

# Lecture 5

## How to Conduct Data Preparation Phase?

Kim, Yang Sok  
Dept. of MIS, Keimyung University

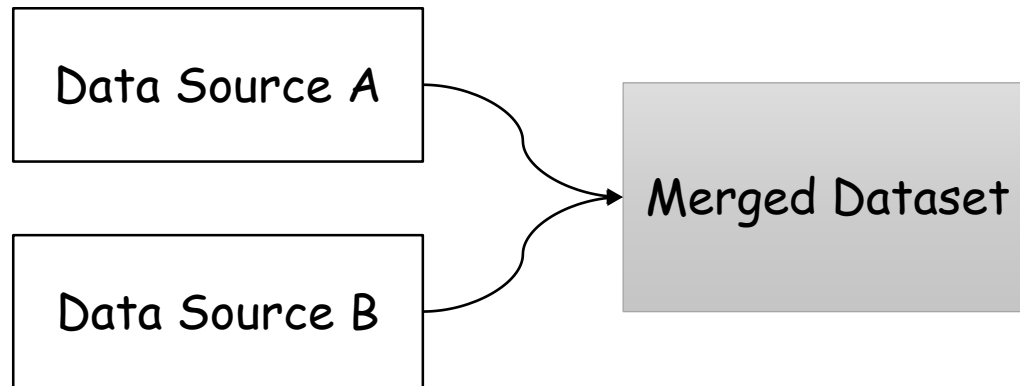
- **Introduction**
- **Task & Deliverables**
- **Exercises**
  - Exercise 1. Integrating Data with Rapidminer
  - Exercise 2. Selecting Data with Rapidminer
  - Exercise 3. Cleansing Data with Rapidminer
- **Conclusion**

# Introduction

- **Data miners spend most of their time on Data Preparation. Most data used for data mining was originally collected and preserved for other purposes and needs some refinement before it is ready to use for modeling.**
- **The data preparation phase includes five tasks. These are**
  - Selecting data
  - Cleaning data
  - Constructing data
  - Integrating data
  - Formatting data
- **Rapidminer supports data preparation through Data Blending and Data Cleansing packages.**

# Task & Deliverables

- Your data may now be in several disparate datasets.
- You'll need to merge some or all of those disparate datasets together to get ready for the modeling phase.
- The deliverable for this task is the merged dataset.



- Merging dataset refers to **joining together two or more datasets** that have different information about the same objects.
- At this stage, it may also be advisable to **generate new records**.
- It may also be recommended to **generate aggregate values**. **Aggregation** refers to operations where new values are computed by summarizing information from multiple records and/or tables.

- Now you will decide which portion of the data that you have is actually going to be used for data mining.
- The criteria you might use to make this decision include the relevance of the data to your data mining goals, the quality of the data, and also technical constraints such as limits on data volume or data types.
- Note that data selection covers selection of attributes (columns) as well as selection of records (rows) in a table.
- The deliverable for this task is the rationale for inclusion and exclusion. In it, you'll explain what data will, and will not, be used for further data-mining work.

**Task & Deliverables****Cleaning Data**

- **The data that you've chosen to use is unlikely to be perfectly clean (error-free). You need to raise the data quality to the level required by the selected analysis techniques.**
- **This may involve the selection of clean subsets of the data, the insertion of suitable defaults, or more ambitious techniques such as the estimation of missing data by modeling.**
- **The deliverable for this task is the data-cleaning report, which documents, in excruciating detail, every decision and action used to clean your data.**
  - This report should cover and refer to each data quality problem that was identified in the verify data quality task in the data-understanding phase of the process.
  - Your report should also address the potential impact on results of the choices you have made during data cleaning.



- **This task includes constructive data preparation operations such as the production of derived attributes, complete new records, or transformed values for existing attributes.**
- **Deliverables for this task include two reports:**
  - Derived attributes: A report that describes what new fields (columns) you have constructed, how you did it, and why.
  - Generated records: A report that describes what new cases (rows) you have constructed, how you did it, and why.

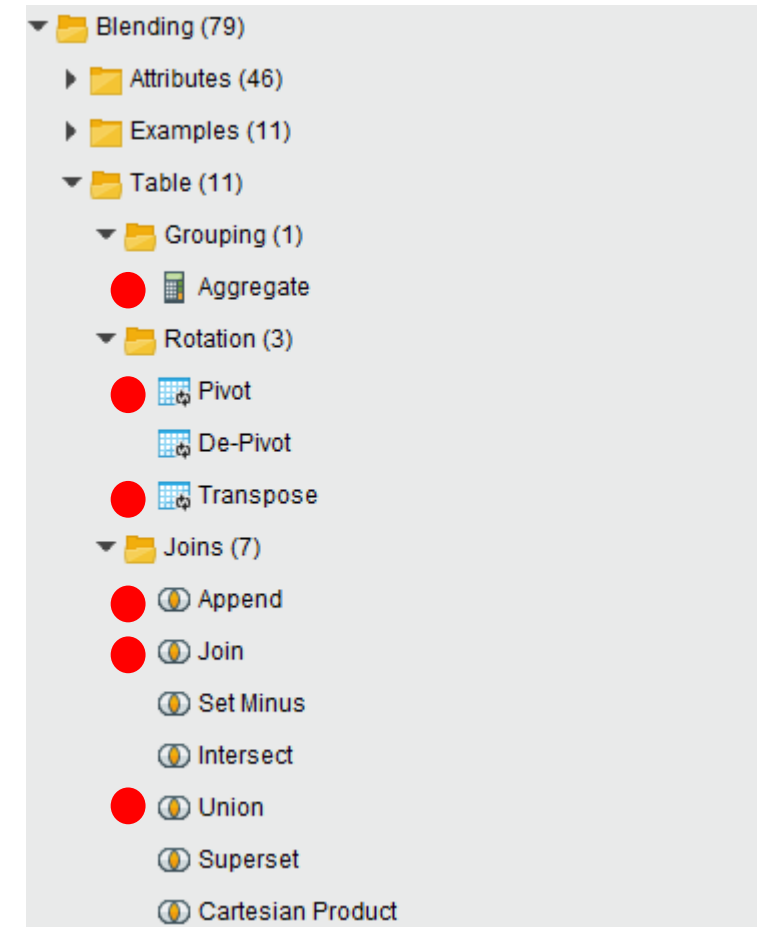
- Data often comes to you in formats other than the ones that are most convenient for modeling. (Format changes are usually driven by the design of your tools.) So convert those formats now.
- Formatting transformations refers primarily to syntactic modifications made to the data that do not change its meaning, but might be required by the modeling tool.
- Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict.
- The deliverable for this task is your reformatted data. (And a little report describing the changes you have made would be a smart thing to include.)

# Exercise 1: Integrating Data with Rapidminer

## Exercise 1: Integration Data with Rapidminer

### Data Blending

- Data can be exists multiple sources, so it is necessary to integrate them to conduct data analysis.
- Rapidminer provides operators related to data integration in Blending > Table package.
- The package has sub-package that support data integration, such as grouping, rotation, and joins.



## Exercise 1: Integration Data with Rapidminer

### Task & Steps

- **Context**

- Wine quality data and wine chemical data stored separate in two different tables and we wish to integrate them.

- **Tasks**

- Load two datasets and integrate them into single dataset

- **Steps**

1. Load red wine chemical properties data
2. Load red wine quality data
3. Combine two datasets into single data
4. Do steps 1 ~ 3 for white wine data
5. Combine red and white wine data

# Load chemical data of red wine

The screenshot displays the RapidMiner Studio interface in the 'Design' view. The main workspace shows a process diagram with two 'Read CSV' operators. The first operator, 'Read CSV-Chem', is connected to the 'Process' input and its output is connected to the 'Read CSV-Quality' operator. The second operator, 'Read CSV-Quality', is connected to the 'Read CSV-Chem' output and its output is connected to the 'Process' output. Two callout boxes provide instructions: Box 1 points to the 'Read CSV-Chem' operator with the text 'Load red wine chemical dataset (winequality-red-chem.csv)'. Box 2 points to the 'Read CSV-Quality' operator with the text 'Load red wine chemical dataset (winequality-red-quality.csv)'. The left sidebar contains the 'Repository' and 'Operators' panels. The 'Repository' panel shows a tree structure with 'Training Resources (connected)', 'Samples', 'Community Samples (connected)', 'Keras Samples', 'DB (Legacy)', and 'Local Repository (admin)'. The 'Operators' panel shows a search for 'Rea' and a list of operators under 'Data Access (23)' > 'Files (15)' > 'Read (15)', including 'Read CSV', 'Read Excel', and 'Read Excel with Format'. The right sidebar contains the 'Parameters' and 'Help' panels. The 'Parameters' panel shows settings for the 'Process' operator, including 'logverbosity' (init), 'logfile', 'resultfile', 'random seed' (2001), 'send mail' (never), and 'encoding' (SYSTEM). The 'Help' panel shows the 'Process' operator's synopsis: 'The root operator which is the outer most operator of every process.' The bottom status bar shows 'Recommended Operators' with a list of operators and their usage percentages: 'Select Attributes' (35%), 'Set Role' (32%), 'Apply Model' (25%), and 'Filter Examples' (21%).

Views: Design Results Turbo Prep Auto Model

Find data, operators...etc All Studio

Repository

- Import Data
- Training Resources (connected)
- Samples
- Community Samples (connected)
- Keras Samples
- DB (Legacy)
- Local Repository (admin)

Operators

Rea

- Data Access (23)
  - Files (15)
    - Read (15)
      - Read CSV
      - Read Excel
      - Read Excel with Format

We found "Spreadsheet Table Extraction", "Projects" and 5 more results in the Marketplace. [Show me!](#)

Process

Process

Read CSV-Chem

Read CSV-Quality

Load red wine chemical dataset (winequality-red-chem.csv)

Load red wine chemical dataset (winequality-red-quality.csv)

Parameters

Process

logverbosity: init

logfile:

resultfile:

random seed: 2001

send mail: never

encoding: SYSTEM

[Hide advanced parameters](#)

[Change compatibility \(9.3.001\)](#)

Help

Process

RapidMiner Studio Core

Synopsis

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

Description

Recommended Operators

- Select Attributes 35%
- Set Role 32%
- Apply Model 25%
- Filter Examples 21%

Business Data Mining ©Kim Yang Sok

ALT + click to disconnect.

