MATEMATICA III

Trabajo Practico – Potabilidad del Agua

Resumen

El acceso al agua potable es un derecho fundamental y un factor clave para la salud pública. Sin embargo, garantizar la calidad del agua sigue siendo un desafío en muchas partes del mundo. En esta tesis, se analiza la potabilidad del agua utilizando un conjunto de datos de muestras, aplicando técnicas de procesamiento de datos y redes neuronales para predecir si el agua es apta para el consumo humano. A través de Python, se trabajó en el análisis y limpieza de los datos, para luego entrenar una red neuronal capaz de clasificar la potabilidad de manera automatizada.

Docentes a cargo: Bompensieri Josefina & Prudente Tomas
Integrantes: Nicolas Cernadas & Catalina Correa

Primeros pasos

Para esta investigación, decidimos utilizar un Data Set llamado ‘Water Potability’ tomado de la página ‘Kaggle’, en el cual hay 3276 muestras de agua, algunas potables, y otras no.

A continuación, se describirán cada una de las columnas del archivo, y se dará una estimación de cuando una muestra tiende a ser más potable.

Variables categóricas

Potability: Potabilidad del agua.

Descripción: Indica si el agua es potable con valores 1 (potable) y 0 (no potable).

Variables continuas

pH: Mide el nivel de acidez o alcalinidad del agua.

Descripción: En el Data Set, podemos ver como el pH en las muestras potables, es del rango establecido como "saludable", que va desde 6.5 hasta 8.5

Hardness: Minerales disueltos en el agua.

Descripción: Las mismas son encontradas en las muestras por el roce con las piedras, arenas, plantas...El tiempo que el agua está en contacto con estas sales/minerales, determina su dureza.

Solids: Sólidos orgánicos/no orgánicos totales disueltos en el agua.

Descripción: Estos minerales producen efectos de sabor/coloración en el agua. Un alto nivel de sólidos, indica una muestra demasiado mineralizada. Lo deseable se encuentra entre 500mg/l y 1000mg/l.

Chloramines: Desinfectantes (cloros).

Descripción: La cloración tiene como objetivo principal eliminar o inactivar microorganismos patógenos, como bacterias, virus y parásitos, que podrían estar presentes en el agua y que podrían causar enfermedades si son ingeridos. Los limites saludables son de hasta 4mg/l.

Sulfate: Sustancias encontradas en minerales, piedras, plantas...

Descripción: Los sulfatos son una mezcla de oxígeno y azufre y son partes de las sustancias existentes en algunas formaciones de rocas y suelos que incorporan agua subterránea. El mineral se convierte gradualmente en una solución y se libera al agua subterránea.

Conductivity: Conductividad de la muestra.

Descripción: Las sales disueltas en agua se descomponen en iones cargados positiva y negativamente. La conductividad se define como la capacidad del agua para conducir una corriente eléctrica a través de los iones disueltos. Los iones más positivos son sodio (Na+), calcio (Ca+2), potasio (K+) y magnesio (Mg+2). No se debería exceder de los 400µS/cm

Organic carbon: Representa el contenido de carbono orgánico en el agua.

Descripción: El carbono orgánico total es la medida de todo el carbono que contiene la materia orgánica disuelta o particulada presente en la muestra de agua. El origen de esta materia orgánica es, principalmente, sustancias vegetales o animales, contaminaciones que se producen en el sistema de distribución. Su presencia por encima de los valores establecidos, al reaccionar con los agentes desinfectantes utilizados en la potabilización, como el cloro, pueden dar

lugar a subproductos de la desinfección nocivos para la salud. Lo recomendable se encuentra en < 2mg/l cuando el agua es potable, y < 4mg/l cuando el agua está en tratamiento

Trihalomethanes: Concentración de trihalometanos en el agua.

Descripción: Los trihalometanos son unos subproductos de la desinfección que se forman cuando se emplea el cloro como desinfectante. Los trihalometanos que se encuentran en el agua de consumo humano son el cloroformo, el bromo diclorometano, el dibromoclorometano y el bromoformo. Hasta 80 ppm, se consideran seguros.

Turbidity: Grado de transparencia que pierde el agua.

Descripción: La turbidez es una medida de la cantidad de partículas en suspensión en el agua. Cuanto mayor sea la cantidad de sólidos suspendidos en el líquido, como las algas, sedimentos, materia orgánica y los contaminantes, mayor será el grado de turbidez.

EJEMPLO DE MUESTRAS DE AGUA:

AGUA POTABLE

pH: 9.45 (bastante alcalino).

Hardness: 145.80 (pocos minerales disueltos).

Solids: 13,168 mg/L (niveles bajos de sólidos disueltos).

Chloramines: 9.44 mg/L (alto nivel de cloraminas).

Sulfate: 310.58 mg/L (nivel moderado de sustancias).

Conductivity: 592.66 µS/cm (alta conductividad => buena cantidad de sales).

Organic carbon: 8.61 mg/L (nivel moderado de carbono orgánico).

Trihalomethanes: 77.58 µg/L (dentro del rango aceptable).

Turbidity: 3.87 NTU (turbiedad aceptable moderada).

AGUA NO POTABLE:

pH: No disponible (dato faltante).

Hardness: 204.89 (minerales disueltos moderados).

Solids: 20,791 mg/L (nivel bastante alto de sólidos disueltos).

Chloramines: 7.30 mg/L (nivel relativamente alto).

Sulfate: 368.52 mg/L (nivel alto de sustancias en el agua).

Conductivity: 564.31 µS/cm (nivel alto => buena cantidad de sales).

Organic carbon: 10.38 mg/L (nivel elevado de carbono orgánico).

Trihalomethanes: 86.99 µg/L (por encima del límite superior).

Turbidity: 2.96 NTU (dentro de los límites aceptables).

Determinación de valores clave.

En base en las características de las muestras, ciertos valores parecen ser determinantes:

pH: Aunque en este caso el agua potable tiene un pH elevado (9.45), en general un pH fuera del rango neutro (6.5-8.5) podría indicar no potabilidad.

Solids: Niveles muy altos de sólidos disueltos (>20,000 mg/L) parecen estar asociados con agua no potable, ya que superan ampliamente el promedio de 22,014 mg/L.

Sulfate: El agua no potable tiene niveles más altos de sulfato (>360 mg/L), mientras que el agua potable está en un rango más moderado (300-350 mg/L).

Organic carbon: El agua no potable tiende a tener niveles elevados de carbono orgánico (>10 mg/L), lo que puede indicar contaminación orgánica.

Trihalomethanes: Niveles superiores a 80 µg/L pueden ser un indicador de no potabilidad.

Reglas aproximadas basadas en los datos:

Si 6.5 > pH > 8.5, ‘Solids’ > 20,000 mg/L, ‘Sulfate’ > 360 mg/L y ‘Organic carbon’ > 10 mg/L: El agua tiende a ser menos potable.

