PRACTICOS DOMICILIARIOS INDIVIDUALES #1

Ejercicio 1

Dentro del Pre-procesamiento de datos podemos encontrar **dos grandes categorías**:
La mezcla y la limpieza.
En este primer tutorial se trabajan tareas básicas de **limpieza**, particularmente en este caso, la **limpieza de valores nulos**.

”Handling Missing Values”

Parto de un dataset que cuenta con varios **atributos con valores nulos**. A continuación voy a proceder a ejecutar diversas acciones para el tratamiento de dichos valores.

Lo primero que voy a hacer es **descartar** dos atributos que cuentan con muchos valores nulos **por estar fuertemente relacionados con el label del dataset.**

Para ello utilizo el **operador select attributes** conectado al dataset y me quedo con todos los atributos menos esos dos.

Para resolver los valores faltantes del atributo *age*, que tiene bastantes, utilizo otra de las técnicas normalmente usadas: **reemplazar valores nulos por el promedio del resto de los valores**.

Para realizar esto, voy a utilizar el operador **replace missing values**, teniendo en cuenta como parámetro ***single*** que significa remplazar por el promedio.

Por último con los restantes valores nulos que aún me quedan, como se trata de una cantidad muy pequeña y despreciable los voy a descartar.

Voy a tomar el operador **Filter Examples** y en las opciones avanzadas selecciono ***no missing valúes*** lo que borra las tuplas sin valores.

”Normalization and Outlier detection”

Si bien puede que haya algunos casos para los cuales los valores anómalos o outliers resulten interesantes (como el caso de tarjetas de crédito fraudulentas). En general los outliers son malas mediciones y debemos removerlos.

En este caso, volvemos a utilizar el operador **select attributes** para seleccionar aquellos atributos del dataset que creemos que contribuyen a la presentación de outliers.

Una vez filtrada la información con los atributos deseados, vamos a normalizar el dataset obtenido hasta el momento.

El motivo de la normalización es porque vamos a detectar los outliers del conjunto de datos, mediante el algoritmo de distancias euclideanas. Es una buena práctica normalizar los datos antes de cualquier tratamiento con algoritmos basados en distancias. Visto lo anterior, utilizamos el operador **Normalize** previo a procesar los datos.

Ahora lo siguiente es la detección de outliers con las distancias euclidianas previamente mencionadas. Para ello utilizamos el operador **Detect Outlier (distances)** estableciendo como función ***euclidian distances***

Ahora que ya contamos con el conjunto de datos con sus outliers conocidos, simplemente vamos a filtrar de manera de mostrar nuevo conjunto de datos, eliminando aquellos outliers detectados. Usamos el operador **Filter Examples** con la propiedad ***outliers equals false***

Conclusiones Ejercicio 1:

En estos dos ejemplos vimos algunas de las técnicas utilizadas al procesar los datos con la herramienta RapidMiner. Tanto para el tratamiento de valores nulos como outliers , es importante ver que, si bien el software tiene operadores puntuales para el tratamiento de valores nulos y anomalías, son necesarias operaciones previas y posteriores para poder realizar una correcta limpieza del conjunto de datos.

Ejercicio 2

Problema a resolver

El dataset consta de un análisis quimico realizado sobre vinos de 3 diferentes cultivos. El anásis determina 13 diferentes propiedades encontradas de los 3 tipos diferentes de vinos.

Con esta información, el objetivo es determinar mediante las propiedades de un determinado vino, a qué cultivo o variante pertenece.

Atributos y características

1. Graduación Alcohólica
2. Ácido málico
3. Ceniza
4. Alcalinidad de ceniza
5. Magnesio
6. Fenoles totales
7. Flavonoides
8. Fenoles no flavonoides
9. Proanthocyanins
10. intensidad del color
11. Hue
12. OD280 / OD315 de vinos diluidos
13. Proline

El dataset cuenta 177 registros divididos en **3 clases** (correspondientes a los 3 tipos de vinos)

- 59 registros para la clase 1
- 71 registros para la clase 2
- 48 registros para la clase 3

