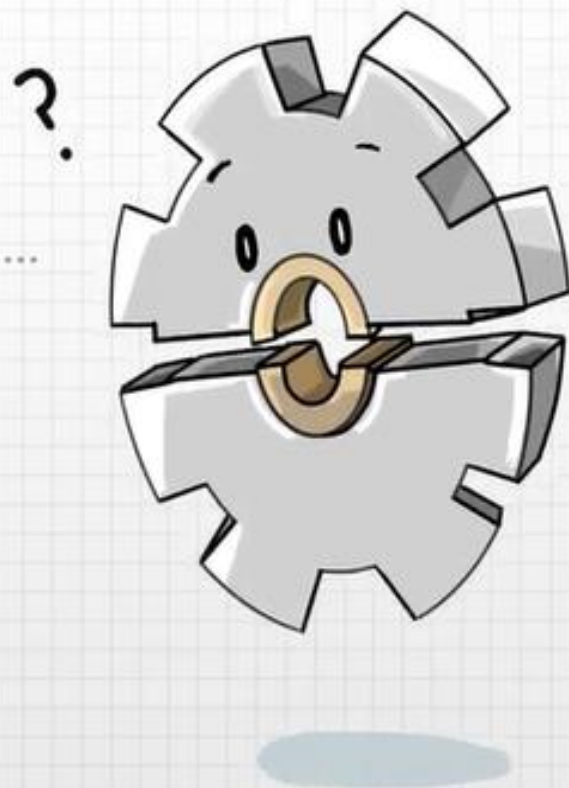


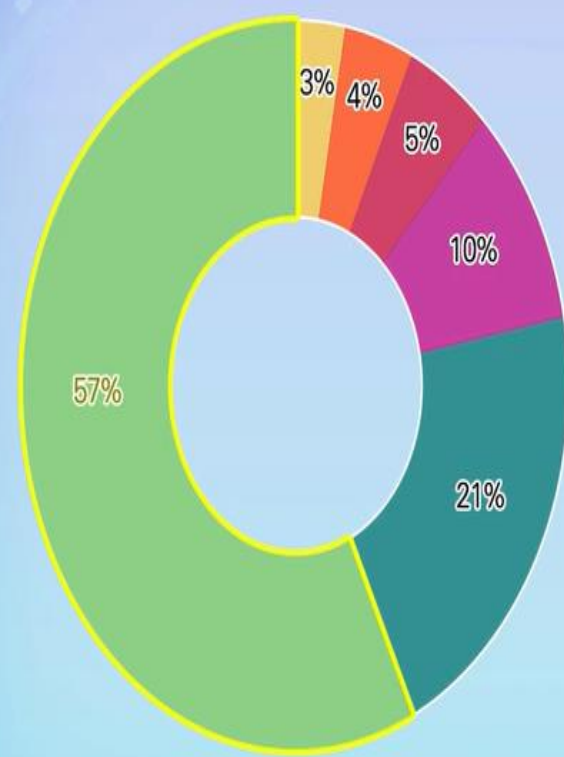


- ◆ 데이터 전처리과정 알기
- ◆ 데이터 전처리하기(실습)



- ◆ 데이터 전처리과정의 필요성에 대해 설명할 수 있다.
- ◆ 데이터 전처리 작업을 수행할 수 있다.





What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

## ◆ 데이터 나누기

학습 데이터 세트(Training Dataset)

- ◆ 머신러닝 인풋

테스트 데이터 세트(Test Dataset)

- ◆ 모델의 성과 측정

## 데이터 나누기

### 선형 표본화(linear sampling)

- ◆ 데이터를 순차적으로 선별

### 무작위 표본화(random sampling)

- ◆ 무작위로 랜덤하게 분류

### 층화 표본화(stratified sampling)

- ◆ 각각의 레이블별로 무작위 추출

## ❖ 값이 없는 경우: 결측치(Missing value)

- ◆ 속성 사용 여부
- ◆ 해당 사례만 제거
- ◆ 결측치 예측



## 2. 데이터 전처리하기(실습)

The screenshot displays the RapidMiner Studio interface. The main window shows the 'Process' tab with a 'Read Excel' operator selected. The 'Parameters' panel on the right is configured for 'Import Configuration Wizard...'. The 'Read Excel' operator is set to 'Import Configuration Wizard...'. The 'Sheet' is set to 'customer data', 'Cell range' is 'A:F', and 'Define header row' is checked with '1' selected. The 'Imported cell range' is 'A1'. The 'Encoding' is 'SYSTEM', 'First row as names' is checked, 'File format' is 'Enter value...', 'File zone' is 'SYSTEM', and 'File' is 'English (United States)'. The 'Description' panel at the bottom right explains that the operator reads data from Microsoft Excel spreadsheets.