14

# Load chemical data of red wine

Name	Type	Missing	Statistics			Filter (12 / 12 attributes): <input type="text" value="Search for Attributes"/>
fixed acidity	Real	0	Min 4.600	Max 15.900	Average 8.320	A statistical summary of chemical data of red wine
volatile acidity	Real	21	Min 0.120	Max 1.580	Average 0.527	
citric acid	Real	0	Min 0	Max 1	Average 0.271	
residual sugar	Real	0	Min 0.900	Max 15.500	Average 2.539	
chlorides	Real	0	Min 0.012	Max 0.611	Average 0.087	
free sulfur dioxide	Integer	19	Min 1	Max 72	Average 15.818	
total sulfur dioxide	Integer	0	Min 6	Max 289	Average 46.467	
density	Real	0	Min 0.990	Max 1.004	Average 0.997	
pH	Real	0	Min 2.740	Max 4.010	Average 3.311	
sulphates	Real	0	Min 0.330	Max 2	Average 0.658	
alcohol	Real	0	Min 8.400	Max 14.900	Average 10.423	
id	Integer	0	Min 1	Max 1599	Average 800	

# Load quality data of red wine

Name		Type	Missing	Statistics			Filter (2 / 2 attributes):	Search for Attributes	
▼	quality	Integer	0	Min 3	Max 8	Average 5.636			
▼	id	Integer	0	Min 1	Max 1599	Average 800			

A statistical summary  
of quality data of red  
wine



# Combine two datasets into single data

Views: Design Results

Add <<Join>> Operator 1

Repository

- Import Data
- Training Resources (connected)
- Samples
- Community Samples (connected)
- Keras Samples
- DB (Legacy)
- Local Repository (admin)

Process

Process

Read CSV-Chem

Read CSV-Quality

Join

Parameters

Join

- ☒ remove double attributes
- join type ☒ inner
- ☐ use id attribute as key
- key attributes [Edit List \(0\)...](#)
- ☐ keep both join attributes
- [Hide advanced parameters](#)

Operators

Join

Blending (7)

Table (7)

Joins (7)

- Append
- Join
- Set Minus
- Intersect

Edit Parameter List: key attributes

Edit Parameter List: key attributes

The attributes which shall be used for join. Attributes which shall be matched must be of the same type.

left key attributes	right key attributes
id	id

Set "id" as left key attributes and right key attributes 3

Click "Edit List" button 2

Click 'Apply' button 4

Apply

Cancel

Retrieve 62%

Select Attributes 53%

Set Role 35%

Filter Examples 35%

Business Data Mining ©Kim Yang Sok

Tags: Combine, Merge, Match, Left, Right, Outer, Inner, Joins

Synopsis

This Operator joins two ExampleSets using one or more Attributes from the input ExampleSets as keys

17

# Combine two datasets into single data

Views: Design Results Turbo Prep Auto Model

Find data, operators...etc All Studio ▾

Result History ExampleSet (Join) ×

Data

Statistics

Visualizations

Annotations

Open in  Turbo Prep  Auto Model

Filter (1,599 / 1,599 examples): all ▾

lity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH	sulphates	alcohol	id	quality
	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400	1	5
	0.880	0	2.600	0.098	25	67	0.997	3.200	0.680	9.400	2	5
	0.760	0.040	2.300	0.092	15	54	0.997	3.260	0.650	9.400	3	5
	0.280	0.560	1.900	0.075	17	60	0.998	3.160	0.580	9.400	4	5
	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400	5	5
	0.660	0	1.800	0.075	13	40	0.998	3.510	0.560	9.400	6	5
	0.600	0.060	1.600	0.069	15	59	0.996	3.300	0.460	9.400	7	5
	0.650	0	1.200	0.065	15	21	0.995	3.390	0.470	10	8	7
	0.580	0.020	2	0.073	9	18	0.997	3.360	0.570	9.500	9	7
	0.500	0.360	6.100	0.071	17	102	0.998	3.350	0.800	10.500	10	5
	0.580	0.080	1.800	0.097	15	65	0.996	3.280	0.540	9.200	11	5
	0.500	0.360	6.100	0.071	17	102	0.998	3.350	0.800	10.500	12	5
	0.615	0	1.600	0.089	16	59	0.994	3.580	0.520	9.900	13	5
	0.610	0.290	1.600	0.114	9	29	0.997	3.260	1.560	9.100	14	5
	0.620	0.180	3.800	0.176	52	145	0.999	3.160	0.880	9.200	15	5
	0.620	0.190	3.900	0.170	51	148	0.999	3.170	0.930	9.200	16	5

'quality' is combined with chemical attributes

ExampleSet (1,599 examples, 0 special attributes, 13 regular attributes)

Business Data Mining ©Kim Yang Sok

18

# Do steps 1 ~ 3 for white wind data

Views: Design Results Turbo Prep Auto Model

Find data, operators...etc All Studio

**Repository**

- Import data
- Training Resources (connected)
- Samples
- Community Samples (connected)
- Keras Samples
- DB (Legacy)
- Local Repository (admin)

**Operators**

- Blending (17)
- Attributes (17)
  - Names & Roles (1)
    - Rename by Generic Names
  - Generation (16)
    - Generate Attributes
    - Generate ID

**Process**

inp

Read CSV-Chem

Join for Red

Generate Attributes

Read CSV-Quality

Read CSV-White\_Ch...

Join for White

Generate Attributes (2)

res

res

res

**Parameters**

Generate Attributes (2) (Generate Attrib...

function descriptions

Edit List (0)...

☒ keep all

[Hide advanced parameters](#)

[Change compatibility \(9.3.001\)](#)

**Edit Parameter List: function descriptions**

Edit Parameter List: function descriptions

List of functions to generate.

attribute name	function expressions
type	"white"

Add Entry Remove Entry Apply

**1** Load chemical and quality data of white wine.

**2** Add <<Generate Attributes>>

**3** Click "Edit List" button

**4** Type in 'type' for 'attribute name' and "'white'" for function expression

ang Sok

19

File Edit Process View Connections Settings Extensions Help

Repository

- Import Data
- Keras Samples
- DB (Legacy)
- Local Repository (admin)
  - Connections (admin)
  - 2019 (admin)
  - Business Data Mining (admin)
    - Lecture 5 Data Preparation (admin)

Operators

sSTORE

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

Process

Process

100%

Read CSV-Chem

Join for Red

Generate Attributes

Append

Store

Read CSV-Quality

Read CSV-White\_Ch...

Join for White

Generate Attributes...

Parameters

Store

repository entry

../data/Integra

Repository Browser

Select a repository location.

- Lecture 5 Data Preparation (admin)
  - data (admin)
    - Integrating Data Result (admin - v1, 8/10/19 4:37 PM - 590 kB)
    - Selecting Data Result (admin - v1, 8/8/19 9:30 PM - 143 kB)

Name

Integrating Data Result2

Location

../data/Integrating Data Result2

relative to //Local Repository/Business Data Mining/Lecture 5 Data Preparation/pr...

OK Cancel

20

1 Add 'type' attribute with 'white' as value

2 Append white wine dataset to red wine dataset

3 Add <<Store>>

4 Click here to activate "Repository Browser"

5 Select 'data' folder and type in name in 'Name'



Views:

Design

Results

Turbo Prep

Auto Model

Find data, operators...etc



All Studio ▾

Result History

ExampleSet (Append) ×



Data



Statistics



Visualizations



Annotations

Open in



Turbo Prep



Auto Model

Filter (6,497 / 6,497 examples):

all ▾

acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH	sulphates	alcohol	id	quality	type
	0.150	1.800	0.077	26	35	0.993	3.320	0.820	11.600	1591	6	red
	0.090	1.700	0.089	16	26	0.994	3.670	0.560	11.600	1592	6	red
	0.130	2.300	0.076	29	40	0.996	3.420	0.750	11	1593	6	red
	0.080	1.900	0.068	28	38	0.997	3.420	0.820	9.500	1594	6	red
	0.080	2	0.090	32	44	0.995	3.450	0.580	10.500	1595	5	red
	0.100	2.200	0.062	39	51	0.995	3.520	0.760	11.200	1596	6	red
	0.130	2.300	0.076	29	40	0.996	3.420	0.750			6	red
	0.120	2	0.075	32	44	0.995	3.570	0.710			6	red
	0.470	3.600	0.067	18	42	0.995	3.390	0.660			6	red
	0.360	20.700	0.045	45	170	1.001	3	0.450			6	white
	0.340	1.600	0.049	14	132	0.994	3.300	0.490	9.500	2	6	white
	0.400	6.900	0.050	30	97	0.995	3.260	0.440	10.100	3	6	white
	0.320	8.500	0.058	47	186	0.996	3.190	0.400	9.900	4	6	white
	0.320	8.500	0.058	47	186	0.996	3.190	0.400	9.900	5	6	white
	0.400	6.900	0.050	30	97	0.995	3.260	0.440	10.100	6	6	white
	0.160	7	0.045	30	136	0.995	3.180	0.470	9.600	7	6	white

'type' attribute has  
been created with  
'red' or 'white' value

ExampleSet (6,497 examples, 0 special attributes, 14 regular attributes)

