**Trabajo integrador entrega 3**

Minería de datos

**Grupo 8**

Integrantes:

* Lautaro Ezequiel Aria
* Fermin Moreno

Fecha de entrega: 7/8/21

[**Dominio**](#_sdipt2lt5vq1) **2**

[**Objetivos**](#_1684v7rtdkol) **4**

[**Potabilidad del agua**](#_x2b9l3yreo83) **4**

[Conclusiones](#_ejf070cx4mvh) 5

[**Agua de riego**](#_sciozgriahgr) **5**

[Conclusiones](#_8jamx24gwo21) 5

[**Dureza para un mejor sabor**](#_djixfilpiaup) **6**

[Conclusiones](#_e8zoxudzteqn) 7

[**Producción de alimento Daphnia magna (acuacultura)**](#_4uq1lbudttbb) **8**

[Conclusiones](#_f4hxrftggk57) **8**

[**Conclusiones finales**](#_odh8e1icmhu5) **9**

# Dominio

El conjunto de datos trata sobre propiedades de aguas para saber si las mismas son potables.

Cantidad de ejemplos: 3276

Atributos: 10

* **PH (Cuantitativo continuo):** Es el indicador de la condición ácida o alcalina del estado del agua. La OMS ha recomendado el límite máximo permisible de pH de 6,5 a 8,5.  
  *Rango de valores (0 - 13,99).*
* **Hardness (Cuantitativo continuo):** Es causada principalmente por sales de calcio y magnesio. Estas sales se disuelven a partir de depósitos geológicos a través de los cuales viaja el agua. El tiempo que el agua está en contacto con el material que produce dureza ayuda a determinar cuánta dureza hay en el agua cruda.   
  *Rango de valores (47,432 - 323,124).*
* **Solids (Total dissolved solids - TDS) (Cuantitativo continuo):** El agua tiene la capacidad de disolver una amplia gama de minerales o sales inorgánicos y algunos orgánicos, como potasio, calcio, sodio, bicarbonatos, cloruros, magnesio, sulfatos, etc. Estos minerales producen un sabor no deseado y un color diluido en apariencia de agua. Este es el parámetro importante para el uso del agua. El agua con alto valor de TDS indica que el agua está altamente mineralizada. El límite deseable de TDS es de 500 mg / ly el límite máximo es de 1000 mg / l que se prescribe para beber.   
  *Rango de valores (320,9426 - 61227,1960)*
* **Chloramines (Cuantitativo continuo):** El cloro y la cloramina son los principales desinfectantes que se utilizan en los sistemas públicos de agua. Las cloraminas se forman con mayor frecuencia cuando se agrega amoníaco al cloro para tratar el agua potable. Los niveles de cloro de hasta 4 miligramos por litro (mg / L o 4 partes por millón (ppm)) se consideran seguros en el agua potable.  
  *Rango de valores (0.352 - 13,127).*
* **Sulfate (Cuantitativo continuo):** Los sulfatos son sustancias naturales que se encuentran en minerales, suelo y rocas. Están presentes en el aire ambiente, el agua subterránea, las plantas y los alimentos. El principal uso comercial del sulfato es en la industria química. La concentración de sulfato en el agua de mar es de aproximadamente 2700 miligramos por litro (mg / L). Varía de 3 a 30 mg / L en la mayoría de los suministros de agua dulce, aunque se encuentran concentraciones mucho más altas (1000 mg / L) en algunas ubicaciones geográficas.  
  *Rango de valores (129 - 481,0306)*
* **Conductivity (Cuantitativo continuo):** El aumento de la concentración de iones mejora la conductividad eléctrica del agua. Generalmente, la cantidad de sólidos disueltos en el agua determina la conductividad eléctrica. La conductividad eléctrica (EC) en realidad mide el proceso iónico de una solución que le permite transmitir corriente. Según los estándares de la OMS, el valor de CE no debe exceder los 400 μS / cm.   
  *Rango de valores (181,4837 - 753,3426).*
* **Organic\_carbon (Cuantitativo continuo):** Carbono orgánico total es la cantidad de carbono unido a un compuesto orgánico y se usa frecuentemente como un indicador no específico de calidad del agua o del grado de limpieza de los equipos de fabricación de medicamentos. Se mide por la cantidad de dióxido de carbono que se genera al oxidar la materia orgánica en condiciones especiales.  
  *Rango de valores (2,1999 - 28,3).*
* **Trihalomethanes (Cuantitativo continuo):** Los THM son sustancias químicas que se pueden encontrar en el agua tratada con cloro. La concentración de THM en el agua potable varía según el nivel de material orgánico en el agua, la cantidad de cloro necesaria para tratar el agua y la temperatura del agua que se está tratando. Los niveles de THM de hasta 80 ppm se consideran seguros en el agua potable.  
  *Rango de valores (0,7379 - 124).*
* **Turbidity (Cuantitativo continuo):** La turbidez del agua depende de la cantidad de materia sólida presente en estado suspendido. Es una medida de las propiedades emisoras de luz del agua y la prueba se utiliza para indicar la calidad de la descarga de desechos con respecto a la materia coloidal. Según la OMS el valor de turbidez no debe pasar de 5,00 NTU y lo ideal sería que esté por debajo de 1.  
  *Rango de valores (1,45 - 6,379).*
* **Potability (Cualitativa nominal):** Indica si el agua es segura para el consumo humano, donde 1 significa potable y 0 significa no potable.

