Revista Innovación y Software
Vol. 3, No. 2, Mes Septiembre-Febrero, 2022
ISSN: 2708-0935
Pág. 121-131
https://revistas.ulasalle.edu.pe/innosoft

Tipo de artículo: Artículos originales
Temática: Inteligencia artificial
Recibido: 28/06/2022 | Aceptado: 02/08/2022 | Publicado: 30/09/2022

Identificadores persistentes:
ARK: ark:/42411/s9/a72
PURL: 42411/s9/a72

Modelo predictivo de la potabilidad del agua mediante un árbol de decisión en Inteligencia Artificial

Predictive model of water potability through a decision tree in Artificial Intelligence

Angel Alexis Zevallos Apaza [1], Sofía Sair Onque Gárate [2], Arian Eduardo Javier Canaza Cuadros [3], Paulina Miriam Choqueneira Ccasa [4]

[1] Universidad Nacional de San Agustín. azevallosa@unsa.edu.pe
[2] Universidad Nacional de San Agustín. sonque@unsa.edu.pe
[3] Universidad Nacional de San Agustín. acanazacua@unsa.edu.pe
[4] Universidad Nacional de San Agustín. pchoqueneira@unsa.edu.pe

[*] Autor para correspondencia: azevallosa@unsa.edu.pe

Resumen

En este trabajo se planteó como objetivo utilizar la técnica de árbol de decisión para definir un modelo capaz de predecir la potabilidad del agua. Para evaluar el rendimiento de la clasificación del árbol de decisión se utilizó un dataset extraído de Kaggle que cuenta con 3276 muestras de agua divididas por la variable de potabilidad. Aplicando las librerías Pandas y Scikit Learn se logró definir un modelo basado en un árbol de decisión evaluado con las métricas de precisión, exactitud, exhaustividad y puntuación F1 logrando 0.77, 0.80, 0.85 y 0.81 respectivamente.

Palabras clave: Agua potable, inteligencia artificial, árbol de decisión.

Abstract

The objective of this work was to use the decision tree technique to define a model capable of predicting water potability. To evaluate the performance of the decision tree classification, a dataset extracted from Kaggle was used, which has 3276 water samples divided by the potability variable. Applying the Pandas and Scikit Learn libraries, a model based on a decision tree evaluated with the metrics of precision, accuracy, completeness, and F1 score was defined, achieving 0.77, 0.80, 0.85, and 0.81, respectively.

Keywords: *Drinking water, artificial intelligence, decision tree.*

Revista Innovación y Software
Vol. 3, No. 2, Mes Septiembre-Febrero, 2022
ISSN: 2708-0935
Pág. 121-131
https://revistas.ulasalle.edu.pe/innosoft

Introducción

La vida del hombre y su existencia se la debe al agua, existe una alta necesidad de esta y cada vez más por el incremento de la población, por consecuencia ocurre una mayor demanda de agua. Existen desigualdades entre las zonas urbanas y rurales, puesto que el 96% de la población mundial urbana utiliza fuentes de agua potable frente al 84% de la población rural, tal como lo dicen en [1], por lo que el poseer agua potable es una necesidad primaria, como lo menciona la Asamblea General de las Naciones Unidas [2].

"Todas las personas tienen derecho a disponer de forma continuada de agua suficiente, salubre, físicamente accesible, asequible y de una potabilidad del agua aceptable, para uso personal y doméstico". Entonces podemos reafirmar que el consumo de agua se debe garantizar, el agua no debe estar contaminada o con sustancias que puedan producir enfermedades ya que este sería un problema muy grave.

Entrando en un contexto cercano, en el Perú más de un 70% de las aguas residuales no tienen tratamiento, la contaminación en el agua puede ser una gran preocupación para nosotros, porque pone en peligro la salud pública. Los principales lugares que superan los límites recomendados por la OMS para el consumo humano de agua son Lima, La Oroya y Juliaca de la cual puede ver más información en [3].

