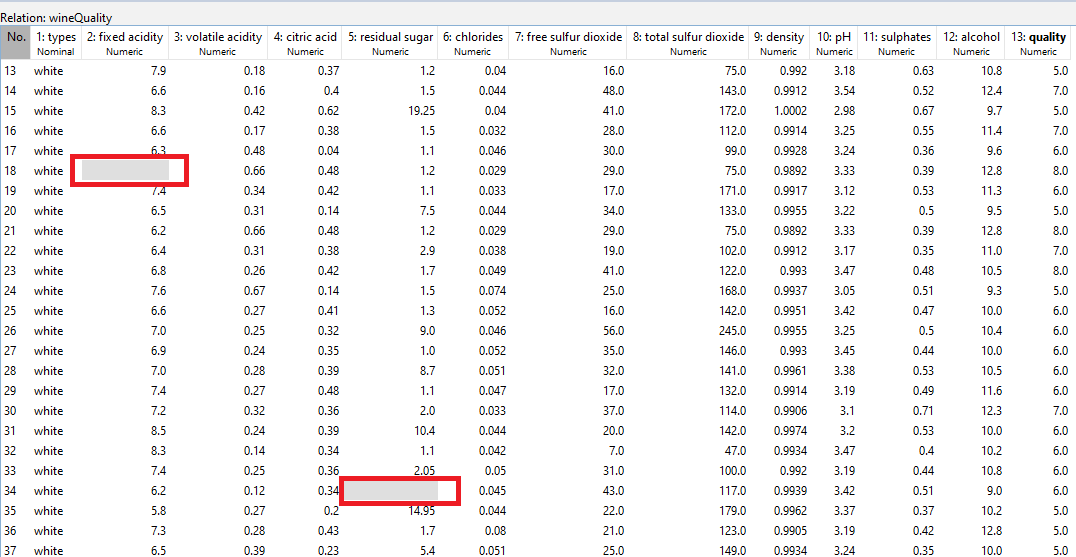
|  |  |  |  |
| --- | --- | --- | --- |
| Nina | Huacani | Steve Brandom | N |
| 1. **Paterno** | 1. **Materno** | **Nombres** |
| Inteligencia Artificial | PhD. Moises Silva | 9990778 |
| **Materia** | **Docente** | **CI** | **INICIAL A.P.** |

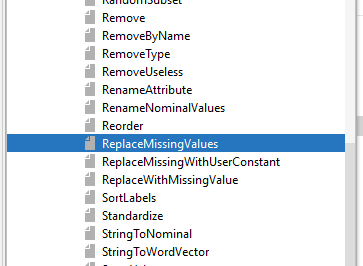
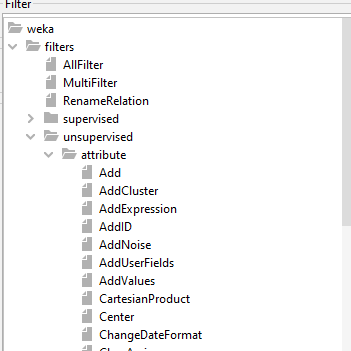
**5. Del Dataset elegido, migre el mismo a WEKA y utilice cuatro técnicas de preprocesamiento (realice la captura de pantallas de estos por fases). Explique la razón de usar estas técnicas.**

**Preprocesamiento N°1: Remplazar valores faltantes**

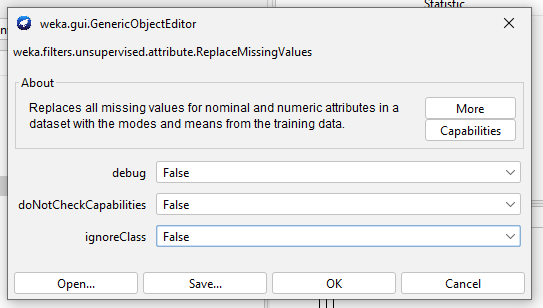
**Paso 1. Abrimos el dataset y damos click a “Edit” para poder ver el dataset**