Primeras impresiones del Data Set

Drescripcion del DF sin procesar

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2785.000000	3276.000000	3276.000000	3276.000000	2495.000000	3276.000000	3276.000000	3114.000000	3276.000000	3276.000000
mean	7.080795	196.369496	22014.092526	7.122277	333.775777	426.205111	14.284970	66.396293	3.966786	0.390110
std	1.594320	32.879761	8768.570828	1.583085	41.416840	80.824064	3.308162	16.175008	0.780382	0.487849
min	0.000000	47.432000	320.942611	0.352000	129.000000	181.483754	2.200000	0.738000	1.450000	0.000000
25%	6.093092	176.850538	15666.690297	6.127421	307.699498	365.734414	12.065801	55.844536	3.439711	0.000000
50%	7.036752	196.967627	20927.833607	7.130299	333.073546	421.884968	14.218338	66.622485	3.955028	0.000000
75%	8.062066	216.667456	27332.762127	8.114887	359.950170	481.792304	16.557652	77.337473	4.500320	1.000000
max	14.000000	323.124000	61227.196008	13.127000	481.030642	753.342620	28.300000	124.000000	6.739000	1.000000

15 Primeras filas

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0
5	5.584087	188.313324	28748.687739	7.544869	326.678363	280.467916	8.399735	54.917862	2.559708	0
6	10.223862	248.071735	28749.716544	7.513408	393.663396	283.651634	13.789695	84.603556	2.672989	0
7	8.635849	203.361523	13672.091764	4.563009	303.309771	474.607645	12.363817	62.798309	4.401425	0
8	NaN	118.988579	14285.583854	7.804174	268.646941	389.375566	12.706049	53.928846	3.595017	0
9	11.180284	227.231469	25484.508491	9.077200	404.041635	563.885481	17.927806	71.976601	4.370562	0
10	7.360640	165.520797	32452.614409	7.550701	326.624353	425.383419	15.586810	78.740016	3.662292	0
11	7.974522	218.693300	18767.656682	8.110385	NaN	364.098230	14.525746	76.485911	4.011718	0
12	7.119824	156.704993	18730.813653	3.606036	282.344050	347.715027	15.929536	79.500778	3.445756	0
13	NaN	150.174923	27331.361962	6.838223	299.415781	379.761835	19.370807	76.509996	4.413974	0
14	7.496232	205.344982	28388.004887	5.072558	NaN	444.645352	13.228311	70.300213	4.777382	0

NaNs por columna

```
ph                491
Sulfate           781
Trihalomethanes   162
dtype: int64
```

Al traer las primeras 15 filas del Data Set, podemos darnos cuenta de 2 cosas: Tenemos valores en escalas muy diferentes, y tenemos celdas incompletas (‘NaNs’).

Las columnas de ‘pH’, ‘Sulfate’ y ‘Trihalomethanes’, son las únicas con valores no indicados, pero ‘Sulfate’, en particular, tiene demasiados. Podemos ver abajo a la izquierda, cuantas celdas faltantes contiene cada columna.

Tenemos 2 opciones:

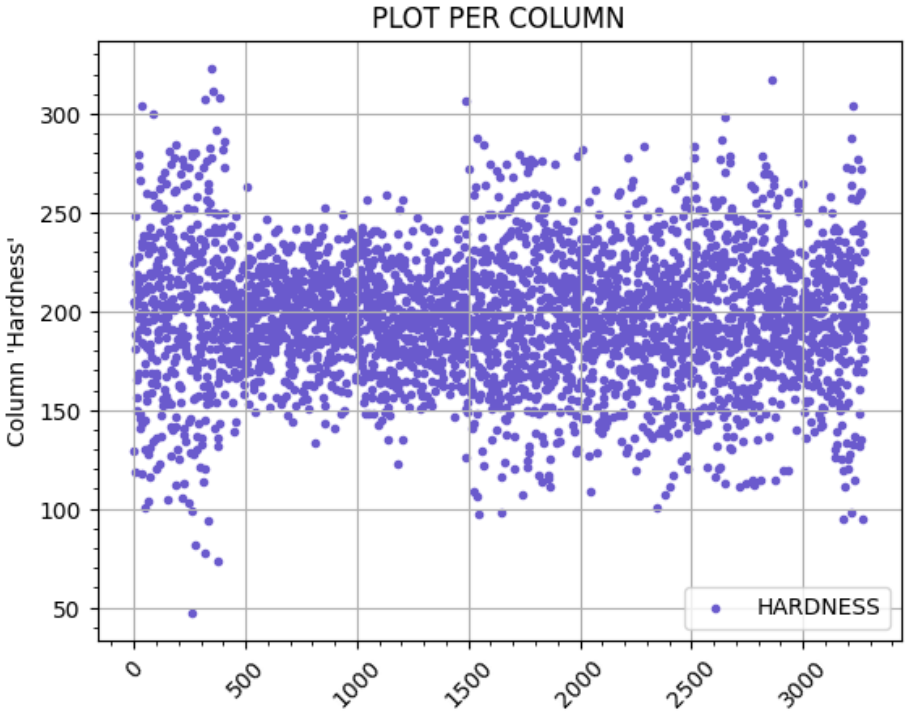
a) Eliminamos completamente las filas que tengan celdas sin algún dato (¡Estaríamos perdiendo 1434 datos!)

b) Llenamos esos ‘espacios en blanco’ por la mediana de cada columna.

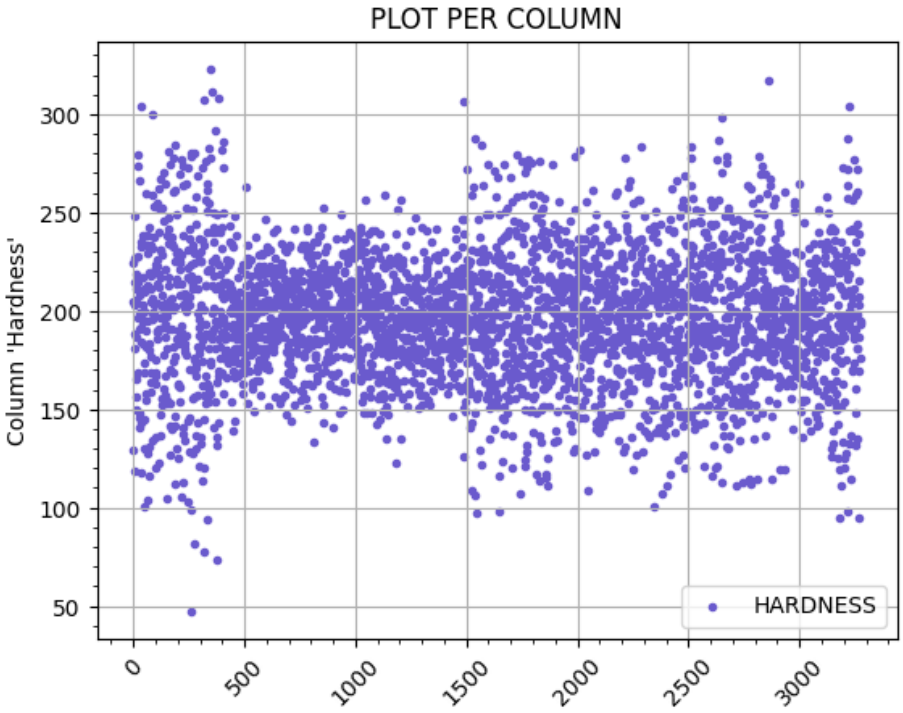
Después de realizar pruebas con ambas 2 opciones, nos decidimos por la segunda, solo que con una pequeña modificación. Si, las celdas vacías se rellenan con la mediana de cada columna, pero dependiendo de si la muestra, es o no potable. (Osea: así como armamos el primero, generamos 2 Data Frames nuevos a partir de este, pero uno contiene todas las muestras no potables, y el otro las potables. Tomamos la mediana de la columna de ‘pH’ de los valores potables, y recorrimos el Data Frame original, donde hubiese datos faltantes en la columna ‘pH’ y la muestra resultara potable, ahí introduciríamos el valor tomado. Idem para las muestras no potables con el Data Frame donde la potabilidad resulta 0. Y así para las 3 columnas con datos ‘NaN’)

Estos son algunos de los gráficos de los datos en función de las filas (índices). Como se puede apreciar en el gráfico de los datos de ‘pH’, están bastante dispersos, siendo 14 el más alto, y 0 el menor.