Incluso aquí en Arequipa mediante estudios se determinó que se superaron los parámetros establecidos en bacterias coliformes, que se evalúan en [4]. Por lo que para evitar que esta situación empeore se necesita un buen control del agua verificando que esta sea apta para consumo humano.

“La forma como se mide la contaminación química, los límites que se toleran y las decisiones que se toman al respecto de las afluentes de agua, depende de procesos de monitoreo y vigilancia.” [5]. Con esta perspectiva, este trabajo plantea la implementación de inteligencia artificial para determinar las condiciones de la potabilidad del agua a partir de datos que han sido obtenidos de un trabajo de análisis. Con el modelo se espera determinar con un error mínimo la potabilidad del agua, para beneficiar a la población que obtiene el líquido de afluentes, además de aportar a los procesos que determinan su potabilidad. Ello se plantea hallar en función a los distintos parámetros que influyen en el resultado de la potabilidad [6], los cuales incluyen el pH, dureza, sólidos disueltos totales, cloraminas, sulfato, conductividad, carbono orgánico, trihalometanos y turbidez.

Revista Innovación y Software
Vol. 3, No. 2, Mes Septiembre-Febrero, 2022
ISSN: 2708-0935
Pág. 121-131
https://revistas.ulasalle.edu.pe/innosoft

Materiales y métodos o Metodología computacional

Trabajos relacionados

En [7] se analizó el problema de la contaminación del agua un tema tratado también por nosotros, pero con un análisis en tiempo real lo cual tuvo resultados positivos, ya que mediante esta detección se pudo prevenir seguir usando agua contaminada, lo que demuestra que el modelo utilizado en este artículo puede ser muy factible y aplicarse a cualquier variable física que pueda ser medida con un elemento sensor y requiere cierto monitoreo.

En [8] se propuso una solución para determinar el índice de potabilidad del agua del río Utcubamba utilizando redes neuronales artificiales, la RNA planteada fue entrenada usando el algoritmo de Levenberg-Marquardt determinando la distribución óptima consistente en seis neuronas en la capa de entrada, doce neuronas en la capa oculta y una neurona en la capa de salida. La evaluación de la RNA se realizó mediante el coeficiente de correlación y el error de raíz cuadrada media. Los resultados para el coeficiente de correlación para los tres conjuntos de datos, entrenamiento, validación y prueba fueron 0.979, 1 y 0.940 respectivamente, para el error de raíz cuadrada media para los tres conjuntos de datos fueron 2.562 para entrenamiento, 1.546 para validación y 1.997 para la prueba.

En [9] se mantuvo el problema debido a la población, ya que a más población estas requerirán más necesidades con respecto al agua, pero controla las condiciones apropiadas para asegurar la potabilidad del agua. En muchas situaciones no es suficiente las actividades de monitorización que se brindan, por lo que se requiere contar con modelos o mecanismos que permitan anticiparse a la materialización del riesgo con el suficiente rango de tiempo para prevenir los efectos negativos que afecten la calidad del recurso hídrico. Los resultados obtenidos al final demostraron que la utilización de diferentes técnicas aplicadas, permiten obtener mejores resultados respecto a las técnicas que son utilizadas de forma independiente

En [10] se propone la revisión de las técnicas de aprendizaje automático y su aplicación para la estimación de la potabilidad del agua en ríos, cuencas y lagos, entre otros, debido a su importancia para la sobrevivencia de los seres vivos. Tras el desarrollo del trabajo, se evidenciaron en los resultados que, pese a que se puede encontrar una gran variedad de estrategias, las redes neuronales han abarcado este campo con buenos resultados, pero aún concentra su atención en los desafíos de combinar sus propiedades para modelar sistemas con características no lineales y no estacionarias. Finalmente, como conclusiones se afirma que las técnicas de aprendizaje automático de mayor aplicabilidad en el recurso hídrico son la redes neuronales, las máquinas de vectores de soporte y los sistemas de

inferencia neuro difusa con porcentajes del 36 %, 24 % y 16 %, respectivamente; el 24% restante corresponde a la implementación de otro tipo de estrategias. Con ello se evidencia que el modelado híbrido es una herramienta mejorada que arroja buenos resultados en comparación con técnicas predictivas tradicionales.