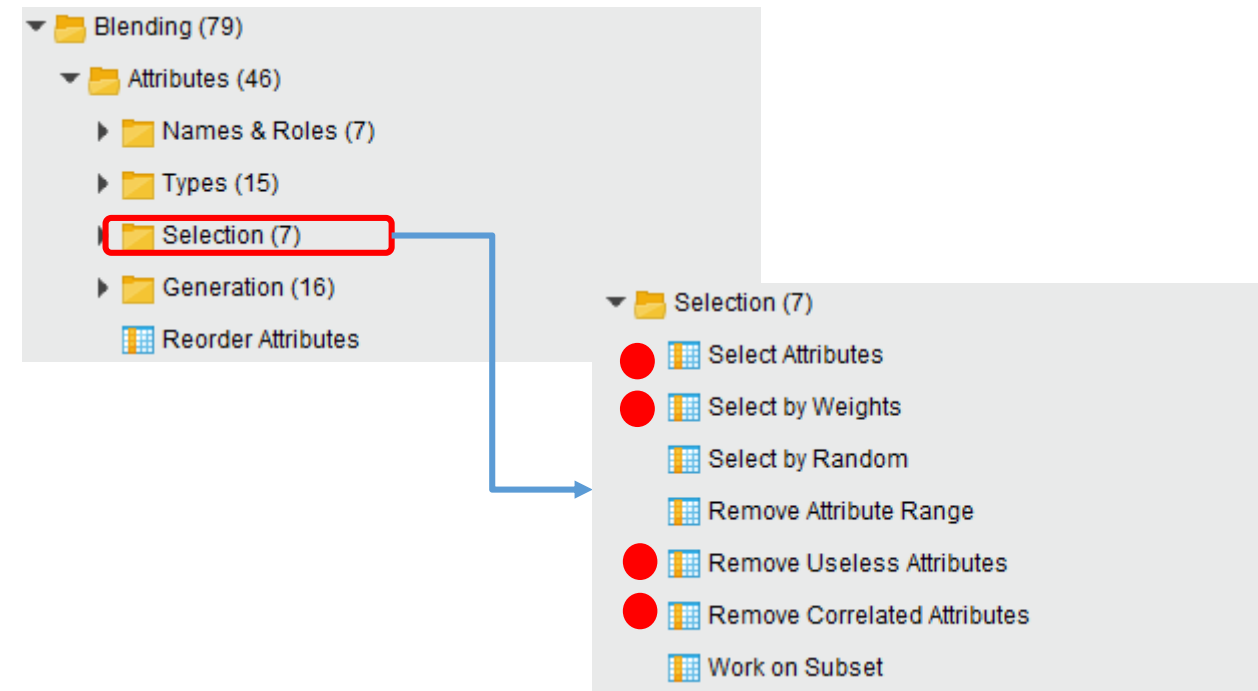
# Exercise 2:

## Selecting Data with Rapidminer

## Exercise 2: Selecting Data with Rapidminer

### Data Selection

- Data selection covers selection of attributes (columns) as well as selection of examples (rows) in a dataset.
- You can select attributes manually (e.g., Select Attributes) or automatically (e.g., Select by Weight, Remove Useless Attributes, Remove Correlated Attributes).

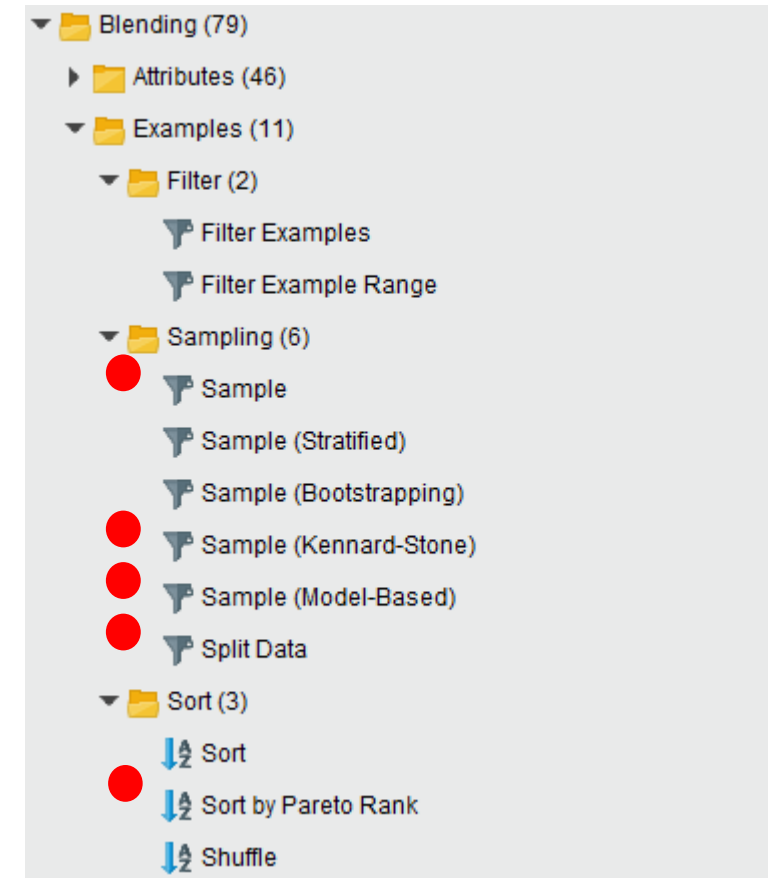




## Exercise 2: Selecting Data with Rapidminer

## Data Selection

- You can select examples using filtering (e.g., Filter Examples and sampling operators (e.g., Sample)
- Sampling is used to select subsets of examples for analysis. Sampling types are as follows:
  - Linear Sampling: Select subset of examples sequentially(linearly)
  - Random Sampling: Select subset of examples randomly
  - Stratified Sampling: Select subset of examples randomly but class distribution unchanged.
- Filtering is used to select subsets of examples using filtering condition/s.





## Exercise 2 : Selecting Data with Rapidminer

### Task & Steps

- **Tasks**

- Select useful subset of dataset from the merged dataset using attribute / example selection operators

- **Steps**

1. Select attributes except 'id'
2. Change 'type' into numeric using dummy coding
3. Remove useless attributes
4. Remove correlated attributes
5. Filter examples with certain attribute's threshold
6. Get subset of examples using sampling techniques
7. Save "Selecting Data" Result

# Select attributes except 'id'.

The screenshot displays the RapidMiner Studio interface with the 'Design' view selected. The 'Repository' pane on the left shows a project structure with 'data' and 'process' folders. The 'Operators' pane at the bottom left lists various data transformation operators, with 'Nominal to Numerical' highlighted. The 'Process' pane in the center shows a workflow with two operators: 'Retrieve Integrating ...' and 'Select Attributes'. A blue circle with the number '1' points to the 'Select Attributes' operator. A callout box points to its parameters, stating: 'Set 'attribute filter type' as 'single' and attribute is 'id'. Check 'invert selection''. The 'Parameters' pane on the right shows the 'Select Attributes' operator configured with 'attribute filter type' set to 'single', 'attribute' set to 'id', and 'invert selection' checked. A white callout box states: 'id' has been removed because it is useless. The 'Help' pane at the bottom right provides a synopsis of the 'Select Attributes' operator. The status bar at the bottom indicates 'Business Data Mining ©Kim Yang Sok'.

Views: Design

Repository

- Import Data
- Business Data Mining (admin)
  - Lecture 5 Data Preparation (admin)
    - data (admin)
      - Integrating Data Result (admin)
      - Integrating Data Result2 (admin)
      - Selecting Data Result (admin)
    - process (admin)
      - Cleansing Data (admin - v1, 8/)
      - Integrating Data (admin - v1, 8/)
      - Integrating Data2 (admin - v1, 8/)

Operators

Nomi

- Numerical to Binominal
- Numerical to Polynomial
- Nominal to Binominal
- Nominal to Text
- Nominal to Numerical
- Nominal to Date
- Text to Nominal

We found "RapidMiner Finance and Econom..." and "Edda - Extensions for Binomin..." in the Marketplace. [Show me!](#)

Process

Process

Retrieve Integrating ...

Select Attributes

Parameters

Select Attributes

attribute filter type single

attribute id

invert selection

include special attributes

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 Example

Recommended Operators

Business Data Mining ©Kim Yang Sok

26

Select attributes except 'id'.

Views: DesignResultsTurbo PrepAuto Model

Find data, operators...etcAll Studio

Result HistoryExampleSet (Append)

Open in Turbo PrepAuto Model

Filter (6,497 / 6,497 examples): all

acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH	sulphates	alcohol	id	quality	type
0.150	1.800	0.077	26	35	0.993	3.320	0.820	11.600	1591	6	red	
0.090	1.700	0.089	16	26	0.994	3.670	0.560	11.600				
0.130	2.300	0.076	29	40	0.996	3.420	0.750	11				
0.080	1.900	0.068	28	38	0.997	3.420	0.820	9.500				
0.080	2	0.090	32	44	0.995	3.450	0.580	10.500	1595	5	red	

Before attributes are selected

'id' has not been selected

<new process\*> - RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ

Views: DesignResultsTurbo PrepAuto Model

Find data, operators...etcAll Studio

Result HistoryExampleSet (Select Attributes)

Open in Turbo PrepAuto Model

Filter (6,497 / 6,497 examples): all

ity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH	sulphates	alcohol	quality	type
0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400	5	red	
0.880	0	2.600	0.098	25	67	0.997	3.200	0.680	9.800	5	red	
0.760	0.040	2.300	0.092	15	54	0.997	3.260	0.650	9.800	5	red	
0.280	0.560	1.900	0.075	17	60	0.998	3.160	0.580	9.800	6	red	
0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400	5	red	
0.660	0	1.800	0.075	13	40	0.998	3.510	0.560	9.400	5	red	

After attributes are selected

Business Data Mining © Kim YoungSok

27

# Change 'type' into numeric using dummy coding

The screenshot displays the RapidMiner Studio interface with the following components and annotations:

- Repository:** Shows a project structure with 'Lecture 5 Data Preparation' containing 'data' and 'process' folders. The 'process' folder contains operators like 'Cleansing Data', 'Integrating Data', and 'Integrating Data2'.
- Process Panel:** Contains a workflow with three operators: 'Retrieve Integrating ...', 'Select Attributes', and 'Nominal to Numerical'. The 'Nominal to Numerical' operator is highlighted with a blue circle and labeled '1'.
- Parameters Panel:** Shows the configuration for the 'Nominal to Numerical' operator. The 'attribute filter type' is set to 'single' (labeled '2'), and the 'attribute' is set to 'type'. The 'invert selection' checkbox is checked. The 'coding type' is set to 'dummy coding'.
- Operators Panel:** Shows a search for 'Nomi' with 'Nominal to Numerical' selected.
- Help Panel:** Provides information about the 'Nominal to Numerical' operator, including tags like 'Categorical', 'Ordinal', 'Qualitative', 'Quantitative', 'Dummy', 'Coding', 'One hot', 'Encoding', 'Index', 'Continuous', and 'Types'.