**Observamos campos vacíos**

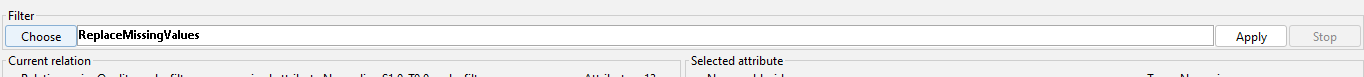
****

**Paso 2. Vamos a la opción de *Choose/filters/unsupervised/ReplaceMissingValues***

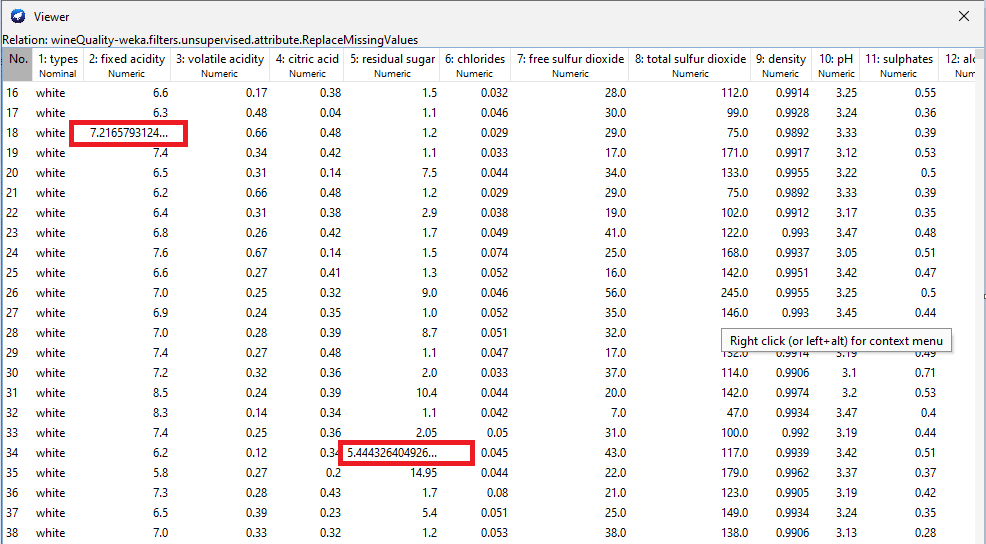
**Paso 3. Luego hacemos clic derecho sobre la opción elegida para poder ver las opciones que nos ofrece weka**



**Paso 4. Finalmente apretamos la opción de Apply para poder ver los cambios**

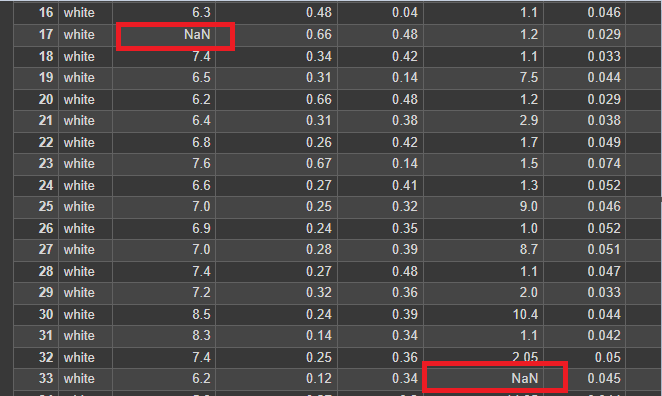


**Paso 5. Damos click en la pestaña de “Edit” para poder observar los cambios**

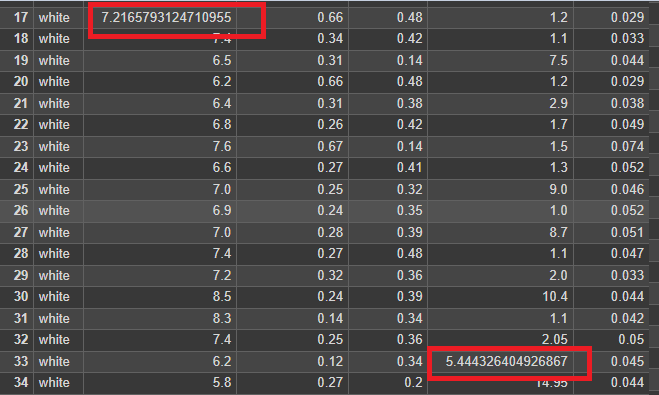


**Compararemos la técnica aplicada con el resultado obtenido en Python**

**Antes**

****

**Después**

****

**Como podemos evidenciar obtenemos los mismos resultados**

**Missing Values**

La imputacion de valores faltantes es una tecnica de preprocesamiento muy importante ya que los algoritmos de aprendiza automatico no pueden manejar valores faltantes.

Habitualmente se asocian tres tipos de problemas con los valores faltantes:

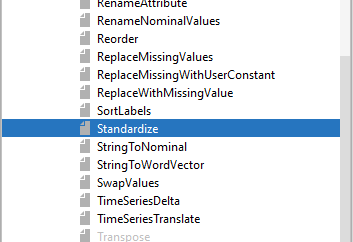
* Perdida de eficiencia
* Complicaciones en el manejo y análisis de los datos
* Sesgo resultante de las diferencias entre los datos faltantes y los que están completos

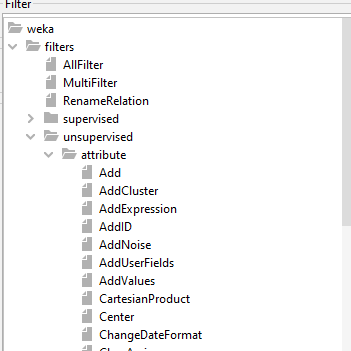
Existe una amplia variedad de métodos de imputación como, por ejemplo: la sustitución por medias, medianas o modas, el vecindario K más cercano, procedimientos de máxima verosimilitud entre otras. Algunos de los beneficios de utilizar la imputación de valores faltantes son las siguientes.

