



Estudiante: Díaz Padilla Danny Sebastián

Tema: Trabajo en clase y Taller

## Trabajo en clase

### 1. Reemplazamiento de valores 'perdidos' en la base de datos de Titanic.

Process

100%

Process

Retrieve Titanic

Select Attributes

Drag here

Parameters

Process

logverbosity: init

logfile

resultfile

random seed: 2001

send mail: never

encoding: SYSTEM

[Hide advanced parameters](#)

[Change compatibility \(9.4.001\)](#)

<new process> - RapidMiner Studio Educational 9.4.001 @ DESKTOP-U6QA500

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments

Find data, operators, etc.

Result History

ExampleSet (Select Attributes)

Open in Turbo Prep Auto Model

Filter (1,309 / 1,309 examples): all

Row No.	Passenger ...	Name	Sex	Age	No of Sibling...	No of Parent...	Ticket Num...	Passenger F...	Port of Emb...	Survived
1	First	Allen, Miss. E...	Female	29	0	0	24160	211.338	Southampton	Yes
2	First	Allison, Mast...	Male	0.917	1	2	113781	151.550	Southampton	Yes
3	First	Allison, Miss. ...	Female	2	1	2	113781	151.550	Southampton	No
4	First	Allison, Mr. H...	Male	30	1	2	113781	151.550	Southampton	No
5	First	Allison, Mrs. ...	Female	25	1	2	113781	151.550	Southampton	No
6	First	Anderson, Mr...	Male	48	0	0	19952	26.550	Southampton	Yes
7	First	Andrews, Mis...	Female	63	1	0	13502	77.958	Southampton	Yes
8	First	Andrews, Mr. ...	Male	39	0	0	112050	0	Southampton	No
9	First	Appleton, Mrs...	Female	53	2	0	11789	51.479	Southampton	Yes
10	First	Artagaveytia, ...	Male	71	0	0	PC 17609	49.504	Cherbourg	No
11	First	Astor, Col. Jo...	Male	47	1	0	PC 17757	227.525	Cherbourg	No
12	First	Astor, Mrs. Jo...	Female	18	1	0	PC 17757	227.525	Cherbourg	Yes
13	First	Aubart, Mme. ...	Female	24	0	0	PC 17477	69.300	Cherbourg	Yes
14	First	Barber, Miss. ...	Female	26	0	0	19877	78.850	Southampton	Yes
15	First	Barkworth, Mr...	Male	80	0	0	27042	30	Southampton	Yes

<new process> - RapidMiner Studio Educational 9.4.001 @ DESKTOP-U6QA500

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Deployments

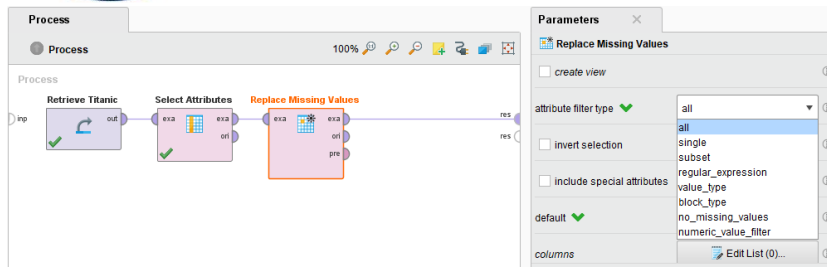
Find data, operators, etc.

