Artículo científico

Prospectiva Universitaria
p-ISSN: 1990-2409 / e-ISSN 1990-7044
Vol. 18, Número 1, Enero – Diciembre 2021, pp. 161 - 166
https://doi.org/10.26490/uncp.prospectivauniversitaria.2021.18.1421

Aplicación de Data Science para la evaluación de la turbidez en el tratamiento de agua potable de la ciudad de Huancayo

Data Science application for the evaluation of turbidity in the drinking water treatment of the city of Huancayo

Anieval Cirilo Peña Rojas[1] / Helar Iván Fernández Véliz[1] / Gerson Yovanni Orihuela Maita[2]
0000-0001-9853-7532 / ---------- / 0000-0002-3701-5404

Autor corresponsal: acpenia@uncp.edu.pe
hveliz@uncp.edu.pe / e_2014200089d@uncp.edu.pe

Cómo citar:
Peña Rojas, A. C.; Fernández Véliz, H. I. & Orihuela Maita, G. Y. (2021). *Aplicación de Data Science para la evaluación de la turbidez en el tratamiento de agua potable de la ciudad de Huancayo.* Prospectiva Universitaria, revista de la UNCP. 18(1), 161-166. https://doi.org/10.26490/uncp.prospectivauniversitaria.2021.18.1421

Resumen

La turbidez es un contaminante físico de mayor presencia en el agua a tratarse, cuando su producción está destinada, sobre todo, a uso doméstico. Un método muy práctico y económico de reducir este contaminante es utilizando coagulantes químicos o polímeros, los cuales, bajo el principio de iones bipolares, pueden desestabilizar los coloides y propiciar la precipitación de sólidos solubles totales. Posteriormente, a la desestabilización se forman los flocs, que vienen a ser agrupaciones de material fino en suspensión que, para su rápida precipitación, son acelerados con la adición de floculantes especiales. La finalidad, de la presente investigación, está relacionada a la aplicación de algoritmos computacionales para crear un modelo de predicción de la dosis óptima de coagulante en la reducción de la turbidez de agua de una planta de tratamiento de la ciudad de Huancayo. Para ello, se tomó datos del laboratorio de la planta, con una antiguedad de nueve meses, los que fueron filtradas y procesadas para la aplicación de la ciencia de datos en dos fases principales importantes: la de entrenamiento, con el 70 % de datos y; de prueba, con el 30 %. El principal hallazgo fue que, los resultados de predicción llegaron a un 70 % de similitud con los resultados verdaderos, teniendo en cuenta las variables independientes de turbidez inicial, turbidez final, pH, color, entre otras. Se concluye que, el entrenamiento y validación del algoritmo más eficiente es el de Random forest con 82 % y 72 %, respectivamente; asimismo, los factores más relevantes son: turbidez, color de los sólidos disueltos totales y conductividad para predecir la dósis óptima de coagulante con el modelo generado. En ese sentido, el modelo podría servir para propósitos de mejora en el tratamiento de agua.

Palabras clave: evaluación de turbidez, agua potable, Data Sciencie, contaminantes, coagulante químico

Abstract

Turbidity is a physical pollutant with a greater presence in the water to be treated, when its production is intended, above all, for domestic use. A very practical and economical method of reducing this pollutant is by using chemical coagulants or polymers, which, under the principle of bipolar ions, can destabilize colloids and promote the precipitation of total soluble solids. Subsequently, after destabilization, flocs are formed, which become clusters of fine material in suspension which, for their rapid precipitation, are accelerated by the addition of special flocculants. The purpose of this research is related to the application of computational algorithms to create a prediction model of the optimal coagulant dose in the reduction of water turbidity in a treatment plant in the city of Huancayo. For this purpose, data was taken from the plant laboratory, which was nine months old, which were filtered and processed for the application of data science in two major important phases: training, with 70 % of data and; test, with 30 %. The main finding was that the prediction results were 70 % similar to the true results, taking into account the independent variables of initial turbidity, final turbidity, pH, color, among others. It is concluded that the training and validation of the most efficient algorithm is that of Random Forest with 82 % and 72 %, respectively; likewise, the most relevant factors are: turbidity, color of total dissolved solids and conductivity to predict the optimal coagulant dose with the generated model. In that sense, the model could serve purposes of improvement in water treatment.

