Ejercicio 1

Link Repositorio git@github.com:Cdogonza/InteligenciaArtificial.git

TUTORIAL " Handling Missing Values"

Este tutorial mostrará los enfoques mas comunes para eliminar o remplazar los valores perdidos.

PASOS

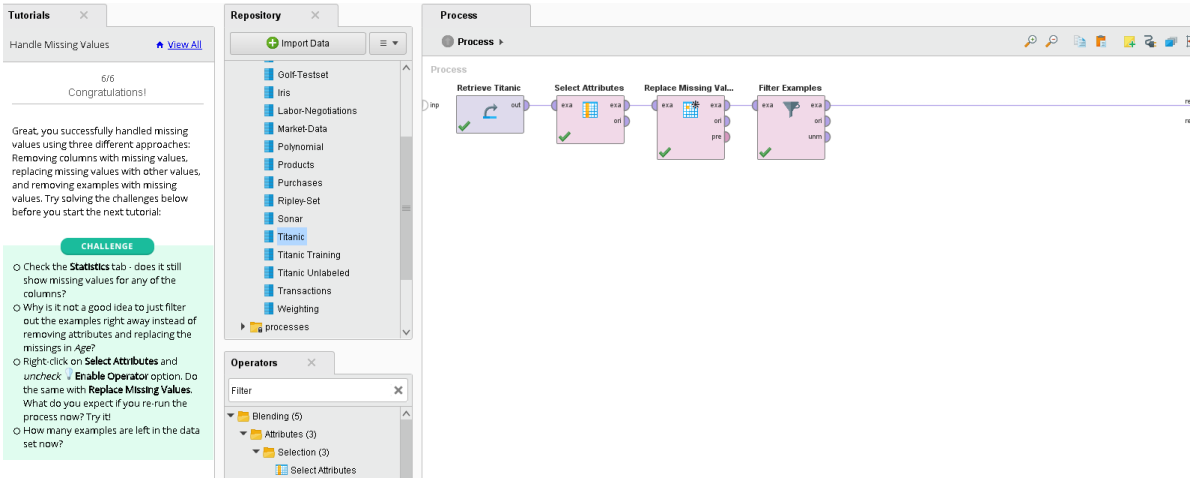
1. Se utilizó el conjunto de datos de ejemplo del Titánic.
2. Se realizó una selección de atributos, seleccionando todos excepto bote salvavidas y cabina. Esto se debió a que la cabina contenía muchos valores missing y los pocos valores que contenía no tenían información relevante probablemente. Y bote salvavidas por la misma razón, ademas de que tiene una fuerte correlacion con el dateo label o etiqueta, que son los supervivientes.
3. Se realizó el remplazo de los valores perdidos que quedaban de la columna edad, por el promedio de edad. Esto es una practica común para tratamiento de muchos valores perdidos de un atributo.

PREGUNTAS FINALES

¿sigue mostrando valores perdidos para alguna de las columnas?

- No, se han filtrado de forma correcta las columnas con valores missing

En caso de que no se haga el remplazo de los valores missing del atributo Age, y posteriormente se filtre valores missing quedan un resultado de 1043, mientras que haciendo el remplazo de los valores missing del atributo edad, y despues hacer el filtrado, nos queden una cantidad de ejemplos de 1306. En el caso de que se haga el remplazo de los valores missing del atributo edad, y no se aplique el filtro, nos quedarian una cantidad de 1309 ejemplos



TUTORIAL " Normalization and Outlier detection"

Este tutorial se va a realizar un trabajo sobre la eliminacion de los datos atipicos. Si bien estos en determinados casos pueden ser interesantes para alguna circunstancia puntual, en este ejemplo y como en un gran numero de casos, estos datos representan un resultado de medicion incorrecta, y deben ser eliminados.

1. Como paso inicial, agregamos el data de Titanic.
2. Agregamos el operador Select Attributes, que lo que haremos es seleccionar únicamente los atributos que inicialmente viendo los datos, van a contribuir a la detección de valores atípicos basados en la distancia. Se utilizará algoritmo basado en la distancia que calcula distancia euclídea que básicamente trata del cálculo de una distancia recta finita entre dos puntos en un espacio tridimensional o en dimensiones superiores. Es una distancia derivada del teorema de Pitágoras, extendido a múltiples dimensiones.
3. Agregamos el operador Normalize, debido a que cuando se trata del trabajo con distancias, se deben de normalizar los datos previamente. Por defecto, este operador realizará una estandarización que da como resultado un valor medio de 0 y una desviación estándar de 1 para cada atributo. Lo que significa que todos los atributos se encuentran ahora en una misma escala posterior a la normalización, por lo que ahora sí pueden ser comparados todos los valores entre sí.
4. Se agrega el operador "Detect Oulier (Distances)" lo que nos proporciona una nueva columna con con datos tipo boleanos donde podemos identificar como verdaderos aquellos que mediante el algoritomo son seleccionados como atipicos, y falsos aquellos que no.
5. Agregamos el operador "Filter" donde relalizamos un filtros de todos aquellos valores con valor false, o sea que no son considerados como atipicos.
6. El resultado de ejecutar lo que tenemos hasta el momento es haber realizado un filtrado de todos aquellos valores que son atipicos, en este caso 10, posterior a haber sido

