Nina	Huacani	Steve Brandom	N
A. Paterno	**A. 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

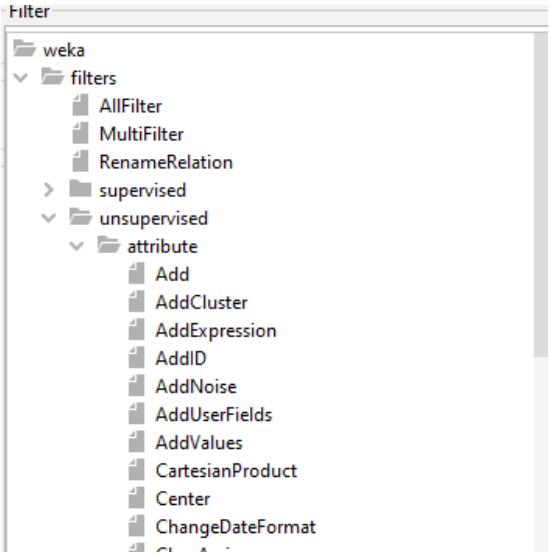
Paso 1. Abrimos el dataset y damos click a "Edit" para poder ver el dataset

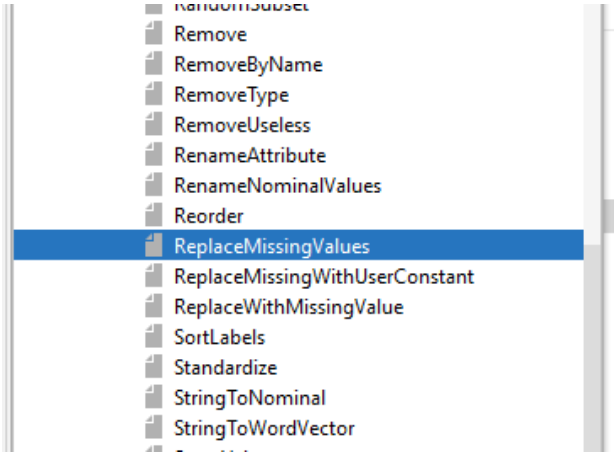
Observamos campos vacíos

Relation: wineQuality

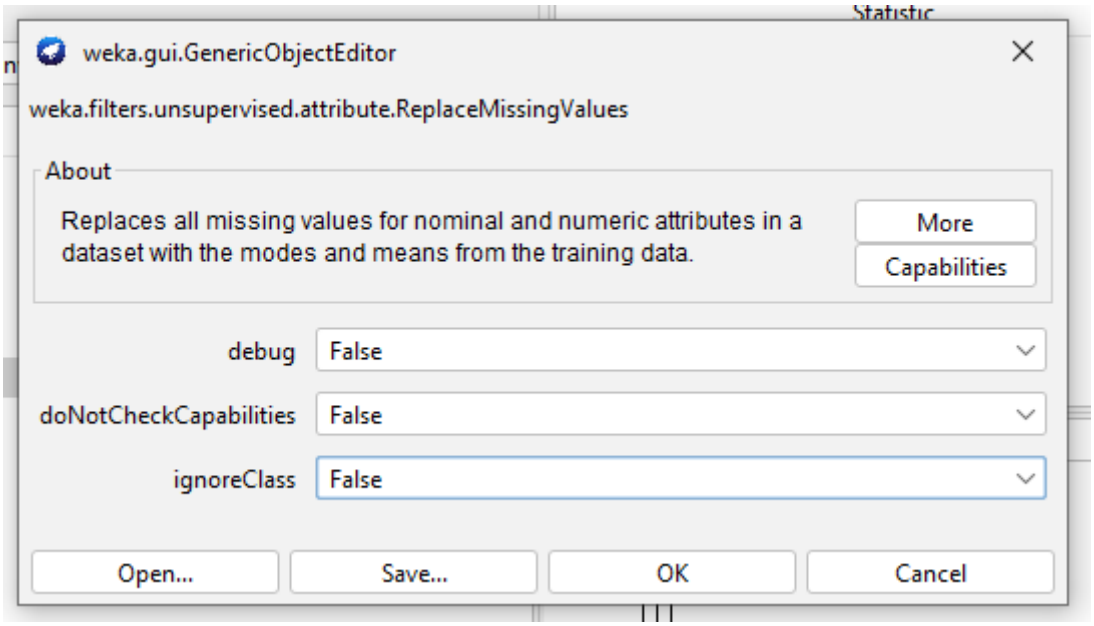
No.	1: types Nominal	2: fixed acidity Numeric	3: volatile acidity Numeric	4: citric acid Numeric	5: residual sugar Numeric	6: chlorides Numeric	7: free sulfur dioxide Numeric	8: total sulfur dioxide Numeric	9: density Numeric	10: pH Numeric	11: sulphates Numeric	12: alcohol Numeric	13: **quality** Numeric
13	white	7.9	0.18	0.37	1.2	0.04	16.0	75.0	0.992	3.18	0.63	10.8	5.0
14	white	6.6	0.16	0.4	1.5	0.044	48.0	143.0	0.9912	3.54	0.52	12.4	7.0
15	white	8.3	0.42	0.62	19.25	0.04	41.0	172.0	1.0002	2.98	0.67	9.7	5.0
16	white	6.6	0.17	0.38	1.5	0.032	28.0	112.0	0.9914	3.25	0.55	11.4	7.0
17	white	6.3	0.48	0.04	1.1	0.046	30.0	99.0	0.9928	3.24	0.36	9.6	6.0
18	white		0.66	0.48	1.2	0.029	29.0	75.0	0.9892	3.33	0.39	12.8	8.0
19	white	7.4	0.34	0.42	1.1	0.033	17.0	171.0	0.9917	3.12	0.53	11.3	6.0
20	white	6.5	0.31	0.14	7.5	0.044	34.0	133.0	0.9955	3.22	0.5	9.5	5.0
21	white	6.2	0.66	0.48	1.2	0.029	29.0	75.0	0.9892	3.33	0.39	12.8	8.0
22	white	6.4	0.31	0.38	2.9	0.038	19.0	102.0	0.9912	3.17	0.35	11.0	7.0
23	white	6.8	0.26	0.42	1.7	0.049	41.0	122.0	0.993	3.47	0.48	10.5	8.0
24	white	7.6	0.67	0.14	1.5	0.074	25.0	168.0	0.9937	3.05	0.51	9.3	5.0
25	white	6.6	0.27	0.41	1.3	0.052	16.0	142.0	0.9951	3.42	0.47	10.0	6.0
26	white	7.0	0.25	0.32	9.0	0.046	56.0	245.0	0.9955	3.25	0.5	10.4	6.0
27	white	6.9	0.24	0.35	1.0	0.052	35.0	146.0	0.993	3.45	0.44	10.0	6.0
28	white	7.0	0.28	0.39	8.7	0.051	32.0	141.0	0.9961	3.38	0.53	10.5	6.0
29	white	7.4	0.27	0.48	1.1	0.047	17.0	132.0	0.9914	3.19	0.49	11.6	6.0
30	white	7.2	0.32	0.36	2.0	0.033	37.0	114.0	0.9906	3.1	0.71	12.3	7.0
31	white	8.5	0.24	0.39	10.4	0.044	20.0	142.0	0.9974	3.2	0.53	10.0	6.0
32	white	8.3	0.14	0.34	1.1	0.042	7.0	47.0	0.9934	3.47	0.4	10.2	6.0
33	white	7.4	0.25	0.36	2.05	0.05	31.0	100.0	0.992	3.19	0.44	10.8	6.0
34	white	6.2	0.12	0.34		0.045	43.0	117.0	0.9939	3.42	0.51	9.0	6.0
35	white	5.8	0.27	0.2	14.95	0.044	22.0	179.0	0.9962	3.37	0.37	10.2	5.0
36	white	7.3	0.28	0.43	1.7	0.08	21.0	123.0	0.9905	3.19	0.42	12.8	5.0
37	white	6.5	0.39	0.23	5.4	0.051	25.0	149.0	0.9934	3.24	0.35	10.0	5.0