Fundamentación teórica

La inteligencia artificial

La Inteligencia Artificial (IA) es la combinación de algoritmos planteados con el fin de crear máquinas que poseen las mismas capacidades que el ser humano, una ansiada tecnología que todavía resulta lejana y misteriosa, pero que hace algunos años está presente en nuestra vida cotidiana.

Stuart Russell y Peter Norvig diferenciaban la inteligencia artificial en varios tipos, de forma que se tienen los sistemas que piensan como humanos, los que actúan como humanos, los que piensan racionalmente y los que actúan racionalmente.

Los campos de aplicación de la inteligencia artificial ciertamente son muchos y variados, por ejemplo, algunos de los principales son el uso de la IA para los asistentes virtuales, en el campo de la climatología, las finanzas, la agricultura, la educación, la logística y el sistema de transporte, el comercio y los sistemas de sanidad [11].

Machine Learning

Machine Learning y el Procesamiento del Lenguaje Natural son campos que convierten a los datos en la información necesaria para alcanzar la Inteligencia Artificial requerida para la extracción de conclusiones, realización de predicciones o comunicación con los usuarios. Las técnicas de Machine Learning permiten a los algoritmos identificar patrones complejos entre grandes cantidades de datos, infiriendo así sus reglas para detectar patrones similares en nuevos conjuntos de datos. Cuando se crean sistemas inteligentes que mejoran de forma autónoma viendo datos, se permite en sí la creación de sistemas que pueden aprender a predecir comportamientos mediante ejemplos, detectar similitudes o anomalías automáticamente o tomar las decisiones adecuadas [12].

Por ellos se dice que mientras dispongamos de más datos, más fiable y representativa será la muestra del proceso que buscamos automatizar, y mejor va a funcionar el software que buscamos generar.

Análisis de datos

El análisis de datos es el proceso de limpieza, cambio y procesamiento de datos en estado bruto, y la extracción de información relevante y procesable que ayuda a tomar decisiones informadas en base a estos. El preprocesamiento ayuda a reducir los riesgos inherentes a la toma de decisiones al proporcionar información y estadísticas útiles, que a menudo se presentan en cuadros o tablas [13].

Árbol de decisión

Un árbol de decisión es un mapa de los posibles resultados de una serie de decisiones relacionadas. Este mapa permite que un individuo o una organización comparen posibles acciones entre sí según sus costos, probabilidades y beneficios. También puede ser usado para dirigir un intercambio de ideas informal o trazar un algoritmo que anticipe matemáticamente la mejor opción.

Un árbol de decisión, por lo general, comienza con un único nodo y luego se ramifica en resultados posibles. Cada uno de esos resultados crea nodos adicionales, que se ramifican en otras posibilidades. Esto le da una forma similar a la de un árbol.

Algunas de las ventajas de este concepto son su fácil uso y comprensión, la mínima preparación necesaria, la posibilidad de agregar nuevas opciones a árboles ya existentes, así como la facilidad de combinación con otras herramientas de decisión. Sin embargo, también maneja ciertas desventajas, como el hecho de que el árbol se vuelva repentinamente complejo y que demande necesariamente la propuesta de una nueva solución [14].

Ganancia de información

Durante la construcción de un árbol de decisión iremos haciendo varias divisiones y la ganancia de información vendrá a ser precisamente esa información que puede aumentar el nivel de certeza después de una división. Es la entropía de un árbol antes de la división menos la entropía ponderada después de la división por un atributo, por lo que podemos pensar en la ganancia de la información y en la entropía como opuestos [15].

Preprocesamiento de datos

Para comenzar la elaboración de un árbol de decisión, en primera instancia, requerimos de un conjunto de datos (dataset) para trabajar en base a ello. Esta extracción comprende un conjunto de pasos que buscan preparar los datos sin limitarse con la integración u homogeneización, sino que abarca también otras tareas más, bajo la denominación genérica de preprocesamiento de datos [16].