Annotations and workflow details:

- A blue circle with the number '1' points to the 'Nominal to Numerical' operator in the Process Panel.
- A blue circle with the number '2' points to the 'attribute filter type' dropdown in the Parameters Panel.
- A text box states: 'Set 'attribute filter type' as 'single' and attribute is 'type'. Check 'invert selection''.
- A text box states: 'Add <<Nominal to Numerical>> into Process Panel'.
- The workflow in the Process Panel shows data flowing from 'inp' through 'Retrieve Integrating ...' and 'Select Attributes' to the 'Nominal to Numerical' operator, which then outputs to 'res'.

# Change 'type' into numeric using dummy coding

Views: Design Results Turbo Prep Auto Model

Find data, operators...etc All Studio ▾

Result History ExampleSet (Nominal to Numerical) ×

Data

Statistics

Visualizations

Annotations

Open in  Turbo Prep  Auto Model

Filter (6,497 / 6,497 examples): all ▾

Row No.	type = red	type = white	fixed acidity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH	sulphur
1	1	0	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560
2	1	0	7.800	0.880	0	2.600	0.098	25	67	0.997	3.200	0.680
3	1	0	7.800	0.760	0.040	2.300	0.092	15	54	0.997	3.260	0.650
4	1	0	11.200	0.280	0.560	1.900	0.075	17	60	0.998	3.160	0.580
5	1	0	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560
6	1	0	7.400	0.660	0	1.800	0.075	13	40	0.998	3.510	0.560
7	1	0	7.900	0.600	0.060	1.600	0.069	15	59	0.996	3.300	0.460
8	1	0	7.300	0.650	0	1.200	0.065	15	21	0.995	3.390	0.470
9	1	0	7.400	0.580	0.020	2	0.073	9	18	0.997	3.360	0.570
10	1	0	7.400	0.600	0	6.100	0.071	17	102	0.998	3.350	0.800
11	1	0	7.400	0.600	0	1.800	0.097	15	65	0.996	3.280	0.540
12	1	0	7.400	0.600	0	6.100	0.071	17	102	0.998	3.350	0.800
13	1	0	7.400	0.600	0	1.600	0.089	16	59	0.994	3.580	0.520
14	1	0	7.800	0.610	0.290	1.600	0.114	9	29	0.997	3.260	1.560
15	1	0	8.900	0.620	0.180	3.800	0.176	52	145	0.999	3.160	0.880
16	1	0	8.900	0.620	0.190	3.900	0.170	51	148	0.999	3.170	0.930

'type' has been changed into numeric by using dummy coding

ExampleSet (6,497 examples, 0 special attributes, 14 regular attributes)

Business Data Mining ©Kim Yang Sok

29

# Remove useless attributes based on deviation

The screenshot displays the RapidMiner Studio interface with the following components:

- Repository:** A tree view on the left showing a project structure with folders like 'data' and 'process'. The 'Integrating Data Result2' file is selected.
- Process Panel:** The central workspace showing a workflow. It starts with an input 'inp' connected to a 'Retrieve Integrating ...' operator. This is followed by a 'Select Attributes' operator, then a 'Nominal to Numerical' operator, and finally the 'Remove Useless Attributes' operator (highlighted with a red border). The workflow ends with two 'res' output ports.
- Operators:** A panel on the bottom left with a search bar containing 'remove'. The 'Remove Useless Attributes' operator is highlighted in the list.
- Parameters:** A panel on the right titled 'Remove Useless Attributes' showing configuration options:
  - 'numerical min deviat...': 0.1
  - 'nominal useless ab...': 1.0
  - 'nominal remove id like': ☐
  - 'nominal useless bel...': 0.0
- Help:** A panel on the bottom right showing the 'Remove Useless Attributes' operator's documentation, including tags and a synopsis.

Two callout boxes provide instructions:

- 1 Find <<Remove Useless Attributes>> and place it on Process Panel
- 2 Set 'numerical min deviation' as '0.1'

At the bottom of the interface, the text 'Recommended Operators' and 'Business Data Mining ©Kim Yang Sok' are visible.

# Remove useless attributes based on deviation

Views: DesignResultsTurbo PrepAuto Model

Find data, operators...etc

All Studio

Result History

ExampleSet (Nominal to Numerical)

Open in

Turbo Prep

Auto Model

chloride' is removed from dataset

density' is removed from dataset

Row No.	type = red	type = white	fixed acidity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH	sulph
1	1	0	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560
2	1	0	7.800	0.880	0	2.600	0.098	25	67	0.997	3.200	0.680
3	1	0	7.800	0.760	0.040	2.300	0.092	15	54	0.997	3.260	0.650
4	1	0	11.200	0.280	0.560	1.900				0.998	3.160	0.580
5	1	0	7.400	0.700	0	1.900				0.998	3.510	0.560
6	1	0	7.400	0.660	0	1.800				0.998	3.510	0.560
7	1	0	7.000	0.600	0.060	1.600	0.050	15	50	0.998	3.200	0.450

After removing useless attributes

<new process\*> - RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ

FileEditProcessViewConnectionsSettingsExtensionsHelp

Views: DesignResultsTurbo PrepAuto Model

Find data, operators...etc

All Studio

Result History

ExampleSet (Remove Useless Attributes)

Open in

Turbo Prep

Auto Model

Filter (6,497 / 6,497 examples):

all

Row No.	type = red	type = white	fixed acidity	volatile acidity	citric acid	residual sug...	free sulfur d...	total sulfur d...	pH	sulphates	alcohol	quali
1	1	0	7.400	0.700	0	1.900	11	34	3.510	0.560	9.400	5
2	1	0	7.800	0.880	0	2.600	25	67	3.200	0.680	9.800	5
3	1	0	7.800	0.760	0.040	2.300	15	54	3.260	0.650	9.800	31
4	1	0	11.200	0.280	0.560	1.900	17	60	3.160	0.580	9.800	6

Business Data Mining © Kim Yang Sok



# Remove correlated attributes

The screenshot displays the RapidMiner Studio interface with the 'Design' view selected. The 'Repository' panel on the left shows a project structure with 'Lecture 5 Data Preparation' containing 'data' and 'process' folders. The 'Operators' panel on the bottom left shows a search for 'remove', with 'Remove Correlated Attributes' highlighted under the 'Filter' category. The 'Process' panel in the center shows a workflow: 'Retrieve Integrating ...' → 'Select Attributes' → 'Nominal to Numerical' → 'Remove Useless Att...' → 'Remove Correlated Attributes'. The 'Remove Correlated Attributes' operator is highlighted with a red border. A callout box with a blue circle containing the number '2' points to the operator, containing the text: 'Set 'correlation' as '0.65' and 'filter relation' as 'greater''. Another callout box with a blue circle containing the number '1' points to the 'Remove Correlated Attributes' operator in the 'Parameters' panel on the right, containing the text: 'Find <<Remove Correlated Attributes>> and place it on Process Panel'. The 'Parameters' panel shows the following settings: 'correlation' is set to '0.65', 'filter relation' is set to 'greater', 'attribute order' is set to 'original', 'use absolute correlation' is checked, and 'use local random seed' is unchecked. The 'Help' panel on the bottom right shows the 'Remove Correlated Attributes' operator's synopsis: 'This operator removes correlated attributes from an ExampleSet'.

Views: Design

Find data, operators...etc

All Studio

Repository

Import Data

Business Data Mining (admin)

Lecture 5 Data Preparation (admin)

data (admin)

Integrating Data Result (admin)

Integrating Data Result2 (admin)

Selecting Data Result (admin)

process (admin)

Cleansing Data (admin - v1, 8/)

Integrating Data (admin - v1, 8/)

Integrating Data2 (admin - v1, 8/)

Operators

remove

Select Attributes

Remove Attribute Range

Remove Useless Attributes

Remove Correlated Attributes

Examples (2)

Filter (2)

Filter Examples

Filter Example Range

No results were found.

Process

Process

Retrieve Integrating ...

Select Attributes

Nominal to Numerical

Remove Useless Att...

Remove Correlated Attributes

Set 'correlation' as '0.65' and 'filter relation' as 'greater'

Find <<Remove Correlated Attributes>> and place it on Process Panel

Parameters

Remove Correlated Attributes

correlation 0.65

filter relation greater

attribute order original

☒ use absolute correlation

☐ use local random seed

[Hide advanced parameters](#)

[Range compatibility \(9.3.001\)](#)

Help

Remove Correlated Attributes

RapidMiner Studio Core

Tags: Filter, Keep, Remove, Drop, Delete, Correlations, Selection

Synopsis






This operator removes correlated attributes from an ExampleSet

Business Data Mining ©Kim Yang Sok



32




# Remove correlated attributes




Views: Design Results Turbo Prep Auto Model

Find data, operators...etc  All Studio 


Result History ExampleSet (Remove Useless Attributes)





Data



Statistics



Visualizations

Open in  Turbo Prep  Auto Model

Filter (6,497 / 6,497 examples): all






Row No.	type = red	type = white	fixed acidity	volatile acidity	citric acid	residual sug...	free sulfur d...	total sulfur d...	pH	sulphates	alcohol	quality
1	1	0	7.400	0.700	0	1.900	11	34	3.510	0.560	9.400	5
2	1	0	7.800	0.880		2.600	25	67	3.200	0.680	9.800	5
3	1	0					15	54	3.260	0.650	9.800	5
4	1	0					17	60	3.160	0.580	9.800	6
5	1	0					11	34	3.510	0.560	9.400	5
6	1	0	7.400	0.660	0	1.800	13				10	5

'type=white' and 'total sulfur dioxide' is removed from dataset



After removing correlated attributes

<new process\*> - RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ

File Edit Process View Connections Settings Extensions Help



Views: Design Results Turbo Prep Auto Model

Find data, operators...etc  All Studio 

Result History ExampleSet (Remove Useless Attributes) ExampleSet (Remove Correlated Attributes)



Data



Statistics

Open in  Turbo Prep  Auto Model

Filter (6,497 / 6,497 examples): all

Row No.	type = red	fixed acidity	volatile acidity	citric acid	residual sug...	free sulfur d...	pH	sulphates	alcohol	quality
1	1	7.400	0.700	0	1.900	11	3.510	0.560	9.400	5
2	1	7.800	0.880	0	2.600	25	3.200	0.680	9.800	5
3	1	7.800	0.760	0.040	2.300	15	3.260	0.650	9.800	5
4	1	11.200	0.280	0.560	1.900	17	3.160	0.580	9.800	6

