Nombre: Alegre Flores Samuel Alejandro

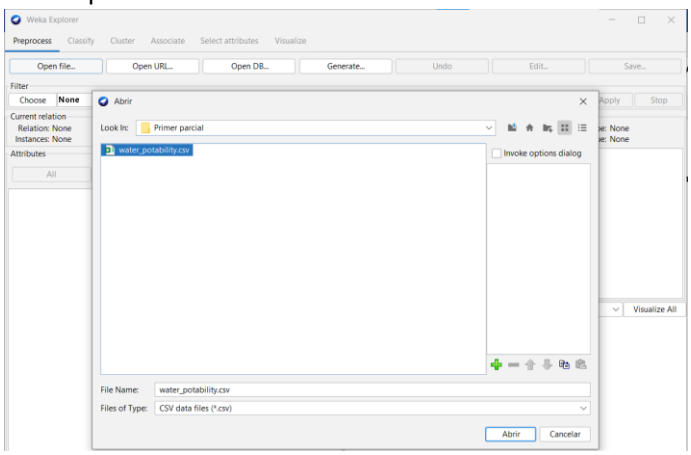
CI: 12391101

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 aplicar estas técnicas.

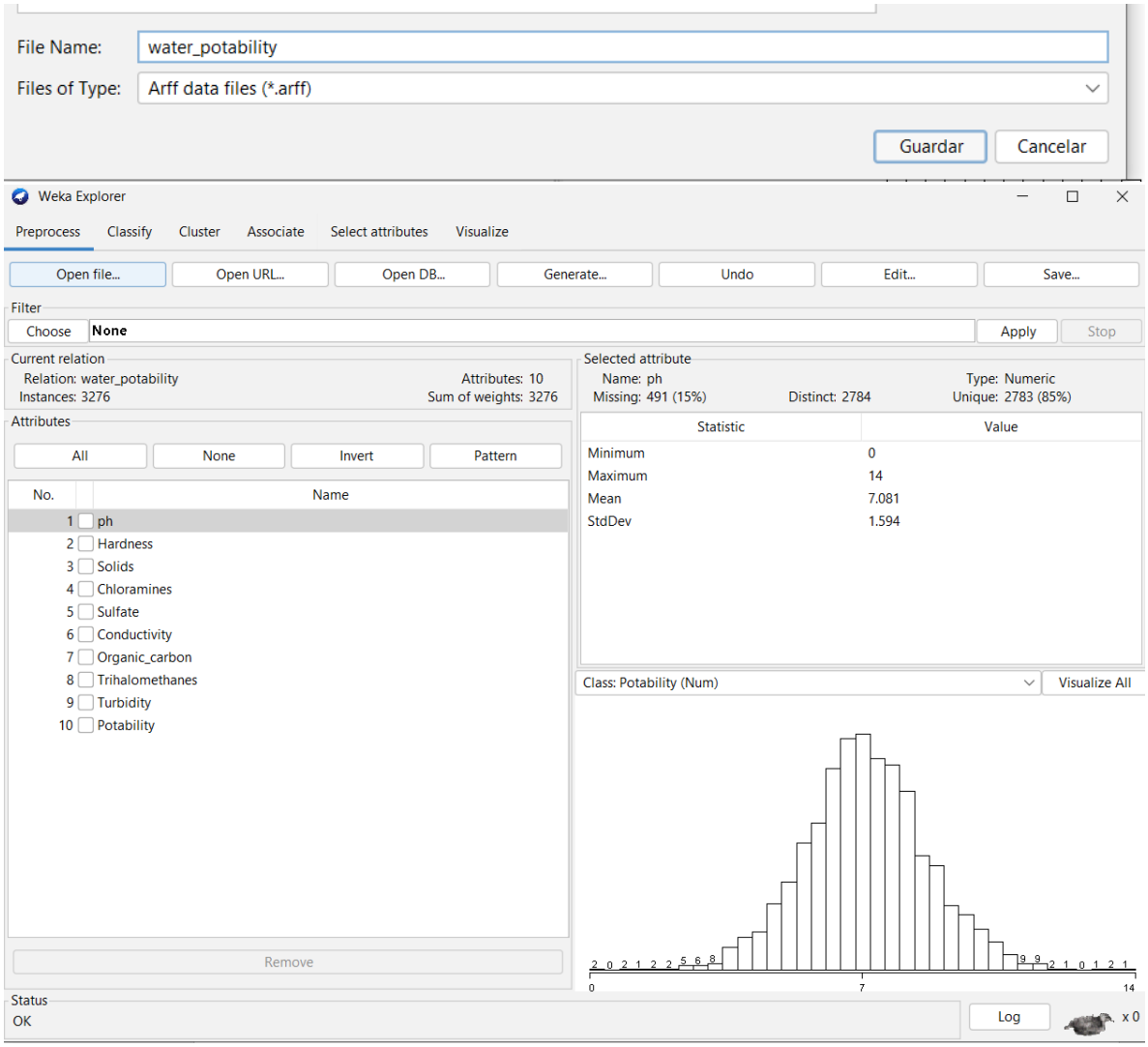
Índice:

Paso 1: Abrimos weka y seleccionamos el dataset en formato .csv

- Entramos a la opción Explorer



- Seleccionamos el archivo y damos click en abrir, una vez abierto guardamos otra vez en formato .arff y volvemos a abrir este archivo para trabajar con este.



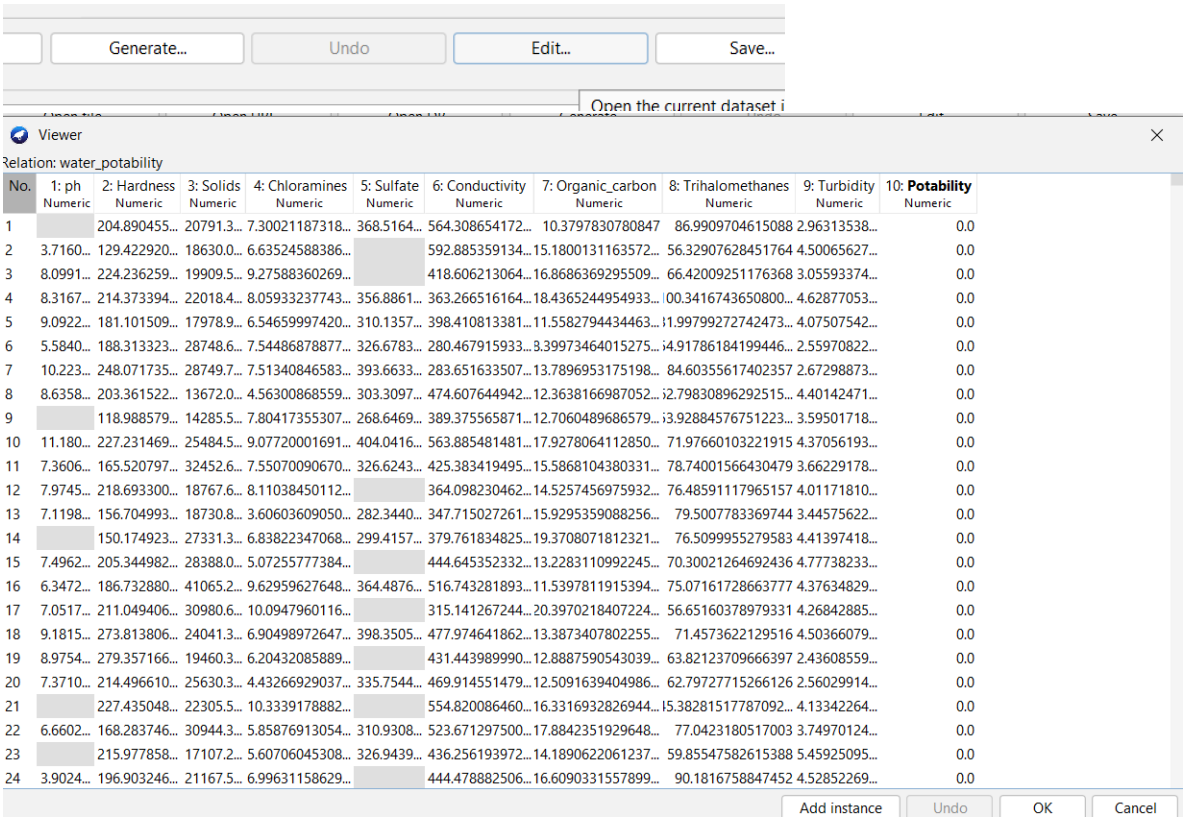
Preprocesamiento 1: ReplaceMissingValues

Este preprocesamiento trata sobre la presencia de valores faltantes (MVs) en conjuntos de datos industriales y de investigación. Se destaca la necesidad de limpiar y preprocesar los datos debido a factores como errores de entrada manual y mediciones incorrectas que generan MVs. La presencia de MVs puede dificultar el análisis de datos y llevar a conclusiones sesgadas.