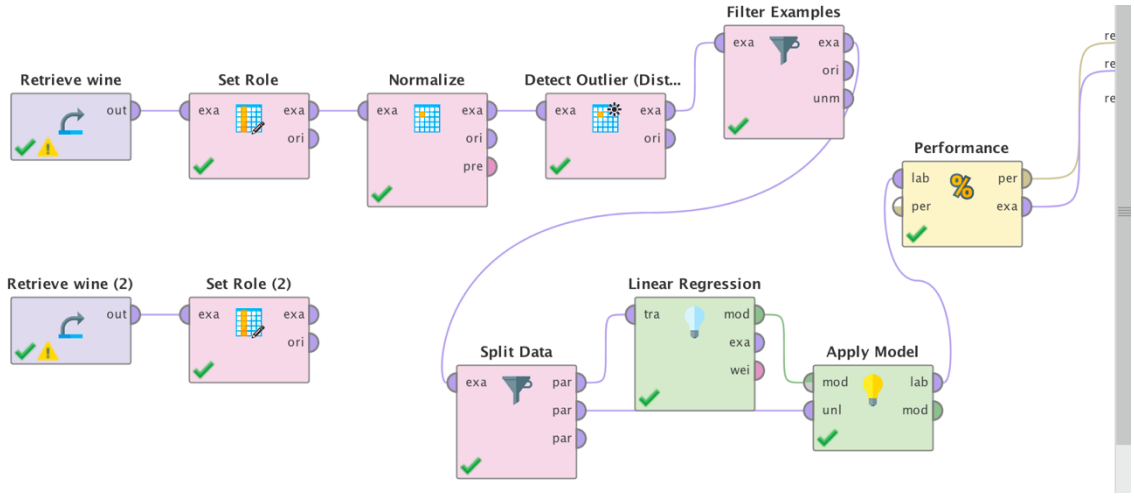
El dataset **no tiene atributos con valores nulos**

En cuanto a los outliers, aplicando lo aprendido en el ejercicio 1, detecto 10 outliers, por lo cual mi nuevo dataset tendrá 10 registros menos (167 total)

La importancia del procesamiento de la data

A continuación, utilizo dos canales en los que voy a clasificar los datos y evaluar la perfomance de predicción el primer canal procesando los datos y en el segundo sin procesarlo para evaluar los resultados:

Canal con la data **procesada:**



Resultados:

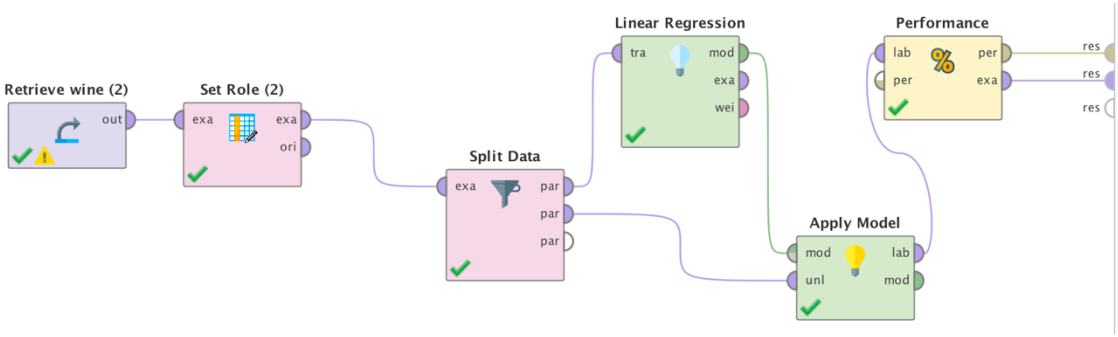
squared_error

```
squared_error: 0.061 +/- 0.075
```

root_mean_squared_error

```
root_mean_squared_error: 0.247 +/- 0.000
```

Canal con data **sin procesar**



Resultados:

squared_error

```
squared_error: 0.076 +/- 0.151
```

root_mean_squared_error

```
root_mean_squared_error: 0.277 +/- 0.000
```

Si comparamos los resultados de los datos procesados y sin procesar, observamos un menor porcentaje de error al utilizer datos procesados, lo que da a entender que cuanto más trabajo se haga en el preprocesamiento de datos, mejor serán los resultados obtenidos.