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

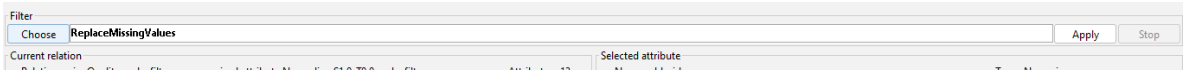




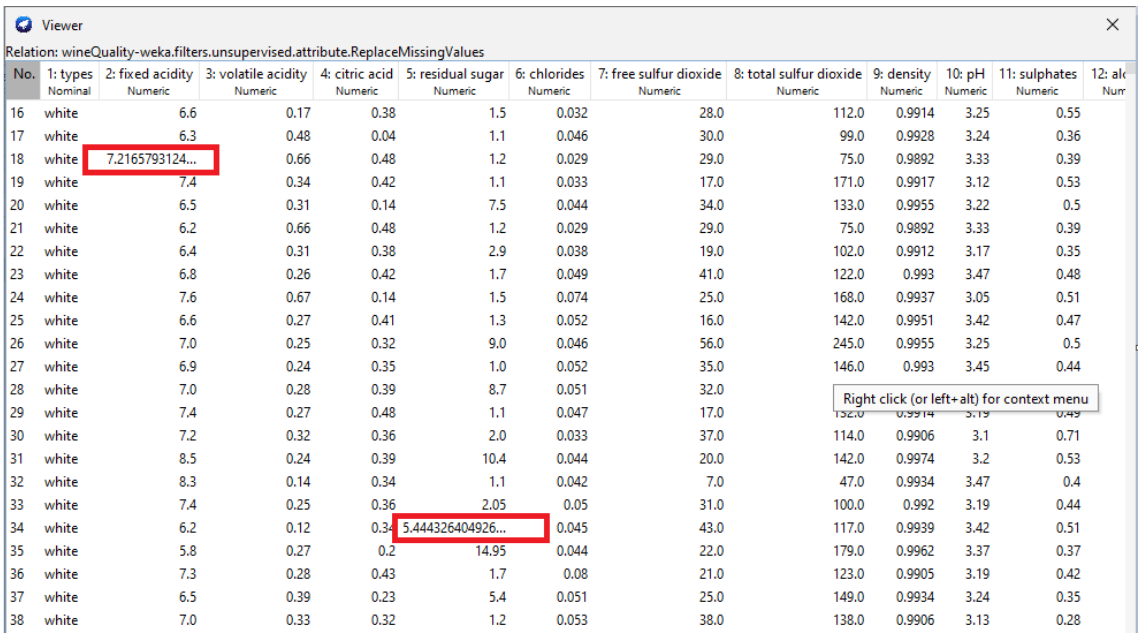
Paso 3. Luego hacemos clic derecho sobre la opción elegida para poder ver las opciones que nos ofrece weka



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



No.	1: types Nominal	2: fixed acidity Numeric	3: volatile acidity Numeric	4: citric acid Numeric	5: residual sugar Numeric	6: chlorides Numeric	7: free sulfur dioxide Numeric	8: total sulfur dioxide Numeric	9: density Numeric	10: pH Numeric	11: sulphates Numeric	12: al Num
16	white	6.6	0.17	0.38	1.5	0.032	28.0	112.0	0.9914	3.25	0.55	
17	white	6.3	0.48	0.04	1.1	0.046	30.0	99.0	0.9928	3.24	0.36	
18	white	7.2165793124...	0.66	0.48	1.2	0.029	29.0	75.0	0.9892	3.33	0.39	
19	white	7.4	0.34	0.42	1.1	0.033	17.0	171.0	0.9917	3.12	0.53	
20	white	6.5	0.31	0.14	7.5	0.044	34.0	133.0	0.9955	3.22	0.5	
21	white	6.2	0.66	0.48	1.2	0.029	29.0	75.0	0.9892	3.33	0.39	
22	white	6.4	0.31	0.38	2.9	0.038	19.0	102.0	0.9912	3.17	0.35	
23	white	6.8	0.26	0.42	1.7	0.049	41.0	122.0	0.993	3.47	0.48	
24	white	7.6	0.67	0.14	1.5	0.074	25.0	168.0	0.9937	3.05	0.51	
25	white	6.6	0.27	0.41	1.3	0.052	16.0	142.0	0.9951	3.42	0.47	
26	white	7.0	0.25	0.32	9.0	0.046	56.0	245.0	0.9955	3.25	0.5	
27	white	6.9	0.24	0.35	1.0	0.052	35.0	146.0	0.993	3.45	0.44	
28	white	7.0	0.28	0.39	8.7	0.051	32.0					
29	white	7.4	0.27	0.48	1.1	0.047	17.0					
30	white	7.2	0.32	0.36	2.0	0.033	37.0	114.0	0.9906	3.1	0.71	
31	white	8.5	0.24	0.39	10.4	0.044	20.0	142.0	0.9974	3.2	0.53	
32	white	8.3	0.14	0.34	1.1	0.042	7.0	47.0	0.9934	3.47	0.4	
33	white	7.4	0.25	0.36	2.05	0.05	31.0	100.0	0.992	3.19	0.44	
34	white	6.2	0.12	0.34	5.444326404926...	0.045	43.0	117.0	0.9939	3.42	0.51	
35	white	5.8	0.27	0.2	14.95	0.044	22.0	179.0	0.9962	3.37	0.37	
36	white	7.3	0.28	0.43	1.7	0.08	21.0	123.0	0.9905	3.19	0.42	
37	white	6.5	0.39	0.23	5.4	0.051	25.0	149.0	0.9934	3.24	0.35	
38	white	7.0	0.33	0.32	1.2	0.053	38.0	138.0	0.9906	3.13	0.28	