‘pH’

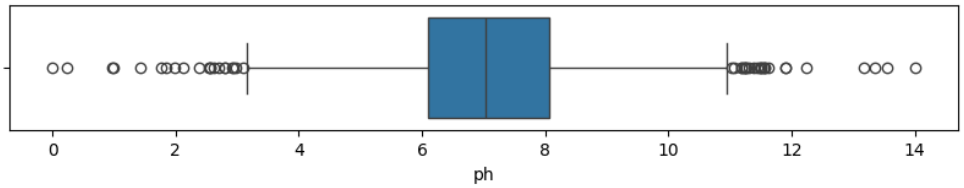


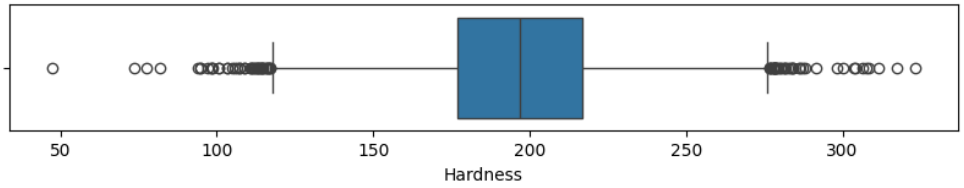
‘Hardness’

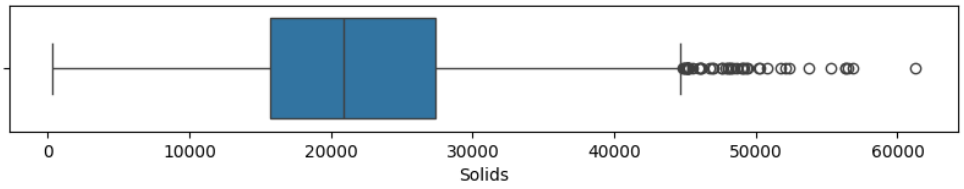


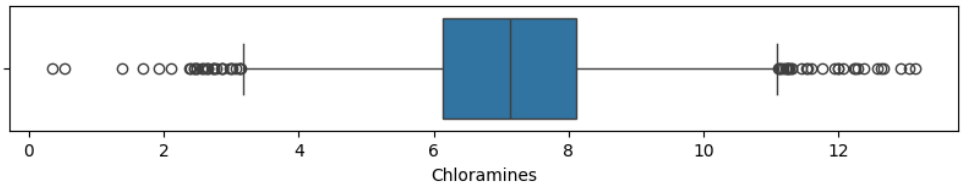
Estos son solo 2 de los 9 gráficos de tipo ‘Scatter’ que se encuentran en el JupyterNotebook.

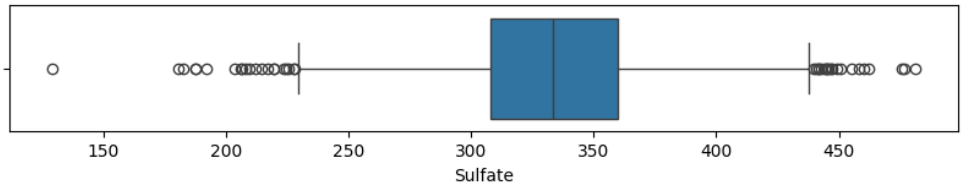
Podemos también apreciarlo de forma gráfica como ‘Boxplot’, gráficos donde se muestra la dispersión de los puntos de manera horizontal, con una cajita azul que marca el rango intercuartil (rango de valores para determinar atípicos)