Forma de recolección de los datos:

Los datos son generados por una simulación por computadora, es decir, son datos sintéticos. Se generan por completo algorítmicamente.

Durante el transcurso del tp nos dimos cuenta que el dataset está creado con datos falsos, por eso hay datos faltantes y los solids están mal, abajo explicamos cómo encaramos estos problemas.

El link con la información de cómo fue generado el dataset es el siguiente:

<https://www.kaggle.com/adityakadiwal/water-potability/metadata>

Información relevante:

Existen datos faltantes en los atributos ph (491), sulfate(781) y trihalomethanes(162). Decidimos sacar estas filas con datos faltantes, ya que cada propiedad del agua es importante y como tenemos 3276 ejemplos optamos por sacarlos.

Se arregló el atributo Solids, el cual cuenta con valores erróneos, se generó un nuevo atributo Solidos que es Solids/100. En todas las pruebas el atributo original no se tiene en cuenta y se utiliza este nuevo atributo en su lugar.

Nueva cantidad de ejemplos:

Luego de eliminar las filas con datos faltantes el dataset quedó con 2011ejemplos.

# Objetivos

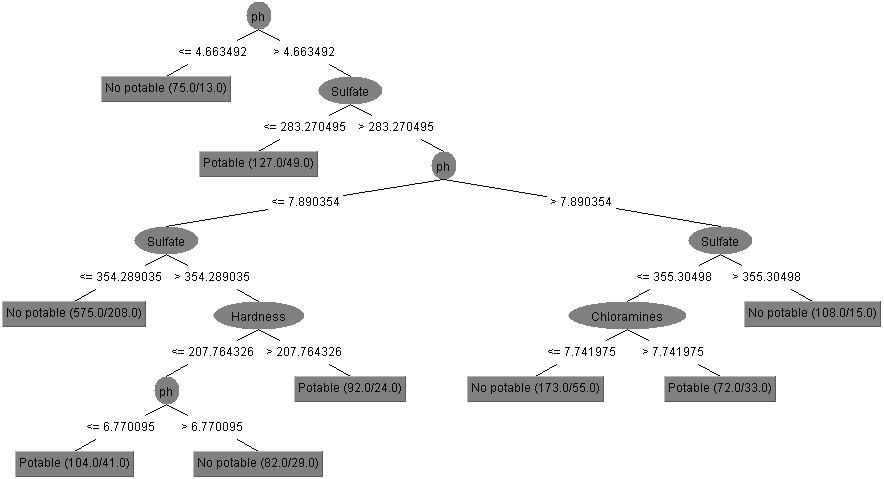
* Para una empresa potabilizadora le es útil predecir con qué propiedades el agua es potable. Para esto vamos a hacer un árbol de clasificación (modelo predictivo).
* Para una empresa de riego le resultaría útil predecir cuando el agua es recomendable para el riego. Para esto se utilizó reglas de clasificación (modelo predictivo).
* Para una empresa que vende agua para un consumidor final, le resultaría útil saber con qué propiedades el agua obtiene el mejor sabor. Para esto se utilizó el modelo de naive bayes (modelo predictivo y descriptivo).
* Para una empresa de acuacultura, le es útil predecir si el agua es óptima para la producción del alimento vivo Daphnia magna. Para esto vamos a utilizar una red neuronal. (Modelo predictivo).