**Select the cells to import.**

Sheet: customer data Cell range: A:F Select All Define header row: 1

	A	B	C	D	E	F
1	Name	Gender	Age	Payment Method	Churn	LastTransaction
2	Nicolas Garrett	male	64.000	credit card	loyal	98.000
3	Isaac Reyes	male	35.000	cheque	churn	118.000
4	Jaime Sullivan	female	25.000	credit card	loyal	107.000
5	Geraldine Miller	female	39.000	credit card	loyal	177.000
6	Curtis Frazer	m	39.000	credit card	loyal	90.000
7	Jeannie Palmer	female	28.000	cheque	churn	189.000
8	Phyllis Romero	female	21.000	credit card	loyal	102.000
9	Maxine Edwards	female		cheque	loyal	111.000
10	Marty Cohen		32.000	cheque	churn	50.000
11	Lionel Mendoza	male	48.000	credit card	loyal	141.000
12	Maureen Norman	female	70.000	credit card	churn	153.000
13	Santiago Cruz	male	36.000	credit card	loyal	46.000
14	Santiago Cruz	male	36.000	credit card	loyal	46.000
15	Nelson Davis	male	22.000	credit card	loyal	51.000
16	Josephine Owens	female	53.000	cash		183.000
17	Clarence Vaughn	male	27.000	cash	loyal	137.000
18	Jon Griffin	male	22.000	cash	loyal	147.000
19	Nettie Neal	female	49.000	credit card	churn	158.000
20	Belinda Reeves	female	24.000	cash	churn	162.000
21	Taylor Murphy	male	45.000	credit card	loyal	55.000

**Read Excel**  
RapidMiner Studio Core  
Load, Imports, Read, Data, Files, Xls, Yxls, Microsoft, Spreadsheets, Imports  
This operator reads an ExampleSet from the specified Excel file.  
[Go to Tutorial Process](#)

**Description**  
This operator can be used to load data from Microsoft Excel spreadsheets. This operator is able to read data from Excel 95, 97, 2000.

필요한 데이터 가져오기

## 2. 데이터 전처리하기(실습)

The screenshot displays the RapidMiner Studio interface with the 'Read Excel' operator selected in the process canvas. The interface is divided into several panels:

- Repository:** Shows a tree view with 'Samples', 'DB (Legacy)', and 'Local Repository (xls)'.
- Process:** The central canvas showing the 'Read Excel' operator connected to the input and output ports.
- Parameters:** A panel on the right showing the configuration for the 'Read Excel' operator. The parameters are:
  - excel file: 2-2 customer\_churn\_data.xlsx
  - sheet selection: sheet number
  - sheet number: 1
  - imported cell range: A1
  - encoding: SYSTEM
  - first row as names: ☒
  - date format: Enter value...
  - time zone: SYSTEM
  - locale: English (United States)
- Operators:** A panel on the bottom left showing a list of operators under the 'read' category. The 'Read Excel' operator is highlighted.
- Help:** A panel on the bottom right showing the documentation for the 'Read Excel' operator, including its synopsis and description.

read excel의 아웃풋과 패널의 result port 연결

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments

Find data, operators, etc. All Studio

Result History ExampleSet (Read Excel) X

Open in Turbo Prep Auto Model

Filter (999 / 999 examples) all

Row No.	Name	Gender	Age	Payment Me...	Churn	LastTransa...
1	Nicolas Garrett	male	64	credit card	loyal	98
2	Isaac Reyes	male	35	cheque	churn	118
3	Jaime Sullivan	female	25	credit card	loyal	107
4	Geraldine Mill...	female	39	credit card	?	177
5	Curtis Fraser	m	39	credit card	loyal	90
6	Jeanne Pal...	female	28	cheque	churn	189
7	Phyllis Romero	female	21	credit card	loyal	102
8	Maxine Edwa...	female	?	cheque	loyal	111
9	Marty Cohen	?	32	cheque	churn	50
10	Lionel Mendo...	male	48	credit card	loyal	141
11	Maureen Nor...	female	70	credit card	churn	153
12	Santiago Cruz	male	36	credit card	loyal	46
13	Santiago Cruz	male	36	credit card	loyal	46
14	Nelson Davis	male	22	credit card	loyal	51
15	Josephine O...	female	53	cash	?	183
16	Clarence Vau...	male	27	cash	loyal	137
17	Jon Griffin	male	22	cash	loyal	147
18	Nettie Neal	female	49	credit card	churn	158
19	Belinda Reev...	female	24	cash	churn	162
20	Taylor Murphy	male	45	credit card	loyal	55
21	Emmett James	male	45	credit card	loyal	160
22	Paula Murray	female	66	cash	churn	156
23	Penny Reese	female	82	cash	churn	177
24	Janis Heman...	female	35	credit card	loyal	176
25	Dianne Wolfe	female	17	credit card	loyal	133