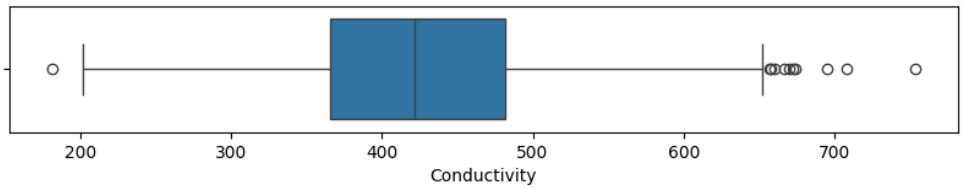

0
2
4
6
8
10
12
14
ph

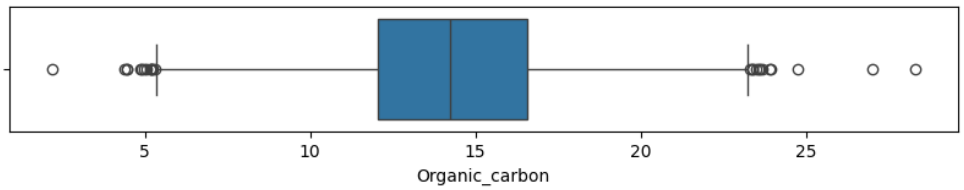

50
100
150
200
250
300
Hardness

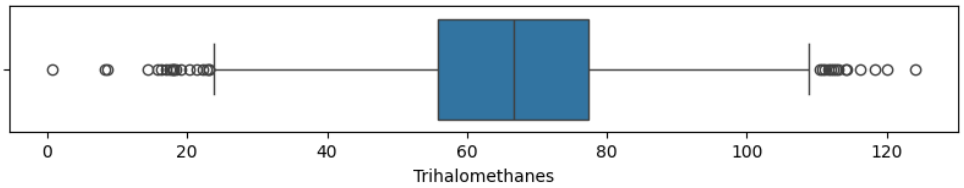

0
10000
20000
30000
40000
50000
60000
Solids

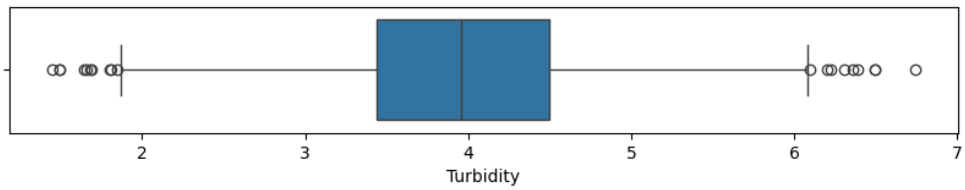

0
2
4
6
8
10
12
Chloramines


150
200
250
300
350
400
450
Sulfate


200
300
400
500
600
700
Conductivity

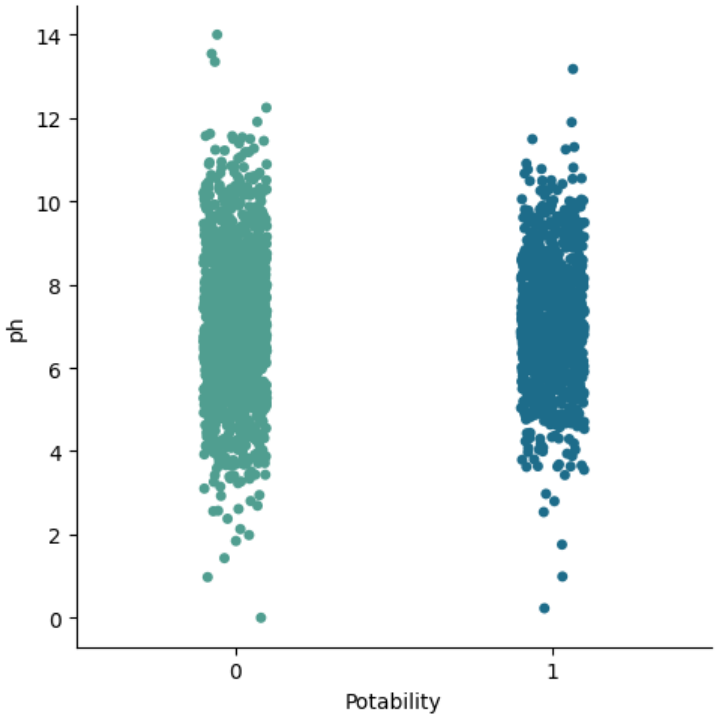

5
10
15
20
25
Organic_carbon


0
20
40
60
80
100
120
Trihalomethanes

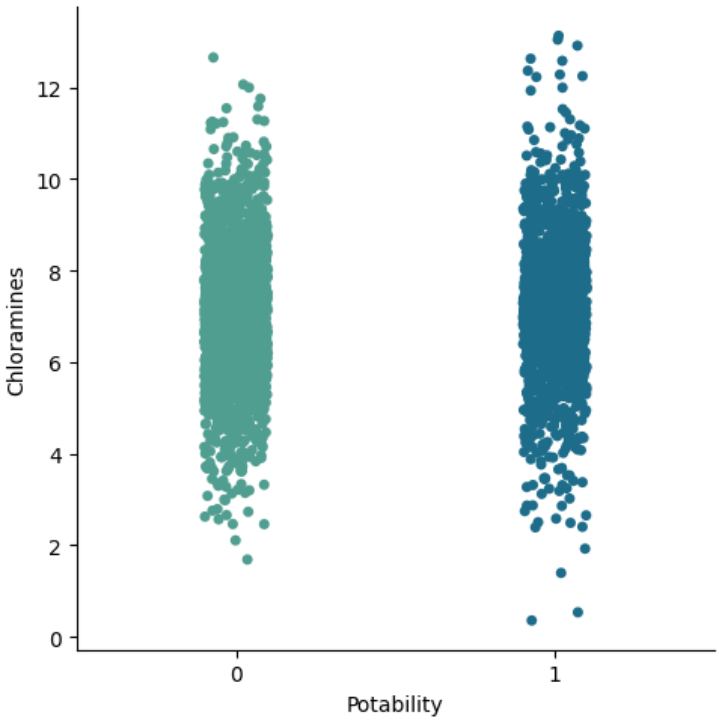

2
3
4
5
6
7
Turbidity

Y los valores de los datos en función de si terminan siendo o no, potables.

‘pH’

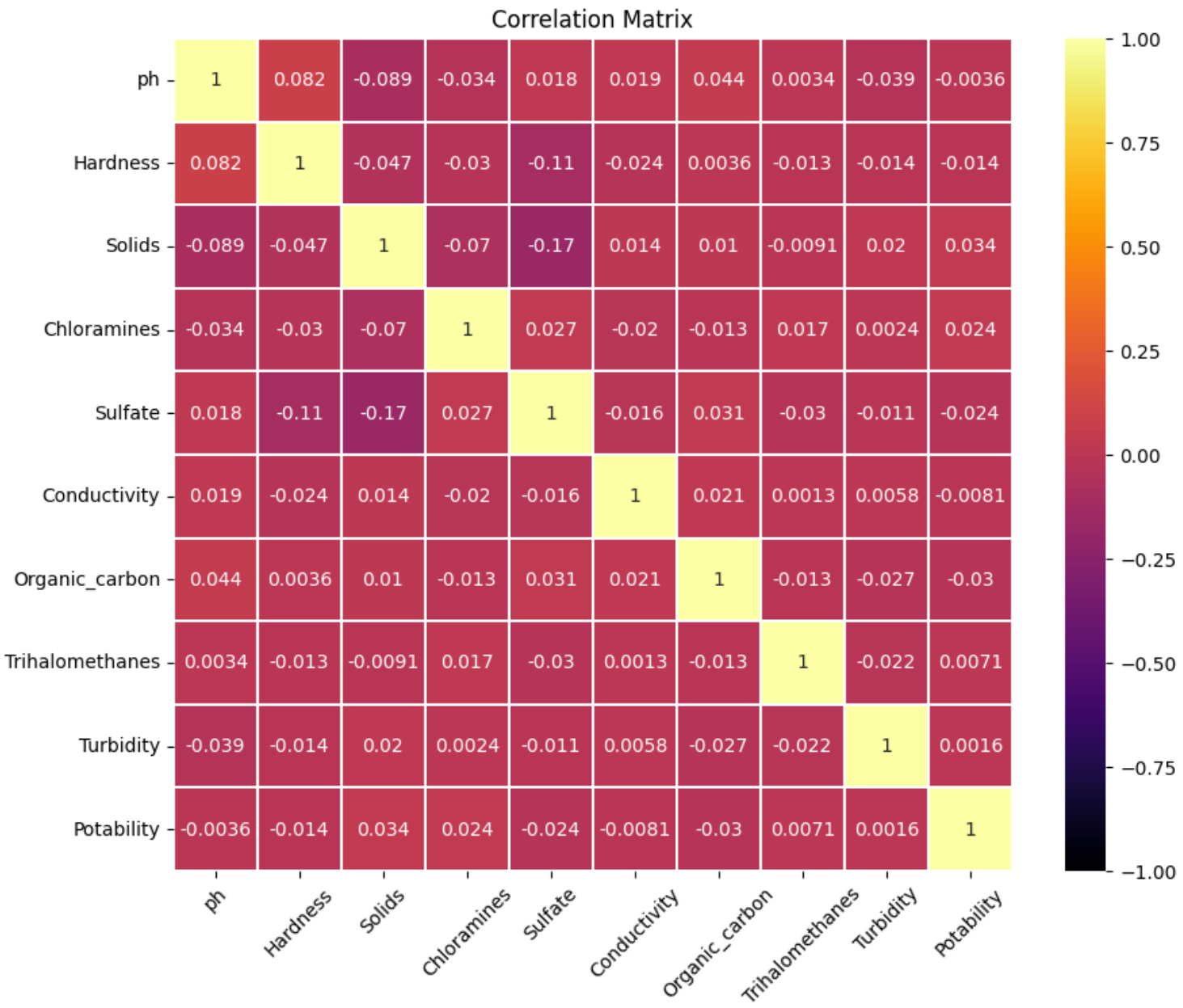


‘Chloramines’



Otra vez, como podemos apreciar, no es como si un dato solo pudiese determinar un factor tan crucial como la potabilidad, podemos ver que, por ejemplo, la columna ‘Chloramines’ tiene muchísimos 6 < valores < 8, sin hacer distinción de si la muestra es potable o no. ¿De qué nos damos cuenta? De que la potabilidad del agua no es dependiente de un solo valor, incluso algunas muestras de agua que terminan siendo potables, tienen valores

faltantes. Esto nos da a entender que los datos, tampoco tienen mucha correlación. Es decir, no son dependientes unos de otros. Aunque una muestra tenga mas solidos disueltos en el agua, el pH no necesariamente va a crecer también. Para poder ver esto un poco mejor, armamos una matriz de correlatividad.