**la preservación de los datos**: Ya que si se eliminara una tupla que contiene el valor faltante se pierde la información de las otras características de la fila o columna, la imputación de valores faltantes nos permite mantener la máxima cantidad de información posible.

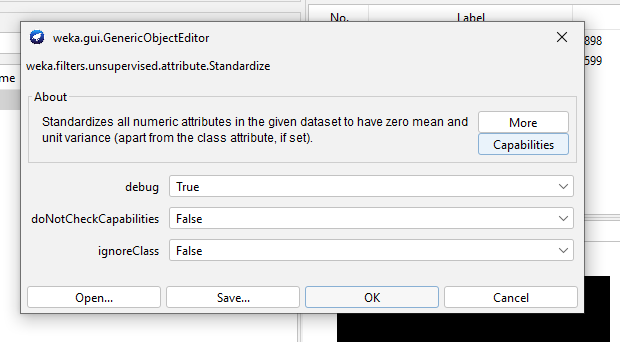
**Consistencia de los datos:** Ayuda a mantener la consistencia de un conjunto de datos ya que es importante mantener el mismo número de muestras en todas las categorías para que el análisis sea el más optimo posible.

**Preprocesamiento N°2: Estandarización**

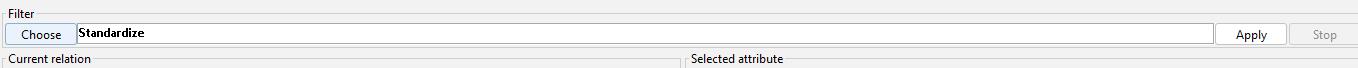
****Paso 1. Vamos a la opción de *Choose/filters/unsupervised/Standarize***