Keywords: assessment of turbidity, drinking water, Data Science, contaminants, chemical coagulant

[1]Docentes de la Facultad de Ingeniería de Sistema / [2]Investigador invitado

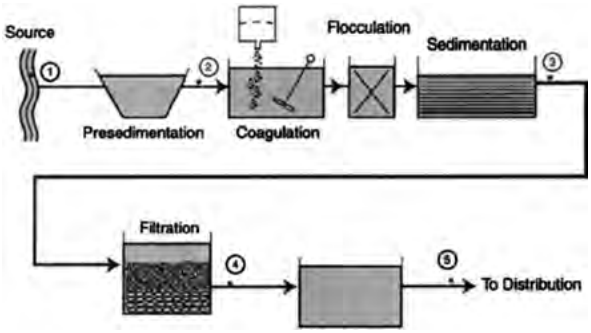
Fecha de recepción: 10/11/2021 *Fecha de evaluación: 17/12/2021* *Fecha de aprobación: 23/12/2021*

Introducción

El flujograma que más se utiliza en el tratamiento de agua es el referido a American Water Works (2003), que en su publicación de calidad de agua, como se puede observar en la Figura 1, la coagulación es uno de los procesos iniciales más importantes:

Figura 1

Esquema básico de tratamiento de agua por coagulación.



Fuente: American Water Works (2003)

El consumo de agua potable por la población de Huancayo fue evidenciada en muchas oportunidades, al no ser de calidad adecuada, siendo observado en los indicadores de vigilancia, tales como: turbidez, cloración, análisis bacteriológico y parasitológico (Mera & Coronel, 2016).

El agua a tratarse es captada en el nevado Huaytapallana, a más de 4 000 msnm, represado en las lagunas de Lazohuntay y Chuspicocha, allí comienza el recorrido del río Shullcas que, a lo largo de 10 km, aproximadamente, llega hasta la planta de tratamiento ubicada en el Centro Poblado de Vilcacoto.

Figura 2

Cuenca del río Shullcas, en la zona de Junín.

Fuente: Google Maps (2019)

A lo largo del recorrido del río, se incorporan sólidos de carácter muy fino por el deslizamiento de sedimentos del suelo, lo cuales generan la turbidez crítica, sobre todo en época de lluvias, entre los meses de setiembre a marzo.

Marco teórico

Antecedentes

Según Villagra et al. (2016), "los algoritmos evolutivos computacionales fueron utilizados para proponer soluciones de optimización en el tratamiento del agua o aguas residuales". En esta investigación, el autor propone los algoritmos: Genético Celular – cGA- y CHC, (Crossover elitism population, Half uniform crossover combination, Cataclysm mutation). Ello a conllevado a una propuesta muy positiva en la distribución del agua en la localidad de Caleta Olivia a través de la red optimizada.

Los investigadores Khan & See (2016) proponen un estudio que permita desarrollar "la predicción de la calidad del agua basado en factores de calidad del agua utilizando la red neural artificial (ANN) y el análisis de series de tiempo". Dicha investigación utiliza los datos históricos de la calidad del agua del año 2014, con un intervalo de tiempo de 6 minutos. La data se obtuvo del Servicio Geológico de los Estados Unidos (USGS), llamado Sistema Nacional de Información del Agua (NWIS). Fue propuesta basado en cuatro factores que influyen en la calidad del agua; para ello, se tuvo modelos de regresión calculadas con métodos computacionales.

Kim et al. (2014) investigan "las aguas costeras contaminadas, proponiéndose un modelo de monitoreo de la calidad de agua marina. El estudio se enfoca a la clorofila-a (chl-a); asimismo, de sólidos suspendidos (SPM), en las costas de Corea del Sur. Para ello, se utilizó los datos satelitales de imágenes geográficas del color del océano (GOCI). Para el aprendizaje del algoritmo, se propusieron los siguientes: el automático, que incluían bosque aleatorio, cubista y regresión de vectores de soporte (SVR) para la estimación de la calidad del agua costera.

Los controles in situ (63 muestras) recolectadas durante cuatro días en 2011 y 2012 se usaron como datos de referencia. Por la poca cantidad de muestras, utilizaron validación cruzada (CV), para proponer el modelo más óptimo de calidad de agua. "Los resultados muestran que SVR superó a los otros dos enfoques de aprendizaje automático, produciendo una calibración R2 de 0.91 y un error de raíz cuadrática media (RMSE) de 1.74 mg/m^3 (40.7 %) para chl-a, y una calibración R2 de 0.98 y CV RMSE de 11.42 g/m^3 (63.1 %) para SPM, cuando se usan datos de radiancia derivados de GOCI.