Se mencionan tres enfoques comunes para tratar los MVs en la minería de datos:

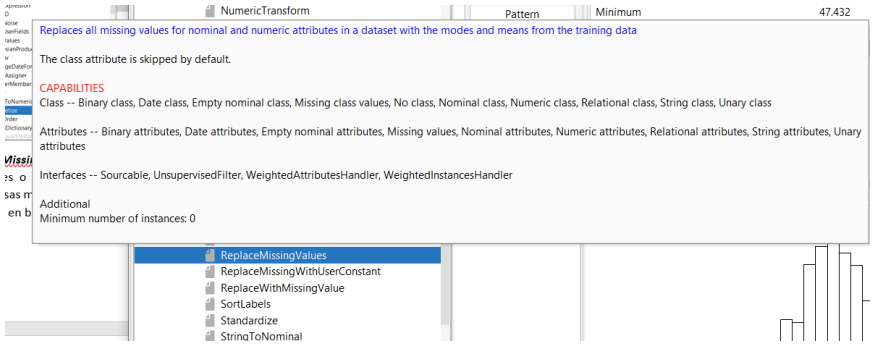
- Descartar ejemplos con MVs, incluso eliminando atributos con una alta cantidad de MVs.
- Utilizar procedimientos de máxima verosimilitud para estimar parámetros a partir de datos completos y luego imputar MVs mediante muestreo.
- Emplear métodos de imputación para estimar y completar MVs, teniendo en cuenta las relaciones entre atributos.

Paso 1: Al dar click en el botón edit podemos percatarnos que el dataset tiene varios valores nulos en sus columnas:



Paso 2: Ahora nos dirigimos a choose>filters>unsupervised>replaceMissingValues

Elegimos no supervisado porque en el aprendizaje no supervisado, tienes un conjunto de datos sin etiquetas, es decir, solo tienes datos de entrada sin información sobre las salidas deseadas.



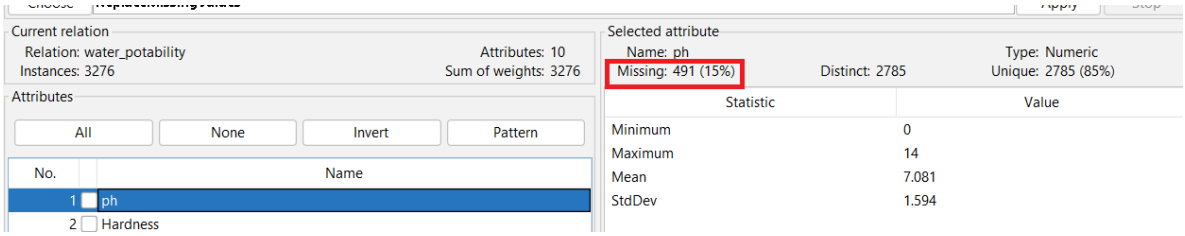
La función ***replaceMissingValues*** se utiliza en el preprocesamiento de datos para manejar y tratar los valores faltantes o ausentes en un conjunto de datos. Estos valores faltantes pueden ser denotados de diversas maneras en un conjunto de datos, como "NaN" (no es un número), "N/A" (no aplicable), espacios en blanco, valores nulos, etc.

La razón principal para usar replaceMissingValues o técnicas similares es que los algoritmos de aprendizaje automático pueden no funcionar correctamente si se les alimenta con datos que contienen valores faltantes. Estos valores faltantes pueden causar problemas durante el

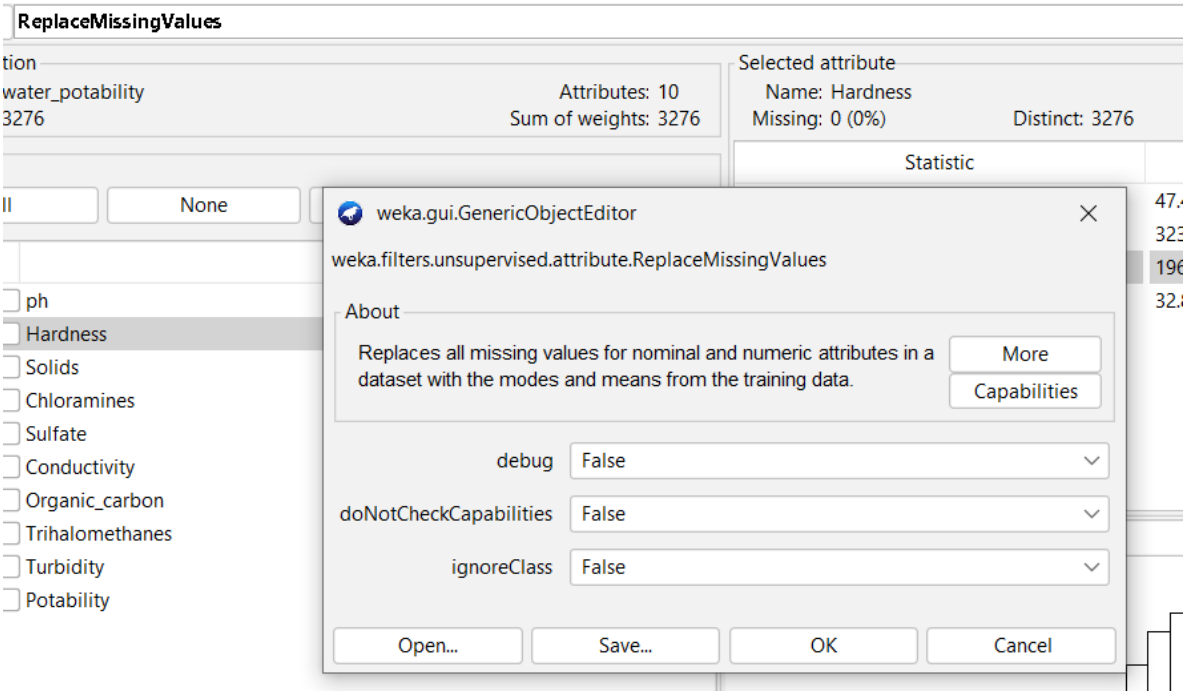
entrenamiento y la predicción de modelos, ya que los algoritmos pueden no saber cómo manejarlos y pueden llevar a resultados incorrectos.

Paso 3: Configuramos el preprocesamiento replaceMissingValues dando click en la barra superior donde aparece el nombre del preprocesamiento:

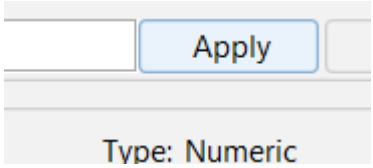
Podemos ver que el atributo ph tiene 15% de valores nulos:



Configuramos el preprocesamiento para que reemplace con la media los valores perdidos:



Le damos en apply



Vemos nuestros datos y observamos que reemplazo los valores vacíos con la media en cada columna:

Viewer

elation: water_potability-weka.filters.unsupervised.attribute.ReplaceMissingValues

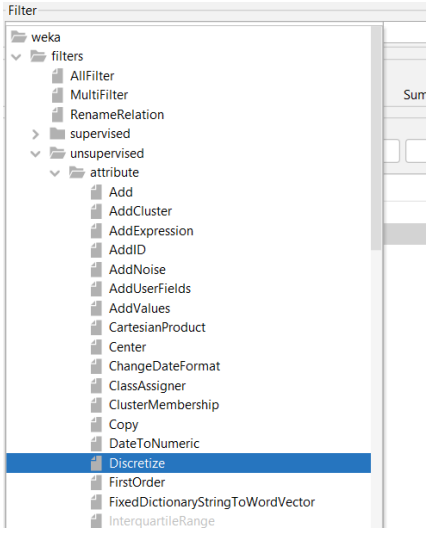
No.	1: ph Numeric	2: Hardness Numeric	3: Solids Numeric	4: Chloramines Numeric	5: Sulfate Numeric	6: Conductivity Numeric	7: Organic_carbon Numeric	8: Trihalomethanes Numeric	9: Turbidity Numeric	10: **Potability** Numeric
1	7.080794504276819	204.890455...	20791.3...	7.30021187318...	368.5164...	564.308654172...	10.3797830780847	86.9909704615088	2.96313538...	0.0
2	3.71608007538699	129.422920...	18630.0...	6.63524588386...	333.7757...	592.885359134...	15.1800131163572...	56.32907628451764	4.50065627...	0.0
3	8.099124189298397	224.236259...	19909.5...	9.27588360269...	333.7757...	418.606213064...	16.8686369295509...	66.42009251176368	3.05593374...	0.0
4	8.316765884214679	214.373394...	22018.4...	8.05933237743...	356.8861...	363.266516164...	18.4365244954933...	100.3416743650800...	4.62877053...	0.0
5	9.092223456290965	181.101509...	17978.9...	6.54659997420...	310.1357...	398.410813381...	11.5582794434463...	31.99799272742473...	4.07507542...	0.0
6	5.584086638456089	188.313323...	28748.6...	7.54486878877...	326.6783...	280.467915933...	3.39973464015275...	54.91786184199446...	2.55970822...	0.0
7	10.223862164528773	248.071735...	28749.7...	7.51340846583...	393.6633...	283.651633507...	13.7896953175198...	84.60355617402357	2.67298873...	0.0
8	8.635848718500734	203.361522...	13672.0...	4.56300868559...	303.3097...	474.607644942...	12.3638166987052...	62.79830896292515...	4.40142471...	0.0
9	7.080794504276819	118.988579...	14285.5...	7.80417355307...	268.6469...	389.375565871...	12.7060489686579...	53.92884576751223...	3.59501718...	0.0
10	11.180284470721592	227.231469...	25484.5...	9.07720001691...	404.0416...	563.885481481...	17.9278064112850...	71.97660103221915	4.37056193...	0.0
11	7.360640105838258	165.520797...	32452.6...	7.55070090670...	326.6243...	425.383419495...	15.5868104380331...	78.74001566430479	3.66229178...	0.0
12	7.974521648923869	218.693300...	18767.6...	8.11038450112...	333.7757...	364.098230462...	14.5257456975932...	76.48591117965157	4.01171810...	0.0
13	7.119824384264552	156.704993...	18730.8...	3.60603609050...	282.3440...	347.715027261...	15.9295359088256...	79.5007783369744	3.44575622...	0.0
14	7.080794504276819	150.174923...	27331.3...	6.83822347068...	299.4157...	379.761834825...	19.3708071812321...	76.5099955279583	4.41397418...	0.0
15	7.49623220797336	205.344982...	28388.0...	5.07255777384...	333.7757...	444.645352332...	13.2283110992245...	70.30021264692436	4.77738233...	0.0
16	6.347271760539316	186.732880...	41065.2...	9.62959627648...	364.4876...	516.743281893...	11.5397811915394...	75.07161728663777	4.37634829...	0.0
17	7.051785800016845	211.049406...	30980.6...	10.0947960116...	333.7757...	315.141267244...	20.3970218407224...	56.65160378979331	4.26842885...	0.0
18	9.181560007151536	273.813806...	24041.3...	6.90498972647...	398.3505...	477.974641862...	13.3873407802255...	71.4573622129516	4.50366079...	0.0
19	8.975464347533963	279.357166...	19460.3...	6.20432085889...	333.7757...	431.443989990...	12.8887590543039...	63.82123709666397	2.43608559...	0.0