# Potabilidad del agua

Resultaría útil predecir y conocer reglas a seguir de cuando un agua es potable o no, para esto decidimos utilizar W-J48 porque tenemos todos atributos numéricos, y el algoritmo ID3 no soporta atributos numéricos.

Para realizar el árbol se creó un nuevo atributo “Potabilidad”, el cual tiene los valores “Potable” y “No potable” cuando el valor Potability es 1 o 0 respectivamente. Se quitaron tanto el atributo Potability como Potabilidad, y se usó Potabilidad como label.

Se utilizó un split para separar 70% para entrenamiento y 30% para testeo (No se utilizó 80-20 como en el resto de análisis ya que fuimos probando distintos porcentajes para encontrar el mejor árbol posible, y con 70-30 obtuvimos los mejores resultados para dar un análisis más certero).

El modelo W-J48 se utilizó con los valores C=0.5 y M=70.

## Conclusiones

**Accuracy del árbol(67.00%)**

Como propiedad más determinante tenemos el ph, el cual debe estar entre 6.5 y 8.5 según la OMS, en este caso si es menor a 4.6 no es potable.

Luego con un valor de ph más apropiado (mayor a 4.6) y con un sulfato bajo (menor a 283 mg/L; 200 es el valor promedio que suelen tener las aguas potables, con un límite tolerable hasta 400) el agua es potable.

Luego nos quedan aguas con relativamente un alto grado de sulfato, este alto grado se subsana con cloro, con más de 7.7 ppm de cloro alcanza para potabilizar un agua con sulfato mayor a 283 mg/L y menor a 355 mg/L.

También se soporta un alto grado de sulfato con una dureza menor a 207 mg/L y ph menor a 6.7. O simplemente con una dureza mayor a 207 mg/L.

# Agua de riego

Resultaría útil conocer reglas para predecir si un agua es recomendable para la agricultura y para observar los aspectos más importantes a tener en cuenta a la hora de utilizar el agua para agricultura.

Para realizar las reglas se generó un nuevo atributo “Riego” el cual tiene los valores “Recomendado” si el ph del agua está entre 5.5 y 6.5, y “No recomendado” en caso contrario. Este atributo es utilizado como label, y el ph es quitado. (El ph es un atributo fácil de medir, por lo que en la vida real no tendría mucho sentido quitar el ph para predecir si el agua es apta para riego o no. Pero se hizo desde el punto de vista matemático, para la realización del trabajo práctico).

Se utilizó un split que separa un 80% para entrenamiento y un 20% para testeo.

El modelo W-PART se utilizó con los valores C=0.6 y M=10.

Las reglas son las siguientes:

* Hardness > 193.43213 AND Potability <= 0 AND Sulfate > 319.494952: No recomendado (360.0/34.0)
* Potability > 0: No recomendado (738.0/148.0)
* Chloramines <= 7.691012 AND Sulfate > 352.229961 AND Turbidity <= 4.014682: No recomendado (60.0)

## Conclusiones

**Accuracy del modelo (80.10%)**

Lo más importante es si la dureza del agua es mayor a 193 mg/L, el agua no es potable (este aspecto no tiene mucho para analizar) y el sulfato es mayor a 319 mg/L no es recomendable para riego:

Esto se debe a que las aguas duras disminuyen la vida media de los agroquímicos.

Es decir, un insecticida o fungicida se degradará más rápido cuando se aplica usando aguas duras, que cuando se usan hablas de calidad. Las aguas duras reducen la eficiencia de los agroquímicos al reducir su vida media.

Las aguas duras también poseen una conductividad eléctrica elevada. Este factor puede limitar el uso de estas aguas en el riego de ciertos cultivos susceptibles a conductividades eléctricas elevadas.

Y un alto nivel de sulfato genera una deficiencia de Mg (magnesio) y de Ca (calcio), las cuales son necesarios para algunos tipos de cultivos.