Se examinó la importancia relativa de las variables predictoras. Cuando se utilizaron datos de radiancia derivados de GOCI, la relación de la banda

2 a la banda 4 y, las bandas 6 y 5 fueron las variables de entrada más influyentes en la predicción de las concentraciones de chl-a y SPM, respectivamente. Las imágenes GOCI, disponibles por hora, fueron útiles para analizar las distribuciones espacio-temporales de los parámetros de calidad del agua con fases de marea en la costa oeste de Corea" (Kim et al., 2014)

Menezes et al. (2009) plantearon que la "coagulación es un proceso crítico para reducir la turbidez en el tratamiento del agua, se estima la coagulación óptima basada en la prueba de jarras y que consiste en poner diferentes dosis de coagulante para niveles de turbidez heterogénea con procesos de mezcla rápida y lenta, llegándose a establecer la de mayor eficiencia, controlando, asimismo, los otros parámetros físicos como el pH, color y conductividad. Debido a que ello consume mucho tiempo y las respuestas no son inmediatas, se proponen realizar un modelo de predicción; para ello, se configuraron topologías de redes neuronales y sus procesos de entrenamiento y prueba de perceptrón multicapa que calculan las dosis de hidróxido de aluminio y sodio, como coagulante y alcalinizante, respectivamente. Las conclusiones de estas son correctas con alta probabilidad de asertividad en la predicción para nuevos casos" (Menezes et al., 2009).

Los autores, Iriondo & Mota (2004) proponen que "la turbidez y oxígeno disuelto constituye una variable fundamental en la evaluación de la calidad del agua de los ríos, ya que determina la diversidad de organismos presentes en su seno. Para ello, desarrollaron un sensor software haciendo uso de la técnica de redes neuronales. Esta configuración a permitido reducir los costos de tratamiento en esa planta de aguas residuales (E.D.A.R.), a partir de los sensores para el control de los procesos de desnitrificación y decantación. Pruebas de validación, realizadas sobre datos tomados de las estaciones medioambientales de Navarra, prueban la efectividad de la solución propuesta" (Iriondo & Mota, 2004).

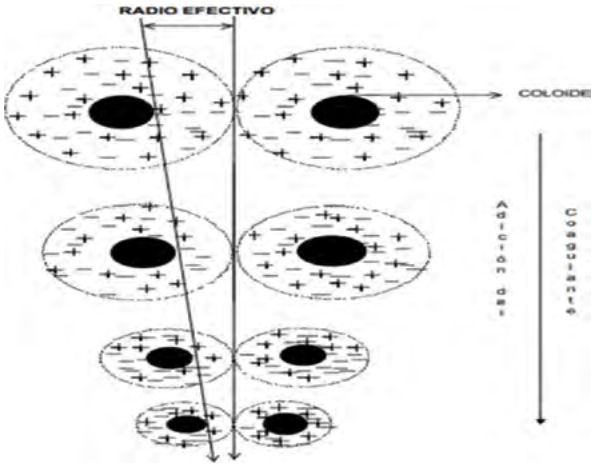
Marco conceptual

A. Coagulación

La coagulación, es un proceso fisicoquímico por la cual las partículas coloidales son desestabilizadas por efecto de un agente de carga opuesta, formándose los flocs y tenderán a precipitarse produciendo un colapso de la "nube de iones que rodean los coloides, de modo que puedan aglomerarse" (Andía, 2000). En la Figura 3, se muestra esa propiedad.

Figura 3

Principio teórico de la coagulación.



Fuente: Andía (2000).

La desestabilización puede obtenerse por los mecanismos fisicoquímicos siguientes:

- Compresión de la doble capa
- Adsorción y neutralización de cargas
- Atrapamiento de partículas en un precipitado
- Adsorción y puente

B. Data Science

La ciencia de datos o Data Science, según Ozdemir (2016), es "el arte de adquirir conocimiento mediante datos. Tiene que ver con la manera en cómo se toman los datos o, en otros términos, la manera en cómo se usa los datos para adquirir conocimiento que servirán en la toma de decisiones, predecir el futuro y/o comprender el pasado/presente".

Es decir, esta ciencia está basada en el análisis de datos históricos establecidas en una base de datos en las cuales se trata de encontrar patrones o comportamientos de respuesta a variables a estudiarse. "Están basadas en algoritmos computacionales basados en la estadística inferencial, el algebra lineal, lingüística computacional, análisis de gráficos, aprendizaje automático, inteligencia de negocios y almacenamiento y recuperación de datos" (Boschetti & Massaron, 2015).