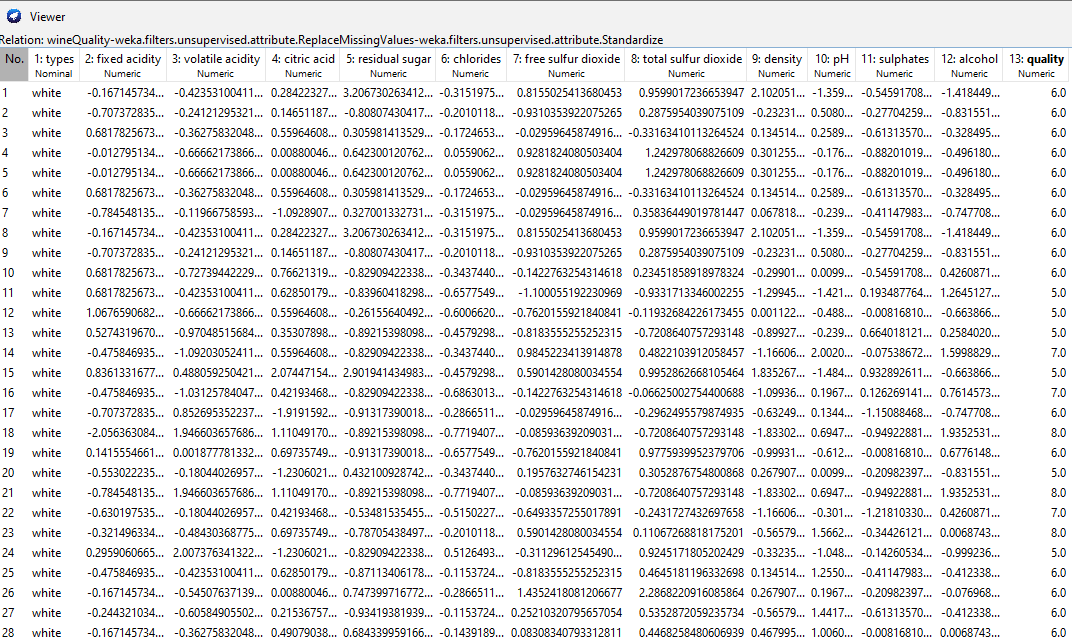
****

**Paso 2. Luego hacemos clic derecho sobre la opción elegida para poder ver las opciones que nos ofrece weka**

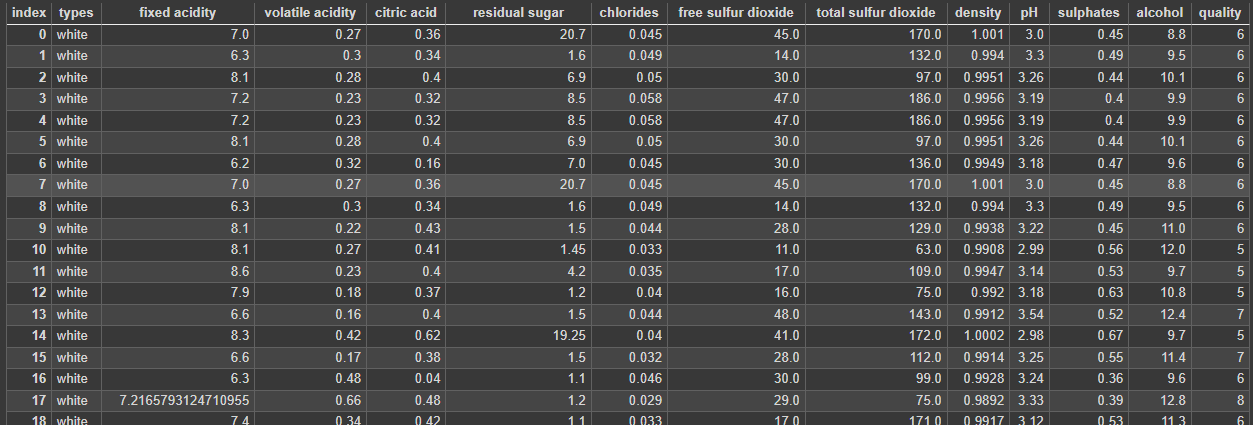


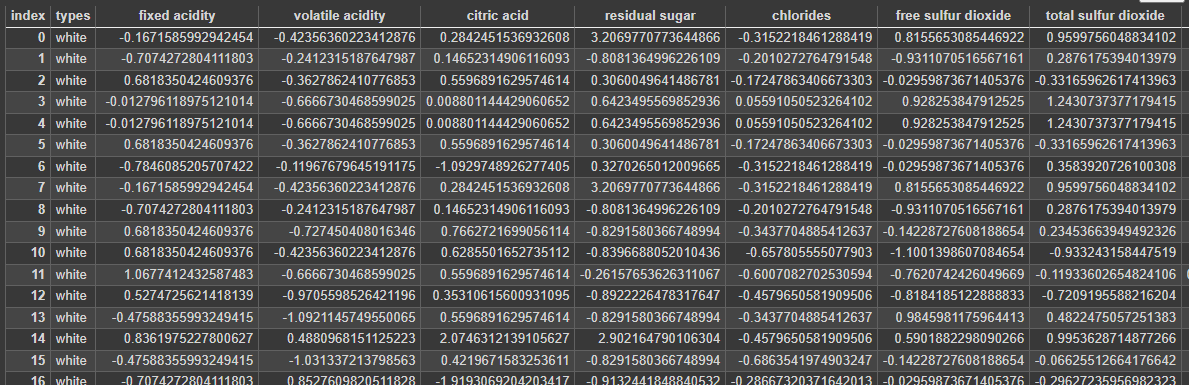
**Paso 3. Finalmente apretamos la opción de Apply para poder ver los cambios**



**Paso 4. Damos click en la pestaña de “Edit” para poder observar los cambios**

**Compararemos la técnica aplicada con el resultado obtenido en Python**

**Antes**

**Despues**

**Como podemos evidenciar obtenemos los mismos resultados**

**Estandarización**

La estandarización es una técnica de preprocesamiento de datos que se utiliza para transformar las características de un conjunto de datos de forma que tengan una media en 0 y una desviación estándar de 1.

Algunas de las desventajas de trabajar con características no estandarizadas son las siguientes:

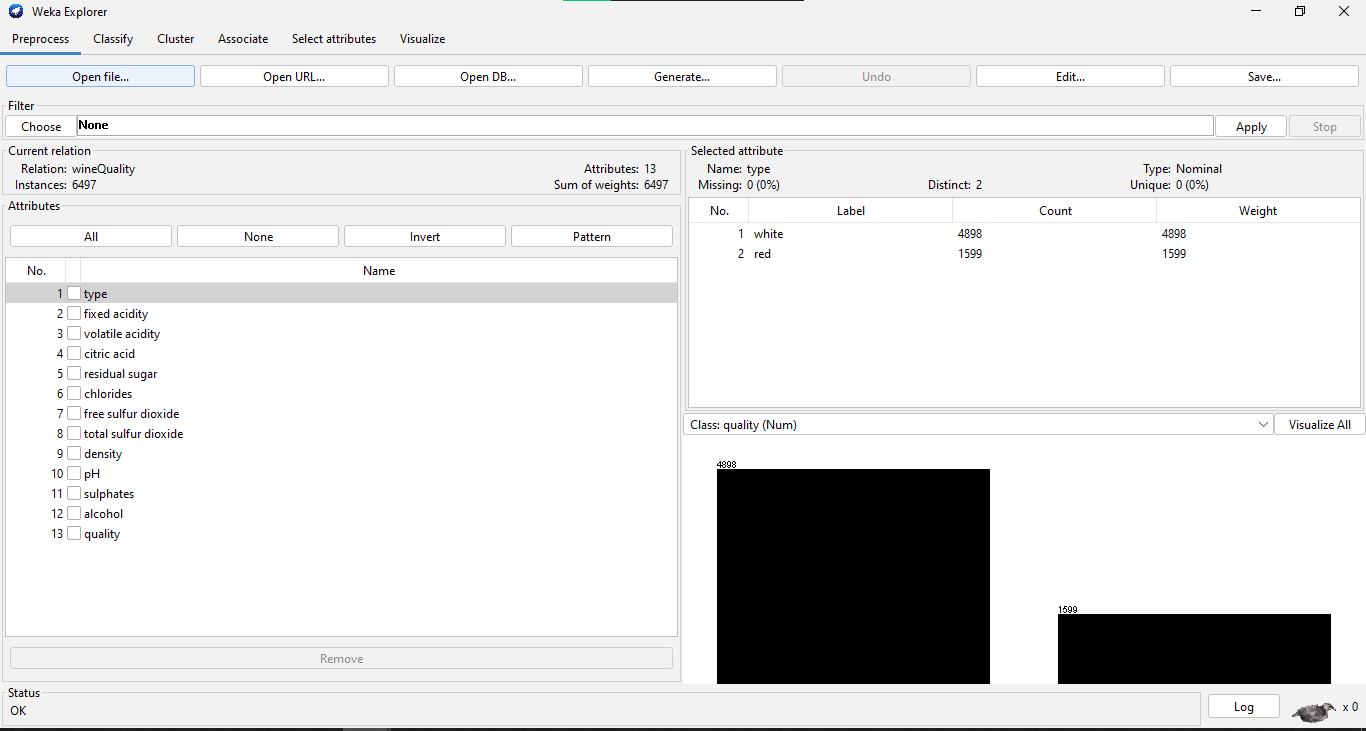
* Mayor sensibilidad a los valores atípicos, ya que pueden tener mayor impacto si es que la característica no esta estandarizada.
* Problemas de interpretación, puede ser difícil interpretar la importancia de una característica si no están en la misma escala.
* Dificultades en la visualización de los datos, Al graficar datos no estandarizados las diferencias en las escalas pueden dar origen a datos difíciles de interpretar

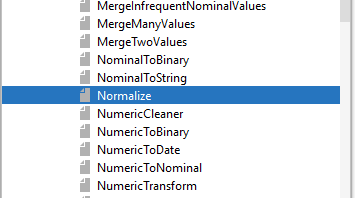
Algunos de los beneficios de estandarizar los datos son:

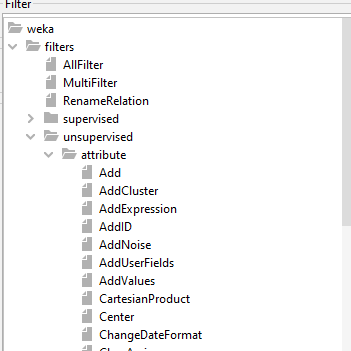
* Los algoritmos basados en la distancia como el k-NN reducen la desproporcionalidad que puedan existir en las características
* Convergencia rápida, algunos algoritmos de optimización convergen mucho más rápido cuando las características están en la misma escala, como en los algoritmos de descenso de gradiente.

**Preprocesamiento N°3 Normalización**

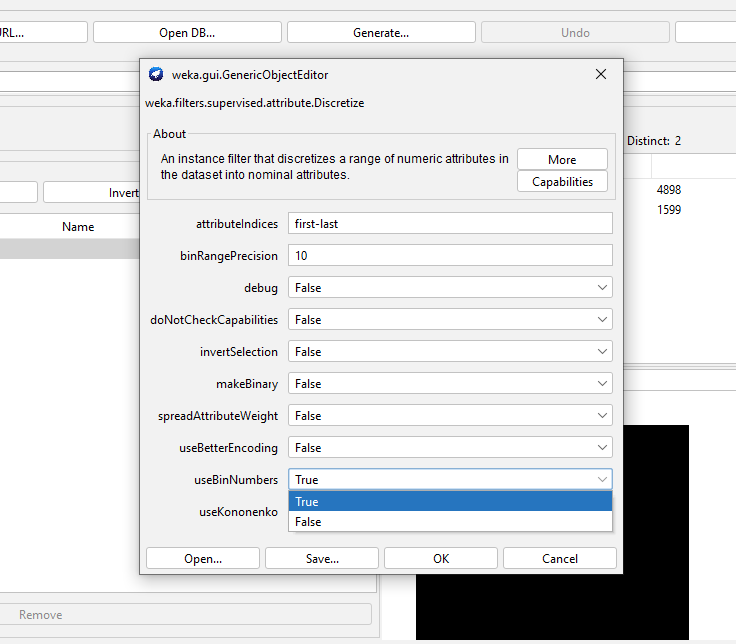
**Paso 1. Primero cargamos nuestro dataset al entorno de WEKA**