Además si el agua es potable tampoco es recomendable para riego, debido a que esta agua suele contener cloro y cal, los cuales son elementos nocivos para el crecimiento de las plantas.

Por último tenemos una combinación de propiedades, las cuales sucede en muy pocos ejemplos, por lo que no es muy determinante. Dice tener menos de 7 ppm de cloro, más de 352 mg/L de sulfato (el cual es incluso más alto que en la primer regla), y una turbiedad menor a 4 NTU. El punto determinante acá es el sulfato, el cual es muy elevado, las otras 2 propiedades no determinan mucho.

# Dureza para un mejor sabor

El agua blanda es la más recomendable para beber o cocinar, porque:

1. Muchas personas se quejan de que cuando beben agua dura, pueden sentir la "textura" del agua en la boca. Con agua blanda se elimina esta sensación y la experiencia de beberla será mucho más suave y placentera.
2. El agua dura a veces puede tener un sabor químico Otra queja común entre quienes han probado el agua dura es que sabe a tiza o químicos. Este es el resultado del mayor nivel de minerales disueltos que contiene. El agua blanda se trata de manera diferente y contiene menos de estos minerales, lo que garantiza un sabor más claro y libre de químicos.
3. Es menos probable que tenga trozos de cal en su bebida. El agua dura puede hacer que los grifos y los hervidores de agua acumulen cal, que a menudo pueden desprenderse de su bebida si llena un vaso directamente del grifo. El agua blanda no produce cal en la misma medida, por lo que es mucho menos probable que encuentre trozos calcáreos desagradables en su bebida.
4. El agua blanda tiene un sabor más neutro Además del sabor químico, la gente suele decir que el agua dura tiene un "sabor" diferente, mientras que el agua blanda no tiene ningún sabor. El hecho de que el agua blanda tenga un sabor más neutro significa que puede ser disfrutada por una amplia gama de personas, independientemente de sus gustos y preferencias. También significa que no afectará el sabor de otras bebidas a base de agua, como té, café o calabaza.

Resultaría útil conocer cuando un agua es blanda, ya que la misma suele ser más recomendable para beber o cocinar, vamos a utilizar Naive Bayes para predecir las aguas que cumplan con las condiciones, y además aprovecharemos la tabla de distribución para sacar algunas conclusiones en cuanto a sus propiedades.

Para realizar Naive Bayes generamos un nuevo atributo “Dureza” el cual tiene los valores “BLANDA” si está por debajo de 150 ppm, y “NO BLANDA” en caso contrario. Este nuevo atributo se setea como label, y se quita el valor Hardness original.

Se utilizó un split el cual separó un 80% para entrenamiento y un 20% para testeo.

El modelo de naive bayes se utilizó con corrección de laplace ya que trabajamos con relativamente pocos registros.

Tabla de distribución:

Se marcaron los atributos que se analizaron en la conclusión.

| **Atributo** | **Parámetro** | **No blanda** | **Blanda** |
| --- | --- | --- | --- |
| ph | media | 7.129 | 6.920 |
| ph | Desviación estándar | 1.530 | 2.155 |
| Chloramines | media | 7.135 | 7.344 |
| Chloramines | Desviación estándar | 1.579 | 1.777 |
| Sulfate | media | 333.244 | 343.238 |
| Sulfate | Desviación estándar | 40.513 | 53.389 |
| Conductivity | media | 426.084 | 429.663 |
| Conductivity | Desviación estándar | 79.802 | 77.277 |
| Organic\_carbon | media | 14.442 | 14.101 |
| Organic\_carbon | Desviación estándar | 3.358 | 3.225 |
| Trihalomethanes | media | 66.623 | 65.929 |
| Trihalomethanes | Desviación estándar | 16.319 | 14.668 |
| Turbidity | media | 3.954 | 4.104 |
| Turbidity | Desviación estándar | 0.781 | 0.770 |
| Potability | media | 0.405 | 0.492 |
| Potability | Desviación estándar | 0.491 | 0.502 |
| Solidos | media | 217.267 | 228.844 |
| Solidos | Desviación estándar | 84.915 | 92.884 |

## Conclusiones

**Accuracy del modelo (92.29%).**

El modelo encontró 29 ejemplos de aguas blandas.