normalizados, lo que se conoce como una limpieza de los datos mejorando la calidad del modelo.

PREGUNTAS

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

- En el caso que se encuentren mas valores atipicos se puede deber a varias razones, dentro de las cuales estan una sensibilidad en el algoritmo utilizado, ya que dependiendo de los parametros proporcionados, la cantidad de valores podria llegar a variar mas o menos significativamente. Por otro lado la caracteristica del conjunto de datos y su distribucion tambien puede llegar a ser significativa a la hora de obtener mas o menos datos atipicos, asi como tambien diferente densidades de los datos. Otra variacion en la cantidad de datos atipicos es el tamano del conjunto de datos, ya que el presentar mayor cantidad de datos, puede influir en la variabilidad del numero de datos atipicos.

¿Cómo puede cambiar el proceso para que sólo muestre los valores atípicos en lugar de eliminarlos?

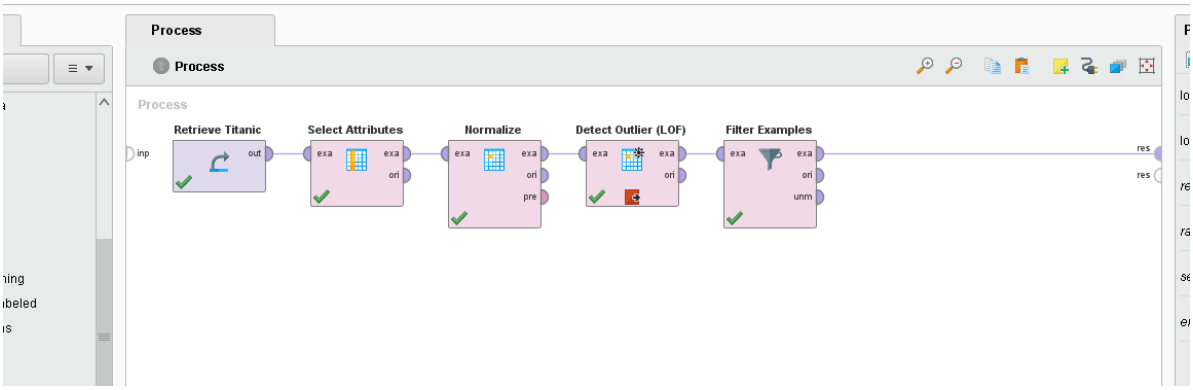
- En el operador de filtrado posterior, en vez de realizar la eliminación de los atípicos (valores true), eliminamos valores false y de ese modo podemos mostrar únicamente los valores atípicos.

Sustituye el operador de detección de valores atípicos por Detectar valores atípicos (LOF) y añade un punto de interrupción después de este operador antes de ejecutarlo. ¿Cuál es la diferencia con respecto a antes?

- La diferencia del resultado con el operador de detector de valores atípicos (LOF) con el anterior, es que el sí, el operador funciona diferente, por lo que el resultado es diferente. Su resultado es el cálculo de valores atípicos, pero mediante un algoritmo que calcula los valores atípicos en función al comportamiento de los puntos más cercanos (estaos parámetros son configurables). Por cada punto, el algoritmo calcula el LOF (Local Outler Factor) Este valor mide la desviación de la densidad local de un punto en comparación con la densidad local de sus vecinos cercanos. Los puntos con este valor calculado más alto en función a un umbral predefinido son considerados los puntos atípicos, agregando en la columna de oultiders el valor calculado y no un valor booleano.

¿Cómo debe cambiar ahora el filtro para mantener solo los valores atípicos más altos?

- Considerando que el valor para el umbral típico es 1, todos aquellos datos con valor de LOF superiores a 1, son considerados atípicos, por lo que el operador "Filter" lo modificamos para que nos muestre únicamente los valores que en la columna de outlaider sean menores que 1.


Process
Process
Process
Retrieve Titanic
Select Attributes
Normalize
Detect Outlier (LOF)
Filter Examples
inp
out
exa
ori
pre
unm
res