C. Tipos de Datos

Los tipos de datos son importantes para implementar el algoritmo, "se pueden presentar en forma estructurada y no estructurada, las primeras son de mayor versatilidad en su organización y adecuado manejo; para ello, se harán filtros o limpieza de datos, las variables podrían ser de tipo categórico o cuantitativos discretos y continuos" (Ozdemir, 2016).

D. Los cinco pasos de la ciencia de datos

Los pasos a seguir, según Ozdemir (2016), son las siguientes:

a) Formular la pregunta u objetivo de la investigación:

De acuerdo con la estructura de los datos, se plantea la pregunta de analizar la factibilidad de procesar los datos; para ello, deben ser los suficientes y suficientemente concisos.

b) Obtener y preparar los datos

Considerar los tipos de datos es vital en ciencia de datos. Por ello, se debe analizar si el dato es tipo texto, booleano, categórico o cuantitativo, una mala propuesta podría caer en algoritmos ineficientes (Ozdemir, 2016). Los datos estructurados mediante tablas, son la de mejor rendimiento frente a los no estructurados que, sin ningún formato, son más difíciles de organizar.

La preparación de los datos, identificando su tipo y su función en el modelo, es muy importante, existen librerías dentro de los distintos aplicativos o lenguajes, que permiten refinar y filtrar adecuadamente dichos datos (Massaron & Boschetti, 2018).

"En la etapa de preparación o limpieza de datos, se pueden aplicar distintas estrategias en función de las necesidades que se requieren con los datos, para este caso, también se usa la librería Pandas. En el procesamiento de los datos o la construcción de las estructuras de datos, si es necesario, se prepara una matriz en cuanto a los datos filtrados, para usarlos en los procedimientos de aprendizaje supervisado y no supervisado, para lo cual se hace uso de la librería NumPy" (Massaron & Boschetti, 2018).

c) Explorar los datos

La visualización adecuada de los datos es muy importante para determinar la importancia de los factores que inciden en la variable dependiente (coagulante) "se pueden utilizar diagramas de barras, histogramas, diagramas de caja o diagramas de dispersión que permiten la visualización del comportamiento de las mismas" (Ozdemir, 2016).

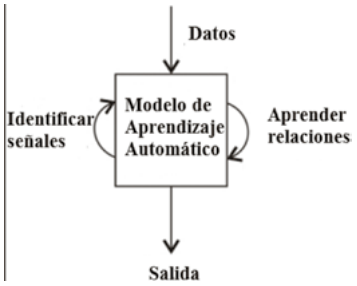
e) Modelar los datos

El modelamiento de datos es importante, pues genera un modelo para optimizar los procesos estudiados; para ello, se puede referir que, según Ozdemir (2016), "el aprendizaje automático difiere de los algoritmos clásicos debido a que a estos se les dice como encontrar la mejor respuesta, por el contrario en el aprendizaje automático, al modelo usado no se le indica la mejor solución sino que se entrena al modelo con ejemplos del problema y este determina la mejor solución"

En la Figura 4, se esquematiza el procedimiento respecto al modelo de aprendizaje con los datos históricos.

Figura 4

Una visión general de los modelos de aprendizaje automático.



Fuente: Ozdemir (2016)

El aprendizaje puede ser supervisado o no supervisado, dependiendo del algoritmo propuesto; estos son eficientes, siempre en cuando la cantidad de datos sea lo suficientemente grande y esté validado o con la limpieza de las mismas; es decir, no debería existir registros en blanco o datos no definidos.