****

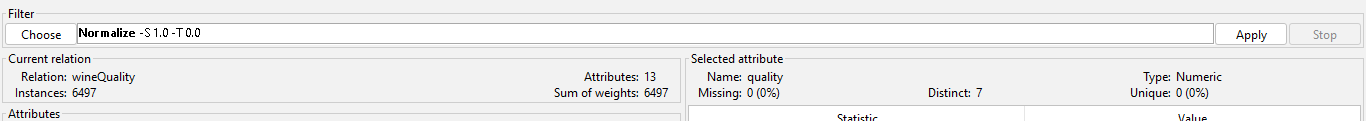
**Paso 2. Después de haber cargado el dataset vamos a la opción de *Choose/filters/unsupervised/Normalize***

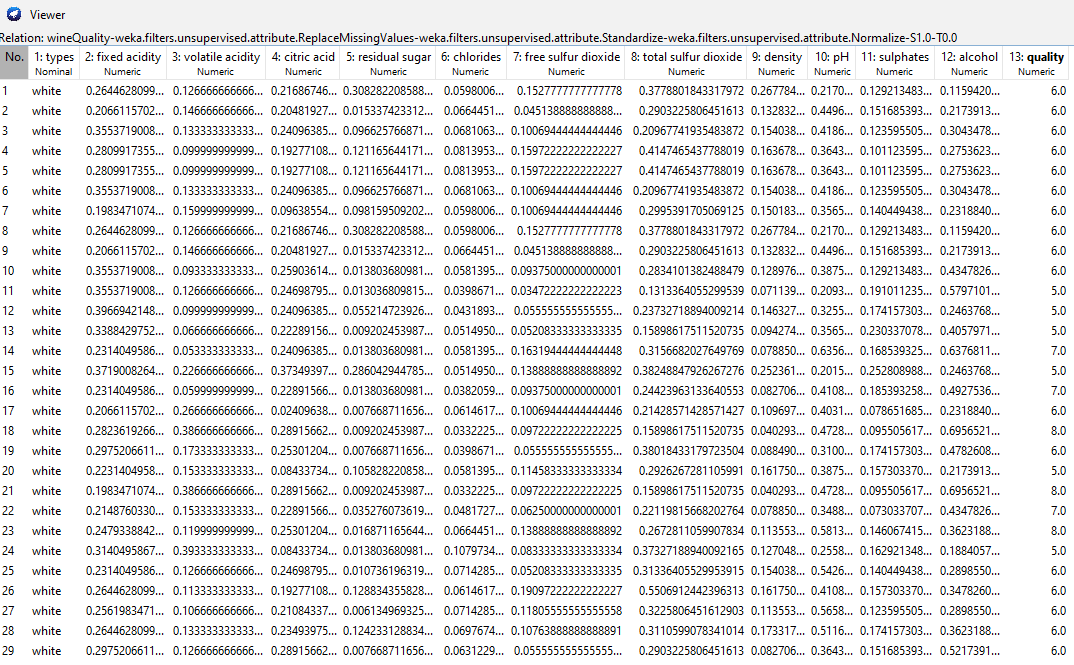
****

**Paso 3.** Luego hacemos clic derecho sobre la opción elegida para poder ver las opciones que nos ofrece weka

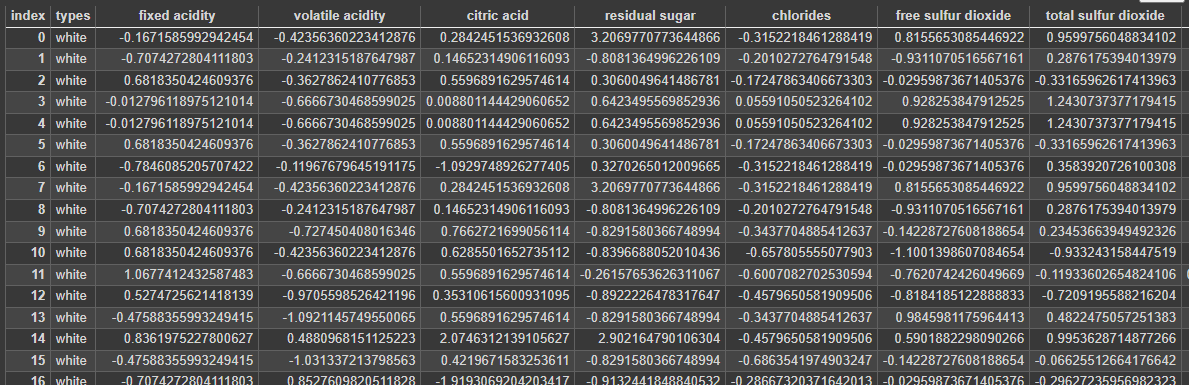
****

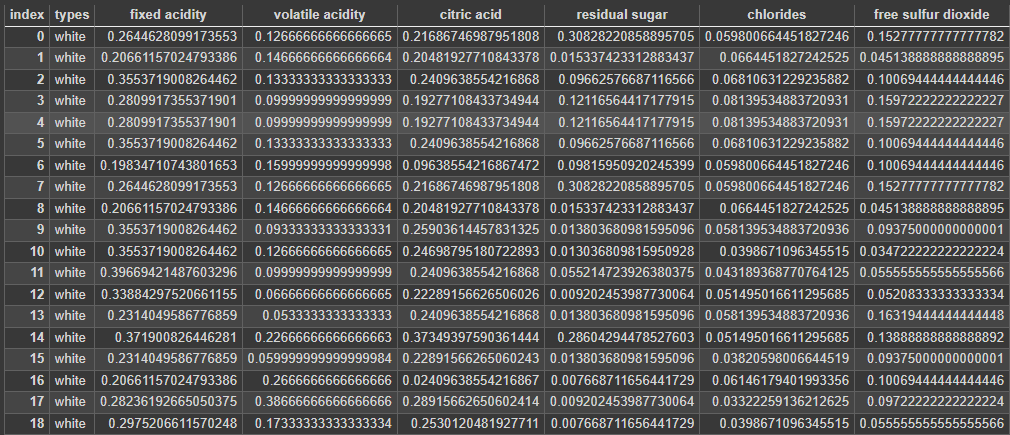
**Paso 4. Finalmente apretamos la opción de Apply para poder ver los cambios**



**Paso 5. Damos click en la pestaña de “Edit” para poder observar los cambios**

**Compararemos la técnica aplicada con el resultado obtenido en Python**

**Antes**

**Despues**

**Como podemos evidenciar obtenemos los mismos resultados**

**Normalización**

La normalización es una técnica de preprocesamiento que ajusta los valores de las características de un conjunto de datos para que se encuentren dentro de un rango especifico, generalmente este es de 0 y 1.