# Filter examples with certain attribute's threshold

Views: Design Results Turbo Prep Auto Model

Find data, operators...etc All Studio

Result History ExampleSet (Remove Correlated Attributes)

Open in Turbo Prep Auto Model

Filter (6,497 / 6,497 examples): all

Sort examples with 'residual sugar' attribute in descending order

These three examples have too high value → They might be outliers

Row No.	type	tritic acid	residual ... ↓	free sulfur d...	pH	sulphates	alcohol	quality
4381	0	600	65.800	8	3.390	0.690	11.700	6
3253	0		31.600	35	3.150	0.380	8.800	6
3263	0	280	31.600	35	3.150	0.380	8.800	6
5219	0	6.800	0.450	0.280				
5223	0	6.800	0.450	0.280				
3208	0	6.900	0.270	0.490				
6080	0	5.900	0.220	0.450				
1782	0	6.800	0.280	0.400				
1791	0	6.800	0.280	0.400				
2044	0	6.900	0.240	0.360				
5330	0	6.200	0.220	0.200				
1600	0	7	0.270	0.360				
1607	0	7	0.270	0.360				
4220	0	6.500	0.280	0.280				
5707	0	6.800	0.300	0.260				
4385	0	6.400	0.240	0.250				

ExampleSet (6,497 examples, 0 special attributes, 10 regular attributes)

Generate ID

residual sugar

id

red white

Threshold > 30

Business Data Mining ©Kim Yang Sok

34

# Filter examples with certain attribute's threshold

Views: Design Results Turb Prep

Repository

- Import Data
- Training Resources (connected)
- Samples
- Community Samples (connected)
- Keras Samples
- DB (Legacy)
- Local Repository (admin)

Process

Process

Retrieve Integrating ...

Select Attributes

Nominal to Numerical

Remove Useless Att...

Remove Correlated ...

Filter Examples

Parameters

Filter Examples

filters

condition class

custom\_filters

invert filter

Hide advanced parameters

Change compatibility (9.3.001)

Operators

Filter Example

Filter (2)

Filter Examples

Filter Example Range

Extensions (1)

Operator Toolbox (1)

Blending (1)

Filter Examples with Missing V

No results were found.

Recommended Op

Set

29%

Process Data Mining © Kim Yang Sok

Add Entry

OK

Cancel

1 Find <<Filter Examples>> and add to Process Panel

2 Click 'Add Filters...' button

3 Set filtering condition

Create Filters: filters

Create Filters: filters

Defines the list of filters to apply.

residual sugar

≤

30

Examples

er Studio Core

Remove, Drop, Delete,

ances, Lines, Observations,

er

lects which Examples of

re kept

35

# Filter examples with certain attribute's threshold

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

Result History ExampleSet (Remove Correlated Attributes)

Open in Turbo Prep Auto Model Filter (6,497 / 6,497 examples): all

These three examples are removed from dataset

Row No.	citric acid	residual ... ↓	free sulfur d...	pH	sulphates	alcohol	quality
4381	600	65.800	8	3.390	0.690	11.700	6
3253	280	31.600	35	3.150	0.380	8.800	6
3263	0	31.600	35	3.150	0.380	8.800	6
5219	0	26.050	27	3.060	0.420	10.600	6
5223	0	26.050	27	3.060	0.420	10.600	6

//Local Repository/Business Data Mining/Lecture 5 Data Preparation/process/Cleansing Data2\* - RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ

File Edit Process View Connections Settings Extensions Help

Result History ExampleSet (Filter Examples)

Open in Turbo Prep Auto Model

After filtering examples with attribute's threshold

Row No.	type = red	fixed acidity	volatile acidity	citric acid	residual ... ↓	free sulfur d...	pH	sulphates
5216	0	6.800	0.450	0.280	26.050	27	3.060	0.420
5220	0	6.800	0.450	0.280	26.050	27	3.060	0.420
3208	0	6.900	0.270	0.490	23.500	59	2.980	0.470
6077	0	5.900	0.220	0.450	22.600	55	3.100	0.350
1782	0	6.800	0.280	0.400	22	48	2.930	36
1791	0	6.800	0.280	0.400	22	48	2.930	0.500

ExampleSet (6,497 examples, 0 special attributes)

# Get subset of examples using stratified sampling techniques

The screenshot displays the RapidMiner Studio interface with a workflow in the Process panel. The workflow consists of the following operators: Retrieve Integrating..., Select Attributes, Nominal to Numerical, Remove Useless Att..., Remove Correlated..., Filter Examples, and Sample (Stratified). The Sample (Stratified) operator is highlighted with an orange border. A blue arrow points from the 'Find <<Sample(Stratified)>> and add it to Process Panel' callout to the operator in the workflow. Another blue arrow points from the 'Set 'sample' as 'absolute' and 'sample size' as '500'' callout to the parameters of the Sample (Stratified) operator in the Parameters panel. The Parameters panel shows 'sample' set to 'absolute' and 'sample size' set to '500'. The Repository panel on the left shows a tree structure with 'Business Data Mining' and 'Lecture 5 Data Preparation' folders. The 'Outliers' folder is expanded, showing 'Detect Outlier (Distances)', 'Detect Outlier (Densities)', 'Detect Outlier (LOF)', and 'Detect Outlier (COF)'. The 'Extensions' folder is also expanded. The bottom status bar indicates 'Business Data Mining ©Kim Yang Sok'.

Views: Design Results

Repository

- Import Data
- Business Data Mining (admin)
  - Lecture 5 Data Preparation (admin)
    - data (admin)
      - Integrating Data Result (admin)
      - Integrating Data Result2 (admin)
      - Selecting Data Result (admin)
    - process (admin)
      - Data Loading Practice 2 (admin - v1)
    - Handl

Operator

- Outlier
  - Cleansing (4)
    - Outliers (4)
      - Detect Outlier (Distances)
      - Detect Outlier (Densities)
      - Detect Outlier (LOF)
      - Detect Outlier (COF)
  - Extensions (1)

We found "Information Selection" in the Marketplace. [Show me!](#)

Process

Process

Retrieve Integrating ...

Select Attributes

Nominal to Numerical

Remove Useless Att...

Remove Correlated ...

Filter Examples

Sample (Stratified)

Parameters

Sample (Stratified)

sample: absolute

sample size: 500

☐ use local random seed

[Hide advanced parameters](#)

Help

Sample (Stratified)

RapidMiner Studio Core






Tags: Subsets, Random, Ratio, Stratified, Stratification, Bootstrap, Population, Downsample, Sampling

Synopsis


This operator creates a stratified sample from an ExampleSet. Str

Business Data Mining ©Kim Yang Sok


# Get subset of examples using sampling techniques





Views: Design Results Turbo Prep Auto Model


Find data, operators...etc  All Studio ▼



Result History ExampleSet (Sample (Stratified)) ✕

 Data

 Statistics

 Visualizations

 Annotations

Open in  Turbo Prep  Auto Model

Filter (500 / 500 examples): all ▼

Row No.	type = red	fixed acidity	volatile acidity	citric acid	residual sug...	free sulfur d...	pH	sulphates	alcohol	ty
1	1	6.700	0.580	0.080	1.800	15	3.280			
2	1	7.500	0.500	0.360	6.100	17	3.350			
3	1	6.900	0.685	0	2.500	22	3.460			
4	1	7.800	0.645	0	5.500	5	3.400	0.550	9.600	6
5	1	8.100	0.380	0.280	2.100	13	3.230	0.730	9.700	7
6	1	7.500	0.630	0.120	5.100	50	3.260	0.770	9.400	5
7	1	7.700	0.630	0.080	1.900	15	3.320	0.540	9.500	6
8	1	6.900	0.550	0.150	2.200	19	3.410	0.590	10.100	5
9	1	7	0.620	0.080	1.800	8	3.480	0.530	9	5
10	1	7	0.690	0.080	1.800	22	3.340	0.540	9.200	6
11	1	9	0.620	0.040	1.900	27	3.160	0.700	9.400	5
12	1	7.300	0.330	0.470	2.100	5	3.330	0.530	10.300	6
13	1	9.200	0.520	1	3.400	32	2.740	2	9.400	4
14	1	7.800	0.630	0.480	1.700	14	3.190	0.620	9.500	5
15	1	6.800	0.640	0.100	2.100	18	3.340	0.520	10.200	5
16	1	8.800	0.610	0.140	2.400	10	3.190	0.590	9.500	5

A total of 500 examples are selected.

ExampleSet (500 examples, 0 special attributes, 10 regular attributes)

Business Data Mining ©Kim Yang Sok

38



# Save Selecting Data Results

The screenshot shows the RapidMiner Studio interface with a workflow diagram and several annotation boxes:

- Annotation 1:** "Find <<Store>> and add it to Process Panel" points to the **Store** operator in the **Operators** panel.
- Annotation 2:** "Click 'File Open' icon" points to the folder icon in the **repository entry** field.
- Annotation 3:** Points to the **data** folder in the **Select a repository location** dialog.
- Annotation 4:** "Type in 'Selecting Data Result'" points to the **Name** field in the **Select a repository location** dialog.

The **Select a repository location** dialog shows the following details:

- Name:** Selecting Data Result
- Location:** ../data/Selecting Data Result
- Resolve relative to //Local Repository/Business Data Mining/Lecture 5 Data Preparation/process** (checked)

The workflow diagram shows the following sequence of operators:

- Nominal to Numerical** (pink box)
- Filter Examples** (pink box)
- Sample (Stratified)** (pink box)
- Store** (orange box)

The **Store** operator is highlighted with a blue arrow from the **Operators** panel.

**Help Panel:**

- Store** (RapidMiner Studio Core)
- Tags: [Save](#), [Export](#), [Write](#), [Datasets](#), [Repository](#), [Data Access](#)
- Synopsis**
- This operator stores an IO Object in the data repository.
- [Jump to Tutorial Process](#)

**Page-Footer:** Business Data Mining ©Kim Yang Sok

# Exercise 3:

# Cleansing Data with

# Rapidminer



- Data cleansing aims to raise the data quality to the level required by the selected analysis techniques.
- Rapidminer supports data cleansing through Normalization, Binning, Missing, Duplicates, Outliers and Dimensionality Reduction packages.