En el grafico se puede ver la relación de una columna con otra. Por ejemplo, el segundo cuadradito de la primera fila de arriba de todo (léase fila 0, posición 1 si se quiere), nos muestra la relación de la columna pH con la columna Hardness: 0,082. ¿Qué quiere decir? La correlación mide la relación entre 2 valores, va entre -1 y 1, indicando la fuerza y dirección de una relación lineal. Si este valor estuviese cerca de -1, significaría que están fuertemente ligados en una relación lineal que decrece, y viceversa si estuviese cerca de 1. Al estar cerca de 0, quiere decir que su relación lineal es casi nula, y son valores independientes. ¿A que columna deberíamos prestarle atención? Y...A la de potabilidad. Ahí nos damos cuenta de que valores influyen más en esto, e igual podemos ver que el valor mas alto, es el de la fila de sólidos, e igual es 0.034, ósea absolutamente nada relacionados.

Tratamiento del Data Set

Una vez visualizados un poco los datos y explicados un par de conceptos, vamos a proceder a tratarlos.

Como ya se dijo, vamos a rellenar los datos faltantes con las medianas de las columnas dependiendo de su potabilidad, pero también vamos a hacer 2 cositas más.

Una vez cargados los datos, se procede a limpiar el Data Frame. ¿Y esto como se hace? Calculamos, por cada columna, el rango intercuartílico, para después ir iterando por cada una, guardando los valores que cayeran fuera de ese rango en una lista. Con esa lista, después determinamos las filas donde se encuentran esos valores atípicos, y se eliminan por completo.

Una vez hecho esto, el Data Frame quedo de la siguiente manera:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2659.000000	2659.000000	2659.000000	2659.000000	2659.000000	2659.000000	2659.000000	2659.000000	2659.000000	2659.000000
mean	7.070749	196.987774	21469.620999	7.112308	333.600048	425.677064	14.323277	66.516661	3.962617	0.373825
std	1.214443	28.284819	7872.666485	1.410906	26.261055	79.928875	3.225671	15.013309	0.761310	0.483909
min	3.902476	118.988579	320.942611	3.194969	268.646941	201.619737	5.362371	26.140863	1.872573	0.000000
25%	6.353274	178.995757	15615.309555	6.192423	319.787497	365.528850	12.093506	56.953332	3.439518	0.000000
50%	7.035456	197.517093	20575.573160	7.117859	333.389426	421.140278	14.241965	66.542198	3.948304	0.000000
75%	7.793309	215.684992	26730.413973	8.053600	347.912069	481.257238	16.582846	76.766091	4.495585	1.000000
max	10.252816	275.679780	44376.187382	10.999995	399.617217	652.537592	23.373265	107.610806	6.083772	1.000000

Como se puede apreciar, todas las columnas tienen la misma cantidad de datos, ya que estamos borrando completamente la fila donde aparezca un valor atípico, y no la celda, ya que esto, llevaría a que los datos se movieran de lugar, es decir, si yo borro la celda 0, la 1 pasa a su posición, y me cambia toda la fila 0, y no queremos que esto pase.

Ahora sí, vamos a hacer el 2do paso del que ya habíamos hablado, el cual es: estandarizar los datos. Con la media (promedio) de cada columna y la desviación estándar (que tanto se alejan los datos de la media), vamos a iterar, otra vez, sobre cada columna, realizando el siguiente calculo en cada uno de los datos: *(dato - media) / desviación estándar*. Esto, va a poner a absolutamente todos los valores de la columna en un mismo rango, ósea, ya no vamos a tener valores como 3.9 y 10.25 en el caso de pH, sino que los valores se van a medir en cuanto a que tan alejados de la media se encuentren (a cuantas desviaciones estándar). Para entender mejor esto, veamos como queda el Data Frame después de este proceso:

```
Drescripcion del DF post procesamiento

                ph      Hardness        Solids   Chloramines       Sulfate  Conductivity  Organic_carbon  Trihalomethanes     Turbidity   Potability
count  2.659000e+03  2.659000e+03  2.659000e+03  2.659000e+03  2.659000e+03  2.659000e+03    2.659000e+03     2.659000e+03  2.659000e+03  2659.000000
mean   3.300189e-16 -1.529845e-16  2.204580e-16  1.282665e-16  6.974489e-16  2.872634e-16    3.059689e-16    -2.859273e-16 -9.833762e-16     0.373825
std    1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00  1.000000e+00    1.000000e+00     1.000000e+00  1.000000e+00     0.483909
min   -2.608828e+00 -2.757635e+00 -2.686342e+00 -2.776471e+00 -2.473362e+00 -2.803209e+00   -2.777997e+00    -2.689334e+00 -2.745326e+00     0.000000
25%   -5.907856e-01 -6.361015e-01 -7.436250e-01 -6.519821e-01 -5.259709e-01 -7.525217e-01   -6.912579e-01    -6.369901e-01 -6.871035e-01     0.000000
50%   -2.906155e-02  1.871392e-02 -1.135635e-01  3.933790e-03 -8.020333e-03 -5.676029e-02   -2.520768e-02     1.700975e-03 -1.880153e-02     0.000000
75%    5.949718e-01  6.610337e-01  6.682352e-01  6.671543e-01  5.449903e-01  6.953704e-01    7.004958e-01     6.826896e-01  7.000670e-01     1.000000
max    2.620186e+00  2.782129e+00  2.909633e+00  2.755454e+00  2.513881e+00  2.838280e+00    2.805614e+00     2.737181e+00  2.786190e+00     1.000000

15 Primeras filas

          ph  Hardness    Solids  Chloramines   Sulfate  Conductivity  Organic_carbon  Trihalomethanes  Turbidity  Potability
0  -0.029062  0.279397 -0.086159     0.133179  1.329588      1.734437       -1.222534         1.363744  -1.312845           0
2   0.846787  0.963361 -0.198164     1.533465 -0.008020     -0.088464        0.789095        -0.006432  -1.190952           0
3   1.025998  0.614663  0.069709     0.671217  0.886716     -0.780826        1.275160         2.253002   0.875009           0
4   1.664528 -0.561653 -0.443387    -0.400954 -0.893502     -0.341131       -0.857185        -2.299205   0.147717           0
5  -1.224152 -0.306682  0.924600     0.306583 -0.263572     -1.816730       -1.836375        -0.772568  -1.842756           0
6   2.596345  1.806056  0.924731     0.284285  2.287164     -1.776898       -0.165417         1.204724  -1.693959           0
7   1.288738  0.225342 -0.990456    -1.806853 -1.153430      0.612177       -0.607458        -0.247670   0.576385           0
8  -0.029062 -2.757635 -0.912529     0.490369 -2.473362     -0.454173       -0.501362        -0.838444  -0.482852           0
10  0.238703 -1.112504  1.395079     0.310717 -0.265629     -0.003674        0.391712         0.814168  -0.394485           0
11  0.744187  0.767391 -0.343208     0.707401 -0.008020     -0.770420        0.062768         0.664028   0.064495           0
12  0.040410 -1.424184 -0.347888    -2.485121 -1.951787     -0.975393        0.497961         0.864841  -0.678910           0
13 -0.029062 -1.655052  0.744569    -0.194262 -1.301710     -0.574451        1.564800         0.665632   0.592869           0
14  0.350352  0.295466  0.878785    -1.445703 -0.008020      0.237315       -0.339454         0.252013   1.070214           0
15 -0.595728 -0.362558  2.489069     1.784164  1.176177      1.139341       -0.862920         0.569825   0.543446           0
16 -0.015615  0.497144  1.208101     2.113881 -0.008020     -1.382927        1.882940        -0.657087   0.401691           0

NaNs por columna

Series([], dtype: float64)
```

¿Qué paso? Ahora todos los datos están exactamente a 1 desviación estándar de la media, ósea: son cercanos entre sí. Si la media de la columna de pH era 7.07, el valor mínimo anterior, nos va a quedar de la siguiente forma:

(3.9 – 7.07) / 1.21 = -2.61...