Compararemos la técnica aplicada con el resultado obtenido en Python

Antes

16	white	6.3	0.48	0.04	1.1	0.046
17	white	NaN	0.66	0.48	1.2	0.029
18	white	7.4	0.34	0.42	1.1	0.033
19	white	6.5	0.31	0.14	7.5	0.044
20	white	6.2	0.66	0.48	1.2	0.029
21	white	6.4	0.31	0.38	2.9	0.038
22	white	6.8	0.26	0.42	1.7	0.049
23	white	7.6	0.67	0.14	1.5	0.074
24	white	6.6	0.27	0.41	1.3	0.052
25	white	7.0	0.25	0.32	9.0	0.046
26	white	6.9	0.24	0.35	1.0	0.052
27	white	7.0	0.28	0.39	8.7	0.051
28	white	7.4	0.27	0.48	1.1	0.047
29	white	7.2	0.32	0.36	2.0	0.033
30	white	8.5	0.24	0.39	10.4	0.044
31	white	8.3	0.14	0.34	1.1	0.042
32	white	7.4	0.25	0.36	2.05	0.05
33	white	6.2	0.12	0.34	NaN	0.045

Después

17	white	7.2165793124710955	0.66	0.48	1.2	0.029
18	white	7.4	0.34	0.42	1.1	0.033
19	white	6.5	0.31	0.14	7.5	0.044
20	white	6.2	0.66	0.48	1.2	0.029
21	white	6.4	0.31	0.38	2.9	0.038
22	white	6.8	0.26	0.42	1.7	0.049
23	white	7.6	0.67	0.14	1.5	0.074
24	white	6.6	0.27	0.41	1.3	0.052
25	white	7.0	0.25	0.32	9.0	0.046
26	white	6.9	0.24	0.35	1.0	0.052
27	white	7.0	0.28	0.39	8.7	0.051
28	white	7.4	0.27	0.48	1.1	0.047
29	white	7.2	0.32	0.36	2.0	0.033
30	white	8.5	0.24	0.39	10.4	0.044
31	white	8.3	0.14	0.34	1.1	0.042
32	white	7.4	0.25	0.36	2.05	0.05
33	white	6.2	0.12	0.34	5.444326404926867	0.045
34	white	5.8	0.27	0.2	14.95	0.044

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

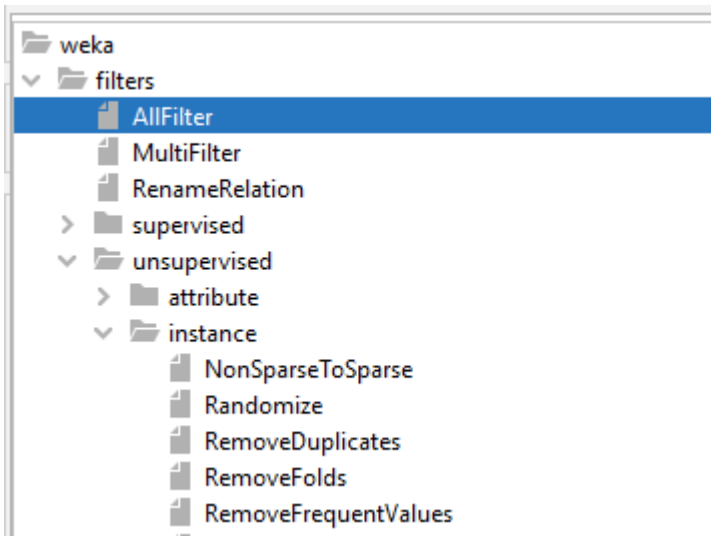
Existe una amplia variedad de métodos de imputación como, por ejemplo: la sustitución por medias, medianas o modas, el vecindario K más cercano, procedimientos de máxima verosimilitud entre otras. Algunos de los beneficios de utilizar la imputación de valores faltantes son las siguientes.

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.

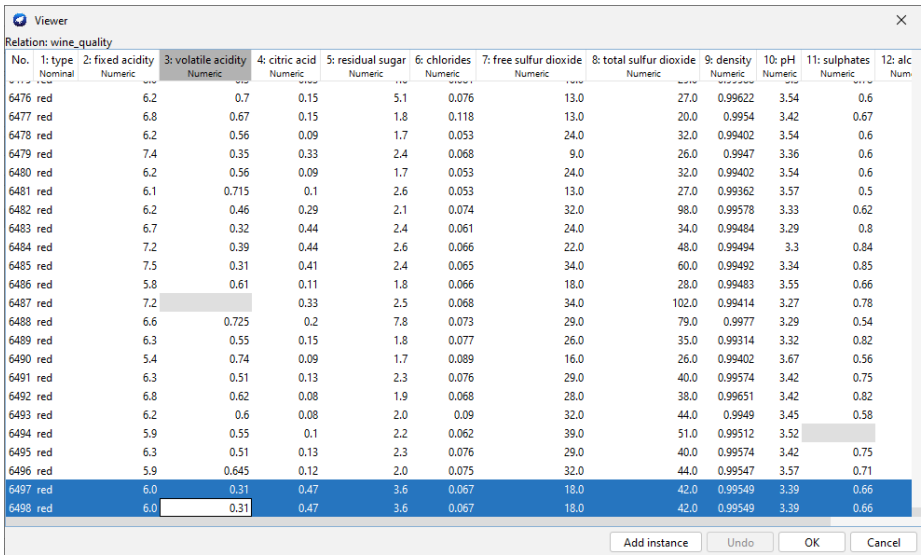
Consistencia de los datos: Ayuda a mantener la consistencia de un conjunto de datos ya que es importante mantener el mismo número de muestras en todas las categorías para que el análisis sea el más optimo posible.

Preprocesamiento N°2 Eliminar valores duplicados

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

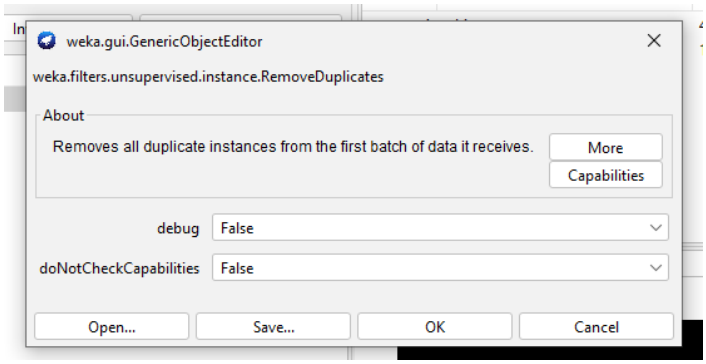


Paso 2. En la pestaña "Edit" podemos verificar que existen dos instancias repetidas