Result History

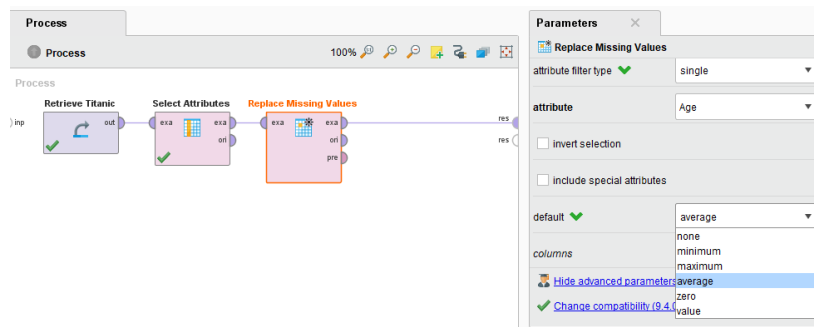
ExampleSet (Select Attributes)

Name Type Missing Statistics Filter (10 / 10 attributes): Search for Attributes

Name	Type	Missing	Statistics	Values
Passenger Class	Polynomial	0	Least: Second (277) Most: Third (709)	Third (709), First (323), ...[1 more]
Name	Polynomial	0	Least: van Melik [...] lemon (1) Most: Connolly, Miss. Kate (2)	Connolly, Miss. Kate (2), Kelly, Mr. James (1), ...[1 more]
Sex	Binomial	0	Least: Female (466) Most: Male (843)	Male (843), Female (466)
Age	Real	263	Min: 0.167 Max: 80	Average: 29.881
No of Siblings or Spouses on Board	Integer	0	Min: 0 Max: 8	Average: 0.499
No of Parents or Children on Board	Integer	0	Min: 0 Max: 9	Average: 0.385
Ticket Number	Polynomial	0	Least: W/C 14208 (1) Most: CA. 2343 (11)	CA. 2343 (11), 1601 (8), ...[927 more]

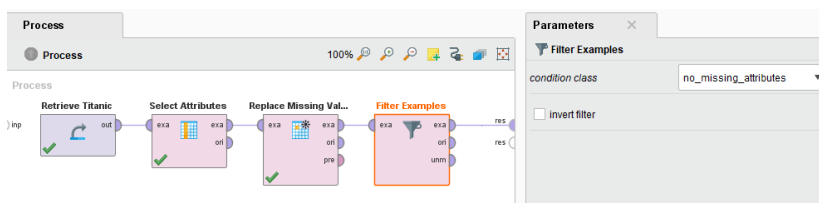


Reemplaza con el promedio de las edades



Name	Type	Missing
Age	Real	0
Passenger Class	Polynomial	0
Name	Polynomial	0
Sex	Binomial	0
No of Siblings or Spouses on B...	Integer	0
No of Parents or Children on B...	Integer	0
Ticket Number	Polynomial	0
Survived	Binomial	1

## 2. Filtrar los elementos con atributos vacíos



Name	Type	Missing
Age	Real	0
Passenger Class	Polynomial	0
Name	Polynomial	0
Sex	Binomial	0
No of Siblings or Spouses on B...	Integer	0
No of Parents or Children on B...	Integer	0
Ticket Number	Polynomial	0
Survived	Binomial	0



### 3. Detectar valores atípicos en la base de datos de Titanic

Process flow: Retrieve Titanic → Select Attributes → Normalize → Detect Outlier (Distances) → Filter Examples.

Filter Examples Parameters:

- filters: Add Filters...
- condition class: custom\_filters
- invert filter: ☐

Create Filters: filters dialog:

- Create Filters: filters
- Defines the list of filters to apply.
- outlier equals false

Result History: ExampleSet (Filter Examples)

