



# Limpieza de datos

Limpieza de datos

Cindy López

# Preparar la data

Process

Process

100%

Process

Retrieve Titanic

inp out

Retrieve Titanic.output (output)  
Meta data: Data Table  
● Source: //Tutorials/99 Data/Titanic

Number of examples = 1309  
12 attributes:  
**Note:** Some of the nominal values in this set were discarded due to performance reasons. You can change this behaviour in the preferences (`rapidminer.general.md_nominal_values_limit`).  
Generated by: [Retrieve Titanic.output](#)

Role	Name	Type	Range	Missings	Comment
	No of Parents...	integer	= [0 - 9]	= 0	
	Ticket Number	polynomial	≥ [110152, 11...	= 0	
	Passenger F...	numeric	= [0 - 512.329]	= 1	
	Cabin	polynomial	≥ [A10, A11, A...	= 1014	
	Port of Embar...	polynomial	= [Cherbourg, ...	= 2	
	Life Boat	polynomial	= [1, 10, 11, 12...	= 823	
	Survived	binominal	= [No, Yes]	= 0	

# Preparar la data – Seleccionar atributos

The screenshot displays the RapidMiner Studio interface in the 'Design' view. The main process canvas shows a workflow starting with 'Retrieve Titanic', followed by 'Select Attributes' (highlighted with an orange border), and then 'Replace Missing Values' and 'Filter Examples'. The 'Select Attributes' operator is currently selected, and its configuration dialog is open.

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

**Attributes**

Search
Cabin
Life Boat

**Selected Attributes**

Search
Age
Name
No of Parents or Children on Board
No of Siblings or Spouses on Board
Passenger Class
Passenger Fare
Port of Embarkation
Sex
Survived
Ticket Number

**Parameters**

**Select Attributes**

attribute filter... subset

attributes Select ...

☐ invert selection

☐ include special attribute:

**Help**

**Select Attributes**  
RapidMiner Studio C

Tags: [Filter](#), [Keep](#), [Remove](#), [Columns](#), [Variables](#), [Feature Selection](#)

**Synopsis**  
This Operator selects a sub

# Preparar la data – Verificar atributos descartados

The screenshot shows a software interface with a sidebar on the left containing icons for Data, Statistics, Charts, Advanced Charts, and Annotations. The main area displays a table of attribute statistics. A red box highlights the 'Missing' column. The table has columns for Name, Type, Missing, Statistics, and Filter (10 / 10 attributes). The data rows are as follows:

Name	Type	Missing	Statistics	Filter (10 / 10 attributes):
Passenger Class	Polynomial	0	Least Second (277) Most Third (709)	Search for Attributes
Name	Polynomial	0	Least van Melk [...] lemon (1) Most Connolly, Miss. Kate (2)	
Sex	Binominal	0	Least Female (466) Most Male (843)	
Age	Real	263	Min 0.167 Max 80	
No of Siblings or Spouses on B...	Integer	0	Min 0 Max 8	
No of Parents or Children on B...	Integer	0	Min 0 Max 9	
Ticket Number	Polynomial	0	Least W/C 14208 (1) Most CA. 2343 (11)	
Passenger Fare	Numeric	1	Min 0 Max 512.329	
Port of Embarkation	Polynomial	2	Least Queenstown (123) Most Southampton (914)	

Showing attributes 1 - 10 Examples: 1,309 Special Attributes: 0 Regular Attributes: 10

# Añadir operador reemplazar valores descartados

The screenshot displays the Orange3 data mining software interface. The main window shows a workflow with four operators: 'Retrieve Titanic', 'Select Attributes', 'Replace Missing Values', and 'Filter Examples'. The 'Replace Missing Values' operator is highlighted with an orange border. A 'Parameters' panel on the right shows the configuration for this operator: 'attribute filter' is set to 'single' and 'attribute' is set to 'Age'. Below the main window, the 'Result History' panel shows the output of the 'Replace Missing Values' operator. It lists various attributes with their types, missing values, and statistics. The 'Age' attribute is highlighted with a red box, showing 0 missing values. The 'Passenger Fare' attribute is also highlighted, showing 1 missing value and a histogram.