Preprocesamiento 2: Discretizer

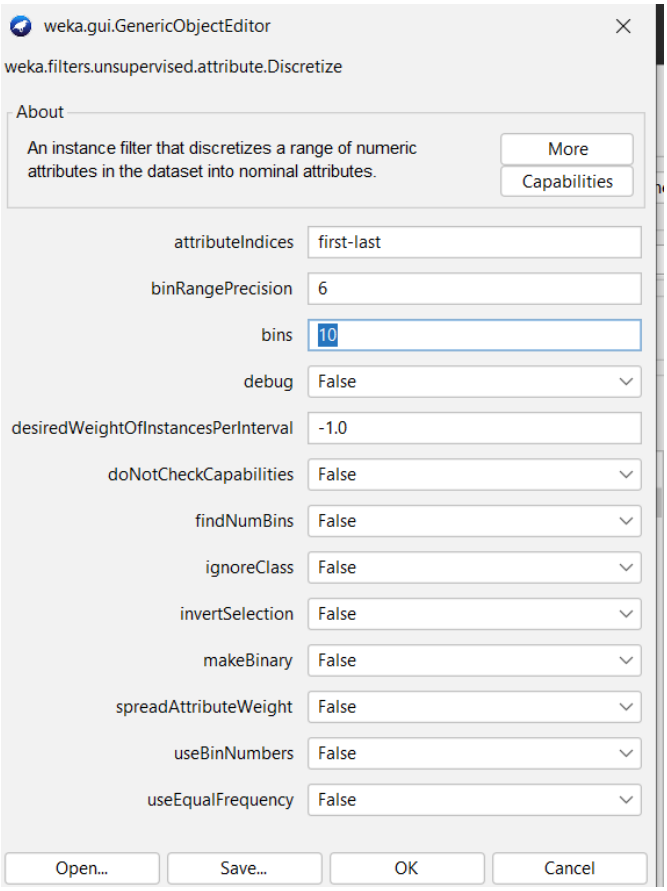
La discretización es una técnica esencial de preprocesamiento utilizada en muchas tareas de descubrimiento de conocimiento y minería de datos. Su principal objetivo es transformar un conjunto de atributos continuos en atributos discretos, asociando valores categóricos a intervalos y, de esta manera, convirtiendo datos cuantitativos en datos cualitativos.

Discretizer se refiere a una técnica de preprocesamiento de datos utilizada en aprendizaje automático y análisis de datos para convertir variables numéricas continuas en variables discretas o categóricas*. Usaremos discretizer porque tenemos datos continuos que casi no se repiten y son más de 3000 datos, para realizar un mejor análisis dividiré los datos de cada columna en 10 rangos.*

Paso 1: Ahora nos dirigimos a choose>filters>unsupervised>discretize

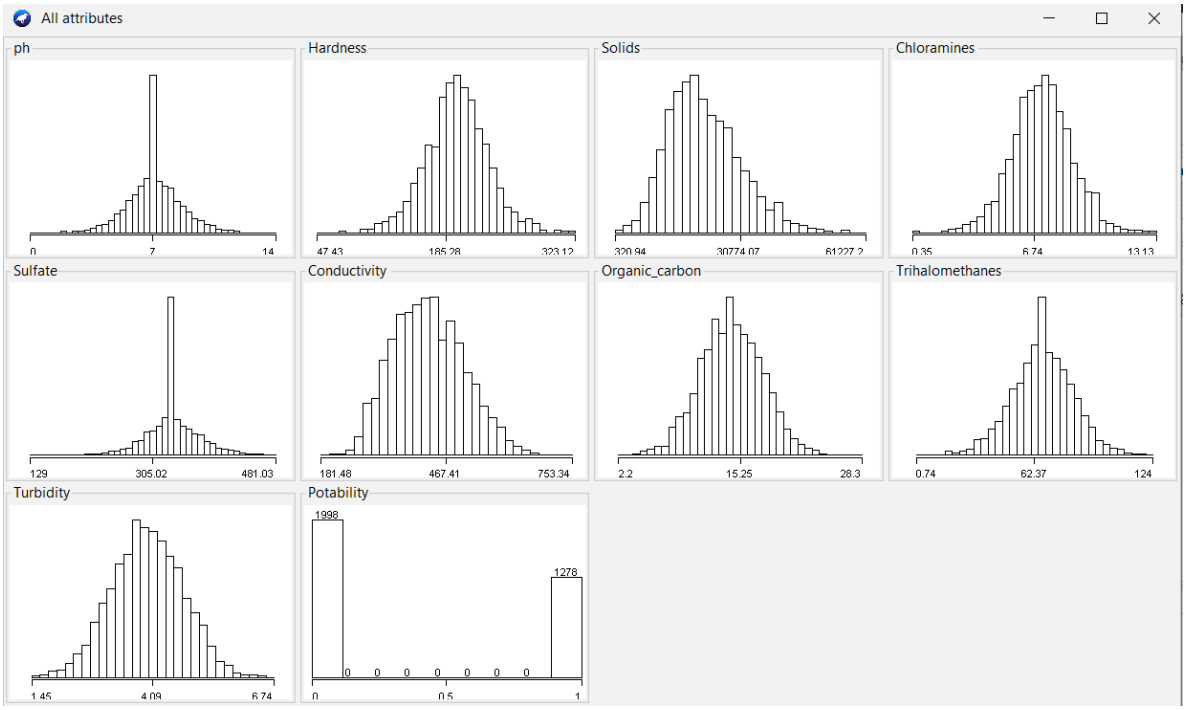


Paso 2: Configuramos el discretize para que coloque 10 rangos a nuestros datos (bins).

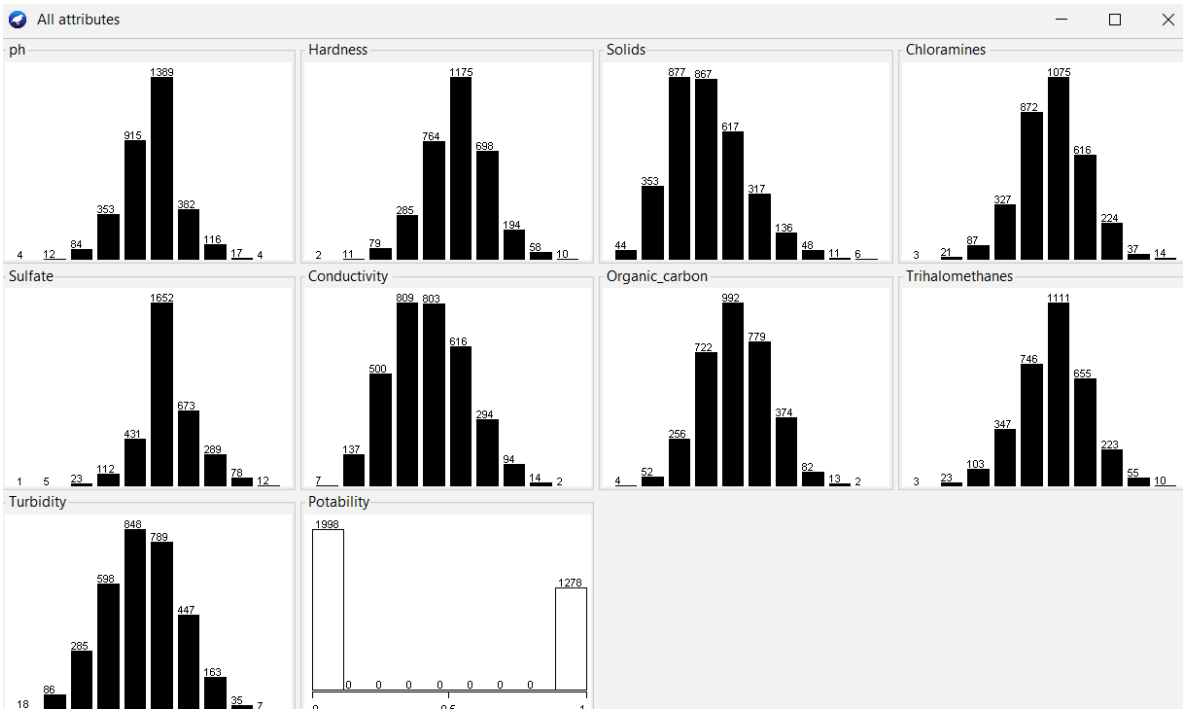


Colocamos bins en 10, que ignore la clase (La última columna que es el resultado) y le damos en ok