### Exercise 3: Cleansing Data with Rapidminer

#### Data Cleansing - Missing Values

- **Many real-world datasets may contain missing values for various reasons.**
- **One way to handle this problem is to get rid of the observations that have missing data. However, you will risk losing data points with valuable information.**
- **A better strategy would be to impute the missing values. In other words, we need to infer those missing values from the existing part of the data.**
- **Note that imputation does not necessarily give better results.**

## Exercise 3: Cleansing Data with Rapidminer

## Data Cleansing - Missing Values

- **In order to handle missing values, it is necessary to understand types of missing values.**
- **Missing at Random (MAR):**
  - Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data
- **Missing Completely at Random (MCAR):**
  - The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.
- **Missing not at Random (MNAR):**
  - Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable's value (e.g. Let's assume that females generally don't want to reveal their ages! Here the missing value in age variable is impacted by gender variable)

Safe to remove the data with missing values

Removing observations with missing values can produce a bias in the model. So we have to be really careful before removing observations.

- **Some algorithms can factor in the missing values and learn the best imputation values for the missing data based on the training loss reduction (ie. XGBoost).**
- **Some others have the option to just ignore them.**
- **However, other algorithms will panic and throw an error complaining about the missing values. In that case, you will need to handle the missing data and clean it before feeding it to the algorithm.**

## Exercise 3: Cleansing Data with Rapidminer

## Data Cleansing - Missing Values

- **Imputation Using (Mean/Median) Values:**

- This works by calculating the mean/median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others. It can only be used with numeric data.

	col1	col2	col3	col4	col5			col1	col2	col3	col4	col5	
0	2	5.0	3.0	6	NaN	mean()		0	2.0	5.0	3.0	6.0	7.0
1	9	NaN	9.0	0	7.0			1	9.0	11.0	9.0	0.0	7.0
2	19	17.0	NaN	9	NaN			2	19.0	17.0	6.0	9.0	7.0

- Pros:
  - Easy and fast.
  - Works well with small numerical datasets.
- Cons:
  - Doesn't factor the correlations between features. It only works on the column level.
  - Will give poor results on encoded categorical features (do NOT use it on categorical features).
  - Not very accurate.
  - Doesn't account for the uncertainty in the imputations.

## Exercise 3: Cleansing Data with Rapidminer

### Data Cleansing - Missing Values

- **Imputation Using (Most Frequent) or (Zero/Constant) Values**
  - Most Frequent is another statistical strategy to impute missing values and YES!! It works with categorical features (strings or numerical representations) by replacing missing data with the most frequent values within each column.
  - Pros:
    - Works well with categorical features.
  - Cons:
    - It also doesn't factor the correlations between features.
    - It can introduce bias in the data.
- Zero or Constant imputation — as the name suggests — it replaces the missing values with either zero or any constant value you specify

	col1	col2	col3	col4	col5			col1	col2	col3	col4	col5	
0	2	5.0	3.0	6	NaN	df.fillna(0)		0	2	5.0	3.0	6	0.0
1	9	NaN	9.0	0	7.0			1	9	0.0	9.0	0	7.0
2	19	17.0	NaN	9	NaN			2	19	17.0	0.0	9	0.0

**Exercise 3: Cleansing Data with Rapidminer****Data Cleansing - Missing Values**

- **We create a predictive model to estimate values that will substitute the missing data.**
- **In this case, we divide our data set into two sets: One set with no missing values for the variable (training) and another one with missing values (test).**
  - Can use various prediction algorithm (e.g., k-NN, logistic regression, linear regression), depending on variables.

- **In conclusion, there is no perfect way to compensate for the missing values in a dataset.**
- **Each strategy can perform better for certain datasets and missing data types but may perform much worse on other types of datasets.**
- **There are some set rules to decide which strategy to use for particular types of missing values, but beyond that, you should experiment and check which model works best for your dataset.**



## Exercise 3: Cleansing Data with Rapidminer

## Data Cleansing - Normalization

- **Normalization** aims to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information.
- **Zscore: Converts all values to a z-score.**
  - The values in the column are transformed using the following formula:
- **MinMax: The min-max normalizer linearly rescales every feature to the [0,1] interval.**
  - Rescaling to the [0,1] interval is done by shifting the values of each feature so that the minimal value is 0, and then dividing by the new maximal value (which is the difference between the original maximal and minimal values).
  - The values in the column are transformed using the following formula:

$$z = \frac{x - \text{mean}(x)}{\text{stdev}(x)}$$

$$z = \frac{x - \min(x)}{[\max(x) - \min(x)]}$$

## Exercise 3: Cleansing Data with Rapidminer

## Data Cleansing - Binning

- **Binning or discretization is the process of transforming numerical variables into categorical counterparts. e.g., an example is to bin values for Age into categories such as 20-39, 40-59, and 60-79.**
- **Unsupervised Binning methods transform numerical variables into categorical counterparts but do not use the target (class) information.**
  - Equal Width Binning
    - The algorithm divides the data into  $k$  intervals of equal size. The width of intervals is:  
$$w = (\text{max} - \text{min}) / k$$
    - And the interval boundaries are:  
$$\text{min} + w, \text{min} + 2w, \dots, \text{min} + (k-1)w$$
  - Equal Frequency Binning
    - The algorithm divides the data into  $k$  groups which each group contains approximately same number of values. For the both methods, the best way of determining  $k$  is by looking at the histogram and try different intervals or groups.

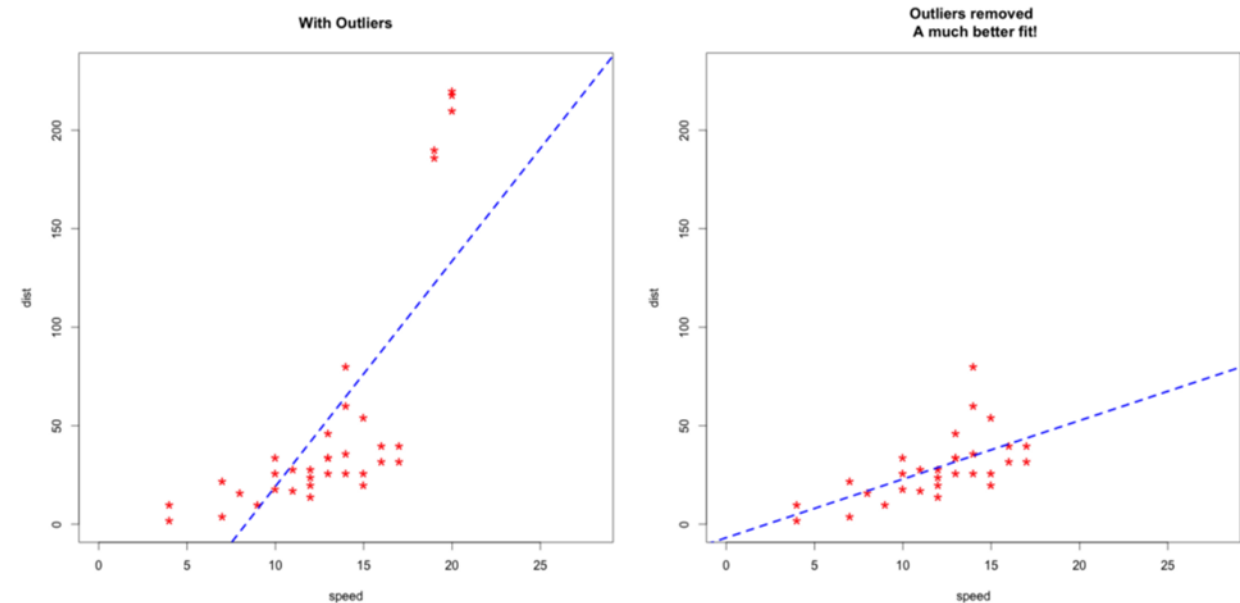
- **Supervised binning methods transform numerical variables into categorical counterparts and refer to the target (class) information when selecting discretization cut points.**
- **Entropy-based binning is an example of a supervised binning method. The entropy (or the information content) is calculated based on the class label.**
  - Intuitively, it finds the best split so that the bins are as pure as possible that is the majority of the values in a bin correspond to have the same class label.
  - Formally, it is characterized by finding the split with the maximal information gain.

- **An outlier is an observation point that is distant from other observations.**
- **Outliers exist due to one of the four following reasons:**
  - Incorrect data entry can cause data to contain extreme cases.
  - A second reason for outliers can be failure to indicate codes for missing values in a dataset.
  - Another possibility is that the case did not come from the intended sample.
  - And finally, the distribution of the sample for specific variables may have a more extreme distribution than normal.

## Exercise 3: Cleansing Data with Rapidminer

### Data Cleansing - Outliers

- While outliers are attributed to a rare chance and may not necessarily be fully explainable, outliers in data can distort predictions and affect the accuracy, if you don't detect and handle them.
- The contentious decision to consider or discard an outlier needs to be taken at the time of building the model.
  - Outliers can drastically bias/change the fit estimates and predictions. It is left to the best judgement of the analyst to decide whether treating outliers is necessary and how to go about it.



## Exercise 3: Cleansing Data with Rapidminer

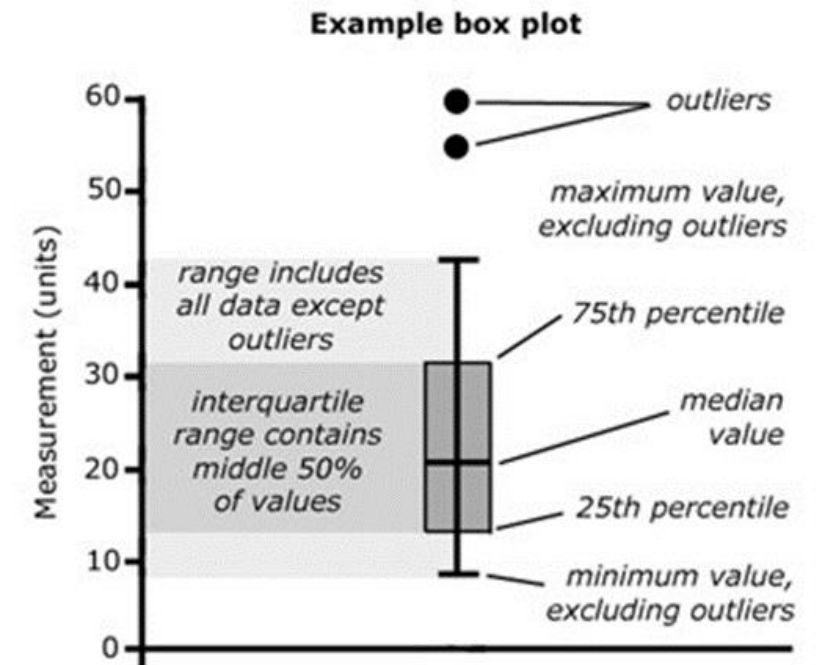
### Data Cleansing - Outlier Detection

- **Univariate Approach:**

- A univariate outlier is a data point that consists of an extreme value on one variable.