**Repository**

- Market-Data (v1)
- Polynomial (v1)
- Products (v1)
- Purchases (v1)
- Ripley-Set (v1)
- Sonar (v1)
- Titanic (v1)
- Titanic Training (v1)
- Titanic Unlabeled (v1)
- Transactions (v1)

**Process**

Process

inp → Retrieve Titanic → Select Attributes → Replace Missing Values → Filter Examples → res

**Parameters**

Replace Missing Values

attribute filter... single

attribute Age

☐ invert selection

☐ include special attribute:

**Operators**

replace

- Blending (3)
- Values (3)
  - Map
  - Replace
  - Replace (Dictionary)
- Cleansing (3)
- Missing (3)
  - Replace Missing Values

No results were found.

**Tutorials**

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

Activate Wisdom of Crowds

**Result History**

Name	Type	Missing	Statistics	Filter (10 / 10 attributes):
Age	Real	0	Min 0.167, Max 80, Average 29.881	
Passenger Class	Polynomial	0	Least Second (277), Most Third (709), Values Third (709)	
Name	Polynomial	0	Least van Melk [...] lemon (1), Most Connolly, Miss. Kate (2), Values Connolly, I	
Sex	Binominal	0	Least Female (466), Most Male (843), Values Male (843)	
No of Siblings or Spouses on B...	Integer	0	Min 0, Max 8, Average 0.499	
No of Parents or Children on B...	Integer	0	Min 0, Max 9, Average 0.385	
Ticket Number	Polynomial	0	Least W/C 14208 (1), Most CA. 2343 (11), Values CA. 2343	
Passenger Fare	Numeric	1	Min 0, Max 512.329	

Open chart

# Filtrar valores descartados

Views: Design Results Auto Model

Find data, operators...etc All Studio ▼

Process

Process

100%

Retrieve Titanic

Select Attributes

Replace Missing Val...

Filter Examples

Parameters

Filter Examples

condition class no\_missing\_attri... ⓘ

☐ invert filter ⓘ

[Hide advanced parameters](#)

[Change compatibility \(8.2.000\)](#)

Name	Type	Missing	Statistics	Filter (10 / 10 attributes)
Age	Real	0	Min 0.167 Max 80 Average 29.827	
Passenger Class	Polynomial	0	Least Second (277) Most Third (708) Values Third (708)	
Name	Polynomial	0	Least Storey, Mr. Thomas (0) Most Connolly, Miss. Kate (2) Values Connolly, Miss. Kate (2)	
Sex	Binomial	0	Least Female (464) Most Male (842) Values Male (842)	
No of Siblings or Spouses on Board	Integer	0	Min 0 Max 8 Average 0.500	
No of Parents or Children on Board	Integer	0	Min 0 Max 9 Average 0.386	
Ticket Number	Polynomial	0	Least 3701 (0) Most CA 2343 (11) Values CA 2343 (11)	
Passenger Fare	Numeric	0	Min 0 Max 512.329 Average 33.224	
Port of Embarkation	Polynomial	0	Least Queenstown (123) Most Southampton (913) Values Southampton (913)	



# Identificar anomalías (Valores atípicos)

Ejemplo: Identificar transacciones fraudulentas de tarjetas de crédito o mediciones incorrectas

The screenshot displays the RapidMiner Studio interface. At the top, a workflow is visible in the 'Process' pane, consisting of the following steps: 'Retrieve Titanic', 'Select Attributes', 'Normalize', 'Detect Outliers (Distances)', and 'Filter Examples'. The 'Select Attributes' step is currently selected.

Below the workflow, the 'Select Attributes: attributes' dialog box is open. It contains two lists of attributes:

- Attributes:** Cabin, Life Boat, Name, Ticket Number (highlighted).
- Selected Attributes:** Age, No of Parents or Children on Board, No of Siblings or Spouses on Board, Passenger Class, Passenger Fare, Port of Embarkation, Sex, Survived.

At the bottom of the dialog are 'Apply' and 'Cancel' buttons.