Repository X

Import Data

Samples

DB (Legacy)

Local Repository (DMS)

실행 후 데이터 확인



The screenshot displays the RapidMiner Studio interface. The main workspace shows a process flow with two operators: 'Read Excel' and 'Trim'. The 'Trim' operator is highlighted with a mouse cursor, indicating it is being added to the process. The 'Parameters' panel on the right shows the 'Trim' operator's settings, including 'attribute filter type' set to 'all'. The 'Operators' panel on the left shows the 'Trim' operator under the 'Values' category. The 'Help' panel on the right provides information about the 'Trim' operator, including its synopsis and description.

**Repository**

- Import Data
- Samples
- DB (Legacy)
- Local Repository (prior)

**Process**

Process

Read Excel

Trim

**Parameters**

Trim

attribute filter type: all

☐ invert selection

☐ include special attributes

Change compatibility (3.5.001)

**Help**

Trim

RapidMiner Studio Core

Tag: Values

**Synopsis**

This operator removes leading and trailing spaces from the values of the selected nominal attributes.

[Jump to Tutorial Process](#)

**Description**

The Trim operator creates new attributes from the selected nominal attributes by removing leading and trailing spaces from the nominal

trim을 가져온 후 read excel 옆에 드래그 및 연결

## 2. 데이터 전처리하기(실습)

The screenshot displays the RapidMiner Studio interface. The main workspace shows a workflow with two operators: 'Read Excel' and 'Trim'. The 'Operators' panel on the left is open, showing the 'Filter' category. The 'Parameters' panel on the right shows the 'Trim' operator settings, including 'attribute filter type' set to 'all'. The 'Help' panel at the bottom right provides a synopsis and description of the 'Trim' operator.

**Operators Panel:**

- Filter
  - Select Attributes
  - Remove Useless Attributes
  - Remove Correlated Attributes

**Parameters Panel:**

- Trim
  - attribute filter type: all
  - invert selection: ☐
  - include special attributes: ☐

**Help Panel:**

**Trim**  
RapidMiner Studio Core

Tags: Values

**Synopsis**  
This operator removes leading and trailing spaces from the values of the selected nominal attributes.  
[Jump to Tutorial Process](#)

**Description**  
The Trim operator creates new attributes from the selected nominal attributes by removing leading and trailing spaces from the nominal

filter example 오퍼레이터를 사용하여 결측치 처리하기

## 2 데이터 전처리하기

### 2. 데이터 전처리하기(실습)

CAU

The screenshot displays the RapidMiner Studio interface. The main window shows a process diagram with a 'Read Excel' operator connected to a 'Filter Examples' operator. A dialog box titled 'Create Filters: filters' is open, allowing the user to define filters. The dialog has a search bar at the top and a list of filters below. The first filter is 'Gender' with the condition 'is not missing'. The 'Match all' radio button is selected, and the 'Preselect comparators' checkbox is checked. The 'Add Entry' button is highlighted. The right sidebar shows the 'Parameters' panel for the 'Filter Examples' operator, with the 'condition class' set to 'custom\_filters'. The bottom right corner shows the 'Help' panel for the 'Filter Examples' operator, which includes a synopsis and description.

add filter → is not missing 선택

## 2 데이터 전처리하기

### 2. 데이터 전처리하기(실습)

CAU

The screenshot displays the RapidMiner Studio interface with a workflow in the 'Process' view. The workflow consists of three operators: 'Read Excel', 'Trim', and 'Filter Examples'. The 'Filter Examples' operator is currently selected, and its parameters are visible in the 'Parameters' panel on the right. The 'Parameters' panel shows 'Filter Examples' with a 'condition class' set to 'custom\_filters' and an 'invert filter' checkbox. The 'Operators' panel on the left shows a search for 'remove', with 'Filter Examples' and 'Filter Example Range' listed. The 'Help' panel at the bottom right provides details for the 'Filter Examples' operator, including its tags, synopsis, and description. The central workspace has a 'Drag here' label.

