UT2 PD1

Ejercicio 1

Handling Missing Values: Con f3 podemos ver los detalles de los datos y ver cuales faltan. Es importante sacar atributos altamente correlacionados con la variable a predecir o que tengan muchos valores faltantes. Una de las políticas más comunes es colocar el promedio en los valores faltantes. Otra forma es sacar todas las filas con valores nulos.

Normalization and Outlier detection: Para detectar outliers se utiliza la distancia euclideana entre las instancias, esto mide la diferencia de valor en los atributos y así se sabe cuales son extremos. Siempre que se calculen distancias entre las instancias se deben normalizar los los atributos. En el operador detect outliers se puede setear la cantidad de outliers que queremos, esto hay que tenerlo en cuenta porque elegir demasiados o muy pocos puede ser problemático.

Ejercicio 2

El problema del dataset wine

El dataset contiene datos provenientes del análisis químico de vinos hechos en la misma región de Italia pero que derivan de distintos cultivadores de uvas. El análisis determinó la cantidad de los 13 componentes encontrados en los tres tipos de vinos.

El problema de esta dataset se basa en construir varios modelos de clasificación para la clase del vino.

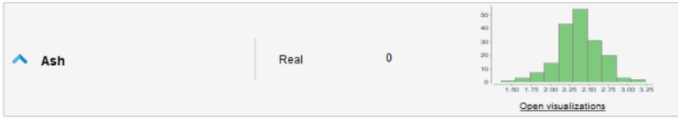
Variable objetivo: consiste en la clase de vino que puede tener valores 1, 2 o 3.

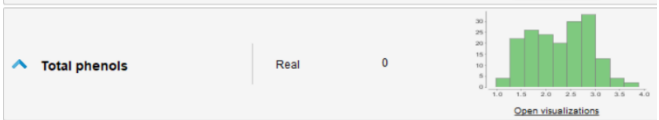
Atributos:

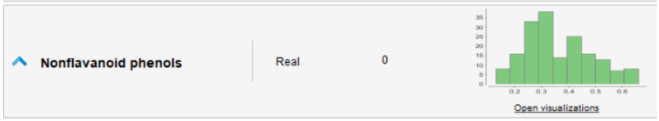
1) Alcohol: Porcentaje de alcohol presente en vino
2) Malic acid: Concentración de ácido málico en el vino. Medida (g/l)
3) Ash: Se define como toda la materia inorgánica que queda tras la ignición del residuo que queda de la evaporación del mosto o del vino. Medida (g/l).
4) Alkalinity of ash: La alcalinidad del ash se define como la suma de cationes, distintos del ion amonio, combinados con los ácidos orgánicos en el vino. Medida: (g/l)
5) Magnesium: La cantidad de magnesio que contiene una botella de vino. Medida (mg)
6) Total phenols: Es la concentración de fenoles en el vino. Medida (mg/l)
7) Flavanoids: Es la concentración de flavanoides en el vino. Medida (mg/l)
8) Nonflavanoid phenols: Es el porcentaje de Fenoles no flavonoides presentes en el vino. Medida (%).
9) Proanthocyanins: Es la concentración de Proantocianidinas en el vino. Medida (mg/l)
10)Color intensity: La intensidad de color del vino.
11)Hue: El tono del vino
12)OD280/OD315 of diluted wines: Es la concentración de proteínas en el vino. Medida (mg/l).
13)Proline: Es un aminoácido que regula el sabor del vino, se mide su concentracón. Medida: (mg/l).

Todos los atributos del dataset son ordinales (reales y enteros).

Características de los atributos

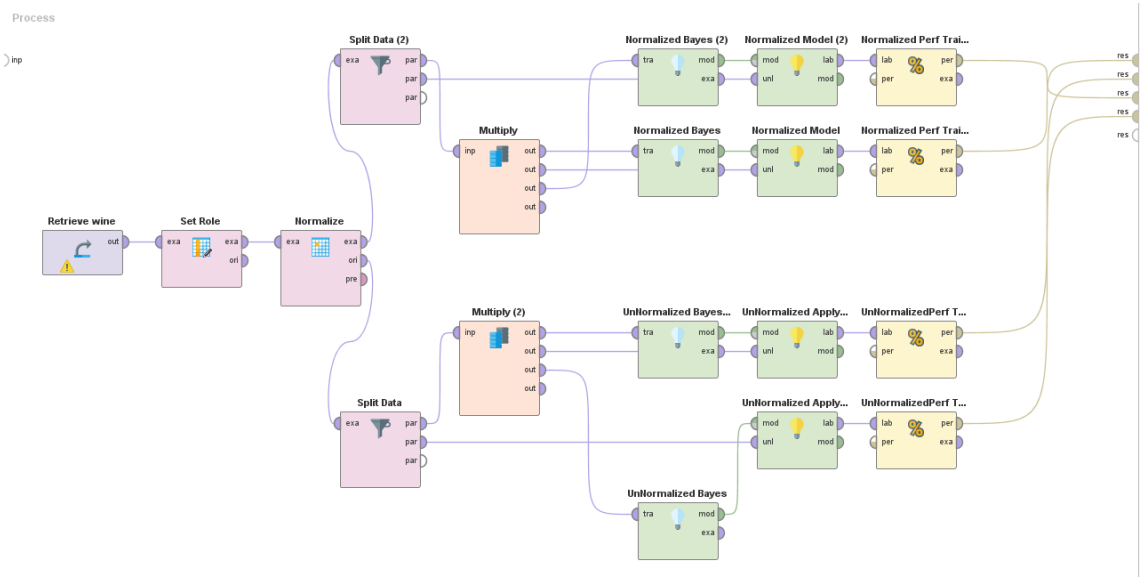






En cuanto a valores faltantes el dataset no presenta nada.

Proceso



La parte superior corresponde al área donde se trabaja con datos normalizados, en cada parte se entrena un modelo con los datos de entramiento y se lo testea con los datos de entrenamiento y abajo con los de test para observar las diferencias.

Normalizado entrenamiento - entrenamiento	99.20%
Normalizado entrenamiento – Test	98.11%
No Normalizado entrenamiento – entrenamiento	99.20%
No Normalizado entrenamiento - Test	94.34%

Se puede ver que cuando se normalizan los datos la diferencia entre las performances es menor.