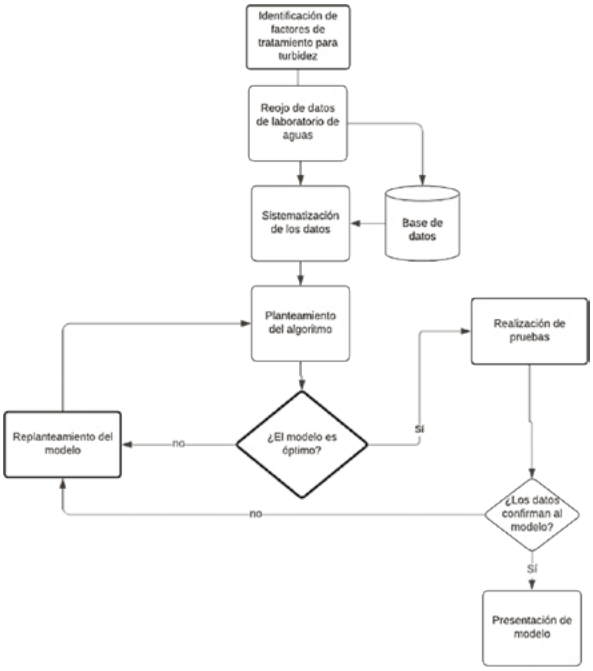
Materiales y métodos

La presente investigación es de tipo aplicado, debido a que se utiliza la teoría y modelo de Data Science enfocado a optimizarla dosificación del coagulante en el tratamiento de agua potable. Se ha seguido un diseño no experimental longitudinal y se obtuvo la muestra con data del laboratorio con registros de casi un año con variables independientes de pH; conductividad, sólidos disueltos totales, fundamentalmente.

La metodología se plantea en el siguiente esquema, en la que se detalla el procesamiento de los datos obtenidos del laboratorio de aguas de la empresa proveedora de ese elemento.

Figura 5

Metodología para el análisis de los datos de turbidez.

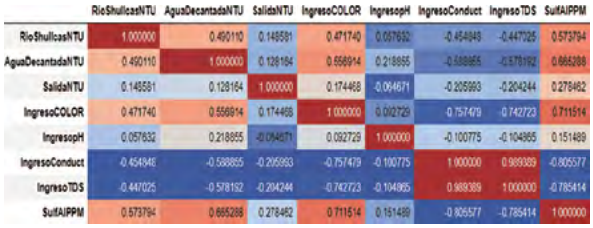


Resultados

El paso de exploración de datos es uno de los más sustanciales, esto se evidencia en la Figura N°6, la cual indica la correlación entre cada una de las variables con las demás. Teniendo en cuenta que la variable objetivo es el SulfAIPPM, las variables con una correlación alta, ya sea positiva o negativa, son: el agua decantada, el color, la conductividad y el TDS.

Figura 6

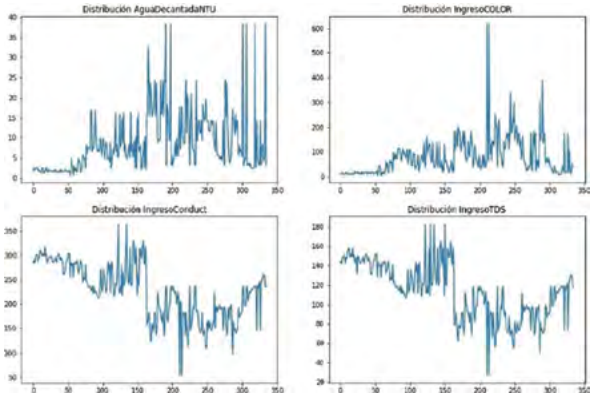
Matriz de correlación entre las variables.



En la Figura 7, se muestra la distribución de los datos de dichas variables a través del tiempo.

Figura 7

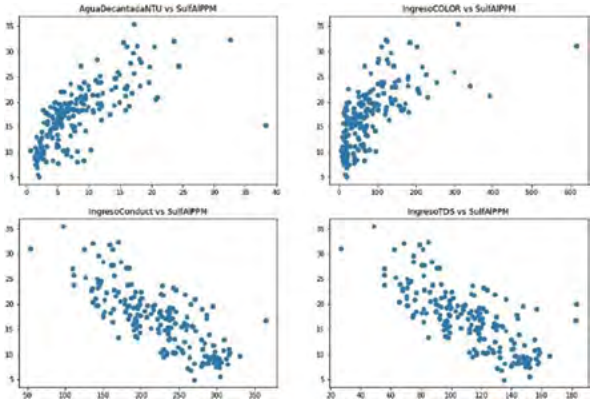
Distribución de los datos temporal longitudinal.



Para tener una mejor visualización de los datos en mención, que tienen una alta correlación con la variable objetivo, en la Figura 8, se muestra la distribución de los datos y se puede afirmar la tendencia obtenida en la tabla de correlaciones.

Figura 8

Diagramas de dispersión de las variables independientes con la dosis de coagulante.



Discusión

Con los resultados obtenidos se plantea el análisis acerca del comportamiento y contrastación de los modelos algorítmicos.

En la Tabla 1, se muestra el R^2 obtenido luego de modelar los datos con los algoritmos de Regresión Lineal y Random Forest.