No.	1: type Nominal	2: fixed acidity Numeric	3: volatile acidity Numeric	4: citric acid Numeric	5: residual sugar Numeric	6: chlorides Numeric	7: free sulfur dioxide Numeric	8: total sulfur dioxide Numeric	9: density Numeric	10: pH Numeric	11: sulphates Numeric	12: alc Num
6476	red	6.2	0.7	0.15	5.1	0.076	13.0	27.0	0.99622	3.54	0.6	
6477	red	6.8	0.67	0.15	1.8	0.118	13.0	20.0	0.9954	3.42	0.67	
6478	red	6.2	0.56	0.09	1.7	0.053	24.0	32.0	0.99402	3.54	0.6	
6479	red	7.4	0.35	0.33	2.4	0.068	9.0	26.0	0.9947	3.36	0.6	
6480	red	6.2	0.56	0.09	1.7	0.053	24.0	32.0	0.99402	3.54	0.6	
6481	red	6.1	0.715	0.1	2.6	0.053	13.0	27.0	0.99362	3.57	0.5	
6482	red	6.2	0.46	0.29	2.1	0.074	32.0	98.0	0.99578	3.33	0.62	
6483	red	6.7	0.32	0.44	2.4	0.061	24.0	34.0	0.99484	3.29	0.8	
6484	red	7.2	0.39	0.44	2.6	0.066	22.0	48.0	0.99494	3.3	0.84	
6485	red	7.5	0.31	0.41	2.4	0.065	34.0	60.0	0.99492	3.34	0.85	
6486	red	5.8	0.61	0.11	1.8	0.066	18.0	28.0	0.99483	3.55	0.66	
6487	red	7.2		0.33	2.5	0.068	34.0	102.0	0.99414	3.27	0.78	
6488	red	6.6	0.725	0.2	7.8	0.073	29.0	79.0	0.9977	3.29	0.54	
6489	red	6.3	0.55	0.15	1.8	0.077	26.0	35.0	0.99314	3.32	0.82	
6490	red	5.4	0.74	0.09	1.7	0.089	16.0	26.0	0.99402	3.67	0.56	
6491	red	6.3	0.51	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	
6492	red	6.8	0.62	0.08	1.9	0.068	28.0	38.0	0.99651	3.42	0.82	
6493	red	6.2	0.6	0.08	2.0	0.09	32.0	44.0	0.9949	3.45	0.58	
6494	red	5.9	0.55	0.1	2.2	0.062	39.0	51.0	0.99512	3.52		
6495	red	6.3	0.51	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	
6496	red	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	
6497	red	6.0	0.31	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	
6498	red	6.0	0.31	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	

Paso 3. Configuramos la técnica de preprocesamiento y la aplicamos



Paso 4. Como podemos verificar se eliminó la instancia duplicada

Antes

No.	Label	Count	Weight
1	white	4898	4898
2	red	1600	1600

Despues

1	white	3970	3970
2	red	1359	1359

Comparacion de la técnica de preprocesamiento con Python

Antes

	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	white	7.0	0.270	0.36	20.7	0.045	45.0	170.0	1.00100	3.00	0.450000	8.8	6
1	white	6.3	0.300	0.34	1.6	0.049	14.0	132.0	0.99400	3.30	0.490000	9.5	6
2	white	8.1	0.280	0.40	6.9	0.050	30.0	97.0	0.99510	3.26	0.440000	10.1	6
3	white	7.2	0.230	0.32	8.5	0.058	47.0	186.0	0.99560	3.19	0.400000	9.9	6
4	white	7.2	0.230	0.32	8.5	0.058	47.0	186.0	0.99560	3.19	0.400000	9.9	6
...	...	...	...	...	...	...	...	...	...	...	...	...	...
6493	red	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.531235	11.2	6
6494	red	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.750000	11.0	6
6495	red	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.710000	10.2	5
6496	red	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.660000	11.0	6
6497	red	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.660000	11.0	6

6498 rows × 13 columns

Despues

	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	white	7.0	0.270	0.36	20.7	0.045	45.0	170.0	1.00100	3.00	0.450000	8.8	6
1	white	6.3	0.300	0.34	1.6	0.049	14.0	132.0	0.99400	3.30	0.490000	9.5	6
2	white	8.1	0.280	0.40	6.9	0.050	30.0	97.0	0.99510	3.26	0.440000	10.1	6
3	white	7.2	0.230	0.32	8.5	0.058	47.0	186.0	0.99560	3.19	0.400000	9.9	6
6	white	6.2	0.320	0.16	7.0	0.045	30.0	136.0	0.99490	3.18	0.470000	9.6	6
...	...	...	...	...	...	...	...	...	...	...	...	...	...
6491	red	6.8	0.620	0.08	1.9	0.068	28.0	38.0	0.99651	3.42	0.820000	9.5	6
6492	red	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.580000	10.5	5
6493	red	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.531235	11.2	6
6495	red	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.710000	10.2	5
6496	red	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.660000	11.0	6

5329 rows × 13 columns

Esta técnica de preprocesamiento se aplica para poder eliminar instancias duplicadas. Esto es especialmente útil cuando se trabaja con conjuntos de datos grandes y se quiere asegurar de que cada entrada sea única. La presencia de duplicados puede afectar negativamente a los análisis y modelos de aprendizaje automatico, ya que podría introducir sesgos o distorsiones en los resultados.

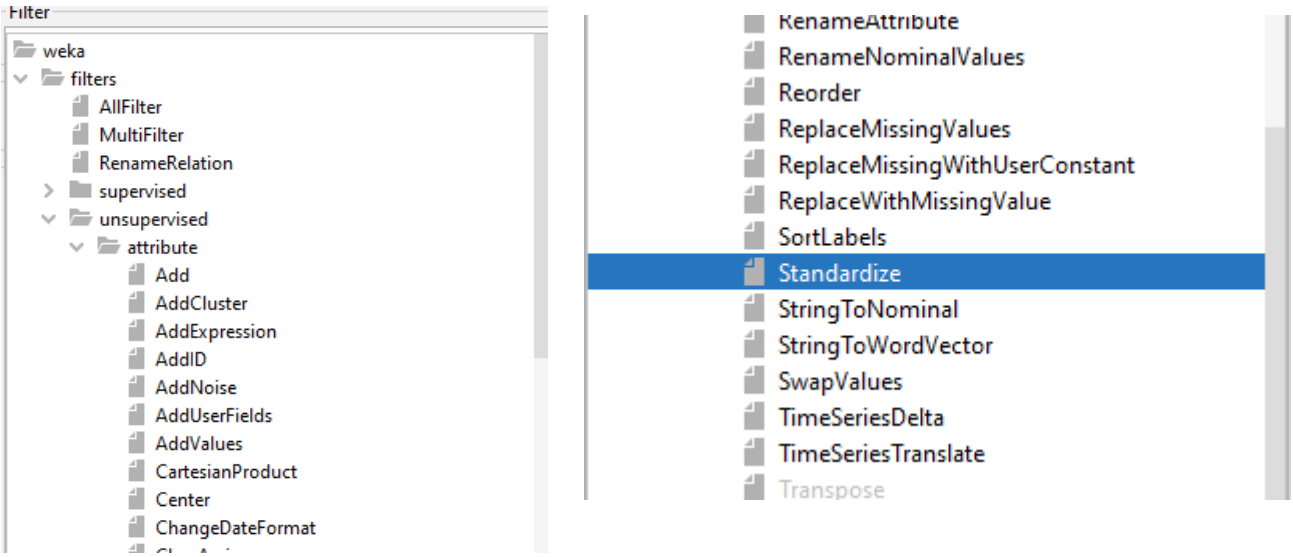
Algunas razones para utilizarlo:

Datos de baja calidad: Los conjuntos de datos pueden contener registros duplicados debido a errores de entrada o problemas en la recopilación de datos. Eliminar duplicados ayuda a limpiar el conjunto de datos y mejorar la calidad de los datos.

