ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS
INGENIERÍA DE SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

**NOMBRE DE ESTUDIANTE:** Danny Sebastián Díaz Padilla

**Laboratorio de:**

ANALÍTICA DE DATOS – BIG DATA

**Práctica No.:** 7

**Tema:** Limpieza de datos

## Objetivos:

- Reemplazar valores 'perdidos' dentro de un conjunto de datos.
- Filtrar datos que no posean valores 'perdidos'.
- Utilizar clasificadores de valores atípicos.

## Marco teórico:

### Valor perdido

Se denomina valor perdido a aquel valor en un atributo que no posee información y tiene un valor por lo general 'null', es decir, se desconoce.

### Valor atípico

Un valor atípico es una observación extrañamente grande o pequeña. Los valores atípicos pueden tener un efecto desproporcionado en los resultados estadísticos, como la media, lo que puede conducir a interpretaciones engañosas. Por ejemplo, un conjunto de datos incluye los valores: 1, 2, 3, y 34. El valor medio, 10, que es mayor que la mayoría de los datos (1, 2, 3), se ve muy afectado por el punto extremo de los datos: 34. En este caso, el valor medio hace que parezca que los valores de los datos son más altos de lo que realmente son. [1]

### Operador de RapidMiner: 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. [2]

### Operador de RapidMiner: 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. [2]

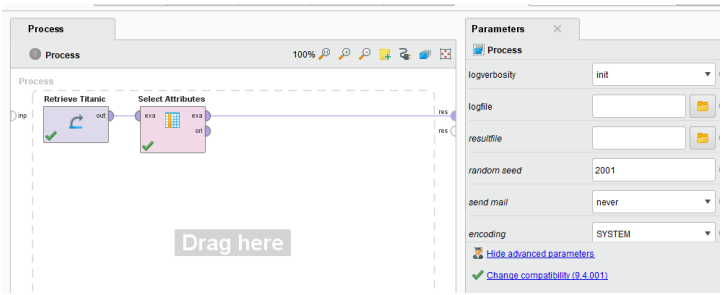
ESCUELA POLITÉCNICA NACIONAL
FACULTAD DE INGENIERÍA DE SISTEMAS
INGENIERÍA DE SISTEMAS INFORMÁTICOS Y DE COMPUTACIÓN

