TA2 - Enzo Cozza - Agustín Fernández

Ejercicio 1

Problema: Se intenta predecir el origen de distintos vinos a partir del análisis químico de los mismos.

Variable objetivo: Class es la variable objetivo y refiere a tres distintos viñedos de Italia.

Atributos

Alcohol: Porcentaje alcohol en el vino.
Malic acid: ácido málico, el principal ácido en las uvas que puede influenciar el gusto del vino.
Ash: cenizas, un indicador de la calidad del vino.
Alcalinity of ash: alcalinidad de las cenizas, propiedad química de las cenizas.
Magnesium: Magnesio, un mineral.
Total phenols: Total de fenoles, una clase de molécula que define el sabor, aroma, beneficios medicinales y diversidad del vino.
Flavanoids: Flavonoide, un tipo de fenol que tiene un gran impacto en el sabor del vino.
Nonflavanoid phenols: Fenoles no flavonoides.
Proanthocyanins: Proantocianidina, un tipo de flavanoide.
Color intensity: Intensidad del color en el vino.
Hue: matiz en la coloración del vino.
OD280/OD315 of diluted wines: una medida del contenido de proteínas.
Proline: Prolina, que cambia según el tipo de uva.

Todas las variables son numéricas. Las únicas del conjunto que son enteras son Magnesium y Proline, el resto son números reales. La variable objetivo es polinomial (1, 2 ó 3).

Normalización

Feature scaling: $X' = (X - Xmin)/(Xmax - Xmin)$
Todos los valores del dataset quedan en un rango [0,1].

Standard score: $(X - \mu)/(\sigma)$
Normaliza errores cuando los parámetros de la población son conocidos. Funciona mejor para datos con distribución normal.

T-student: $(X - \overline{X})/s$
Normalización residual cuando los parámetros de población son desconocidos.

Coeficiente de variación: σ/μ
Normaliza la dispersión, utilizando la media como medida de escala, funciona mejor para la distribución exponencial y distribución de Poisson.

De las técnicas mencionadas previamente, se decide utilizar Feature scaling ya que permite que todos los valores del conjunto queden expresados en un mismo rango. Esto es importante ya que algunos algoritmos de ML requieren esto para poder funcionar de la mejor manera posible.

Ejercicio 2

Los bloques identificados en el RapidMiner para normalización de datos son: Normalize, Scale by Weights y De-Normalize.

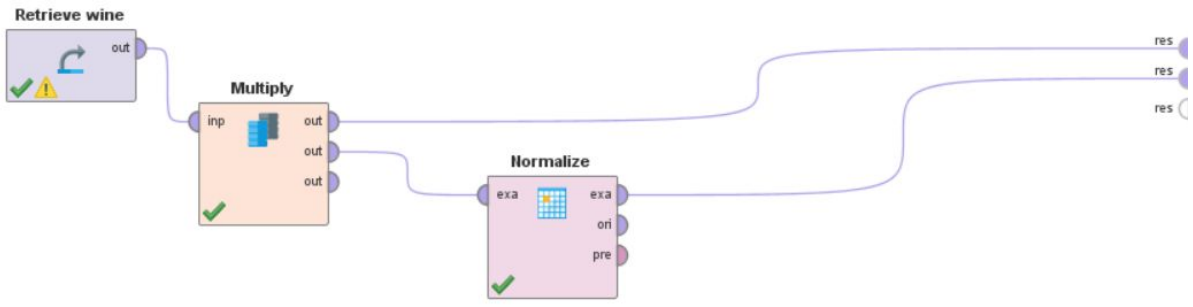
Normalize: este bloque normaliza el valor de los atributos seleccionados.
Parámetros:

- Create view
- Attribute filter type
- Invert selection
- Include special attributes
- Method: hay cuatro métodos disponibles para la estandarización: z-transformation, range transformation, proportion transformation y interquartile range.
- Allow negative values

Scale by Weights: se encarga de darle un peso (precalculado) a los atributos. Los atributos más importantes tendrán un mayor peso en la escala.

De-Normalize: este bloque revierte una normalización previamente aplicada a los datos.
Parámetro:

- Missing attribute handling



Ejercicio 3

Para datos no normalizados:

accuracy: 73.58%

	true 1	true 2	true 3	class precision
pred. 1	16	1	2	84.21%
pred. 2	2	15	2	78.95%
pred. 3	3	4	8	53.33%
class recall	76.19%	75.00%	66.67%	

Para datos normalizados:

accuracy: 96.23%

	true 1	true 2	true 3	class precision
pred. 1	21	1	0	95.45%
pred. 2	0	18	0	100.00%
pred. 3	0	1	12	92.31%
class recall	100.00%	90.00%	100.00%	

Se puede apreciar una notoria diferencia en la precisión (de más del 20%) para cuando los datos fueron normalizados.