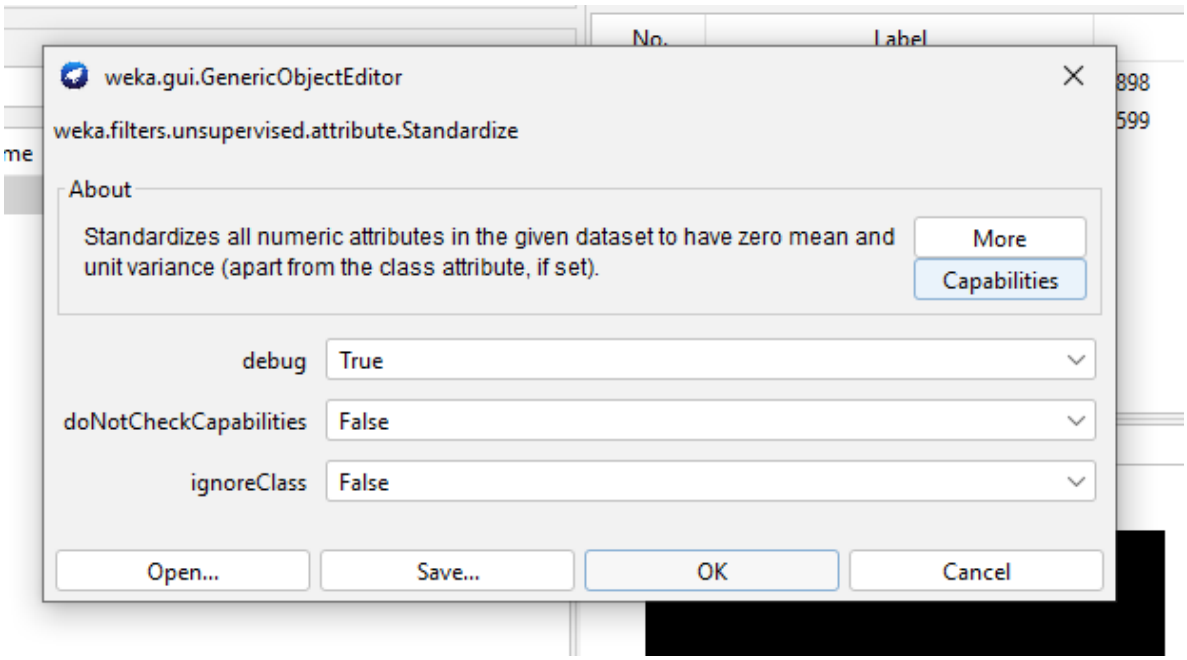
Modelos de aprendizaje automatico: Algunos algoritmos de aprendizaje automatico pueden ser sensibles a la presencia de duplicados. La existencia de registros duplicados puede conducir a un sobreajuste y afectar la capacidad del modelo para generalizar correctamente a nuevos datos.

Preprocesamiento N°3: Estandarización

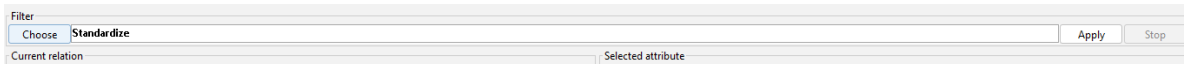
Paso 1. Vamos a la opción de *Choose/filters/unsupervised/Standarize*



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

Viewer

Relation: wineQuality-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Standardize

No.	1: types Nominal	2: fixed acidity Numeric	3: volatile acidity Numeric	4: citric acid Numeric	5: residual sugar Numeric	6: chlorides Numeric	7: free sulfur dioxide Numeric	8: total sulfur dioxide Numeric	9: density Numeric	10: pH Numeric	11: sulphates Numeric	12: alcohol Numeric	13: **quality** Numeric
1	white	-0.167145734...	-0.42353100411...	0.28422327...	3.206730263412...	-0.3151975...	0.8155025413680453	0.9599017236653947	2.102051...	-1.359...	-0.54591708...	-1.418449...	6.0
2	white	-0.707372835...	-0.24121295321...	0.14651187...	-0.80807430417...	-0.2010118...	-0.9310353922075265	0.2875954039075109	-0.23231...	0.5080...	-0.27704259...	-0.831551...	6.0
3	white	0.6817825673...	-0.36275832048...	0.55964608...	0.305981413529...	-0.1724653...	-0.02959645874916...	-0.33163410113264524	0.134514...	0.2589...	-0.61313570...	-0.328495...	6.0
4	white	-0.012795134...	-0.66662173866...	0.00880046...	0.642300120762...	0.0559062...	0.9281824080503404	1.242978068826609	0.301255...	-0.176...	-0.88201019...	-0.496180...	6.0
5	white	-0.012795134...	-0.66662173866...	0.00880046...	0.642300120762...	0.0559062...	0.9281824080503404	1.242978068826609	0.301255...	-0.176...	-0.88201019...	-0.496180...	6.0
6	white	0.6817825673...	-0.36275832048...	0.55964608...	0.305981413529...	-0.1724653...	-0.02959645874916...	-0.33163410113264524	0.134514...	0.2589...	-0.61313570...	-0.328495...	6.0
7	white	-0.784548135...	-0.11966758593...	-1.0928907...	0.327001332731...	-0.3151975...	-0.02959645874916...	0.35836449019781447	0.067818...	-0.239...	-0.41147983...	-0.747708...	6.0
8	white	-0.167145734...	-0.42353100411...	0.28422327...	3.206730263412...	-0.3151975...	0.8155025413680453	0.9599017236653947	2.102051...	-1.359...	-0.54591708...	-1.418449...	6.0
9	white	-0.707372835...	-0.24121295321...	0.14651187...	-0.80807430417...	-0.2010118...	-0.9310353922075265	0.2875954039075109	-0.23231...	0.5080...	-0.27704259...	-0.831551...	6.0
10	white	0.6817825673...	-0.72739442229...	0.76621319...	-0.82909422338...	-0.3437440...	-0.1422763254314618	0.23451858918978324	-0.29901...	0.0099...	-0.54591708...	0.4260871...	6.0
11	white	0.6817825673...	-0.42353100411...	0.62850179...	-0.83960418298...	-0.6577549...	-1.100055192230969	-0.9331713346002255	-1.29945...	-1.421...	0.193487764...	1.2645127...	5.0
12	white	1.0676590682...	-0.66662173866...	0.55964608...	-0.26155640492...	-0.6006620...	-0.7620155921840841	-0.11932684226173455	0.001122...	-0.488...	-0.00816810...	-0.663866...	5.0
13	white	0.5274319670...	-0.97048515684...	0.35307898...	-0.89215398098...	-0.4579298...	-0.8183555255252315	-0.7208640757293148	-0.89927...	-0.239...	0.664018121...	0.2584020...	5.0
14	white	-0.475846935...	-1.09203052411...	0.55964608...	-0.82909422338...	-0.3437440...	0.9845223413914878	0.4822103912058457	-1.16606...	2.0020...	-0.07538672...	1.5998829...	7.0
15	white	0.8361331677...	0.488059250421...	2.07447154...	2.901941434983...	-0.4579298...	0.5901428080034554	0.9952862668105464	1.835267...	-1.484...	0.932892611...	-0.663866...	5.0
16	white	-0.475846935...	-1.03125784047...	0.42193468...	-0.82909422338...	-0.6863013...	-0.1422763254314618	-0.06625002754400688	-1.09936...	0.1967...	0.126269141...	0.7614573...	7.0
17	white	-0.707372835...	0.852695352237...	-1.9191592...	-0.91317390018...	-0.2866511...	-0.02959645874916...	-0.2962495579874935	-0.63249...	0.1344...	-1.15088468...	-0.747708...	6.0
18	white	-2.056363084...	1.946603657686...	1.11049170...	-0.89215398098...	-0.7719407...	-0.08593639209031...	-0.7208640757293148	-1.83302...	0.6947...	-0.94922881...	1.9352531...	8.0
19	white	0.1415554661...	0.001877781332...	0.69735749...	-0.91317390018...	-0.6577549...	-0.7620155921840841	0.9775939952379706	-0.99931...	-0.612...	-0.00816810...	0.6776148...	6.0
20	white	-0.553022235...	-0.18044026957...	-1.2306021...	0.432100928742...	-0.3437440...	0.1957632746154231	0.3052876754800868	0.267907...	0.0099...	-0.20982397...	-0.831551...	5.0
21	white	-0.784548135...	1.946603657686...	1.11049170...	-0.89215398098...	-0.7719407...	-0.08593639209031...	-0.7208640757293148	-1.83302...	0.6947...	-0.94922881...	1.9352531...	8.0
22	white	-0.630197535...	-0.18044026957...	0.42193468...	-0.53481535455...	-0.5150227...	-0.6493357255017891	-0.2431727432697658	-1.16606...	-0.301...	-1.21810330...	0.4260871...	7.0
23	white	-0.321496334...	-0.48430368775...	0.69735749...	-0.78705438497...	-0.2010118...	0.5901428080034554	0.11067268818175201	-0.56579...	1.5662...	-0.34426121...	0.0068743...	8.0
24	white	0.2959060665...	2.007376341322...	-1.2306021...	-0.82909422338...	0.5126493...	-0.31129612545490...	0.9245171805202429	-0.33235...	-1.048...	-0.14260534...	-0.999236...	5.0
25	white	-0.475846935...	-0.42353100411...	0.62850179...	-0.87113406178...	-0.1153724...	-0.8183555255252315	0.4645181196332698	0.134514...	1.2550...	-0.41147983...	-0.412338...	6.0
26	white	-0.167145734...	-0.54507637139...	0.00880046...	0.747399716772...	-0.2866511...	1.4352418081206677	2.2868220916085864	0.267907...	0.1967...	-0.20982397...	-0.076968...	6.0
27	white	-0.244321034...	-0.60584905502...	0.21536757...	-0.93419381939...	-0.1153724...	0.25210320795657054	0.5352872059235734	-0.56579...	1.4417...	-0.61313570...	-0.412338...	6.0
28	white	-0.167145734...	-0.36275832048...	0.49079038...	0.684339959166...	-0.1439189...	0.08308340793312811	0.4468258480606939	0.467995...	1.0060...	-0.00816810...	0.0068743...	6.0