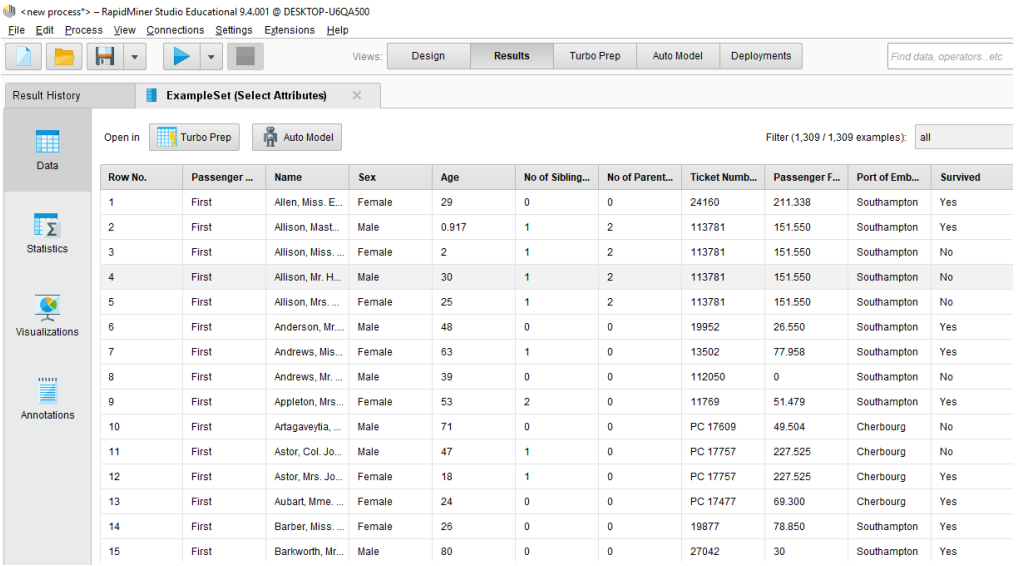
**Desarrollo de la práctica:**

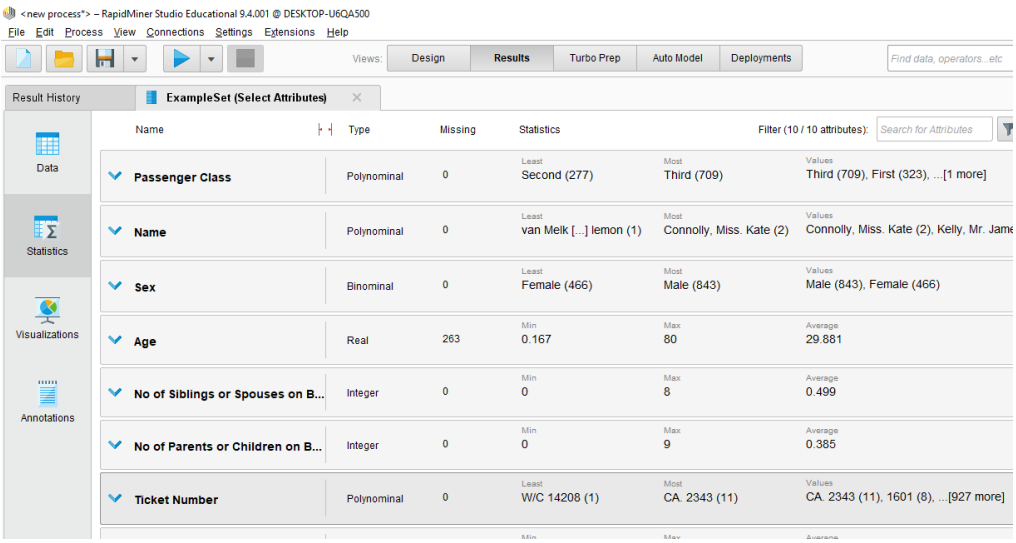
**Primera parte: Trabajo en clase**

1. Reemplazo de valores ‘perdidos’ en la base de datos de Titanic.

Primero se verifica los campos que poseen datos ‘perdidos’.

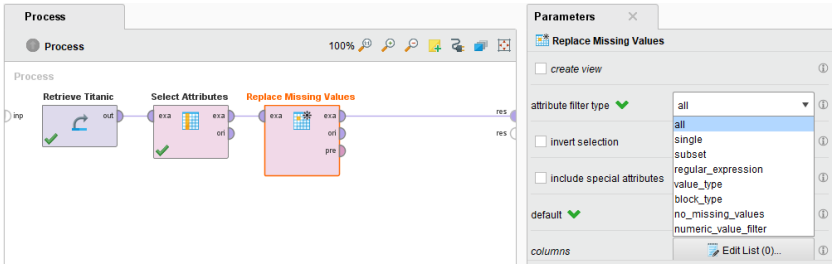


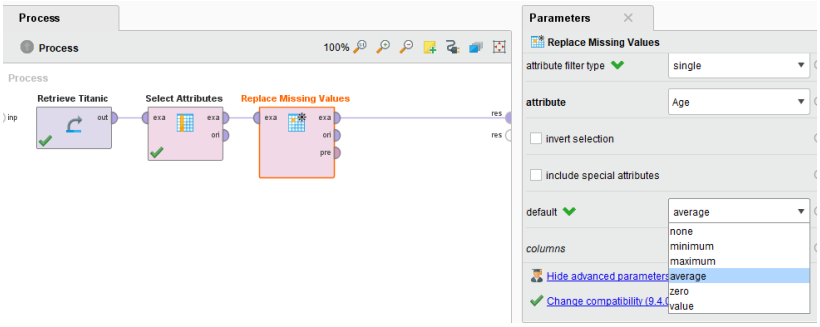




El campo con mayor número de datos perdidos es “Age” con 263 registros.

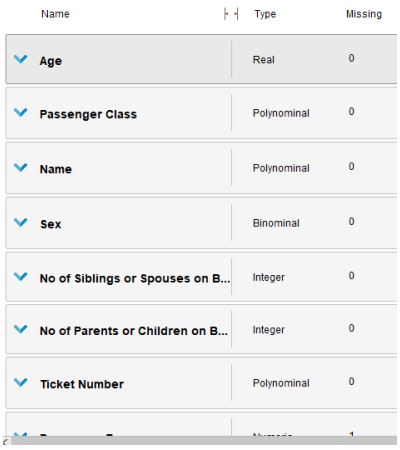
Ahora, se utiliza el operador '**Replace Missing Values'** para el atributo edad y le daremos un valor por defecto.




