# UT2 – TA02

## Ejercicio 1

El problema a resolver es la clasificación del vino en 3 clases diferentes dados atributos como:

1) alcohol
2) ácido málico
3) ceniza
4) Alcalinidad de cenizas
5) magnesio
6) fenoles totales
7) Flavonoides
8) fenoles no flavonoides
9) Proantocianinas
10) intensidad de color
11) Hue
12) OD280 / OD315 de vinos diluidos
13) Prolina

Todos los atributos representan variables continuas, no hay valores faltantes en atributos, sí en predicción de clases faltan valores de 2 instancias.
Técnicas para estandarizar: se consideró pertinente estandarizar utilizando la transformación de rango para que todas las variables quedaran entre 0 y 1 dadas las características de los atrbiutos analizados.

## Ejercicio 2

Bloque de rapid miner utilizado:

**Filter** para valores nulos, que tiene como parámetros:

**filters**: se elige la condición para aplicar el filtro, =,/=, >, < entre otros.

**condition_class** permite seleccionar que no hayan label o atributos con valores missing.

**invert_filter** permite excluir las instancias que cumplen con lo especificado.

**Normalize** para estandarizar atributos, que toma como parámetros:

**crear vista**: si se selecciona la opción, la normalización es demorada hasta que la transformación sea necesaria

**filtro de atributo**: permite seleccionar un filtro para los atributos a normalizar, entre las opciones están seleccionar atributos simples, múltiples, un subset de atributos, expresiones regulares, entre otros.

**atributo:** se selecciona el atributo requerido en caso de aplicar un solo atributo en la opción anterior en caso contrario se despliega desde el botón atributos una lista de selección.

**tipo de valor:** permite filtrar por el tipo de atributo

**excepción a la selección de tipo:** si está habilitado, una excepción al tipo seleccionado puede ser especificada.

**invertir selección**, si se selecciona este parámetro la selección realizada es invertida, todos los atributos que matchean lo especificado son removidos.

**incluir atributos especiales:** atributos con roles especiales, id, label, cluster de predicción, weights o batches. También roles comunes pueden ser asignados a los atributos. Si se selecciona este parámetro los atributos especiales son testeados contra las condiciones especificadas en Seleccionar Atributos.

## Método: cuatro métodos de normalización

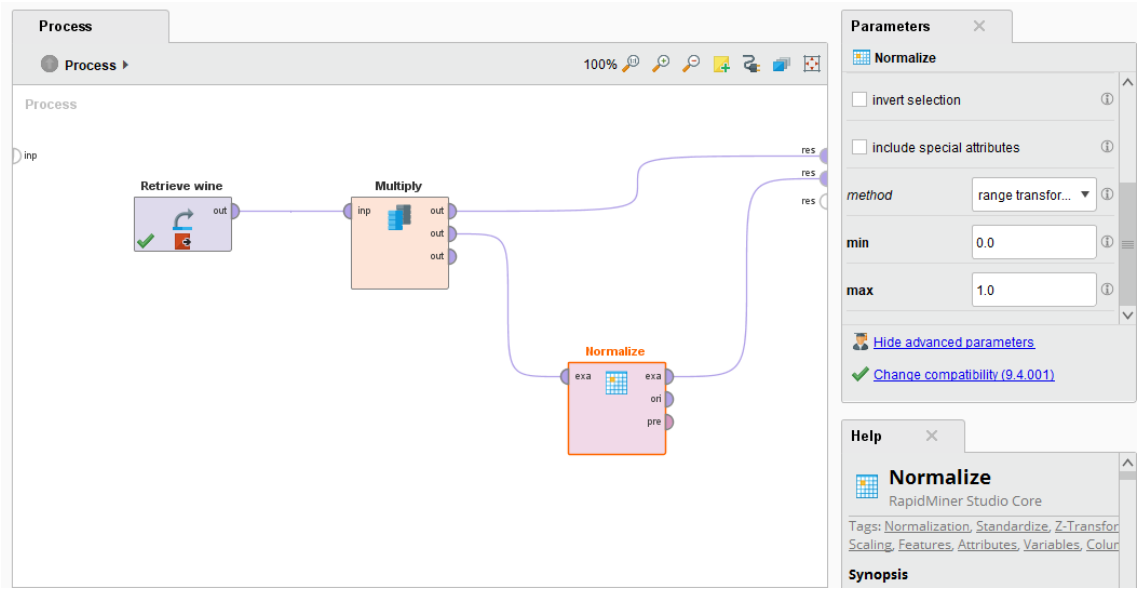
**z_transformation:** También llamado normalización estadística. Sustrae la media de todos los valores y divide por la desviación estándar. Luego de esto la distribución de los datos tiene una media de 0 y varianza de 1. Preserva la distribución original de la información y es menos influenciada por outliers.

**range_transformation:** normaliza todos los atributos a un valor especificado de rango. Deben setearse mínimo y máximo. Todos los valores se escalan a esta medida. Puede ser influenciada por outliers pero mantiene la distribución original. Sirve para ofuscar algunos valorses

**proportion_transformation:** Normalización basada en la proporción que tiene cada valor que toma el atributo sobre el total de valores del atributo. Cada valor es dividido por la suma total de los valores del atributo. La suma sólo considera valores finitos. Cuando se selecciona este método también permite tratar valores negatvios como absolutos, si hay valores negativos y no se tratan como absolutos da un error.

**interquartile_range (rango intercuartílico):** Normalización que usa el rango intercuartílico, es la distancia entre el percentil 25 y el 75, llamados Q1 y Q3. Se calculan ordenando la información y tomando valores que se separan un 25% del primero o último. La mediana es el percentil 50 por lo que separa los valores ordenados en exactamente la mitad. El rango intercuartílico (IQR) es la diferencia entre Q3 y Q1, la fórmula final es media / IQR, siendo este último el rango entre el 50% de la información. Por lo que esta normalización es menos influenciada por outliers. Los valores nulos son ignorados así como los infinitos.

**Captura del proceso de normalización utilizando el método de transformación de rangos en rapidminer;**




## Ejercicio 3

Resultados del ejercicio

Sin normalización:

**accuracy: 77.36%**

| | true 1 | true 2 | true 3 | class precision |
|---|---|---|---|---|
| pred. 1 | 16 | 2 | 0 | 88.89% |
| pred. 2 | 0 | 13 | 2 | 86.67% |
| pred. 3 | 3 | 5 | 12 | 60.00% |
| class recall | 84.21% | 65.00% | 85.71% | |

Con normalización:

**accuracy: 92.80%**

| | true 1 | true 2 | true 3 | class precision |
|---|---|---|---|---|
| pred. 1 | 43 | 1 | 0 | 97.73% |
| pred. 2 | 3 | 41 | 1 | 91.11% |
| pred. 3 | 0 | 4 | 32 | 88.89% |
| class recall | 93.48% | 89.13% | 96.97% | |

La normalización fue realizada de forma posterior a la división del modelo en entrenamiento y test a fines de evitar contaminación de datos.

Se observan los resultados de las predicciones mejoran al normalizar los datos, obteniendo una mejora en la precisión de un 15% en el modelo. Mejorando las predicciones por clase sustancialmente.