Repository

Process

Parameters

Operators

Help

Drag here

remove duplicate 오퍼레이터 사용하여 중복 데이터 없애기



## 2 데이터 전처리하기

### 2. 데이터 전처리하기(실습)

CAU

The screenshot displays the RapidMiner Studio interface with a workflow in the 'Process' view. The workflow consists of four operators: 'Read Excel', 'Trim', 'Filter Examples', and 'Remove Duplicates'. The 'Remove Duplicates' operator is currently selected, and its parameters are shown in the 'Parameters' panel on the right. The 'attribute filter type' is set to 'single', and the 'attribute' is set to 'Name'. The 'Operators' panel on the left shows a search for 'remove' with 'Remove Duplicates' highlighted under the 'Duplications' category. The 'Help' panel at the bottom right provides a synopsis of the 'Remove Duplicates' operator, stating that it removes duplicate examples from an ExampleSet by comparing all examples with each other on the basis of the specified attributes. Two examples are considered duplicate if the selected attributes have the same values in them. A link to the tutorial process is also provided.

**Parameters**

- Remove Duplicates
- attribute filter type: single
- attribute: Name
- ☐ invert selection
- ☐ include special attributes
- ☐ treat missing values as duplicates

**Help**

**Remove Duplicates**  
RapidMiner Studio Core

Tags: Deduplication, Matches, Matching, Replicates, Copies, Cleansing, Quality, Distinct, Equal Filter, Duplicates

**Synopsis**

This operator removes duplicate examples from an ExampleSet by comparing all examples with each other on the basis of the specified attributes. Two examples are considered duplicate if the selected attributes have the same values in them.

[Jump to Tutorial Process](#)

파라미터 설정으로 single 선택 후 중복되는 데이터 없애기

## 2 데이터 전처리하기

### 2. 데이터 전처리하기(실습)

CAU

The screenshot displays the RapidMiner Studio interface with a workflow in the 'Process' view. The workflow consists of the following operators: 'Read Excel', 'Trim', 'Filter Examples', 'Remove Duplicates', and 'Replace'. The 'Replace' operator is currently selected, and its parameters are shown in the 'Parameters' panel on the right. The parameters are: 'attribute filter type' set to 'single', 'attribute' set to 'Gender', 'invert selection' unchecked, 'include special attributes' unchecked, 'replace what' set to 'm', and 'replace by' set to 'male'. The 'Operators' panel on the left shows a search for 'replace' with results under 'Blending (3)', 'Cleansing (4)', and 'Modeling (1)'. The 'Help' panel at the bottom right provides details for the 'Replace' operator, including its synopsis and description.

**Parameters**

- attribute filter type: single
- attribute: Gender
- invert selection: ☐
- include special attributes: ☐
- replace what: m
- replace by: male

[Change compatibility \(9.5.001\)](#)

**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 replacement.  
[Jump to Tutorial Process](#)

**Description**  
This operator allows you to select attributes to make replacements in and to specify a regular expression. Attribute values of selected attributes that

m을 male의 오타로 수정

## 2 데이터 전처리하기

### 2. 데이터 전처리하기(실습)

CAU

The screenshot displays the RapidMiner Studio interface. The main workspace shows a process flow: **Read Excel** → **Trim** → **Filter Examples** → **Remove Duplicates** → **Replace**. The **Replace** operator is selected, and its configuration is shown in the **Parameters** panel on the right. The **attribute filter type** is set to **single**, and the **attribute** is **Gender**. The **replace what** is **m** and the **replace by** is **male**. The **Operators** panel on the left shows the **Replace** operator under the **Values** category. A dialog box titled **Edit Regular Expression** is open, showing the regular expression **/b(m)?** and the replacement **male**. The dialog also includes an **Inline Text Search** tab and a **Result preview** section.