Antes:



Después:



Podemos notar que son los mismos gráficos obtenidos en el ejercicio 1 del examen.

Los datos cambiaron se discretizaron, y se asemejan mucho a la distribución normal:

Viewer

!elation: water_potability-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last-precision6

No.	1: ph Nominal	2: Hardness Nominal	3: Solids Nominal	4: Chloramines Nominal	5: Sulfate Nominal	6: Conductivity Nominal	7: Organic_carbon Nominal	8: Trihalomethanes Nominal	9: Turbidity Nominal	10: **Potability** Numeric
1	'(7-8.4]'	'(185.278-2...	'(18592....	'(6.7395-8.017]'	'(340.218...	'(524.599073-5...	'(10.03-12.64]'	'(74.6952-87.0214]'	'(2.5078-3.0..	0.0
2	'(2.8-4.2]'	'(102.5704-...	'(18592....	'(5.462-6.7395]'	'(305.015...	'(581.78496-63...	'(12.64-15.25]'	'(50.0428-62.369]'	'(4.0945-4.6..	0.0
3	'(7-8.4]'	'(212.8472-...	'(18592....	'(8.017-9.2945]'	'(305.015...	'(410.2273-467....	'(15.25-17.86]'	'(62.369-74.6952]'	'(3.0367-3.5..	0.0
4	'(7-8.4]'	'(212.8472-...	'(18592....	'(8.017-9.2945]'	'(340.218...	'(353.041414-4...	'(17.86-20.47]'	'(99.3476-111.6738]..	'(4.6234-5.1..	0.0
5	'(8.4-9.8]'	'(157.7088-...	'(12502....	'(5.462-6.7395]'	'(305.015...	'(353.041414-4...	'(10.03-12.64]'	'(25.3904-37.7166]'	'(3.5656-4.0..	0.0
6	'(4.2-5.6]'	'(185.278-2...	'(24683....	'(6.7395-8.017]'	'(305.015...	'(238.669641-2...	'(7.42-10.03]'	'(50.0428-62.369]'	'(2.5078-3.0..	0.0
7	'(9.8-11.2]'	'(240.4164-...	'(24683....	'(6.7395-8.017]'	'(375.421...	'(238.669641-2...	'(12.64-15.25]'	'(74.6952-87.0214]'	'(2.5078-3.0..	0.0
8	'(8.4-9.8]'	'(185.278-2...	'(12502....	'(4.1845-5.462]'	'(269.812...	'(467.413187-5...	'(10.03-12.64]'	'(62.369-74.6952]'	'(4.0945-4.6..	0.0
9	'(7-8.4]'	'(102.5704-...	'(12502....	'(6.7395-8.017]'	'(234.609...	'(353.041414-4...	'(12.64-15.25]'	'(50.0428-62.369]'	'(3.5656-4.0..	0.0
10	'(9.8-11.2]'	'(212.8472-...	'(24683....	'(8.017-9.2945]'	'(375.421...	'(524.599073-5...	'(17.86-20.47]'	'(62.369-74.6952]'	'(4.0945-4.6..	0.0
11	'(7-8.4]'	'(157.7088-...	'(30774....	'(6.7395-8.017]'	'(305.015...	'(410.2273-467....	'(15.25-17.86]'	'(74.6952-87.0214]'	'(3.5656-4.0..	0.0
12	'(7-8.4]'	'(212.8472-...	'(18592....	'(8.017-9.2945]'	'(305.015...	'(353.041414-4...	'(12.64-15.25]'	'(74.6952-87.0214]'	'(3.5656-4.0..	0.0
13	'(7-8.4]'	'(130.1396-...	'(18592....	'(2.907-4.1845]'	'(269.812...	'(295.855527-3...	'(15.25-17.86]'	'(74.6952-87.0214]'	'(3.0367-3.5..	0.0
14	'(7-8.4]'	'(130.1396-...	'(24683....	'(6.7395-8.017]'	'(269.812...	'(353.041414-4...	'(17.86-20.47]'	'(74.6952-87.0214]'	'(4.0945-4.6..	0.0
15	'(7-8.4]'	'(185.278-2...	'(24683....	'(4.1845-5.462]'	'(305.015...	'(410.2273-467....	'(12.64-15.25]'	'(62.369-74.6952]'	'(4.6234-5.1..	0.0
16	'(5.6-7]'	'(185.278-2...	'(36864....	'(9.2945-10.572...	'(340.218...	'(467.413187-5...	'(10.03-12.64]'	'(74.6952-87.0214]'	'(4.0945-4.6..	0.0
17	'(7-8.4]'	'(185.278-2...	'(30774....	'(9.2945-10.572...	'(305.015...	'(295.855527-3...	'(17.86-20.47]'	'(50.0428-62.369]'	'(4.0945-4.6..	0.0
18	'(8.4-9.8]'	'(267.9856-...	'(18592....	'(6.7395-8.017]'	'(375.421...	'(467.413187-5...	'(12.64-15.25]'	'(62.369-74.6952]'	'(4.0945-4.6..	0.0
19	'(8.4-9.8]'	'(267.9856-...	'(18592....	'(5.462-6.7395]'	'(305.015...	'(410.2273-467....	'(12.64-15.25]'	'(62.369-74.6952]'	'(1.9789-2.5..	0.0
20	'(7-8.4]'	'(212.8472-...	'(24683....	'(4.1845-5.462]'	'(305.015...	'(467.413187-5...	'(10.03-12.64]'	'(62.369-74.6952]'	'(2.5078-3.0..	0.0
21	'(7-8.4]'	'(212.8472-...	'(18592....	'(9.2945-10.572...	'(305.015...	'(524.599073-5...	'(15.25-17.86]'	'(37.7166-50.0428]'	'(4.0945-4.6..	0.0
22	'(5.6-7]'	'(157.7088-...	'(30774....	'(5.462-6.7395]'	'(305.015...	'(467.413187-5...	'(17.86-20.47]'	'(74.6952-87.0214]'	'(3.5656-4.0..	0.0
23	'(7-8.4]'	'(212.8472-...	'(12502....	'(5.462-6.7395]'	'(305.015...	'(410.2273-467....	'(12.64-15.25]'	'(50.0428-62.369]'	'(5.1523-5.6..	0.0
24	'(2.8-4.2]'	'(185.278-2...	'(18592....	'(6.7395-8.017]'	'(305.015...	'(410.2273-467....	'(15.25-17.86]'	'(87.0214-99.3476]'	'(4.0945-4.6..	0.0