Y el valor máximo:

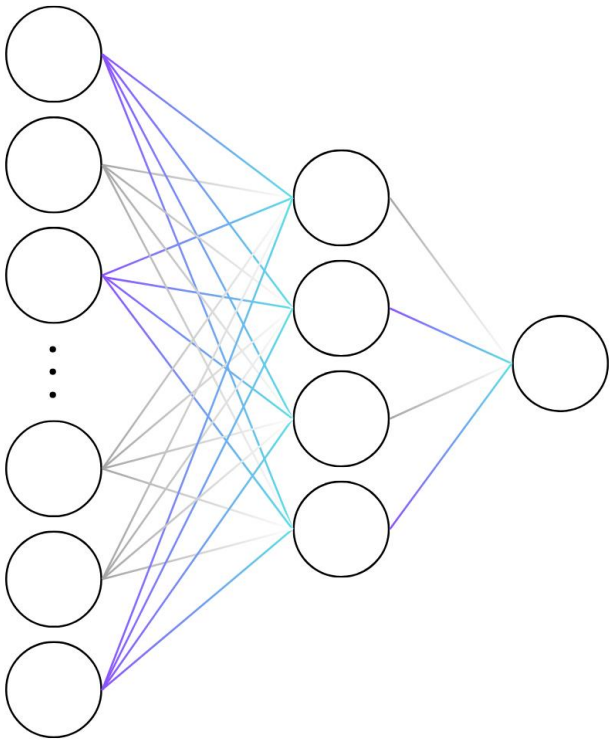
(10.25 – 7.07) / 1.21 = 2.62...

¡Epa! Ahora tenemos todos los valores “reducidos” a un mismo rango. Esto se hace para un optimo entrenamiento de la red, que viene a continuación.

Red Neuronal

Explicaciones y bocetos

Primero vamos a empezar armando un esquema de cómo va a ser nuestra red. De entrada, tenemos 9 datos (las 9 columnas), en la capa oculta, decidimos poner 4 neuronas, ya que fue lo que nos dio resultados mas consistentes. Y solamente 1 neurona de salida, que va a determinar si es, o no potable. Algo de este estilo:



Pensamos también en las multiplicaciones matriciales. Vamos a pasar a explicar este concepto, pero después de un pantallazo visual:

$$Z1 = W_{hidden} \times X + B_{hidden}$$

$$Z1 = \begin{bmatrix} W1 & W2 & W3 & W4 & W5 & W6 & W7 & W8 & W9 \\ W10 & W11 & W12 & W13 & W14 & W15 & W16 & W17 & W18 \\ W19 & W20 & W21 & W22 & W23 & W24 & W25 & W26 & W27 \\ W28 & W29 & W30 & W31 & W32 & W33 & W34 & W35 & W36 \end{bmatrix} \begin{bmatrix} X1 \\ X2 \\ X3 \\ X4 \\ X5 \\ X6 \\ X7 \\ X8 \\ X9 \end{bmatrix} + \begin{bmatrix} B1 \\ B2 \\ B3 \\ B4 \end{bmatrix}$$

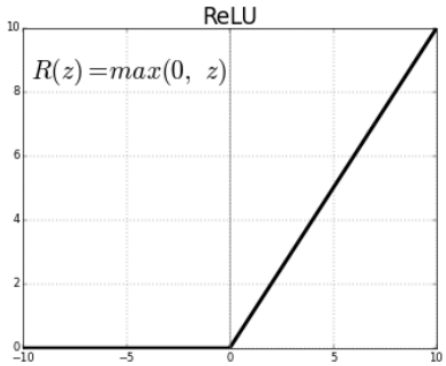
$$Z2 = W_{output} \times A1 + B_{output}$$

$$Z2 = \begin{bmatrix} W82 & W83 & W84 & W85 \end{bmatrix} \begin{bmatrix} A1 \\ A2 \\ A3 \\ A4 \end{bmatrix} + \begin{bmatrix} B10 \end{bmatrix}$$

Bien. Para empezar, decimos que *Z1 = Whidden x X + Bhidden*, ¿Y esto que es? Z1 es todo lo que se encuentra en las neuronas de la capa oculta, justo antes de que se aplique la función de activación (explicamos más adelante).

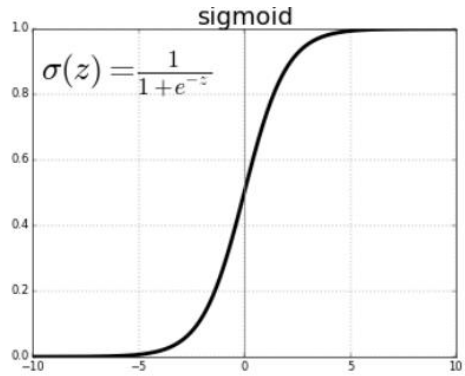
Pensemos cada neurona de entrada como una columna, la de arriba de todo sería la del pH, la de abajo la de Hardness...y así. Osea, a cada neurona de la capa oculta le van a llegar 9 datos, 1 por columna. A esos datos, se les va a aplicar un peso random (0 < random < 1), al cual vamos a llamar W. Entonces no solo le llegan 9 datos, sino que le llegan 9 datos multiplicados cada uno por un valor random, y a eso, se le va a sumar un sesgo. ¿Qué es un sesgo? Es un valor de partida de las neuronas, es decir, si todo lo anterior da cero, todavía tenemos ese valor. Sirve para ajustar y flexibilizar la red para que pueda adaptarse a distintos tipos de patrones.

Ya se hizo toda la multiplicación, ¿Con que seguimos? Aplicamos la función de activación. Hay muchos tipos de función de activación para las capas ocultas, personalmente vamos a utilizar la ReLu. ¿Qué hace la función ReLu? Toma los valores negativos, y los convierte en 0. Los valores positivos, los deja como están.



Ahora que se aplicó la función de activación, la capa pasa a llamarse de otra manera, A1. Y así como dijimos que Z1 eran las multiplicaciones matriciales antes de la función de activación, A1 son las multiplicaciones matriciales antes de la función sigmoide (¡pero posteriores a la función ReLu!), es decir, están justo antes de la salida de la última neurona.

Ahora, cuando se realizan las multiplicaciones matriciales entre A1, los nuevos pesos asignados a los datos y el bias (sesgo) de la ultima neurona, obtenemos Z2. Que como también paso anteriormente, es el dato previo a pasar por la función sigmoide, el cual lo convierte en A2, las predicciones de nuestra red.