On the right side of the interface, the 'Parameters' pane for the 'Select Attributes' step is visible. It shows the 'attribute filter type' set to 'subset' and a 'Select Attributes...' button. Below this, there are checkboxes for 'invert selection' and 'include special attributes', both of which are currently unchecked.

At the bottom right, a 'Help' pane is open, displaying the title 'Select Attributes' and the subtitle 'RapidMiner Studio Core'. It also includes a list of tags: Filter, Keep, Remove, Drop, Delete, Columns, Variables, Features, Feature Set, Selection.

# Normalización

Views: Design Results Auto Model

Find data, operators...etc All Studio

**Repository**

- Market-Data (v1)
- Polynomial (v1)
- Products (v1)
- Purchases (v1)
- Ripley-Set (v1)
- Sonar (v1)
- Titanic (v1)**
- Titanic Training (v1)
- Titanic Unlabeled (v1)
- Transactions (v1)

**Operators**

- Cleansing (3)
  - Normalization (3)
    - Normalize**
    - De-Normalize
    - Scale by Weights

**Process**

Process

100%

inp

Retrieve Titanic

out

exa

ori

exa

ori

pre

**Normalize**

res

res

Filter Examples

**Parameters**

**Normalize**

- ☐ create view
- attribute filter type: all
- ☐ invert selection
- ☐ include special attributes
- method: Z-transformation

[Hide advanced parameters](#)

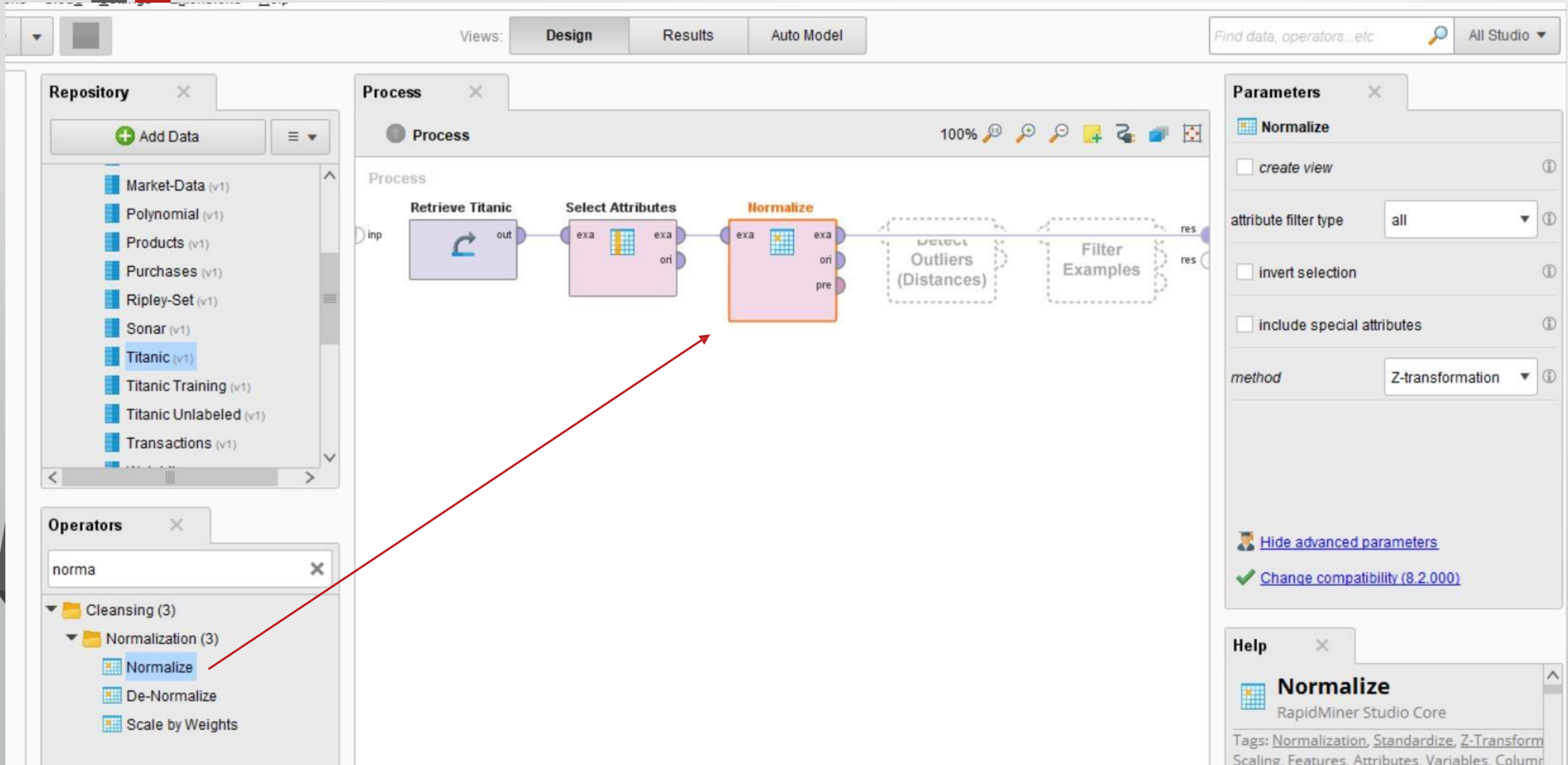
[Change compatibility \(8.2.000\)](#)