**Replace**  
attribute filter type: single  
attribute: Gender  
invert selection: ☐  
include special attributes: ☐  
replace what: m  
replace by: male  
Change compatibility (9.5.001)

**Edit Regular Expression**  
Regular Expression: /b(m)?  
Regular expression valid.  
Replacement (value for 'replace by'): male  
Inline Text Search | Result List (0) | Regexp Options  
Text:  
Result preview:  
Apply | Cancel

**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 replacement.  
Jump to Tutorial Process  
Description  
This operator allows you to select attributes to make replacements in and to specify a regular expression. Attribute values of selected attributes that

단독으로 m이 있는 것만 바꾸기 위해 /b 사용

## 2 데이터 전처리하기

### 2. 데이터 전처리하기(실습)

CAU

The screenshot displays the RapidMiner Studio interface with a workflow designed for data preprocessing. The workflow consists of the following operators in sequence: Read Excel, Trim, Filter Examples, Remove Duplicates, Replace, and Filter Examples (2). The 'Filter Examples (2)' operator is currently selected, and its parameters are visible in the right-hand pane. The parameters for 'Filter Examples (2)' are: filters (Add Filters...), condition class (custom\_filters), and invert filter (unchecked). Below the parameters, there are links for 'Hide advanced parameters' and 'Change compatibility (9.5.001)'. The bottom right pane shows the 'Help' section for the 'Filter Examples' operator, which includes a synopsis and a description. The synopsis states: 'This Operator selects which Examples of an ExampleSet are kept and which Examples are removed.' The description states: 'The Operator returns those Examples that match the given condition. The'.

Repository

Process

Parameters

Filter Examples (2) (Filter Examples)

filters

condition class

invert filter

Help

Filter Examples

RapidMiner Studio Core

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

Synopsis

This Operator selects which Examples of an ExampleSet are kept and which Examples are removed.

[Jump to Tutorial Process](#)

Description

The Operator returns those Examples that match the given condition. The

Churn 데이터가 있는 것만 가져오기



## 2 데이터 전처리하기

### 2. 데이터 전처리하기(실습)

CAU

The screenshot displays the RapidMiner Studio interface with a workflow in the 'Process' view. The workflow consists of the following operators: Read Excel, Trim, Filter Examples, Remove Duplicates, Replace, and Filter Examples (2). A 'Drop here' label is positioned below the workflow. The left sidebar contains the 'Repository' and 'Operators' panels. The 'Operators' panel is expanded to show the 'Filter' category, with 'Select Attributes' highlighted. The right sidebar shows the 'Parameters' panel for the selected 'Filter Examples (2)' operator, with settings for 'filters', 'condition class', and 'invert filter'. Below the parameters is a 'Help' panel for the 'Select Attributes' operator, which includes a synopsis and description.

Repository

Process

Drop here

Parameters

Filter Examples (2) (Filter Examples)

filters

condition class

invert filter

Help

Select Attributes

RapidMiner Studio Core

Tags: Filter, Keep, Remove, Drop, Delete, Columns, Variables, Features, Feature Set, Selection

Synopsis

This Operator selects a subset of Attributes of an ExampleSet and removes the other Attributes.

Jump to Tutorial Process

Description

The Operator provides different filter types to make Attribute selection

select attributes 오퍼레이터 사용

## 2 데이터 전처리하기

### 2. 데이터 전처리하기(실습)

CAU

The screenshot displays the RapidMiner Studio interface with a workflow designed for data preprocessing. The workflow consists of the following operators in sequence: Read Excel, Trim, Filter Examples, Remove Duplicates, Replace, and Filter Examples (2). A 'Select Attributes' operator is positioned below the main workflow, with a line connecting it to the 'Filter Examples (2)' operator, indicating that attribute selection is applied after the final filtering step. The 'Parameters' panel on the right is configured for the 'Select Attributes' operator, showing 'attribute filter type' set to 'single' and 'attribute' set to 'Churn'. The 'Operators' panel on the left shows the 'Select Attributes' operator selected under the 'Blending (9)' > 'Attributes (7)' > 'Selection (7)' category. The 'Help' panel at the bottom right provides details for the 'Select Attributes' operator, including its synopsis and description.

**Repository**

- Import Data
- Samples
- DB (Legacy)
- Local Repository (drivers)

**Process**

Process

100%

Read Excel, Trim, Filter Examples, Remove Duplicates, Replace, Filter Examples (2)

Select Attributes

**Parameters**

Select Attributes

attribute filter type: single

attribute: Churn

☒ invert selection

☐ include special attributes

**Operators**

select

- Blending (9)
- Attributes (7)
- Selection (7)
- Select Attributes
- Select by Weights
- Select by Random
- Remove Attribute Range
- Remove Useless Attributes
- Remove Correlated Attributes
- Work on Subset
- Examples (2)
- Filter (2)
- Filter Examples

**Help**

Select Attributes

RapidMiner Studio Core

Tags: Filter, Keep, Remove, Drop, Delete, Columns, Variables, Features, Feature Set, Selection

**Synopsis**

This Operator selects a subset of Attributes of an ExampleSet and removes the other Attributes.