Hasta ahí, tenemos lo que se conoce como 'Forward Propagation', pero todavía no termina. Cuando llegamos al final, se llama a la función de 'Backward Propagation', que lo que hace es ajustar los pesos y sesgos previamente mencionados (recordemos que empezaron con valores aleatorios) para minimizar el error de la red. Empezando desde la última capa (capa de salida), la red distribuye el error hacia capas anteriores para identificar cómo cada peso contribuyó al error, y utilizando el descenso de gradiente, la retropropagación modifica los pesos y sesgos para reducir el error en cada paso. Esto se hace calculando gradientes (derivadas parciales) respecto al error y ajustando los parámetros en la dirección que lo minimiza.

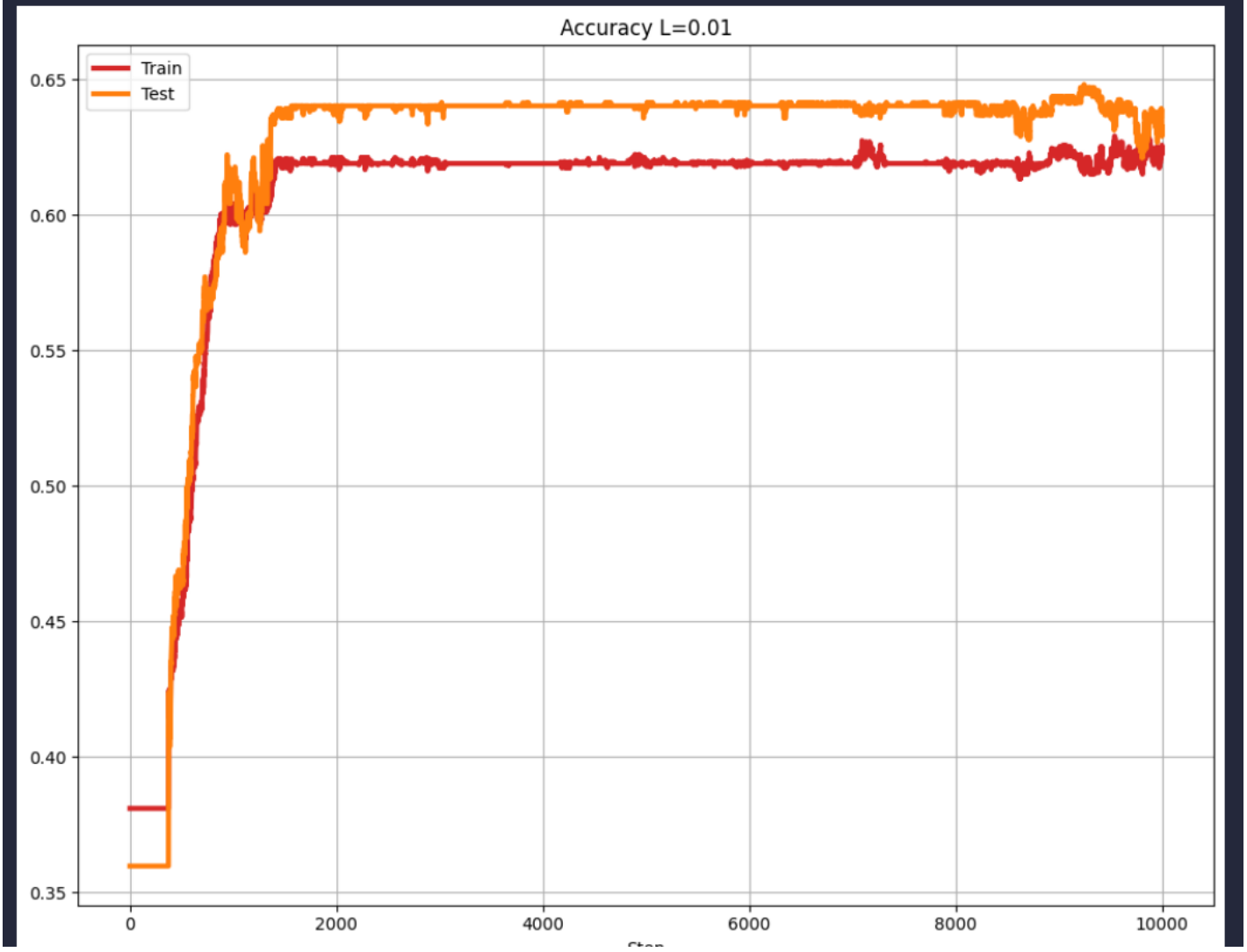
Armado de la Red

Procedimos a armar nuestra Red Neuronal, separando por primera y única vez nuestros datos en entrenamiento y testeo. (Como los nombres bien lo indican, los datos de entrenamiento indican a la red como hacer las cosas, y los de testeo, valga la redundancia, testean el aprendizaje de la red)

Generalmente, se dejan 1/3 de los datos para testeos, que fue lo que hicimos, y 2/3 quedan para el entrenamiento.

Una vez que todo estaba listo, entrenamos y probamos la red. Se le da un determinado paso (step) y una cantidad de repeticiones (epochs). El paso indica que tan grande va a ser el avance por iteración, y claramente los epochs, marcan cuantas veces se va a realizar el proceso.

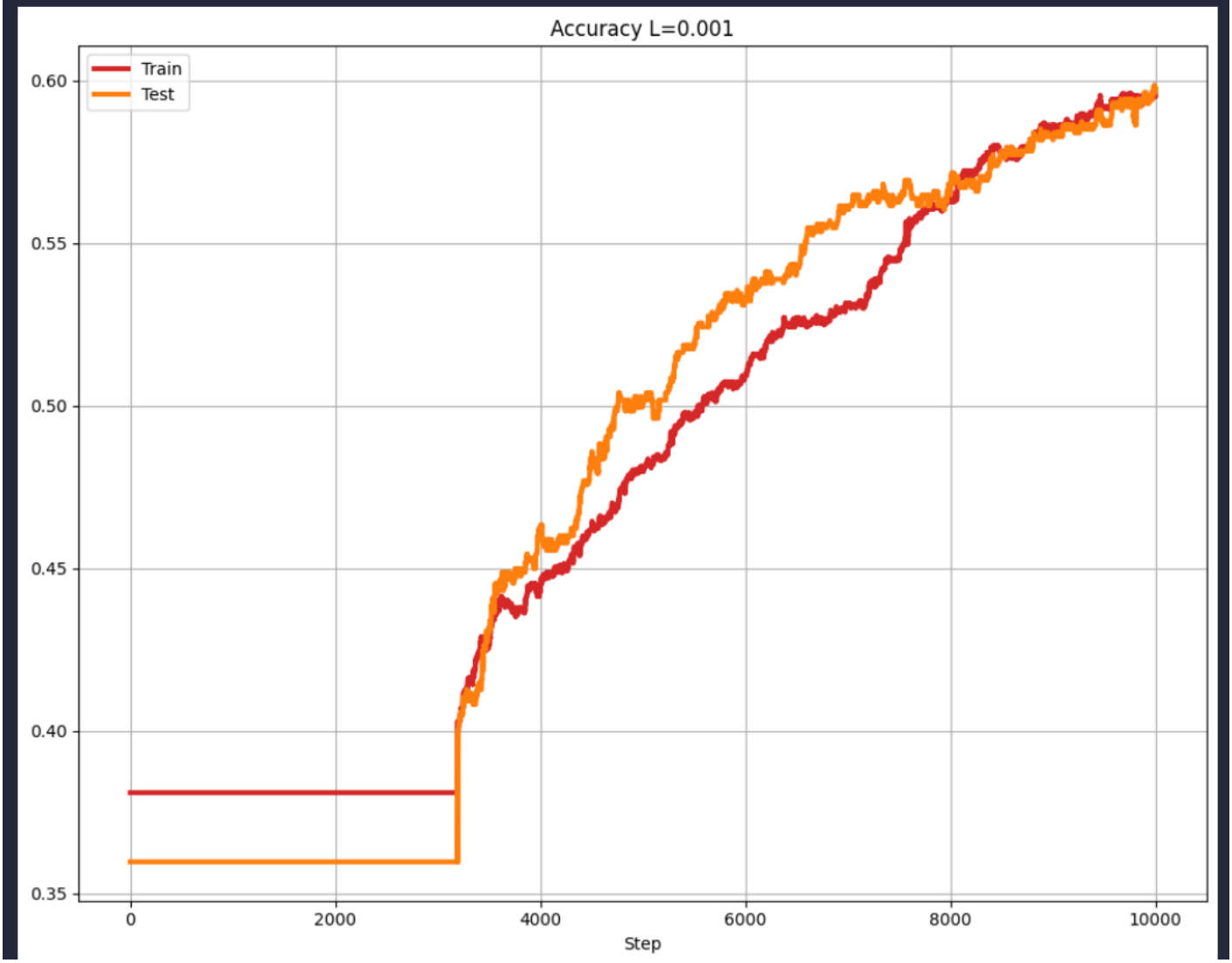
```
TRAIN ACCURACY:  0.6230248306997794
TEST ACCURACY:  0.6324689966178079
Epochs: 10000
Paso: 0.01
```