Add instance | Undo | OK | Cancel

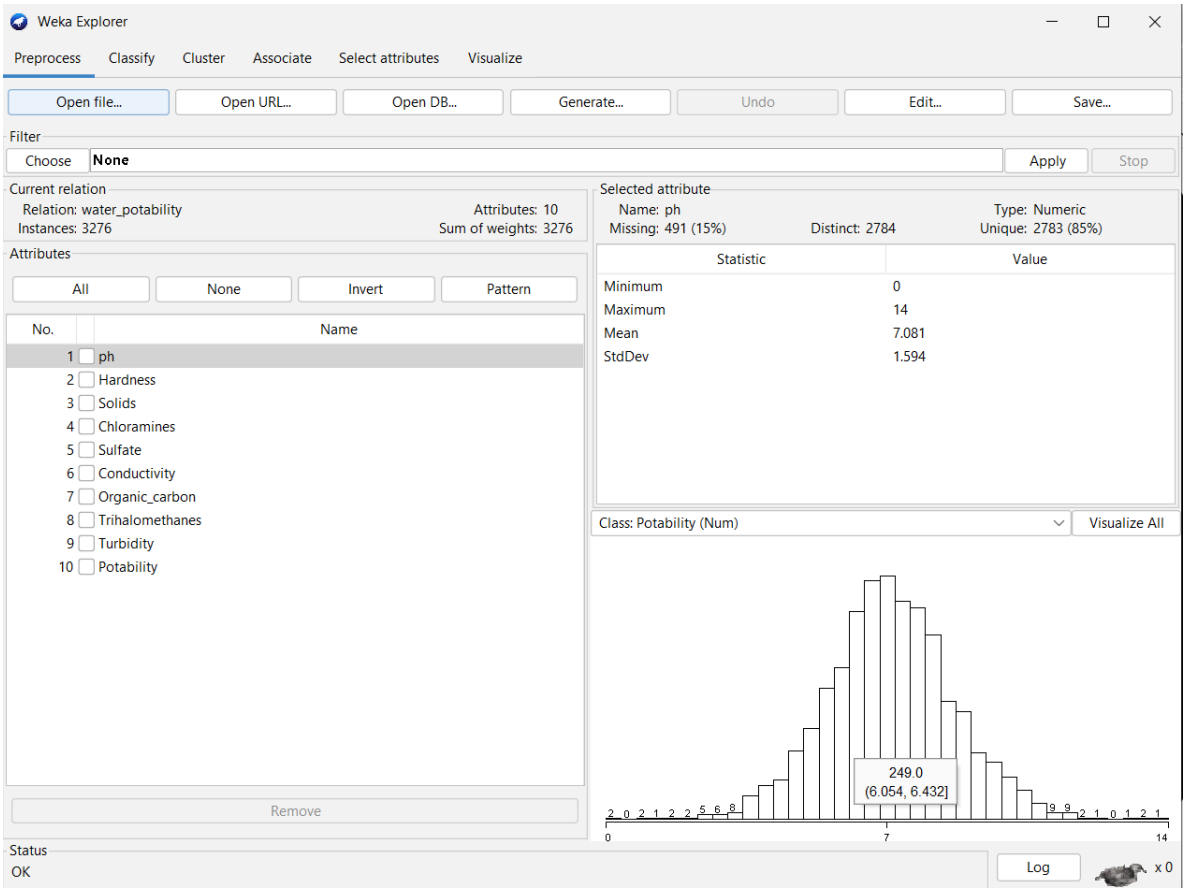
Preprocesamiento 3: Normalize

A veces, los atributos seleccionados son atributos en estado puro que tienen un significado en el dominio original del cual se obtuvieron, o están diseñados para funcionar con el sistema operativo en el que se utilizan actualmente. Por lo general, estos atributos originales no son lo suficientemente buenos para obtener modelos predictivos precisos. Por lo tanto, es común realizar una serie de pasos de manipulación para transformar los atributos originales o generar nuevos atributos con mejores propiedades que mejorarán el poder predictivo del modelo.

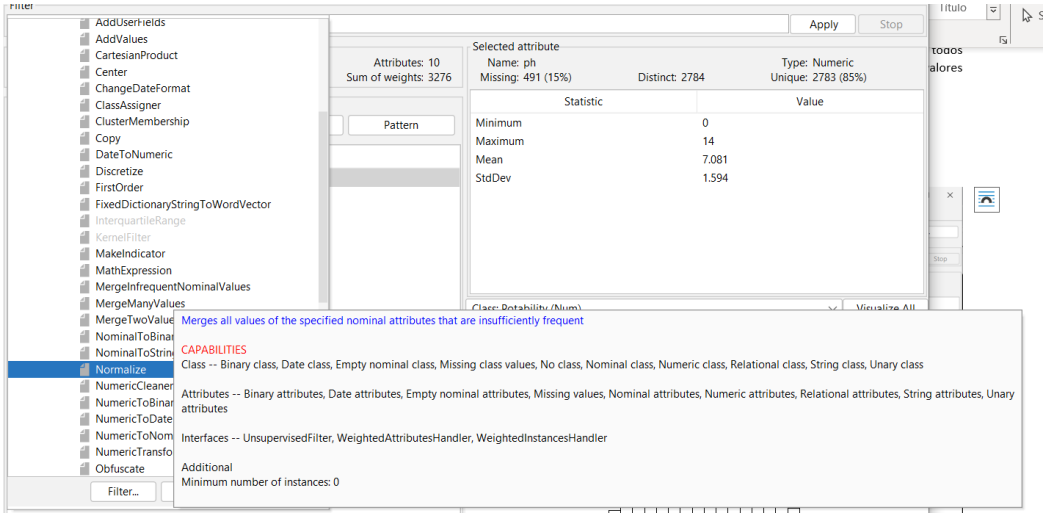
La normalización de datos es esencial en el preprocesamiento de datos para garantizar que todas las características tengan una escala común. Esto ayuda a evitar que características con magnitudes diferentes dominen el proceso de entrenamiento de modelos de aprendizaje automático, asegura la convergencia eficiente de algoritmos y facilita la interpretación de los resultados.

La normalización puede ayudar a reducir el impacto de valores atípicos o extremos al escalar todos los datos dentro de un rango común. Esto puede hacer que el modelo sea más robusto a valores extremos.

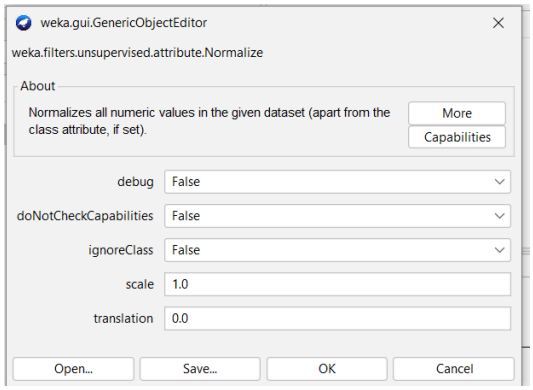
Paso 1: Volvemos al estado inicial del dataset haciendo click en undo:



Paso 2: Nos dirigimos a choose>filters>unsupervised>normalize y seleccionamos el filtro de preprocesamiento



Configuramos para que la escala sea de 0 a 1, es decir nuestros datos estarán dentro de esa escala:



Datos antes de Normalizar:

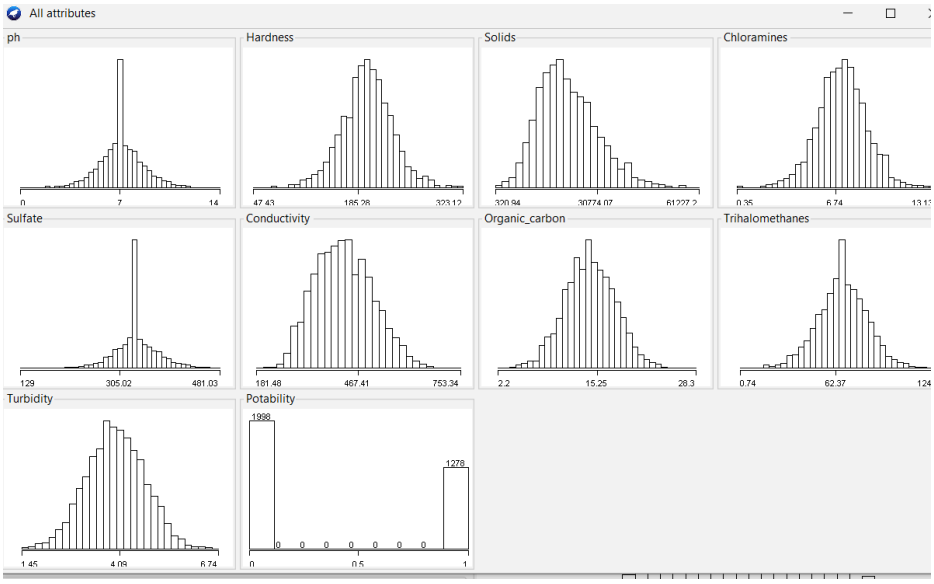
Viewer

Relation: water_potability-weka.filters.unsupervised.attribute.ReplaceMissingValues