- **The Box Plot Rule**

- For a given continuous variable, outliers are those observations that lie outside  $1.5 * IQR$ , where IQR, the 'Inter Quartile Range' is the difference between 75th and 25th quartiles. This is also known as "The Box Plot Rule".
- The box plot rule is the simplest statistical technique that has been applied to detect univariate outliers. Typically, in the Univariate Outlier Detection Approach look at the points outside the whiskers in a box plot.



## Exercise 3: Cleansing Data with Rapidminer

### Data Cleansing - Outlier Detection

- **Multivariate Approach:**
  - Declaring an observation as an outlier based on a just one (rather unimportant) feature could lead to unrealistic inferences. When you have to decide if an individual entity (represented by row or observation) is an extreme value or not, it better to collectively consider the features (X's) that matter.
- **A multivariate outlier is a combination of unusual scores on at least two variables.**
- **Several methods are used to identify outliers in multivariate datasets. Two of the widely used methods are:**
  - Distance-based Outlier Detection
  - Density-based Outlier Detection

- **Task**
  - Improve data quality of dataset by applying data cleansing techniques
- **Steps**
  1. Import dataset from Local Repository (Integrating Data Result)
  2. Handling missing values
  3. Remove outliers using algorithm
  4. Change 'alcohol' attribute into categorical attribute by binning
  5. Normalize numerical attributes



# Import dataset from Local Repository (Integrating Data Result)

**1** Retrieve 'Integrating Data' dataset

Views: Design Results Turbo Prep Auto Model

Find data, operators...etc All Studio

Repository

- Business Data Mining (admin)
  - Lecture 5 Data Preparation (admin)
    - data (admin)
      - Integrating Data Result (admin)
      - Integrating Data Result2 (admin)
      - Selecting Data Result (admin)
    - process (admin)
      - Data Loading Practice 2 (admin - v1)
      - Handling Missing Data (admin - v1)

Process

Process

Retrieve Integrating Data Result2

Parameters

Retrieve Integrating Data Result2 (Retrieve...

repository entry Data Result2

Operators

Store

Data Access (1)

- Store

Utility (1)

- Process Control (1)
- Remember

Extensions (1)

- Text Processing (1)

No results were found.

Recommended Operators

Business Data Mining ©Kim Yang Sok

Help

Retrieve

RapidMiner Studio Core

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

Synopsis

This Operator can access external information in the Repository

57

# Handling missing values

Views: Design Results Turbo Prep Auto Model

Find data, operators...etc

All Studio

Result History

ExampleSet (Retrieve Selecting Data Result)

Data

Statistics

Visualizations

Annotations

Name	Type	Missing	Statistics		
type = red	Integer	0	Min 0	Max 1	Average 0.240
fixed acidity	Real	0	Min 4.800	Max 12.500	Average 7.199
volatile acidity	Real	3	Min 0.100	Max 0.960	Average 0.338
residual sugar	Real	0	Min 0	Max 1	Average 0.319
free sulfur dioxide	Real	2	Min 1	Max 112	Average 30.061
pH	Real	0	Min 2.740	Max 3.900	Average 3.222
sulphates	Real	0	Min 0.260	Max 2	Average 0.530
alcohol	Real	0	Min 8.700	Max 19.700	Average 10.170

Filter (10 / 10 attributes):

Showing attributes 1 - 10

Examples: 500 Special Attributes: 0

These two attributes contain missing values

Business Data Mining ©Kim Yang Sok

58

# Handling missing values

The screenshot displays the RapidMiner Studio interface with the following components and annotations:

- Repository:** Contains a tree view with folders like 'Business Data Mining' and 'Lecture 5 Data Preparation'. A box labeled '1' points to the 'Replace Missing Values' operator in the 'Missing' folder.
- Process:** Shows a workflow starting with 'Retrieve Integrating ...' followed by 'Replace Missing Values'. A box labeled '2' points to the 'Replace Missing Values' operator in the process canvas.
- Parameters:** On the right, the 'Replace Missing Values' operator's parameters are shown. The 'attribute filter type' is set to 'subset'. A box labeled '3' points to the 'Select Attributes' button in the 'attributes' section.
- Select Attributes: attributes dialog:** A central dialog box with two panes: 'Attributes' and 'Selected Attributes'. The 'Attributes' pane lists attributes like 'alcohol', 'chlorides', 'citric acid', 'pH', and 'quality'. The 'Selected Attributes' pane lists 'free sulfur dioxide' and 'volatile acidity'. A box labeled '4' points to the 'Attributes' list.
- Buttons:** At the bottom of the 'Selected Attributes' pane are 'Apply' and 'Cancel' buttons. A box labeled '5' points to the 'Apply' button.

Annotations and steps:

- 1 Add <<Replace Missing Values>>
- 2 Select 'subset'
- 3 Click 'Select Attributes' button
- 4 Click attributes that have missing values
- 5 Click 'Apply' button

# Handling missing values

Views: Design Results Turbo Prep Auto Model

Find data, operators...etc

All Studio

Result History

ExampleSet (Replace Missing Values)

Data

Statistics

Visualizations

Annotations

Name	Type	Missing	Statistics		
✓ volatile acidity	Real	0			
✓ free sulfur dioxide	Real	0			
✓ type = red	Integer	0	Min 0	Max 1	Average 0.240
✓ fixed acidity	Real	0	Min 4.800	Max 12.500	Average 7.199
✓ citric acid	Real	0	Min 0	Max 1	Average 0.319
✓ residual sugar	Real	0	Min 0.600	Max 18.750	Average 5.469
✓ pH	Real	0	Min 2.740	Max 3.900	Average 3.222
✓ sulphates	Real	0	Min 0.260	Max 2	Average 0.530
✓ ...	Real	0	Min 0.700	Max 13.700	Average 10.170

Showing attributes 1 - 10

Missing value has been fixed

Examples: 500 Special Attributes: 0

# Remove outliers using algorithm

The screenshot displays the RapidMiner Studio interface with a workflow designed to detect outliers. The workflow consists of three main operators: 'Retrieve Selecting D...', 'Replace Missing Val...', and 'Detect Outlier (Distances)'. The 'Detect Outlier (Distances)' operator is highlighted with a red border and a green checkmark, indicating it is the current focus. A blue arrow points from the 'Operators' panel to this operator, and another blue arrow points from the 'Parameters' panel to the 'number of outliers' field.

**Repository:** Shows the project structure with folders for 'Business Data Mining', 'Lecture 5 Data Preparation', and 'data'. The 'data' folder contains the 'Detect Outlier (Distances)' operator.

**Process:** The workflow is shown in the 'Process' tab. It starts with 'Retrieve Selecting D...', followed by 'Replace Missing Val...', and then 'Detect Outlier (Distances)'. The 'Detect Outlier (Distances)' operator is highlighted with a red border and a green checkmark.

**Parameters:** The 'Detect Outlier (Distances)' operator has the following parameters set:

- number of neighbors: 10
- number of outliers: 10
- distance function: euclidian distan...

**Operators:** The 'Operators' panel shows the 'Detect Outlier (Distances)' operator selected under the 'Outliers' category.






**Help:** The 'Help' panel provides information about the 'Detect Outlier (Distances)' operator, including its tags and synopsis.

**Annotations:**


- 1: Add <<Detect Outlier(Distance)>>
- 2: Set 'number of neighbor' as '10'
- 3: Set 'number of outliers' as '10'


**Footer:** Business Data Mining ©Kim Yang Sok


# Remove outliers using algorithm




Result History



 Data

 Statistics

 Visualizations

 Annotations

ExampleSet (Detect Outlier (Distance))

Open in  Turbo Prep  Auto Model

Filter (500 / 500 examples): all

Row No.	outlier ↓	volatile acidity	free sulfur d...	type = red	fixed acidity	citric acid	residual sug...	pH	sulphates	alcohol	quality
148	true	0.190	75	0	6.700	0.410	15.600	3.200	0.440	8.800	6
161	true	0.3					6.600	3.290	0.580	9.600	5
185	true	0.2					10	3.250	0.470	9.500	6
220	true	0.					6.700	3.280	0.350	10.200	6
319	true	0.260	110	0	6.800	0.220	4.800	3.290	0.670	10.600	5
336	true	0.260	70	0	6.800	0.240	1.900	3.180	0.520	10.500	5
352	true	0.350	86	0	6.500	0.280	12.400	3.160	0.510	9.900	6
430	true	0.220	112	0	6	0.250	11.100	3.080	0.360	9.400	6
451	true	0.240	81	0	6.900	0.400	15.400	3.200	0.690	9.400	5
484	true	0.335	69	0	4.900	0.140	1.300	3.470	0.460	10.467	5
1	false	0.580	15	1	6.700	0.080	1.800				
2	false	0.500	17	1	7.500	0.360	6.100				
3	false	0.685	22	1	6.900	0	2.500				
4	false	0.645	5	1	7.800	0	5.500				
5	false	0.380	13	1	8.100	0.280	2.100	3.230	0.730	9.700	7
6	false	0.630	50	1	7.500	0.120	5.100	3.260	0.770	9.400	5

ExampleSet (500 examples, 1 special attribute, 10 regular attributes)

Click 'outlier' button to sort examples by outlier attribute in descending order

5

A total of 10 examples are selected as outliers

Try density algorithm using <<Detect Outlier(Densities)>>

# Change 'alcohol' attribute into categorical attribute by binning




The screenshot displays the RapidMiner Studio interface with the following components and annotations:



- Repository:** Shows a project structure with folders like 'Business Data Mining' and 'Lecture 5 Data Preparation'. A blue box with a '1' and the text 'Add <<Discretize by Binning>>' points to the 'Discretize by Binning' operator in the Operators panel.
- Process:** A workflow diagram showing a sequence of operators: 'Retrieve Selecting D...', 'Replace Missing Val...', 'Detect Outlier (Dista...', and 'Discretize'. The 'Discretize' operator is highlighted with a blue box and a '2' and the text 'Select 'single''. A blue arrow points from the 'Discretize' operator in the process to its parameters panel.
- Parameters:** The 'Discretize (Discretize by Binning)' parameters panel is shown on the right. It includes:
  - 'attribute filter ty...': set to 'single'.
  - 'attribute': set to 'alcohol'.
  - 'number of bins': set to '5'.
  - 'range name type': set to 'interval'.
- Operators:** The 'Binning' category is expanded, showing various operators. 'Discretize by Binning' is highlighted.
- Help:** A 'Discretize by Binning' help panel is visible at the bottom right, showing tags like 'Continuous', 'Categorical', 'Nominal', etc.