Para conseguir un agua blanda que es la recomendada, se puede apreciar que el carbono orgánico en el agua y los sólidos tienen que ser bajo, es decir, más limpia y menos contaminante, aproximadamente 14 ppm y 228 ppm respectivamente. La media de sólidos es un poco mayor, por lo que pudimos ver, esto sucede porque hay unos pocos ejemplos de estos 29 que tienen un valor muy alto de sólidos, lo que aumenta la desviación estándar, como se puede apreciar en el cuadro. Lo cual podría indicar que los sólidos en realidad tienden a ser menores.

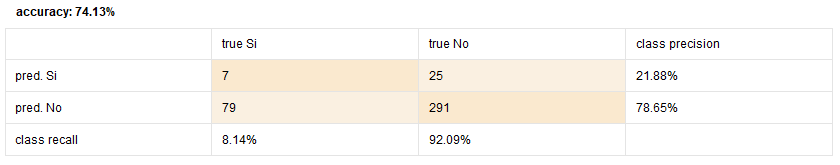
También se puede apreciar que el cloro en el agua es más alto que en las otras dos, esto es de ayuda para potabilizar el agua, de hecho, también se puede ver que suelen ser más potables.

El valor de Trihalometanos se genera durante la desinfección del agua debido a la reacción del cloro con la materia orgánica presente en el agua. El valor de cloro es más alto, pero en el agua se encontraba menor materia orgánica, por lo que este valor es más bajo, esto significa que el agua es más sana.

# Producción de alimento Daphnia magna (acuacultura)

Resultaría útil predecir según las propiedades de un agua si la misma es apta para la producción del alimento vivo Daphnia magna, el cual es utilizado en la acuacultura, para esto vamos a utilizar una red Neuronal.

Para realizar la red neuronal generamos un nuevo atributo “Optimo” (que sería el agua apta para acuacultura) el cual tiene el valor “Si” si tiene “ph” >= 7.1 y <= 8, y si tiene “Hardness” >= 10 y >= 250, y “No” en caso contrario. Este nuevo atributo se selecciona como label.   
Se quitaron ph y hardness. (El ph es un atributo fácil de medir, por lo que en la vida real no tendría mucho sentido quitar el ph para predecir si el agua es apta o no. Pero se hizo desde el punto de vista matemático, para la realización del trabajo práctico).   
Se utilizó un split que se separó un 80% para entrenamiento y un 20% para testeo.

**Accuracy del modelo (74.13%):**

## Conclusiones

Como podemos observar la predicción se encuentra sesgada por la clase “No” y pocos resultados de la clase “Si”, esto sucede ya que solo unas pocas aguas serán aptas para la acuacultura porque tiene unas condiciones en un rango bastante específico de cumplir.

# 

# Conclusiones finales

Como pudimos observar gracias a la recolección de datos obtenidos con nuestro dataset:

En el caso de querer obtener **agua potable** necesitamos un ph mayor a 4.6 y un sulfato menor a 283 mg/L. En el caso de tener un sulfato mayor a 283 y menor a 400 se subsana con cloro, con más de 7.7 ppm alcanza. Otra forma de soportar el alto grado de sulfato es teniendo una dureza menor a 207 mg/L y ph menor a 6.7, o directamente dureza mayor a 207 mg/L.

En el caso de querer obtener **agua apta para riego** se necesita una dureza mayor a 193 mg/L, que no sea potable, sulfato mayor a 319 mg/L y una conductividad eléctrica baja.

Tener bastante cloro (menos de 7.6 ppm) y mucho sulfato (más de 353 mg/L) no es recomendable.

En el caso de querer obtener el **agua más recomendable para beber o cocinar (blanda)** se necesita que el carbono orgánico y los sólidos tiendan a ser bajos, que el cloro sea alto, pocos trihalometanos, y lo ideal es que en general sea potable.

En el caso de querer obtener **agua apta para la producción del alimento del Daphnia Magna** se pudo observar que con “ph” >= 7.1 y <= 8, y “Hardness” >= 10 y >= 250” en un 74,3% de los casos el agua es apta para esta práctica.