Los problemas que se tienen con características no normalizadas son las siguientes:

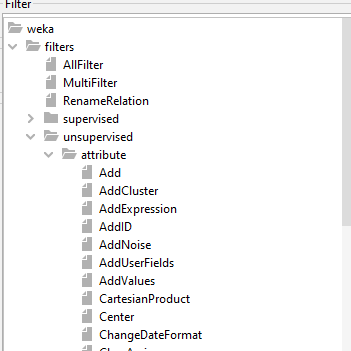
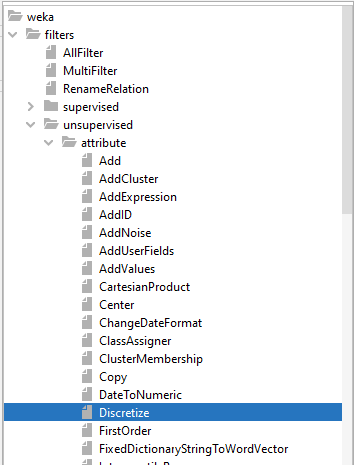
* Diferencia de magnitudes y rangos de las características, ya que puede ocasionar que algunos algoritmos sean menos efectivos y eficientes.
* Problemas de interpretación, resulta mas difíciles interpretar características no normalizadas

Algunos de los beneficios de aplicar la normalización son los siguientes:

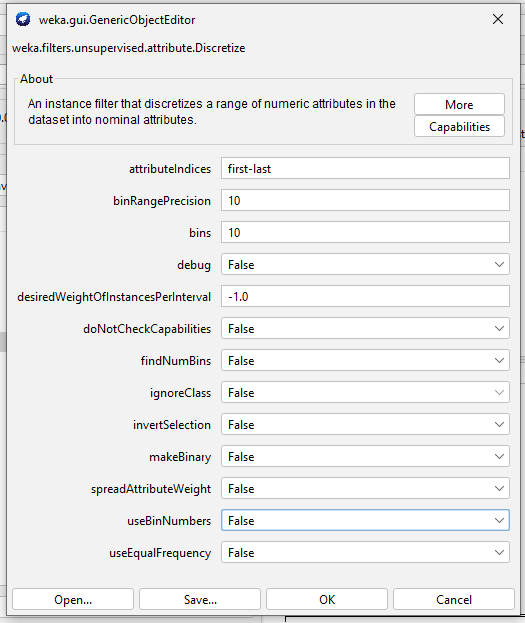
* Reducción de valores atípicos, ya que estos llegaran a tener menor influencia en el modelo después de haberse realizada la normalización
* Aceleración del tiempo de entrenamiento, algunos algoritmos pueden converger más rápido cuando las características están normalizadas