No.	1: ph Numeric	2: Hardness Numeric	3: Solids Numeric	4: Chloramines Numeric	5: Sulfate Numeric	6: Conductivity Numeric	7: Organic_carbon Numeric	8: Trihalomethanes Numeric	9: Turbidity Numeric	10: **Potability** Numeric
1	7.0807...	204.890455	20791.3...	7.300212	368.5164...	564.308654	10.379783	86.99097	2.963135	0.0
2	3.71608	129.422921	18630.0...	6.635246	333.7757...	592.885359	15.180013	56.329076	4.500656	0.0
3	8.0991...	224.236259	19909.5...	9.275884	333.7757...	418.606213	16.868637	66.420093	3.055934	0.0
4	8.3167...	214.373394	22018.4...	8.059332	356.8861...	363.266516	18.436524	100.341674	4.628771	0.0
5	9.0922...	181.101509	17978.9...	6.5466	310.1357...	398.410813	11.558279	31.997993	4.075075	0.0
6	5.5840...	188.313324	28748.6...	7.544869	326.6783...	280.467916	8.399735	54.917862	2.559708	0.0
7	10.223...	248.071735	28749.7...	7.513408	393.6633...	283.651634	13.789695	84.603556	2.672989	0.0
8	8.6358...	203.361523	13672.0...	4.563009	303.3097...	474.607645	12.363817	62.798309	4.401425	0.0
9	7.0807...	118.988579	14285.5...	7.804174	268.6469...	389.375566	12.706049	53.928846	3.595017	0.0
10	11.180...	227.231469	25484.5...	9.0772	404.0416...	563.885481	17.927806	71.976601	4.370562	0.0
11	7.36064	165.520797	32452.6...	7.550701	326.6243...	425.383419	15.58681	78.740016	3.662292	0.0
12	7.9745...	218.6933	18767.6...	8.110385	333.7757...	364.09823	14.525746	76.485911	4.011718	0.0
13	7.1198...	156.704993	18730.8...	3.606036	282.3440...	347.715027	15.929536	79.500778	3.445756	0.0
14	7.0807...	150.174923	27331.3...	6.838223	299.4157...	379.761835	19.370807	76.509996	4.413974	0.0
15	7.4962...	205.344982	28388.0...	5.072558	333.7757...	444.645352	13.228311	70.300213	4.777382	0.0
16	6.3472...	186.732881	41065.2...	9.629596	364.4876...	516.743282	11.539781	75.071617	4.376348	0.0
17	7.0517...	211.049406	30980.6...	10.094796	333.7757...	315.141267	20.397022	56.651604	4.268429	0.0
18	9.18156	273.813807	24041.3...	6.90499	398.3505...	477.974642	13.387341	71.457362	4.503661	0.0
19	8.9754...	279.357167	19460.3...	6.204321	333.7757...	431.44399	12.888759	63.821237	2.436086	0.0
20	7.37105	214.49661	25630.3...	4.432669	335.7544...	469.914551	12.509164	62.797277	2.560299	0.0
21	7.0807...	227.435048	22305.5...	10.333918	333.7757...	554.820086	16.331693	45.382815	4.133423	0.0
22	6.6602...	168.283747	30944.3...	5.858769	310.9308...	523.671298	17.884235	77.042318	3.749701	0.0
23	7.0807...	215.977859	17107.2...	5.60706	326.9439...	436.256194	14.189062	59.855476	5.459251	0.0
24	3.9024...	196.903247	21167.5...	6.996312	333.7757...	444.478883	16.609033	90.181676	4.528523	0.0

Sort view: left click = ascending / Shift + le
Menu: right click (or left+alt)

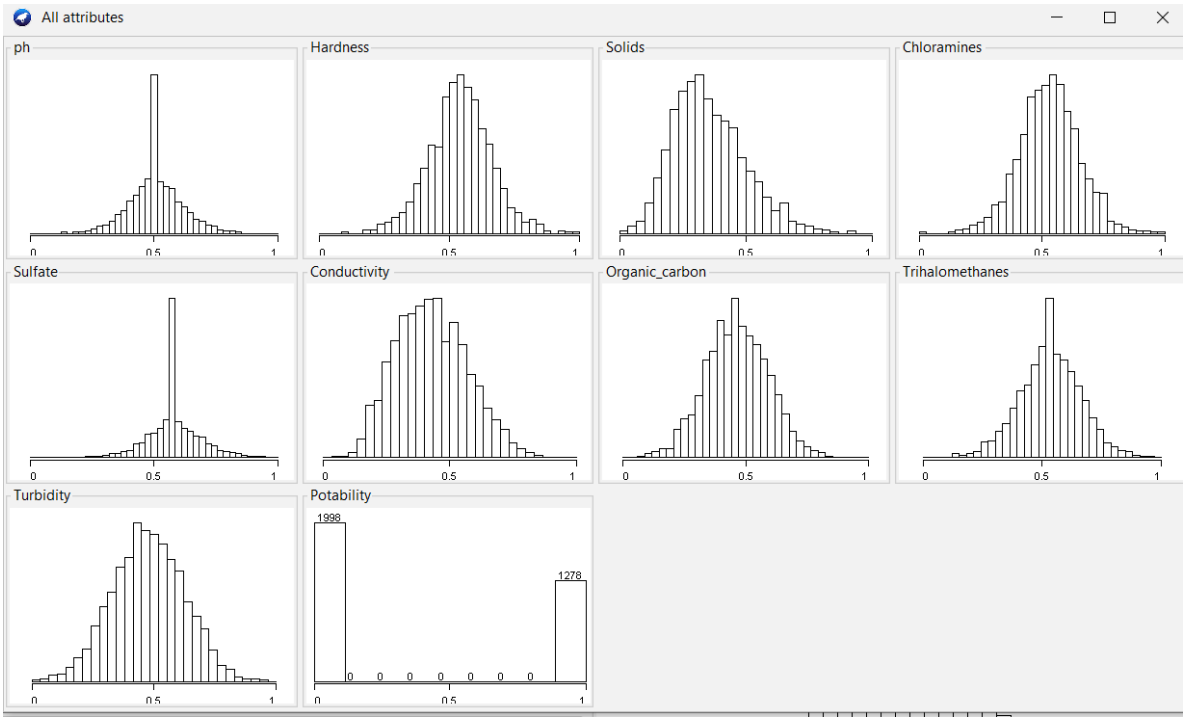
Add instance | Undo | OK | Cancel



Datos después de normalizar:

No.	1: ph Numeric	2: Hardness Numeric	3: Solids Numeric	4: Chloramines Numeric	5: Sulfate Numeric	6: Conductivity Numeric	7: Organic_carbon Numeric	8: Trihalomethanes Numeric	9: Turbidity Numeric	10: **Potability** Numeric
1	0.5057...	0.57113900...	0.33609...	0.54389135029...	0.680385...	0.66943947669...	0.31340164750957...	0.699753127484545...	0.28609094...	0.0
2	0.2654...	0.29740043...	0.30061...	0.49183921722...	0.581698...	0.71941108105...	0.49731850574712...	0.450999302299167...	0.57679258...	0.0
3	0.5785...	0.64131080...	0.32161...	0.69854277886...	0.581698...	0.41465206381...	0.56201674329501...	0.532865708815368...	0.30363660...	0.0
4	0.5940...	0.60553586...	0.35624...	0.60331365949...	0.647347...	0.31788046458...	0.62208904214559...	0.808064723921403...	0.60101550...	0.0
5	0.6494...	0.48485088...	0.28992...	0.48490019569...	0.514545...	0.37933670683...	0.35855475095785...	0.253606082977722...	0.49632728...	0.0
6	0.3988...	0.51100983...	0.46674...	0.56304258317...	0.561537...	0.17309194258...	0.23753773946360...	0.439550404828738...	0.20981433...	0.0
7	0.7302...	0.72776770...	0.46676...	0.56057988258...	0.751819...	0.17865925681...	0.44404961685823...	0.680384514286641...	0.23123255...	0.0
8	0.6168...	0.56559320...	0.21920...	0.32962888454...	0.495155...	0.51258082794...	0.38941827586206...	0.503482898216806...	0.5	0.0
9	0.5057...	0.25955261...	0.22928...	0.58334043052...	0.396689...	0.36353692206...	0.40253061302681...	0.431526715451639...	0.40556192...	0.0
10	0.7985...	0.65217514...	0.41315...	0.68299021526...	0.781300...	0.66869948117...	0.60259793103448...	0.577944549009427	0.55219550...	0.0
11	0.52576	0.42833595...	0.52755...	0.56349909980...	0.561383...	0.42650325019...	0.51290459770114...	0.632814784767406	0.41828171...	0.0
12	0.5696...	0.62120518...	0.30287...	0.60730998043...	0.581698...	0.31933486889...	0.47225080459770...	0.614527680874884...	0.48434826...	0.0
13	0.5085...	0.39635895...	0.30226...	0.25471906066...	0.435598...	0.29068583680...	0.52603586206896...	0.638986695007382...	0.37734089...	0.0
14	0.5057...	0.37267284...	0.44347...	0.50772782778...	0.484093...	0.34672555203...	0.65788532567049...	0.614723077671950...	0.56040347...	0.0
15	0.5354...	0.57278768...	0.46082...	0.36951530332...	0.581698...	0.46018626910...	0.42254065134099...	0.564344347811977...	0.62911363...	0.0
16	0.4533...	0.50527719...	0.66896...	0.72623060665...	0.668940...	0.58626270909...	0.35784601532567...	0.603053795979296...	0.55328946...	0.0
17	0.5036...	0.59347897...	0.50339...	0.76264547945...	0.581698...	0.23372464946...	0.69720390804597...	0.453615907578978...	0.53288504...	0.0
18	0.6558...	0.82114028...	0.38945...	0.51295420743...	0.765133...	0.51846863907...	0.42863375478927...	0.573732066654768	0.57736074...	0.0
19	0.6411...	0.84124735...	0.31424...	0.45810731898...	0.581698...	0.43710126896...	0.40953099616858...	0.511781708880271...	0.18644091...	0.0
20	0.5265...	0.60598279...	0.41554...	0.31942614481...	0.587319...	0.50437409323...	0.39498712643678...	0.503474525806818...	0.20992607...	0.0
21	0.5057...	0.65291357...	0.36095...	0.78136344422...	0.581698...	0.65284697710...	0.54144417624521...	0.362194471937823...	0.50735923...	0.0
22	0.4757...	0.43835783...	0.50279...	0.43105823874...	0.516804...	0.59837761438...	0.60092854406130...	0.619041699793934...	0.43480828...	0.0
23	0.5057...	0.61135563...	0.27560...	0.41135499021...	0.562291...	0.44551628932...	0.45935103448275...	0.479608281546624...	0.75803573...	0.0
24	0.2787...	0.54216751...	0.34227...	0.52010270058...	0.581698...	0.45989516756...	0.55207022988505...	0.725638688322435	0.58206144...	0.0