Compararemos la técnica aplicada con el resultado obtenido en Python

Antes

index	types	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	white	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.001	3.0	0.45	8.8	6
1	white	6.3	0.3	0.34	1.6	0.049	14.0	132.0	0.994	3.3	0.49	9.5	6
2	white	8.1	0.28	0.4	6.9	0.05	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.4	9.9	6
4	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.4	9.9	6
5	white	8.1	0.28	0.4	6.9	0.05	30.0	97.0	0.9951	3.26	0.44	10.1	6
6	white	6.2	0.32	0.16	7.0	0.045	30.0	136.0	0.9949	3.18	0.47	9.6	6
7	white	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.001	3.0	0.45	8.8	6
8	white	6.3	0.3	0.34	1.6	0.049	14.0	132.0	0.994	3.3	0.49	9.5	6
9	white	8.1	0.22	0.43	1.5	0.044	28.0	129.0	0.9938	3.22	0.45	11.0	6
10	white	8.1	0.27	0.41	1.45	0.033	11.0	63.0	0.9908	2.99	0.56	12.0	5
11	white	8.6	0.23	0.4	4.2	0.035	17.0	109.0	0.9947	3.14	0.53	9.7	5
12	white	7.9	0.18	0.37	1.2	0.04	16.0	75.0	0.992	3.18	0.63	10.8	5
13	white	6.6	0.16	0.4	1.5	0.044	48.0	143.0	0.9912	3.54	0.52	12.4	7
14	white	8.3	0.42	0.62	19.25	0.04	41.0	172.0	1.0002	2.98	0.67	9.7	5
15	white	6.6	0.17	0.38	1.5	0.032	28.0	112.0	0.9914	3.25	0.55	11.4	7
16	white	6.3	0.48	0.04	1.1	0.046	30.0	99.0	0.9928	3.24	0.36	9.6	6
17	white	7.2165793124710955	0.66	0.48	1.2	0.029	29.0	75.0	0.9892	3.33	0.39	12.8	8
18	white	7.4	0.34	0.42	1.1	0.033	17.0	171.0	0.9917	3.12	0.53	11.3	6

Despues

index	types	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide
0	white	-0.1671585992942454	-0.42356360223412876	0.2842451536932608	3.2069770773644866	-0.3152218461288419	0.8155653085446922	0.9599756048834102
1	white	-0.7074272804111803	-0.2412315187647987	0.14652314906116093	-0.8081364996226109	-0.2010272764791548	-0.9311070516567161	0.2876175394013979
2	white	0.6818350424609376	-0.3627862410776853	0.5596891629574614	0.3060049641486781	-0.17247863406673303	-0.02959873671405376	-0.33165962617413963
3	white	-0.012796118975121014	-0.6666730468599025	0.008801144429060652	0.6423495569852936	0.0559105052326410 2	0.928253847912525	1.2430737377179415
4	white	-0.012796118975121014	-0.6666730468599025	0.008801144429060652	0.6423495569852936	0.0559105052326410 2	0.928253847912525	1.2430737377179415
5	white	0.6818350424609376	-0.3627862410776853	0.5596891629574614	0.3060049641486781	-0.17247863406673303	-0.02959873671405376	-0.33165962617413963
6	white	-0.7846085205707422	-0.11967679645191175	-1.0929748926277405	0.3270265012009665	-0.3152218461288419	-0.02959873671405376	0.3583920726100308
7	white	-0.1671585992942454	-0.42356360223412876	0.2842451536932608	3.2069770773644866	-0.3152218461288419	0.8155653085446922	0.9599756048834102
8	white	-0.7074272804111803	-0.2412315187647987	0.14652314906116093	-0.8081364996226109	-0.2010272764791548	-0.9311070516567161	0.2876175394013979
9	white	0.6818350424609376	-0.727450408016346	0.7662721699056114	-0.8291580366748994	-0.3437704885412637	-0.14228727608188654	0.23453663949492326
10	white	0.6818350424609376	-0.42356360223412876	0.6285501652735112	-0.8396688052010436	-0.657805555077903	-1.1001398607084654	-0.933243158447519
11	white	1.0677412432587483	-0.6666730468599025	0.5596891629574614	-0.26157653626311067	-0.6007082702530594	-0.7620742426049669	-0.11933602654824106
12	white	0.5274725621418139	-0.9705598526421196	0.35310615600931095	-0.8922226478317647	-0.4579650581909506	-0.8184185122888833	-0.7209195588216204
13	white	-0.47588355993249415	-1.0921145749550065	0.5596891629574614	-0.8291580366748994	-0.3437704885412637	0.9845981175964413	0.4822475057251383
14	white	0.8361975227800627	0.4880968151125223	2.0746312139105627	2.902164790106304	-0.4579650581909506	0.5901882298090266	0.9953628714877266
15	white	-0.47588355993249415	-1.031337213798563	0.4219671583253611	-0.8291580366748994	-0.6863541974903247	-0.14228727608188654	-0.06625512664176642
16	white	-0.7074272804111803	0.8527609820511828	-1.9193069204203417	-0.9132441848840532	-0.28667320371642013	-0.02959873671405376	-0.29627235956982323