En este caso se reemplaza con el promedio (average) de las edades



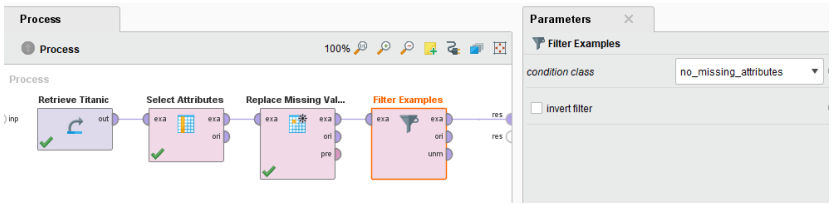

Y como resultado el atributo '**Age'** ya no tiene valores perdidos.



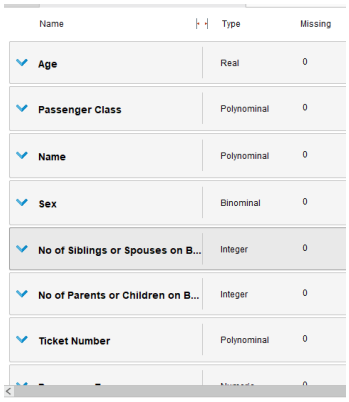
| Name | Type | Missing |
|---|---|---|
| Age | Real | 0 |
| Passenger Class | Polynominal | 0 |
| Name | Polynominal | 0 |
| Sex | Binominal | 0 |
| No of Siblings or Spouses on B... | Integer | 0 |
| No of Parents or Children on B... | Integer | 0 |
| Ticket Number | Polynominal | 0 |

2. Filtrar los elementos con atributos vacios

Además del atributo '**Age'** el atributo que contiene el **'label'** posee un atributo perdido, sin embargo, se utilizar el operador '**Filter examples'** para eliminar cualquier otro campo no contemplado con atributos 'perdidos'.



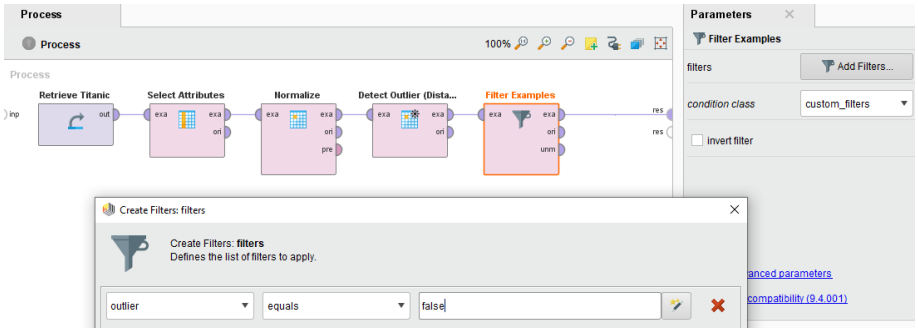

Como resultado ningún atributo dará un dato 'perdido'.