Este es el primer intento. 10.000 repeticiones, con un paso de un tamaño de 0.01. Como podemos ver arriba, también se calculan en la función de la red neuronal, 2 accuracies, uno para entrenamiento, y otro para testeo. Tenemos aproximadamente 63% en los 2, ósea que de un 100% (todos los datos), solo un 63% de las veces, va a ser capaz de determinar si una muestra de agua es o no potable.

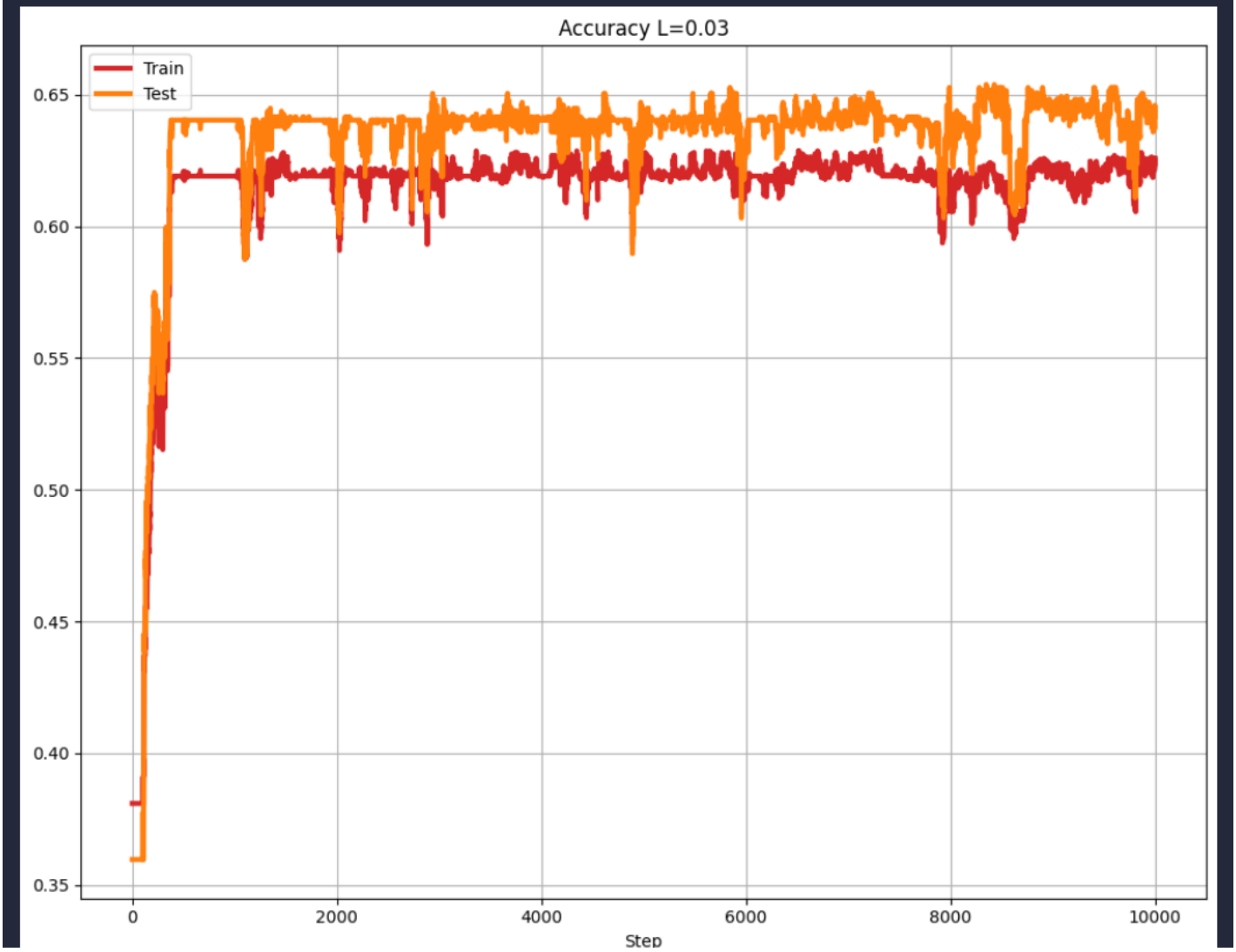
Hicimos lo mismo (en vano) con muchísimas repeticiones y pasos distintos, probamos también modificar cosas de adentro de la función, agregar datos inventados, normalizar de maneras distintas, pero no hubo caso. Con esta cantidad de datos, las redes (en plural porque también se probó realiza redes con los módulos de ScikitLearn y TensorFlow, obteniendo resultados muy similares) parecen no poder llegar a un resultado mucho mejor a este.

```
TRAIN ACCURACY:  0.5953724604966169
TEST ACCURACY:  0.5975197294250237
Epochs: 10000
Paso: 0.001
```



Mas pruebas, 10.000 repeticiones, pero con un paso de 0.001

```
TRAIN ACCURACY:  0.6258465011286736
TEST ACCURACY:  0.6426155580608742
Epochs: 10000
Paso: 0.03
```



10.000 pasos con un L = 0.03

Concluimos en que, con estos datos, no se puede alcanzar un nivel de acierto tanto mayor a este. La usabilidad de los datos tampoco parece ser demasiada, ya que, al tener valores faltantes, y muestras de casos raros, como por ejemplo un pH de 14 en una muestra potable cuando los niveles recomendados son entre 6.5 y 8.5.

Aprendimos mucho, más allá del corto alcance de la red, probamos muchísimas cosas nuevas para hacerla funcionar, así como distintas funciones de activación (sin tantos mejores resultados), más neuronas en la capa oculta, con mas capas ocultas...

Se puede ver el proceso completo en el archivo '.ipynb' del repositorio, así como el Data Set completo y el script de funciones utilizado (creado por nosotros) para el tratamiento de la red y gráficos.

¡Muchas gracias!