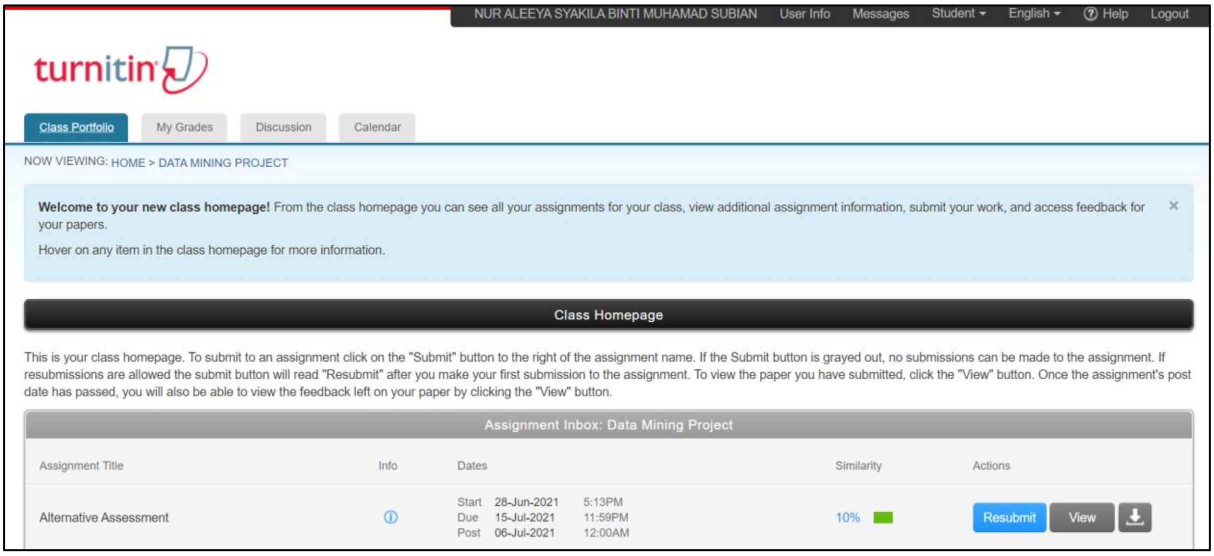
Name	NUR ALEEYA SYAKILA BINTI MUHAMAD SUBIAN
I/C No.	000726100810
Year / Course	2 SECP
Section	01
Lecturer Name	DR ROZILAWATI BINTI DOLLAH

Match Overview

10%

1	docs.rapidminer.com Internet Source	3%
2	Submitted to SASTRA ... Student Paper	1%
3	www.sketchengine.eu Internet Source	1%
4	Submitted to Harrisbur... Student Paper	1%
5	Submitted to The Unive... Student Paper	1%

Page: 1 of 39   Word Count: 3570



NOW VIEWING: HOME > DATA MINING PROJECT

Welcome to your new class homepage! From the class homepage you can see all your assignments for your class, view additional assignment information, submit your work, and access feedback for your papers.

Hover on any item in the class homepage for more information.

Class Homepage

This is your class homepage. To submit to an assignment click on the "Submit" button to the right of the assignment name. If the Submit button is grayed out, no submissions can be made to the assignment. If resubmissions are allowed the submit button will read "Resubmit" after you make your first submission to the assignment. To view the paper you have submitted, click the "View" button. Once the assignment's post date has passed, you will also be able to view the feedback left on your paper by clicking the "View" button.

Assignment Inbox: Data Mining Project

Assignment Title	Info	Dates			Similarity	Actions
Alternative Assessment	ⓘ	Start Due Post	28-Jun-2021 15-Jul-2021 06-Jul-2021	5:13PM 11:59PM 12:00AM	10% 🟩	Resubmit View