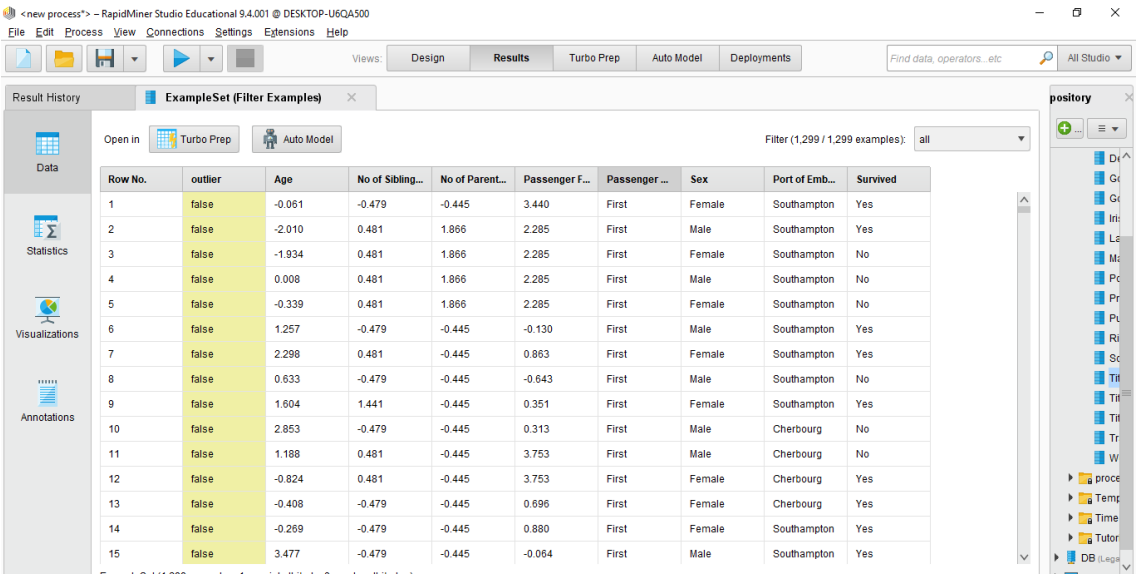

3. Detectar valores atípicos en la base de datos de Titanic

Para esto se utiliza el operador '**Normalize**' primero para escalar los valores a un rango de 0 a 1. Con el fin de luego realizar comparaciones 'justas' entre sus mismos datos.

Para detectar valores atípicos se utiliza el operador '**Detect Outlier (Distances)'** para calcular las distancias entre los datos y detectar los más alejados.



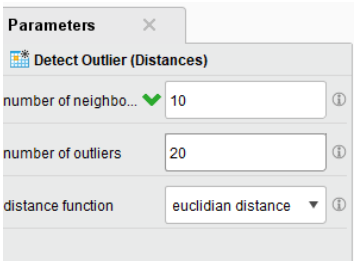
En el resultado solo se filtra aquellos datos que no son atípicos (outlier=false)



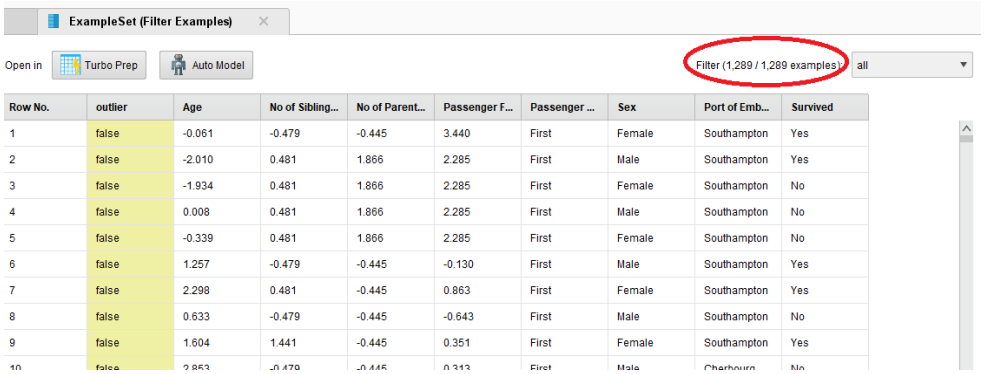

## Segunda parte: Ejercicio en clase

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

Existe una variable encargada de esto (number of outliers) y se la puede modificar en el apartado parámetros.

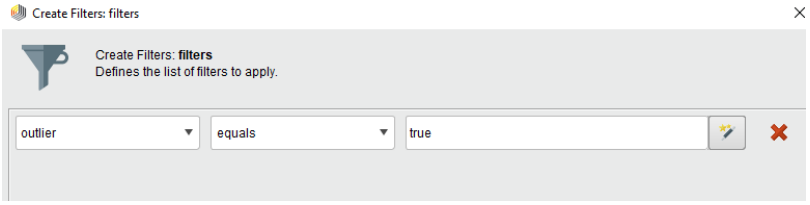



Como resultado se muestran 10 elementos menos porque se filtran los que no son atípicos.



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

Para esto, se cambia la lógica del filtro de datos. En lugar de focalizar el filtro en los que no son atípicos, nos enfocamos en los que si cumplen la condición de atípico.




Como resultado solo se obtienen los 20 datos atípicos en el conjunto de datos.