Annotations 3, 4, and 5 are also present on the right side, pointing to the 'attribute', 'number of bins', and 'range name type' parameters respectively.


Business Data Mining ©Kim Yang Sok

# Change 'alcohol' attribute into categorical attribute by binning







Views: Design Results Turbo Prep Auto Model

Find data, operators...etc  All Studio ▼

Result History

ExampleSet (Discretize) ×

Open in  Turbo Prep  Auto Model

Filter (500 / 500 examples): all ▼

Row No.	outlier	alcohol	volatile acidity	free sulfur d...	type = red	fixed acidity	citric acid	residual sug...	pH	sulphates	quality
1	false	$[-\infty - 9.7]$	0.580	15	1	6.700	0.080	1.800	3.280	0.540	5
2	false	$[9.7 - 10.7]$	0.500	17	1	7.500	0.360	6.100	3.350	0.800	5
3	false	$[9.7 - 10.7]$	0.685	22	1	6.900	0	2.500	3.460	0.570	6
4	false	$[-\infty - 9.7]$	0.645	5	1	7.800	0	5.500	3.400	0.550	6
5	false	$[-\infty - 9.7]$	0.380	13	1	8.100	0.280	2.100	3.230	0.730	7
6	false	$[-\infty - 9.7]$	0.630	50	1	7.500	0.120	5.100	3.260	0.770	5
7	false	$[-\infty - 9.7]$	0.630	15	1	7.700	0.080	1.900	3.320	0.540	6
8	false	$[9.7 - 10.7]$	0.550	19	1	6.900	0.150	2.200	3.410	0.590	5
9	false	$[-\infty - 9.7]$	0.620	8	1	7	0.080	1.800	3.480	0.530	5
10	false	$[-\infty - 9.7]$	0.690	22	1	7	0.080	1.800	3.340	0.540	6
11	false	$[-\infty - 9.7]$	0.620	27	1	9	0.040	1.900	3.160	0.700	5
12	false	$[9.7 - 10.7]$	0.330	5	1	7.300	0.470	2.100	3.330	0.530	6
13	false	$[-\infty - 9.7]$	0.520	32	1	9.200	1	3.400	2.740	2	4
14	false	$[-\infty - 9.7]$	0.630	14	1	7.800	0.480	1.700	3.190	0.620	5
15	false	$[9.7 - 10.7]$	0.640	18	1	6.800	0.100	2.100	3.340	0.520	5
16	false	$[-\infty - 9.7]$	0.610	10	1	8.800	0.140	2.400	3.190	0.590	5

ExampleSet (500 examples, 1 special attribute, 10 regular attributes)

Business Data Mining ©Kim Yang Sok

64



# Normalize numerical attributes

**Repository**

- Business Data Mining (admin)
- Lecture 5 Data Preparation (admin)
  - data (admin)
    - Integrating Data Result (admin)
    - Integrating Data Result2 (admin)
    - Selecting Data Result (admin)
  - process (admin)

**Process**

Process

Retrieve Selecting D... Replace Missing Val... Detect Outlier (Dista... Discretize Normalize

**Parameters**

Normalize

- ☐ create view
- attribute filter type ☒ subset
- attributes
- ☐ invert selection
- ☐ include special attributes
- method ☒ Z-transformation

**Operators**

Nor

- Cleansing (3)
  - Normalization (3)
    - Normalize
    - De-Normalize
    - Scale by Weights
- Modeling (1)
  - Time Series (1)

**Select Attributes: attributes**

The attribute which should be chosen.

**Attributes**

- # citric acid
- # fixed acidity
- # free sulfur dioxide
- # pH
- # residual sugar
- # sulphates

**Selected Attributes**

- # quality
- # type = red

**Normalize**

RapidMiner Studio Core

Tags: [Normalization](#), [Standardize](#), [Z-Transform](#), [Scaling](#), [Features](#), [Attributes](#), [Variables](#), [Color](#)

**Synopsis**

This Operator normalizes the values of the

**Annotations:**

- 1 Add <<Normalize>>
- 2 Select 'subset'
- 3 Click 'Select Attributes' button.
- 4 Select 'quality' and 'type=red' attributes

We will apply normalization except these two attributes. Why we need to exclude them?

# Normalize numerical attributes

The screenshot displays the RapidMiner Studio interface with a workflow in the 'Process' view. The workflow consists of the following operators: Retrieve Selecting D..., Replace Missing Val..., Detect Outlier (Dista..., Discretize, and Normalize. The 'Normalize' operator is highlighted with an orange border. A callout box labeled '1' points to the 'Normalize' operator in the process canvas, containing the text: 'Check 'invert selection' and 'include special attributes''. Another callout box labeled '2' points to the 'Parameters' panel for the 'Normalize' operator, containing the text: 'Set 'method' as 'range transformation''. The 'Parameters' panel shows the following settings: 'create view' is unchecked; 'attribute filter type' is set to 'subset'; 'attributes' is set to 'Select Attribut...'; 'invert selection' is checked; 'include special attributes' is checked; 'method' is set to 'range transfor...'; 'min' is set to '0.0'. The 'Repository' panel on the left shows a tree structure with 'Business Data Mining' > 'Lecture 5 Data Preparation' > 'data' > 'Integrating Data Result' > 'Integrating Data Result2' > 'Selecting Data Result'. The 'Operators' panel on the left shows a search for 'Nor' with results under 'Cleansing' > 'Normalization' > 'Normalize', 'De-Normalize', and 'Scale by Weights'. The 'Help' panel on the right shows the 'Normalize' operator's description: 'RapidMiner Studio Core', 'Tags: Normalization, Standardize, Z-Transform, Scaling, Features, Attributes, Variables, Color', and 'Synopsis: This Operator normalizes the values of the'.

Views: Design Results Turbo Prep Auto Model

Find data, operators...etc

All Studio

Repository

Import Data

Business Data Mining (admin)

Lecture 5 Data Preparation (admin)

data (admin)

Integrating Data Result (admin)

Integrating Data Result2 (admin)

Selecting Data Result (admin)

process (admin)

Data Loading Practice 2 (admin - v1)

Handling Missing Data (admin - v1)

Operators

Nor

Cleansing (3)

Normalization (3)

Normalize

De-Normalize

Scale by Weights

Modeling (1)

Time Series (1)

No results were found.

Process

Process

1

Check 'invert selection' and 'include special attributes'

2

Set 'method' as 'range transformation'

Parameters

Normalize

create view

attribute filter type subset

attributes Select Attribut...

invert selection

include special attributes

method range transfor...

min 0.0

Hide advanced parameters

Change compatibility (9.3.001)

Help

Normalize






RapidMiner Studio Core

Tags: Normalization, Standardize, Z-Transform, Scaling, Features, Attributes, Variables, Color


Synopsis


This Operator normalizes the values of the


# Normalize numerical attributes





Views: Design Results Turbo Prep Auto Model



Find data, operators...etc 


All Studio 

Result History ExampleSet (Discretize) 

  
Data


  
Statistics

Open in  Turbo Prep  Auto Model






Filter (500 / 500 examples): all 

Row No.	outlier	alcohol	volatile acidity	free sulfur d...	type = red	fixed acidity	citric acid	residual sug...	pH	sulphates	quality
1	false	$[-\infty - 9.7]$	0.580	15	1	6.700	0.080				
2	false	$[9.7 - 10.7]$	0.500	17	1	7.500	0.360				
3	false	$[9.7 - 10.7]$	0.685	22	1	6.900	0				
4	false	$[-\infty - 9.7]$	0.645	5	1	7.800	0	5.500	3.400	0.550	6

Before normalization


 //Local Repository/Business Data Mining/Lecture 5 Data Preparation/process/Cleansing Data3\* - RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ


File Edit Process View Connections Settings Extensions Help





Views: Design Results Turbo Prep Auto Model



Try other normalization algorithms by selecting different value for 'method' parameter

Result History ExampleSet (Normalize) 

  
Data

  
Statistics

  
Visualizations

Open in  Turbo Prep  Auto Model

Row No.	outlier	volatile acidity	free sulfur d...	fixed acidity	citric acid	residual sug...	pH	sulphates	alcohol	type = red	quality
1	false	0.558	0.126	0.247	0.080	0.066	3.400	0.161	$[-\infty - 9.7]$	1	5
2	false	0.465	0.144	0.351	0.360	0.303	3.400	0.161	$[-\infty - 9.7]$	1	5
3	false	0.680	0.189	0.273	0	0.110	3.400	0.161	$[-\infty - 9.7]$	1	6
4	false	0.634	0.036	0.390	0	0.270	3.400	0.161	$[-\infty - 9.7]$	1	6
5	false	0.326	0.108	0.429	0.280	0.083	3.400	0.161	$[-\infty - 9.7]$	1	7
6	false	0.616	0.441	0.351	0.120	0.248	3.400	0.161	$[-\infty - 9.7]$	1	5
7	false	0.616	0.126	0.377	0.080	0.072	3.400	0.161	$[-\infty - 9.7]$	1	6

After normalization

# Conclusion

- Various techniques can be applied to data preparation phase such as data integration, data selection, and data cleansing.
- Correct data preparation is essential in data mining in order to get better performance and stable analysis results. If you have data problems in modeling phase, you can return this phase to resolve the problems.
- Now you have prepared dataset, you can go to next phase of CRISP-DM, the Modeling phase.



# QUESTIONS?