Row No.	outlier	Age	No of Sibling...	No of Parent...	Passenger F...	Passenger ...	Sex	Port of Emb...	Survived
1	false	-0.061	-0.479	-0.445	3.440	First	Female	Southampton	Yes
2	false	-2.010	0.481	1.866	2.285	First	Male	Southampton	Yes
3	false	-1.934	0.481	1.866	2.285	First	Female	Southampton	No
4	false	0.008	0.481	1.866	2.285	First	Male	Southampton	No
5	false	-0.339	0.481	1.866	2.285	First	Female	Southampton	No
6	false	1.257	-0.479	-0.445	-0.130	First	Male	Southampton	Yes
7	false	2.298	0.481	-0.445	0.863	First	Female	Southampton	Yes
8	false	0.633	-0.479	-0.445	-0.643	First	Male	Southampton	No
9	false	1.604	1.441	-0.445	0.351	First	Female	Southampton	Yes
10	false	2.853	-0.479	-0.445	0.313	First	Male	Cherbourg	No
11	false	1.188	0.481	-0.445	3.753	First	Male	Cherbourg	No
12	false	-0.824	0.481	-0.445	3.753	First	Female	Cherbourg	Yes
13	false	-0.408	-0.479	-0.445	0.696	First	Female	Cherbourg	Yes
14	false	-0.269	-0.479	-0.445	0.880	First	Female	Southampton	Yes
15	false	3.477	-0.479	-0.445	-0.064	First	Male	Southampton	Yes

## Ejercicio

1. ¿Cómo cambiaría el proceso para que encuentre 20 valores atípicos en lugar de 10?

Detect Outlier (Distances) Parameters:

- number of neighbors: 10
- number of outliers: 20
- distance function: euclidian distance

Result History: ExampleSet (Filter Examples)

Filter (1,289 / 1,289 examples)

Row No.	outlier	Age	No of Sibling...	No of Parent...	Passenger F...	Passenger ...	Sex	Port of Emb...	Survived
1	false	-0.061	-0.479	-0.445	3.440	First	Female	Southampton	Yes
2	false	-2.010	0.481	1.866	2.285	First	Male	Southampton	Yes
3	false	-1.934	0.481	1.866	2.285	First	Female	Southampton	No
4	false	0.008	0.481	1.866	2.285	First	Male	Southampton	No
5	false	-0.339	0.481	1.866	2.285	First	Female	Southampton	No
6	false	1.257	-0.479	-0.445	-0.130	First	Male	Southampton	Yes
7	false	2.298	0.481	-0.445	0.863	First	Female	Southampton	Yes
8	false	0.633	-0.479	-0.445	-0.643	First	Male	Southampton	No
9	false	1.604	1.441	-0.445	0.351	First	Female	Southampton	Yes
10	false	2.853	-0.479	-0.445	0.313	First	Male	Cherbourg	No
11	false	1.188	0.481	-0.445	3.753	First	Male	Cherbourg	No
12	false	-0.824	0.481	-0.445	3.753	First	Female	Cherbourg	Yes
13	false	-0.408	-0.479	-0.445	0.696	First	Female	Cherbourg	Yes
14	false	-0.269	-0.479	-0.445	0.880	First	Female	Southampton	Yes
15	false	3.477	-0.479	-0.445	-0.064	First	Male	Southampton	Yes



## 2. Obtener la lista de valores atípicos encontrados.

Create Filters: filters

×



Create Filters: **filters**

Defines the list of filters to apply.

outlier

equals

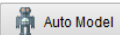
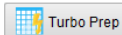
true



ExampleSet (Filter Examples)

×

Open in

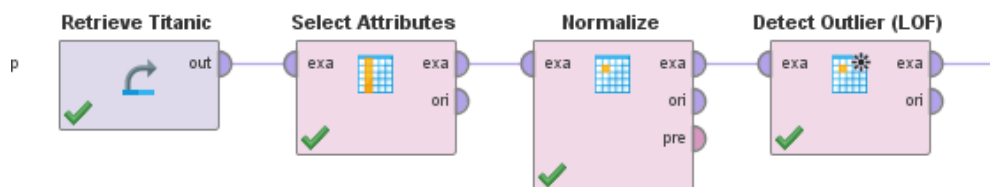


Filter (20 / 20 examples):