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:

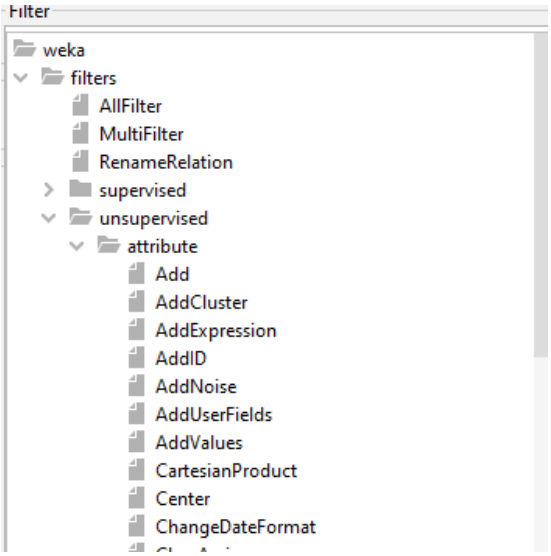
- Mayor sensibilidad a los valores atípicos, ya que pueden tener mayor impacto si es que la característica no esta estandarizada.
- Problemas de interpretación, puede ser difícil interpretar la importancia de una característica si no están en la misma escala.
- Dificultades en la visualización de los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

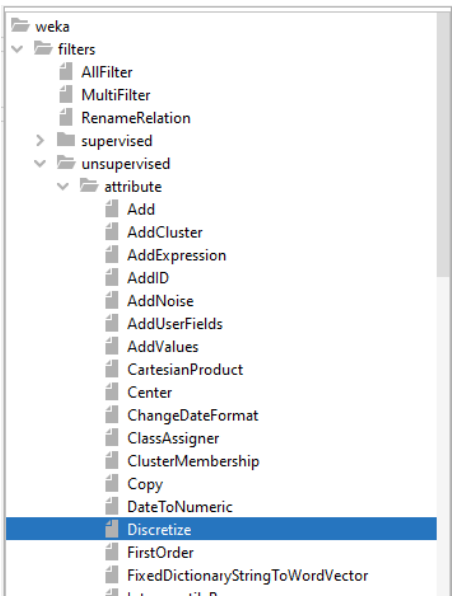
Algunos de los beneficios de estandarizar los datos son:

- 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.

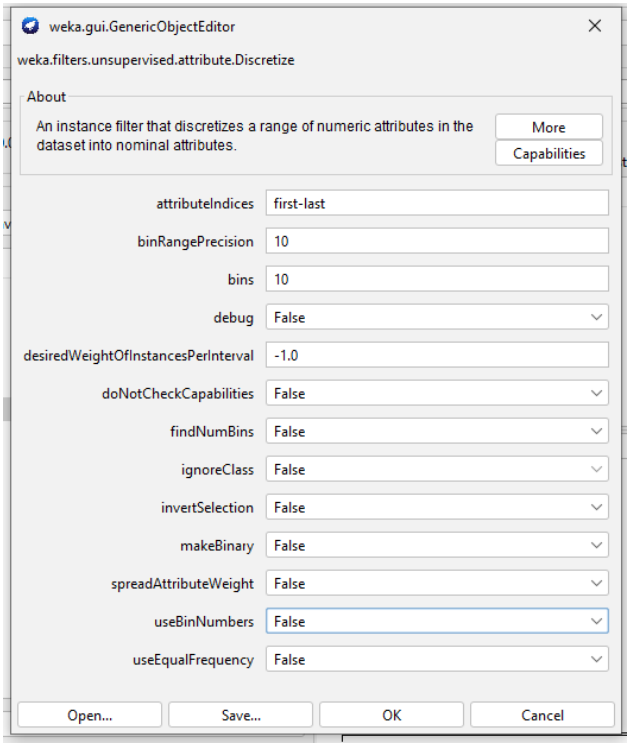
Preprocesamiento N°4: Discretización

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

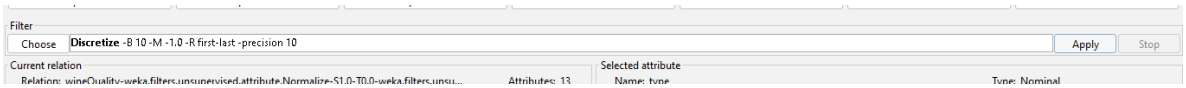




Paso 2. Luego hacemos click derecho sobre la opción elegida y escogemos el número de rango de precisión a 10.



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

Relation: wineQuality-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Standardize-weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0-weka.filters.unsuperv