ExampleSet (Filter Examples)

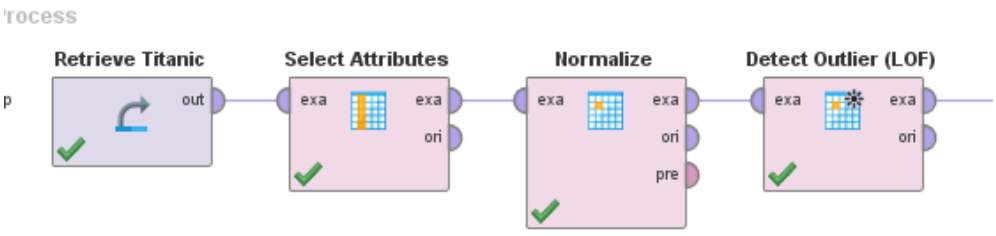
Open in Turbo Prep Auto Model

Filter (20 / 20 examples):

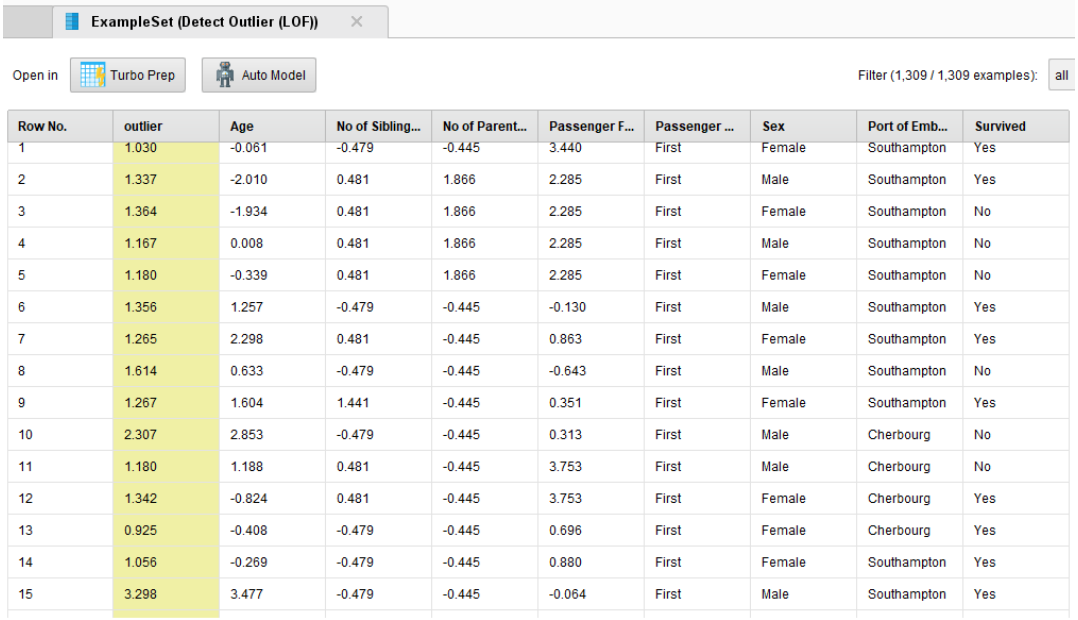
| Row No. | outlier | Age | No of Sibling... | No of Parent... | Passenger F... | Passenger ... | Sex | Port of Emb... | Survived |
|---|---|---|---|---|---|---|---|---|---|
| 2 | true | 0.425 | -0.479 | 0.710 | 9.255 | First | Male | Cherbourg | Yes |
| 3 | true | 1.951 | -0.479 | 0.710 | 9.255 | First | 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.**

Se conserva la estructura de Retrieve->Select Attributes->Normalize y cambiamos el último componente por '**Detect Outlier (LOF)**'