Row No.	outlier	Age	No of Sibling...	No of Parent...	Passenger F...	Passenger ...	Sex	Port of Emb...	Survived
2	true	0.425	-0.479	0.710	9.255	First	Male	Cherbourg	Yes
3	true	1.951	-0.479	0.710	9.255	First	Female	Cherbourg	Yes
4	true	-0.408	2.401	1.866	4.438	First	Female	Southampton	Yes
5	true	-0.131	2.401	1.866	4.438	First	Female	Southampton	Yes
6	true	-0.477	2.401	1.866	4.438	First	Female	Southampton	Yes
7	true	-0.755	2.401	1.866	4.438	First	Male	Southampton	No
8	true	2.367	0.481	4.176	4.438	First	Male	Southampton	No
9	true	2.090	0.481	4.176	4.438	First	Female	Southampton	Yes
10	true	0.355	-0.479	-0.445	9.255	First	Male	Cherbourg	Yes
11	true	-1.171	1.441	1.866	4.426	First	Male	Cherbourg	Yes
12	true	-0.824	1.441	1.866	4.426	First	Female	Cherbourg	Yes
13	true	-0.616	1.441	1.866	4.426	First	Female	Cherbourg	Yes
14	true	2.159	0.481	3.021	4.426	First	Male	Cherbourg	No
15	true	1.257	0.481	3.021	4.426	First	Female	Cherbourg	Yes
16	true	2.575	0.481	-0.445	3.642	First	Male	Southampton	No

## 3. Reemplace el operador de detección de valores atípicos con Detect Outlier (LOF) e identifique la diferencia.

Process





ExampleSet (Detect Outlier (LOF))									
Open in		Turbo Prep	Auto Model	Filter (1,309 / 1,309 examples):					
Row No.	outlier	Age	No of Sibling...	No of Parent...	Passenger F...	Passenger ...	Sex	Port of Emb...	Survived
1	1.030	-0.061	-0.479	-0.445	3.440	First	Female	Southampton	Yes
2	1.337	-2.010	0.481	1.866	2.285	First	Male	Southampton	Yes
3	1.364	-1.934	0.481	1.866	2.285	First	Female	Southampton	No
4	1.167	0.008	0.481	1.866	2.285	First	Male	Southampton	No
5	1.180	-0.339	0.481	1.866	2.285	First	Female	Southampton	No
6	1.356	1.257	-0.479	-0.445	-0.130	First	Male	Southampton	Yes
7	1.265	2.298	0.481	-0.445	0.863	First	Female	Southampton	Yes
8	1.614	0.633	-0.479	-0.445	-0.643	First	Male	Southampton	No
9	1.267	1.604	1.441	-0.445	0.351	First	Female	Southampton	Yes
10	2.307	2.853	-0.479	-0.445	0.313	First	Male	Cherbourg	No
11	1.180	1.188	0.481	-0.445	3.753	First	Male	Cherbourg	No
12	1.342	-0.824	0.481	-0.445	3.753	First	Female	Cherbourg	Yes
13	0.925	-0.408	-0.479	-0.445	0.696	First	Female	Cherbourg	Yes
14	1.056	-0.269	-0.479	-0.445	0.880	First	Female	Southampton	Yes
15	3.298	3.477	-0.479	-0.445	-0.064	First	Male	Southampton	Yes

La principal diferencia es que 'Detect Outlier (Distances)' otorga un booleano indicando si sobrepasó o no una distancia en función de sus 'vecinos' más cercanos mientras que 'Detect Outlier (LOF)' otorga un valor numérico que significa la densidad respecto a sus vecinos más cercanos.