Tabla 1

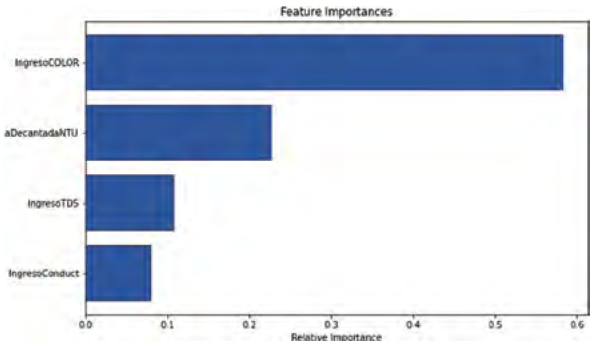
Resumen del índice de determinación para cada fase.

Modelos de machine learning	R^2	
	Datos de entrenamiento	Datos de validación
Regresión Lineal	72%	71%
Random Forest	82%	72%

Como se observa en la tabla anterior, el algoritmo que mejor modela los datos es el Random Forest; por lo que, en la Figura 9 se puede observar las variables mas influyentes en la variable objetivo, según este mismo algoritmo.

Figura 9

Factores mas influyenyes en el modelo para la dosis de coagulante.



Como se observa, se obtuvieron las mismas variables que se hallaron con la matriz de correlación, afirmando que estas variables son las mas influyentes en la variable objetivo.

Conclusiones

- El modelo Random Forest, basado en Data Science presenta un nivel alto de asertividad para la predicción. La fase de entrenamiento da un 82 % de asertividad, frente a un 72 % del modelo de regresión lineal.
- El modelo Random Forest, en la fase de validación, muestra más eficiencia que el modelo de regresión lineal con un 72 % de asertividad.
- Los factores que influyen en el modelo en orden de prioridad vienen a ser la turbidez, el color, los sólidos disueltos totales (TDS) y la conductividad.

- Existe correlación positiva de la dosis de coagulante con los factores color y turbidez y; una correlación inversa con la conductividad y el ingreso de sólidos disueltos totales.

Referencias bibliográficas

American Water Works. (2003). *Water Quality*. American Water Works Association.

Boschetti, A. & Massaron, L. (2015). *Python Data Science Essentials*. Packt Publishing Ltd.

Google Maps. (2019). *Google Maps*. Google Maps. https://www.google.com/maps/place/Capilla+de+Hualahoyo/@-11.9760384,-75.104562,18088m/data=!3m1!1e3!4m5!3m4!1s0x910ebd972336b6bd:0x9a414925675093c0!8m2!3d-11.9995597!4d-75.2203856

Khan, Y. & See, C. S. (2016). *Predicting and analyzing water quality using Machine Learning: A comprehensive model*. 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT), 1-6. https://doi.org/10.1109/LISAT.2016.7494106

Kim, Y. H.; Im, J.; Ha, H. K.; Choi, J.-K. & Ha, S. (2014). *Machine learning approaches to coastal water quality monitoring using GOCI satellite data*. GIScience & Remote Sensing, 51(2), 158-174. https://doi.org/10.1080/15481603.2014.900983

Menezes, F. C. de; Rodriguez Esquerre, K. P. S. O.; Kalid, R. de A.; Kiperstok, A.; Matos, M. C. de O. & Moreira, R. (2009). *Redes neurais artificiais aplicadas ao processode coagulação.* Engenharia Sanitaria e Ambiental, 14(4), 449-454. https://doi.org/10.1590/S1413-41522009000400004

Mera Arrivasplata, Katia, A. & Coronel Guevara, A. del R. (2016). *Propuesta de un método de control de los parámetros de calidad de agua cruda para obtener agua potable de óptima calidad en la empresa Epsel SA*. Tesis para optar el título de Ingeniero Químico. Universidad Nacional Pedro Ruiz Gallo. Lambayeque.

Ozdemir, S. (2016). *Principles of Data Science*. Packt Publishing Ltd.

Iriondo Saint-Gerons, A. & Mota Adrados, J. (2004). *Desarrollo de una red neuronal para estimar el oxígeno disuelto en el agua a partir de instrumentación de E.D.A.R. En: Jornada de Automática. Memorias de Jornada de Automática. España*.

Villagra, A.; Pereyra, G.; Molina, D.; Serón, N.; Goupillat, C. A.; Varas, V.; Montenegro, C.; Lasso, M. G. & Pandolfi, D. (2016, mayo 10). *Algoritmos evolutivos híbridos para el diseño y operación eficiente de una red de distribución de agua potable*. XVIII Workshop de Investigadores en Ciencias de la Computación (WICC 2016, Entre Ríos, Argentina). http://sedici.unlp.edu.ar/handle/10915/52719