No.	1: types Nominal	2: fixed acidity Nominal	3: volatile acidity Nominal	4: citric acid Nominal	5: residual sugar Nominal	6: chlorides Nominal	7: free sulfur dioxide Nominal	8: total sulfur dioxide Nominal	9: density Nominal	10: pH Nominal	11: sulphates Nominal	12: alcohol Nominal	13: **quality** Numeric
1	white	'(0.2-0.3]'	'(0.1-0.2]'	'(0.2-0.3]'	'(0.3-0.4]'	'(-inf-0.1]'	'(0.1-0.2]'	'(0.3-0.4]'	'(0.2-0.3]'	'(0.2-0...	'(0.1-0.2]'	'(0.1-0.2]'	6.0
2	white	'(0.2-0.3]'	'(0.1-0.2]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.2-0.3]'	'(0.1-0.2]'	'(0.4-0...	'(0.1-0.2]'	'(0.2-0.3]'	6.0
3	white	'(0.3-0.4]'	'(0.1-0.2]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.1-0.2]'	'(0.2-0.3]'	'(0.1-0.2]'	'(0.4-0...	'(0.1-0.2]'	'(0.3-0.4]'	6.0
4	white	'(0.2-0.3]'	'(-inf-0.1]'	'(0.1-0.2]'	'(0.1-0.2]'	'(-inf-0.1]'	'(0.1-0.2]'	'(0.4-0.5]'	'(0.1-0.2]'	'(0.3-0...	'(0.1-0.2]'	'(0.2-0.3]'	6.0
5	white	'(0.2-0.3]'	'(-inf-0.1]'	'(0.1-0.2]'	'(0.1-0.2]'	'(-inf-0.1]'	'(0.1-0.2]'	'(0.4-0.5]'	'(0.1-0.2]'	'(0.3-0...	'(0.1-0.2]'	'(0.2-0.3]'	6.0
6	white	'(0.3-0.4]'	'(0.1-0.2]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.1-0.2]'	'(0.2-0.3]'	'(0.1-0.2]'	'(0.4-0...	'(0.1-0.2]'	'(0.3-0.4]'	6.0
7	white	'(0.1-0.2]'	'(0.1-0.2]'	'(-inf-0.1]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.1-0.2]'	'(0.2-0.3]'	'(0.1-0.2]'	'(0.3-0...	'(0.1-0.2]'	'(0.2-0.3]'	6.0
8	white	'(0.2-0.3]'	'(0.1-0.2]'	'(0.2-0.3]'	'(0.3-0.4]'	'(-inf-0.1]'	'(0.1-0.2]'	'(0.3-0.4]'	'(0.2-0.3]'	'(0.2-0...	'(0.1-0.2]'	'(0.1-0.2]'	6.0
9	white	'(0.2-0.3]'	'(0.1-0.2]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.2-0.3]'	'(0.1-0.2]'	'(0.4-0...	'(0.1-0.2]'	'(0.2-0.3]'	6.0
10	white	'(0.3-0.4]'	'(-inf-0.1]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.2-0.3]'	'(0.1-0.2]'	'(0.3-0...	'(0.1-0.2]'	'(0.4-0.5]'	6.0
11	white	'(0.3-0.4]'	'(0.1-0.2]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.1-0.2]'	'(-inf-0.1]'	'(0.2-0...	'(0.1-0.2]'	'(0.5-0.6]'	5.0
12	white	'(0.3-0.4]'	'(-inf-0.1]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.2-0.3]'	'(0.1-0.2]'	'(0.3-0...	'(0.1-0.2]'	'(0.2-0.3]'	5.0
13	white	'(0.3-0.4]'	'(-inf-0.1]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.1-0.2]'	'(-inf-0.1]'	'(0.3-0...	'(0.2-0.3]'	'(0.4-0.5]'	5.0
14	white	'(0.2-0.3]'	'(-inf-0.1]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.1-0.2]'	'(0.3-0.4]'	'(-inf-0.1]'	'(0.6-0...	'(0.1-0.2]'	'(0.6-0.7]'	7.0
15	white	'(0.3-0.4]'	'(0.2-0.3]'	'(0.3-0.4]'	'(0.2-0.3]'	'(-inf-0.1]'	'(0.1-0.2]'	'(0.3-0.4]'	'(0.2-0.3]'	'(0.2-0...	'(0.2-0.3]'	'(0.2-0.3]'	5.0
16	white	'(0.2-0.3]'	'(-inf-0.1]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.2-0.3]'	'(-inf-0.1]'	'(0.4-0...	'(0.1-0.2]'	'(0.4-0.5]'	7.0
17	white	'(0.2-0.3]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.1-0.2]'	'(0.2-0.3]'	'(0.1-0.2]'	'(0.4-0...	'(-inf-0.1]'	'(0.2-0.3]'	6.0
18	white	'(0.2-0.3]'	'(0.3-0.4]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.1-0.2]'	'(-inf-0.1]'	'(0.4-0...	'(-inf-0.1]'	'(0.6-0.7]'	8.0
19	white	'(0.2-0.3]'	'(0.1-0.2]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.3-0.4]'	'(-inf-0.1]'	'(0.3-0...	'(0.1-0.2]'	'(0.4-0.5]'	6.0
20	white	'(0.2-0.3]'	'(0.1-0.2]'	'(-inf-0.1]'	'(0.1-0.2]'	'(-inf-0.1]'	'(0.1-0.2]'	'(0.2-0.3]'	'(0.1-0.2]'	'(0.3-0...	'(0.1-0.2]'	'(0.2-0.3]'	5.0
21	white	'(0.1-0.2]'	'(0.3-0.4]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.1-0.2]'	'(-inf-0.1]'	'(0.4-0...	'(-inf-0.1]'	'(0.6-0.7]'	8.0
22	white	'(0.2-0.3]'	'(0.1-0.2]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.2-0.3]'	'(-inf-0.1]'	'(0.3-0...	'(-inf-0.1]'	'(0.4-0.5]'	7.0
23	white	'(0.2-0.3]'	'(0.1-0.2]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.1-0.2]'	'(0.2-0.3]'	'(0.1-0.2]'	'(0.5-0...	'(0.1-0.2]'	'(0.3-0.4]'	8.0
24	white	'(0.3-0.4]'	'(0.3-0.4]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.1-0.2]'	'(-inf-0.1]'	'(0.3-0.4]'	'(0.1-0.2]'	'(0.2-0...	'(0.1-0.2]'	'(0.1-0.2]'	5.0
25	white	'(0.2-0.3]'	'(0.1-0.2]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.3-0.4]'	'(0.1-0.2]'	'(0.5-0...	'(0.1-0.2]'	'(0.2-0.3]'	6.0
26	white	'(0.2-0.3]'	'(0.1-0.2]'	'(0.1-0.2]'	'(0.1-0.2]'	'(-inf-0.1]'	'(0.1-0.2]'	'(0.5-0.6]'	'(0.1-0.2]'	'(0.4-0...	'(0.1-0.2]'	'(0.3-0.4]'	6.0
27	white	'(0.2-0.3]'	'(0.1-0.2]'	'(0.2-0.3]'	'(-inf-0.1]'	'(-inf-0.1]'	'(0.1-0.2]'	'(0.3-0.4]'	'(0.1-0.2]'	'(0.5-0...	'(0.1-0.2]'	'(0.2-0.3]'	6.0
28	white	'(0.2-0.3]'	'(0.1-0.2]'	'(0.2-0.3]'	'(0.1-0.2]'	'(-inf-0.1]'	'(0.1-0.2]'	'(0.3-0.4]'	'(0.1-0.2]'	'(0.5-0...	'(0.1-0.2]'	'(0.3-0.4]'	6.0

Selected attribute

Name: volatile acidity Type: Nominal
Missing: 0 (0%) Distinct: 10 Unique: 1 (0%)

No.	Label	Count	Weight
1	'(-inf-0.1]'	1799	1799
2	'(0.1-0.2]'	2883	2883
3	'(0.2-0.3]'	934	934
4	'(0.3-0.4]'	583	583
5	'(0.4-0.5]'	203	203
6	'(0.5-0.6]'	69	69
7	'(0.6-0.7]'	20	20
8	'(0.7-0.8]'	3	3
9	'(0.8-0.9]'	2	2

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.