**Preprocesamiento N°4: Discretización**

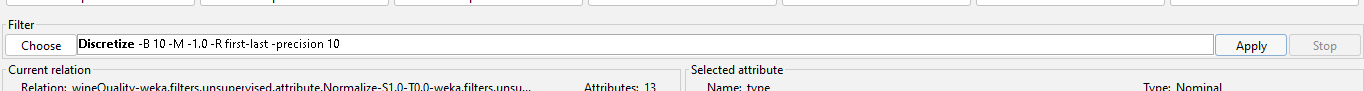
**Paso N°1. Vamos a la opción de *Choose/filters/unsupervised/Discretize***

****

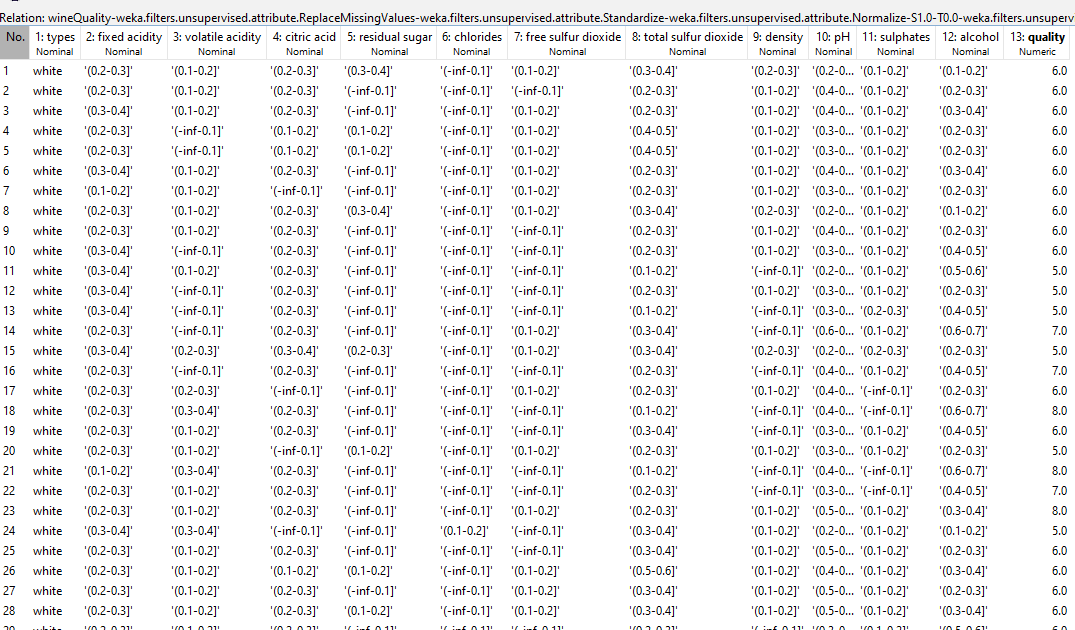
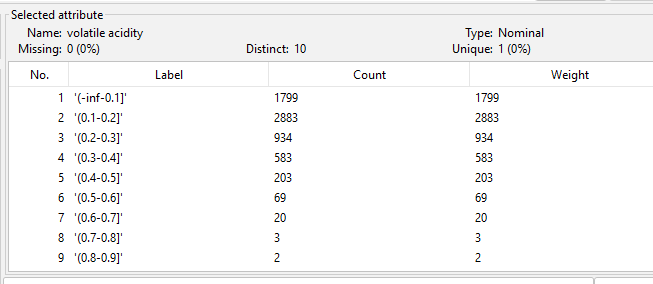
**Paso 2. Luego hacemos click derecho sobre la opción elegida y escogemos el número de rango de precisión a 10.**

****

**Paso 3. Finalmente apretamos la opción de Apply para poder ver los cambios**



**Paso 4. Damos click en la pestaña de “Edit” para poder observar los cambios**

**Discretizacion**

Su principal objetivo es transformar un conjunto de atributos continuos en discretos, asociando valores categóricos a intervalos y transformando así datos cuantitativos en datos cualitativos. Los problemas más comunes que ocurren cuando no se discretizan los datos son los siguientes

* Dificultad en la identificación de patrones: En algunos casos, si las variables son continuas, los patrones pueden ser más difíciles de identificar, especialmente para modelos que funcionan mejor con variables discretas.
* Sensibilidad al ruido: Las variables continuas pueden estar sujetas a ruido en los datos. Esto puede hacer que los modelos sean más sensibles a pequeñas variaciones en los valores.
* Interpretación más difícil: Interpretar los resultados de un modelo que trabaja con variables continuas puede ser más complicado que si se utilizan variables discretas.

Algunos de los beneficios de discretizar son:

* Reducción de complejidad: En algunos casos, trabajar con variables discretas puede simplificar el análisis y hacer que los modelos sean más fáciles de entender y de comunicar.
* Requisitos de algunos modelos: Algunos algoritmos de aprendizaje automático, como árboles de decisión o reglas de asociación, requieren que las variables sean discretas.
* Control del ruido y redundancia: La discretización puede ayudar a reducir el ruido en los datos, ya que se agrupan valores similares. También puede eliminar redundancias al agrupar valores cercanos.