Right click (or left+alt) for context menu



Los datos recopilados en un conjunto de datos pueden no ser lo suficientemente útiles para un algoritmo de minería de datos. A veces, los atributos seleccionados son atributos brutos que tienen un significado en el dominio original de donde se obtuvieron, o están diseñados para funcionar con el sistema operativo en el que se están utilizando actualmente. Por lo general, estos atributos originales

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Formula normalización

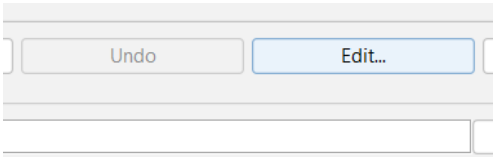
Para que funcionen mejor muchos algoritmos de Machine Learning usados en Data Science, hay que normalizar las variables de entrada al algoritmo. Normalizar significa, en este caso, comprimir o extender los valores de la variable para que estén en un rango definido.

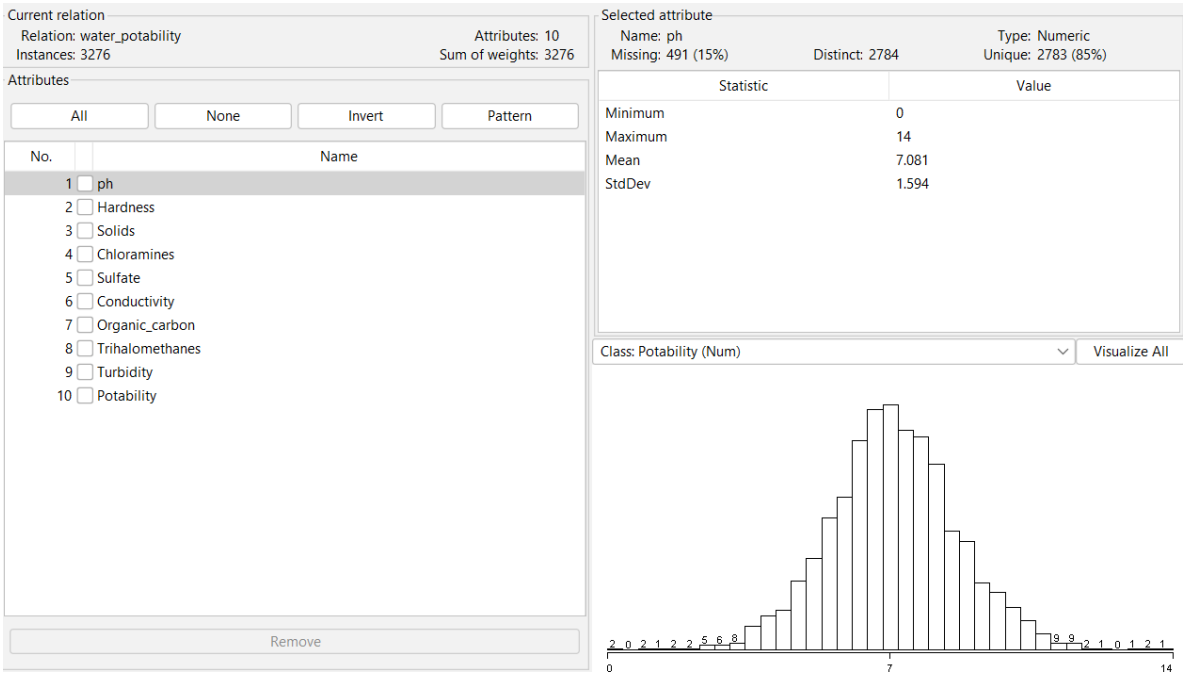
Preprocesamiento 4: RemoveDuplicates

"RemoveDuplicates" (Eliminar duplicados) se aplica para eliminar instancias duplicadas de un conjunto de datos. Esto significa que elimina las filas que tienen los mismos valores en todas sus características. Este filtro se utiliza principalmente para limpiar conjuntos de datos y garantizar que cada instancia sea única.

Eliminación de datos redundantes: Los datos duplicados no aportan información adicional y pueden inflar el tamaño del conjunto de datos innecesariamente. Al eliminar duplicados, se reduce la redundancia y se simplifican los datos.

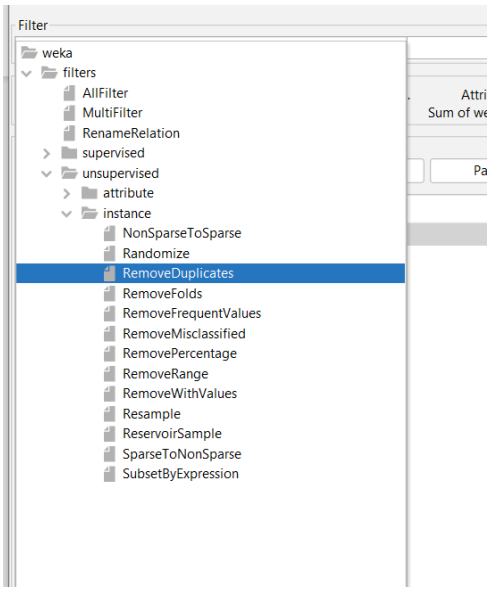
Paso 1: Volvemos al estado inicial del dataset haciendo click en Undo:



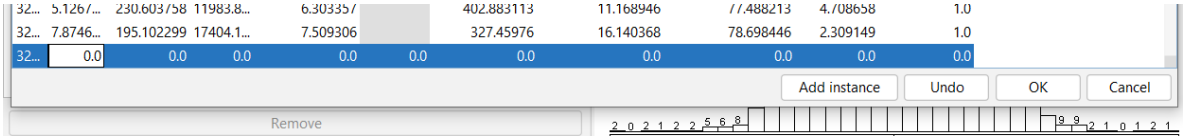


Paso 2: Nos dirigimos a choose>filters>unsupervised>instance>RemoveDuplicates

Instance se refiere a las filas del dataset y attribute a las columnas.



Paso 3: Añadimos una instancia igual al final para mostrar eliminación:

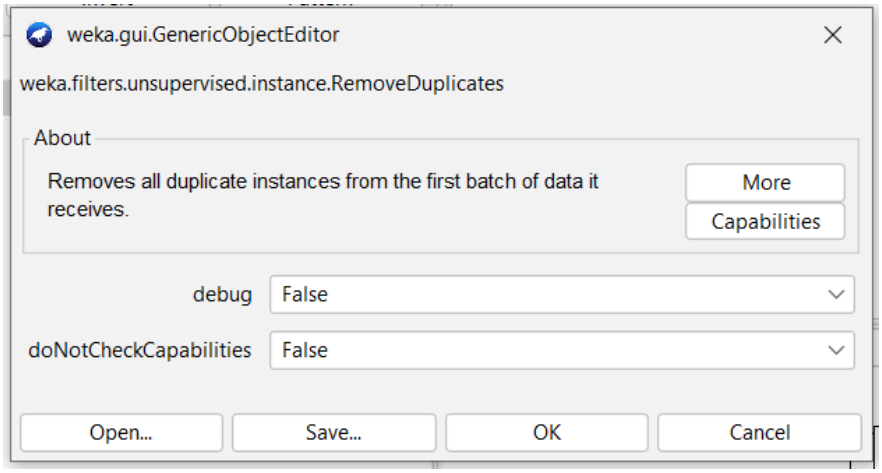


Tenemos dos filas idénticas:

3275	5.1267...	230.603758	11983.8...	6.303357		402.883113	11.168946	77.488213	4.708658	1.0
3276	7.8746...	195.102299	17404.1...	7.509306		327.45976	16.140368	78.698446	2.309149	1.0
3277	7.8746...	195.102299	17404.1...	7.509306		327.45976	16.140368	78.698446	2.309149	1.0

Add instance | Undo | OK

Paso 4: Aplicamos y configuramos el preprocesamiento RemoveDuplicates:



Paso 5: Aplicamos

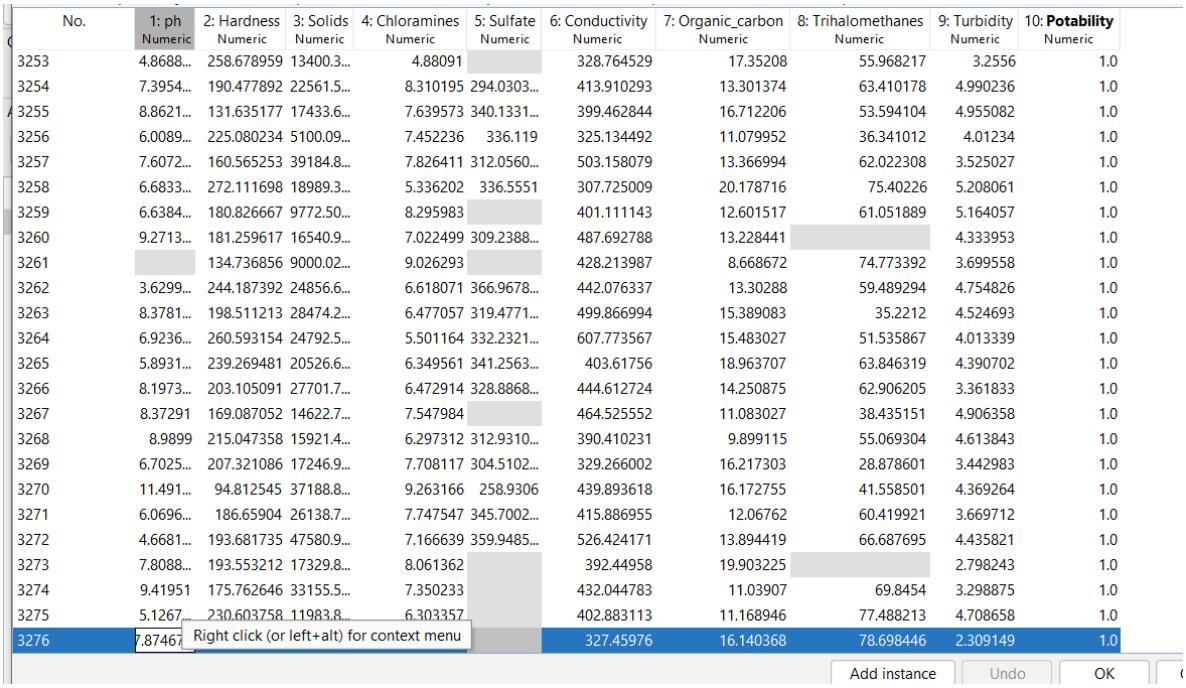
Datos Antes:

Relation: water_potability-weka.filters.unsupervised.instance.RemoveDuplicates

No.	1: ph Numeric	2: Hardness Numeric	3: Solids Numeric	4: Chloramines Numeric	5: Sulfate Numeric	6: Conductivity Numeric	7: Organic_carbon Numeric	8: Trihalomethanes Numeric	9: Turbidity Numeric	10: **Potability** Numeric
32...	7.3954...	190.477892	22561.5...	8.310195	294.0303...	413.910293	13.301374	63.410178	4.990236	1.0
32...	8.8621...	131.635177	17433.6...	7.639573	340.1331...	399.462844	16.712206	53.594104	4.955082	1.0
32...	6.0089...	225.080234	5100.09...	7.452236	336.119	325.134492	11.079952	36.341012	4.01234	1.0
32...	7.6072...	160.565253	39184.8...	7.826411	312.0560...	503.158079	13.366994	62.022308	3.525027	1.0
32...	6.6833...	272.111698	18989.3...	5.336202	336.5551	307.725009	20.178716	75.40226	5.208061	1.0
32...	6.6384...	180.826667	9772.50...	8.295983		401.111143	12.601517	61.051889	5.164057	1.0
32...	9.2713...	181.259617	16540.9...	7.022499	309.2388...	487.692788	13.228441		4.333953	1.0
32...		134.736856	9000.02...	9.026293		428.213987	8.668672	74.773392	3.699558	1.0
32...	3.6299...	244.187392	24856.6...	6.618071	366.9678...	442.076337	13.30288	59.489294	4.754826	1.0
32...	8.3781...	198.511213	28474.2...	6.477057	319.4771...	499.866994	15.389083	35.2212	4.524693	1.0
32...	6.9236...	260.593154	24792.5...	5.501164	332.2321...	607.773567	15.483027	51.535867	4.013339	1.0
32...	5.8931...	239.269481	20526.6...	6.349561	341.2563...	403.61756	18.963707	63.846319	4.390702	1.0
32...	8.1973...	203.105091	27701.7...	6.472914	328.8868...	444.612724	14.250875	62.906205	3.361833	1.0
32...	8.37291	169.087052	14622.7...	7.547984		464.525552	11.083027	38.435151	4.906358	1.0
32...	8.9899	215.047358	15921.4...	6.297312	312.9310...	390.410231	9.899115	55.069304	4.613843	1.0
32...	6.7025...	207.321086	17246.9...	7.708117	304.5102...	329.266002	16.217303	28.878601	3.442983	1.0
32...	11.491...	94.812545	37188.8...	9.263166	258.9306	439.893618	16.172755	41.558501	4.369264	1.0
32...	6.0696...	186.65904	26138.7...	7.747547	345.7002...	415.886955	12.06762	60.419921	3.669712	1.0
32...	4.6681...	193.681735	47580.9...	7.166639	359.9485...	526.424171	13.894419	66.687695	4.435821	1.0
32...	7.8088...	193.553212	17329.8...	8.061362		392.44958	19.903225		2.798243	1.0
32...	9.41951	175.762646	33155.5...	7.350233		432.044783	11.03907	69.8454	3.298875	1.0
32...	5.1267...	230.603758	11983.8...	6.303357		402.883113	11.168946	77.488213	4.708658	1.0
32...	7.8746...	195.102299	17404.1...	7.509306		327.45976	16.140368	78.698446	2.309149	1.0
32...	7.8746...	195.102299	17404.1...	7.509306		327.45976	16.140368	78.698446	2.309149	1.0

Add instance | Undo

Datos Después:



No.	1: ph Numeric	2: Hardness Numeric	3: Solids Numeric	4: Chloramines Numeric	5: Sulfate Numeric	6: Conductivity Numeric	7: Organic_carbon Numeric	8: Trihalomethanes Numeric	9: Turbidity Numeric	10: **Potability** Numeric
3253	4.8688...	258.678959	13400.3...	4.88091		328.764529	17.35208	55.968217	3.2556	1.0
3254	7.3954...	190.477892	22561.5...	8.310195	294.0303...	413.910293	13.301374	63.410178	4.990236	1.0
3255	8.8621...	131.635177	17433.6...	7.639573	340.1331...	399.462844	16.712206	53.594104	4.955082	1.0
3256	6.0089...	225.080234	5100.09...	7.452236	336.119	325.134492	11.079952	36.341012	4.01234	1.0
3257	7.6072...	160.565253	39184.8...	7.826411	312.0560...	503.158079	13.366994	62.022308	3.525027	1.0
3258	6.6833...	272.111698	18989.3...	5.336202	336.5551	307.725009	20.178716	75.40226	5.208061	1.0
3259	6.6384...	180.826667	9772.50...	8.295983		401.111143	12.601517	61.051889	5.164057	1.0
3260	9.2713...	181.259617	16540.9...	7.022499	309.2388...	487.692788	13.228441		4.333953	1.0
3261		134.736856	9000.02...	9.026293		428.213987	8.668672	74.773392	3.699558	1.0
3262	3.6299...	244.187392	24856.6...	6.618071	366.9678...	442.076337	13.30288	59.489294	4.754826	1.0
3263	8.3781...	198.511213	28474.2...	6.477057	319.4771...	499.866994	15.389083	35.2212	4.524693	1.0
3264	6.9236...	260.593154	24792.5...	5.501164	332.2321...	607.773567	15.483027	51.535867	4.013339	1.0
3265	5.8931...	239.269481	20526.6...	6.349561	341.2563...	403.61756	18.963707	63.846319	4.390702	1.0
3266	8.1973...	203.105091	27701.7...	6.472914	328.8868...	444.612724	14.250875	62.906205	3.361833	1.0
3267	8.37291	169.087052	14622.7...	7.547984		464.525552	11.083027	38.435151	4.906358	1.0
3268	8.9899	215.047358	15921.4...	6.297312	312.9310...	390.410231	9.899115	55.069304	4.613843	1.0
3269	6.7025...	207.321086	17246.9...	7.708117	304.5102...	329.266002	16.217303	28.878601	3.442983	1.0
3270	11.491...	94.812545	37188.8...	9.263166	258.9306	439.893618	16.172755	41.558501	4.369264	1.0
3271	6.0696...	186.65904	26138.7...	7.747547	345.7002...	415.886955	12.06762	60.419921	3.669712	1.0
3272	4.6681...	193.681735	47580.9...	7.166639	359.9485...	526.424171	13.894419	66.687695	4.435821	1.0
3273	7.8088...	193.553212	17329.8...	8.061362		392.44958	19.903225		2.798243	1.0
3274	9.41951	175.762646	33155.5...	7.350233		432.044783	11.03907	69.8454	3.298875	1.0
3275	5.1267...	230.603758	11983.8...	6.303357		402.883113	11.168946	77.488213	4.708658	1.0
3276	7.87467					327.45976	16.140368	78.698446	2.309149	1.0

Right click (or left+alt) for context menu

Add instance | Undo | OK

Podemos observar que la fila duplicada se elimino correctamente. Este filtro es de mucha utilidad para eliminar filas duplicadas que pueden deberse a errores humanos a la hora de registrar los datos en el dataset. Y así reducir el número de filas conservando solo los datos importantes.