**Help**

**Normalize**

RapidMiner Studio Core

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





# Detectar valores atípicos

The screenshot displays the RapidMiner Studio interface with a workflow designed for outlier detection. The workflow consists of the following steps:

- Retrieve Titanic**: Retrieves the Titanic dataset.
- Select Attributes**: Selects specific attributes from the dataset.
- Normalize**: Normalizes the selected attributes.
- Detect Outlier (Distances)**: Detects outliers using the distances method. This operator is highlighted with an orange border.
- Filter Examples**: Filters the dataset based on the detected outliers.

The **Parameters** panel for the **Detect Outlier (Distances)** operator is visible on the right, showing the following settings:

- number of neighbors**: 10
- number of outliers**: 10
- distance function**: euclidian distance

The **Repository** panel on the left shows the **Titanic** dataset selected. The **Operators** panel at the bottom left shows the search results for "detect", with **Detect Outlier (Distances)** highlighted. The **Help** panel at the bottom right provides additional information about the **Detect Outlier (Distances)** operator.

# Filtrar

The screenshot displays the Orange3 data mining environment. The main workflow in the 'Process' window includes the following steps: Retrieve Titanic, Select Attributes, Normalize, Detect Outlier (Distance), and Filter Examples. The 'Filter Examples' widget is highlighted with a red box. A callout box points to the 'outlier' filter settings, indicating that 10 values were eliminated.

**Create Filters: filters**  
Defines the list of filters to apply.

outlier equals false

☒ Match all ☐ Match any ☒ Preselect comparators

**Result History**

**ExampleSet (Filter Examples)**

ExampleSet (1299 examples, 1 special attribute, 8 regular attributes)

Row No.	outlier	Age	No of Sibling...	No of Parent...	Passenger F...	Pass
1	false	-0.061	-0.479	-0.445	3.440	First
2	false	-2.010	0.481	1.866	2.285	First
3	false	-1.934	0.481	1.866	2.285	First
4	false	0.008	0.481	1.866	2.285	First
5	false	-0.339	0.481	1.866	2.285	First
6	false	1.257	-0.479	-0.445	-0.130	First
7	false	2.298	0.481	-0.445	0.863	First
8	false	0.633	-0.479	-0.445	-0.643	First
9	false	1.604	1.441	-0.445	0.351	First
10	false	2.853	-0.479	-0.445	0.313	First

# Ejercicio

- ¿Cómo cambiaría el proceso para que encuentre 20 valores atípicos en lugar de 10?
- Reemplace el operador de detección de valores atípicos con **Detect Outlier (LOF)** e identifique la diferencia.
- ¿Cómo cambiar el filtro para mantener solo los valores atípicos superiores?
- Consultar en que consiste la normalización.
- Consultar los diferentes métodos de detección de valores atípicos.