[Jump to Tutorial Process](#)

**Description**

The Operator provides different filter types to make Attribute selection

churn 데이터를 선택해서 속성 자체를 데이터에서 없앴

# 2 데이터 전처리하기



## 2. 데이터 전처리하기(실습)

The screenshot displays the RapidMiner Studio interface with a workflow designed for data preprocessing. The workflow consists of the following operators in sequence:

- Read Excel**: Loads data from an Excel file.
- Trim**: Removes leading and trailing spaces from text attributes.
- Filter Examples**: Filters the data based on specified criteria.
- Remove Duplicates**: Eliminates duplicate rows from the dataset.
- Replace**: Replaces values in a specified attribute.
- Filter Examples (2)**: A second filtering step.
- Select Attributes**: Selects specific attributes for further analysis. The parameters for this operator are:
  - attribute filter type: single
  - attribute: Chum
  - invert selection: checked
  - include special attributes: unchecked
- Store**: Saves the resulting dataset to a specified location.

The **Repository** panel on the left shows the data sources: Samples, DB (Legacy), and Local Repository (Demos). The **Operators** panel at the bottom left lists various operator categories, including Data Access, Utility, Process Control, Extensions, and Text Processing. The **Parameters** panel on the right provides configuration options for the selected **Select Attributes** operator. The **Help** panel at the bottom right offers documentation for the **Select Attributes** operator, including its tags, synopsis, and description.

store 오퍼레이터 사용, 저장 위치 지정

## 2 데이터 전처리하기

### 2. 데이터 전처리하기(실습)

CAU

The screenshot displays the RapidMiner Studio interface with a workflow in the 'Process' view. The workflow consists of the following operators: Read Excel, Trim, Filter Examples, Remove Duplicates, Replace, and Filter Examples (2). A context menu is open over the 'Filter Examples (2)' operator, showing options: Disable 5 Operators, Move into new subprocess (highlighted), Move all currently selected operators into a new subprocess, Copy (Ctrl+C), Paste (Ctrl+V), Delete, Add note, Remove all Breakpoints, and Show ExampleSet Result. The 'Operators' panel on the left shows the 'Store' operator under 'Data Access (1)'. The 'Parameters' panel on the right shows the 'Trim' operator settings. The 'Help' panel at the bottom right provides details for the 'Trim' operator, including its synopsis and description.

**Repository**

- Import Data
- Samples
- DB (Legacy)
- Local Repository (drivers)

**Process**

Process

Read Excel → Trim → Filter Examples → Remove Duplicates → Replace → Filter Examples (2)

**Operators**

- store
- Data Access (1)
  - Store
- Utility (1)
- Process Control (1)
  - Remember
- Extensions (1)
  - Text Processing (1)
    - Process Documents from Mail Store

**Parameters**

Trim

attribute filter type: all

☐ invert selection

☐ include special attributes

[Change compatibility \(9.5.001\)](#)

**Help**

Trim

RapidMiner Studio Core

Tags: Values

**Synopsis**

This operator removes leading and trailing spaces from the values of the selected nominal attributes.

[Jump to Tutorial Process](#)

**Description**

The Trim operator creates new attributes from the selected nominal attributes by removing leading and trailing spaces from the nominal values. This operation can be extended through operators.

묶고자 하는 서브프로세스를 선택 → 우클릭 → move into new sub-process 클릭



## 2 데이터 전처리하기

### 2. 데이터 전처리하기(실습)

CAU

The screenshot displays the RapidMiner Studio interface with a workflow designed for data preprocessing. The workflow consists of the following operators:

- Read Excel**: The starting operator for loading data from an Excel file.
- preprocess**: An operator used for initial data cleaning and preparation.
- Select Attributes (2)**: An operator to select specific features from the dataset.
- Set Role**: An operator used to assign roles (e.g., label, input, output) to the selected attributes.
- Select Attributes**: A second operator for further attribute selection.
- Store**: The final operator to save the processed data into a repository.

The **Parameters** panel on the right shows settings for the **Process** operator, including:

- logverbosity: init
- logfile: (empty field)
- resultfile: (empty field)
- random seed: 2001
- send mail: never
- encoding: SYSTEM

The **Help** panel at the bottom right provides information about the **Process** operator, stating it is the root operator of every process and provides parameters for logging and random number generation.

set role 오퍼레이터를 사용하여 레이블로 지정