En el resultado se muestran números en lugar de valores booleanos




| 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 |

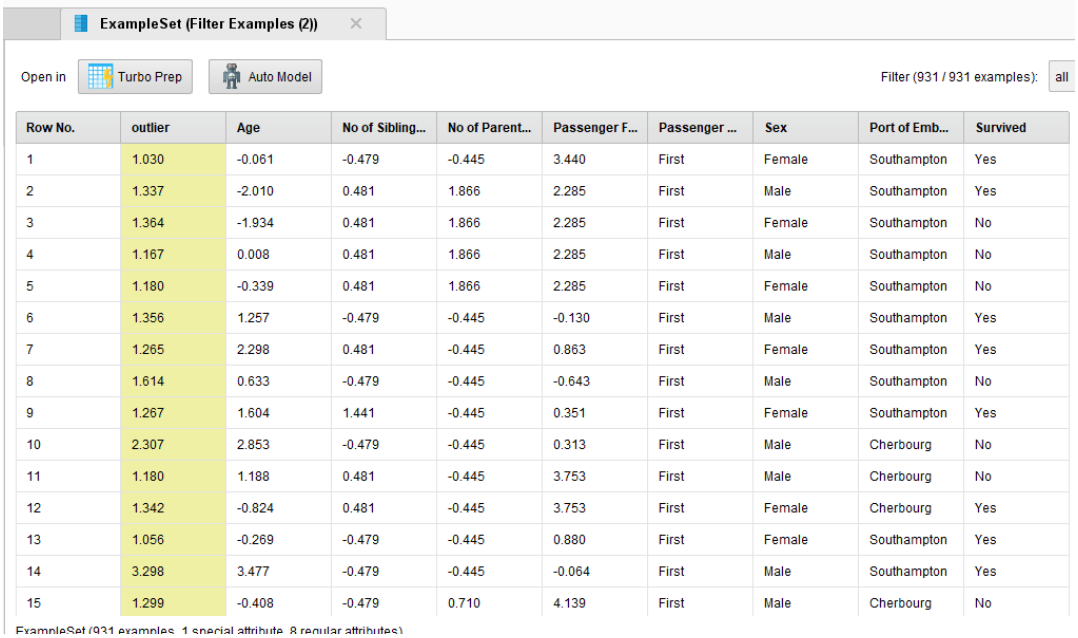
La principal diferencia es que ‘Detect Outlier (Distances)’ otorga un booleano indicando si sobrepasó o no una distancia en función de sus ‘vecinos’ más cercanosmientras que ‘Detect Outlier (LOF)’ otorga un valor numérico que significa la densidad respecto a sus vecinos más cercanos.

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

En este caso se asume que los valores atípicos superiores son más grandes que el valor máximo de la norma (1).

En el operador ‘**Filter Examples’** se agrega la condición de solo obtener aquellos valores atípicos con un valor mayor a 1.

De esta forma todos los valores de outlier consultados superan la base de 1



ExampleSet (Filter Examples (2))

Open in Turbo Prep Auto Model Filter (931 / 931 examples): all

| 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.

**Análisis de resultados:**

Los datos atípicos pueden afectar el comportamiento de un algoritmo de predicción y de clasificación en general (por categoría o por regresión). Aunque al excluirlos, el algoritmo pierde la capacidad de entender ese conjunto de datos, sin embargo, al trabajar siempre con la mayoría es mejor excluirlos.

El número de 'vecinos' que se deba considerar para detectar outliers debe ser relativamente grande, es decir, un porcentaje considerable desde un 10% del número de data hasta un 30%, de forma que se puede quedar con los datos de mayor calidad.

Dependiendo del problema será más conveniente utilizar outliers que den una salida booleana o una salida numérica.

**Conclusiones y recomendaciones:**

- Se reemplazó valores 'perdidos' dentro de un conjunto de datos por el promedio de esos datos para no causar sesgos en la analítica.
- Se logró discriminar datos que poseían valores 'perdidos' y se aumentó la calidad de los datos.
- Se utilizó clasificadores de valores atípicos para detectar que tan lejos se encontraban unos datos de otros.
- Para medir que tan atípico es un dato se realiza un calculo en base a las densidades y distancias euclideanas de los mismos datos y se establece un umbral para el filtro.
- La normalización es útil para realizar una comparación 'justa' entre un conjunto de datos.
- Los datos perdidos y atípicos pueden afectar en la calidad final de los datos.

- Es recomendable usar el operador **Detect Outlier (LOF)** cuando la relación numérica entre los datos sea crucial en los cálculos.
- Se recomienda usar el operador **Normalize** siempre antes de realizar un cálculo de valores atípicos.
- Se recomienda usar el operador **Detect Outlier (Distances)** cuando se desee realizar una clasificación rápida y la relación numérica entre valores no sea crucial.

**Bibliografía**

[1]"Identificar valores atípicos - Minitab", *Support.minitab.com*. [Online]. Available: https://support.minitab.com/es-mx/minitab/18/help-and-how-to/statistics/basic-statistics/supporting-topics/data-concepts/identifying-outliers/. [Accessed: 27- Nov- 2019].

[2]R. GmbH, "Operator Manual - RapidMiner Documentation", *Docs.rapidminer.com*. [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/. [Accessed: 27- Nov- 2019].