#### 4. ¿Cómo cambiar el filtro para mantener solo los valores atípicos superiores?

Create Filters: filters

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

outlier > 1.0

ExampleSet (Filter Examples (2))									
Open in		Turbo Prep	Auto Model	Filter (931 / 931 examples):					
Row No.	outlier	Age	No of Sibling...	No of Parent...	Passenger F...	Passenger ...	Sex	Port of Emb...	Survived
1	1.030	-0.061	-0.479	-0.445	3.440	First	Female	Southampton	Yes
2	1.337	-2.010	0.481	1.866	2.285	First	Male	Southampton	Yes
3	1.364	-1.934	0.481	1.866	2.285	First	Female	Southampton	No
4	1.167	0.008	0.481	1.866	2.285	First	Male	Southampton	No
5	1.180	-0.339	0.481	1.866	2.285	First	Female	Southampton	No
6	1.356	1.257	-0.479	-0.445	-0.130	First	Male	Southampton	Yes
7	1.265	2.298	0.481	-0.445	0.863	First	Female	Southampton	Yes
8	1.614	0.633	-0.479	-0.445	-0.643	First	Male	Southampton	No
9	1.267	1.604	1.441	-0.445	0.351	First	Female	Southampton	Yes
10	2.307	2.853	-0.479	-0.445	0.313	First	Male	Cherbourg	No
11	1.180	1.188	0.481	-0.445	3.753	First	Male	Cherbourg	No
12	1.342	-0.824	0.481	-0.445	3.753	First	Female	Cherbourg	Yes
13	1.056	-0.269	-0.479	-0.445	0.880	First	Female	Southampton	Yes
14	3.298	3.477	-0.479	-0.445	-0.064	First	Male	Southampton	Yes
15	1.299	-0.408	-0.479	0.710	4.139	First	Male	Cherbourg	No

ExampleSet (931 examples, 1 special attribute & 8 regular attributes)



## 5. Consultar en que consiste la normalización.

La normalización se usa para escalar valores para que quepan en un rango específico. Ajustar el rango de valores es muy importante cuando se trata de atributos de diferentes unidades y escalas. Por ejemplo, cuando se usa la distancia euclidiana, todos los atributos deben tener la misma escala para una comparación justa. La normalización es útil para comparar atributos que varían en tamaño. Este operador realiza la normalización de los atributos seleccionados. Se proporcionan cuatro métodos de normalización. Estos métodos se explican en los parámetros.

## 6. Consultar los diferentes métodos de detección de valores atípicos.

En la documentación oficial (<https://docs.rapidminer.com/latest/studio/operators/>) se encuentran 4 operadores en la sección Cleansing -> Outliers.

### 1. Detect Outlier (COF)

Este operador identifica valores atípicos en el conjunto de ejemplos dado en función de los factores de valor atípico de clase (COF).

$$COF = PCL(T, K) - norma(desviación(T)) + norma(kDist(T))$$

**PCL (T, K)** es la probabilidad de la etiqueta de clase de la instancia T con respecto a las etiquetas de clase de sus K vecinos más cercanos.

**norma (Desviación (T))** y **norma (KDist (T))** son los valores normalizados de Desviación (T) y KDist (T) respectivamente y sus valores caen en el rango [0 - 1].

**La desviación (T)** es cuánto se desvía la instancia T de las instancias de la misma clase. Se calcula sumando las distancias entre la instancia T y cada instancia que pertenece a la misma clase.

**KDist (T)** es la suma de la distancia entre la instancia T y sus K vecinos más cercanos.

### 2. Detect Outlier (LOF)

Este operador identifica valores atípicos en el conjunto de ejemplos dado en función de factores locales atípicos (LOF). El LOF se basa en un concepto de densidad local, donde la localidad está dada por los k vecinos más cercanos, cuya distancia se usa para estimar la densidad. Al comparar la densidad local de un objeto con las densidades locales de sus vecinos, uno puede identificar regiones de densidad similar y puntos que tienen una densidad sustancialmente menor que sus vecinos. Se consideran valores atípicos.

### 3. Detect Outlier (Distancias)

Este operador identifica n valores atípicos en el conjunto de ejemplos dado en función de la distancia a sus k vecinos más cercanos. Las variables n y k pueden especificarse a través de parámetros.

### 4. Detect Outlier (Densidades)

Este operador identifica valores atípicos en el conjunto de ejemplos dado en función de la densidad de datos. Todos los objetos que tienen al menos una proporción de todos los objetos más alejados que la distancia D se consideran valores atípicos.