Una vez seleccionado el dataset, se procede con la limpieza de datos o data cleaning, pues antes de la aplicación de una técnica para procesar y/o analizar un conjunto de datos determinado, es necesario aplicar una limpieza de datos mediante técnicas que permitan filtrar el dataset de los datos vacíos, nulos o incongruentes.

Seguidamente, se tiene el proceso de selección o extracción de datos relevantes, debido a que dos de las tareas más usuales en esta fase son la selección de variables y la selección de patrones. Puntualmente consisten en eliminar aquellos datos que, por estar repetidos o pueden estimarse a partir de otros, no aportan mejora en la extracción de conocimiento.

Finalmente, se tiene la transformación de los datos, teniendo lugar una vez que los datos ya se encuentran limpios y no contienen redundancias, aspectos de los que se ocupan las operaciones previas. En este punto, podría pensarse que ya pueden usarse para el aprendizaje de un modelo, sin embargo, hay acciones que podrían mejorar los datos de modo que haga más efectivo ese aprendizaje. Entre ellos están la normalización, el escalado y la discretización.

Herramientas y elementos

Herramientas

Las herramientas que usamos para el desarrollo del modelo predictivo para asegurar la potabilidad del agua fueron en su mayoría colaborativas para que podamos trabajar en equipo a distancia como Google Collab, el cual nos permite escribir y ejecutar código arbitrario de Python en el navegador. Lo tomamos como adecuado para realizar este tipo de tareas y análisis de datos.

Para el almacenamiento de nuestra dataset y los archivos, ya que Google Collab puede trabajar con los documentos de Google, usamos Google Drive, que nos sirvió como sitio de alojamiento de archivos el cual también nos ayudó para trabajar de forma colaborativa. Y para acabar Google Sheet, ya que es ahí en conjunto con Google Drive donde guardamos la dataSet, porque maneja un sistema de celdas, lo cual es permitido para poder arrastrar los datos con Python.

Librerías

Usamos librerías para facilitar el desarrollo de nuestro sistema predictivo. Una de ellas fue Numpy, la cual tiene soporte para vector y matrices grandes multidimensionales junto con funciones matemáticas de alto nivel. Para complementar esto y añadirle gráficas a nuestro proyecto, que en nuestro caso fueron árboles de decisión, usamos la biblioteca

Matplotib la cual nos sirve para graficar a partir de datos contenidos en listas o arrays una extensión de esta es NumPy que lo mencionamos antes, también usamos la librería Pandas ya que al nosotros usar dataset necesitábamos tener un mejor control de esta estructura de datos y esta librería es especialista en el manejo y análisis de estructura de datos, como ejemplo de las funcionalidades que nos ofrece son que puede definir nuevas estructuras de datos basadas en los arrays de la librería NumPy pero con nuevas funcionalidades, leer y escribir fácilmente ficheros en formato CSV, Excel y bases de datos SQL, acceder a los datos mediante índices o nombres para filas y columnas realizando todas estas operaciones y otras de manera muy eficiente, y como último recurso usamos Scikit Learn una librería para aprendizaje automático la cual suma más ya que esta posee herramientas eficientes usadas para el aprendizaje automático y modelado estadístico incluyendo clasificación, regresión, agrupación, y reducción de dimensionalidad.

Sobre nuestro dataset

Nuestra dataset proviene de [17] en la cual analizamos la potabilidad del agua, teniendo nueve variables de entrada y tres mil doscientos setenta y siete instancias, nuestras entradas para poder obtener un correcto análisis del agua son el ph que posee de 0 a 14, hardness que es la capacidad del agua para precipitar jabón (numérico), solids que nos indica los sólidos disueltos totales (numérico), chloramines que nos dice la cantidad de cloraminas que posee en ppn (numérico), sulfate que nos muestra la cantidad de sulfatos disueltos (numéricos), conductivity que nos dice la cantidad eléctrica del agua (numérico), organic carbon que nos muestra la cantidad de carbono orgánico en ppm (numérico), trihalomethanes que nos dice la cantidad de trihalometanos en ug (numérico), y por último a turbidity que nos dice la medida de la propiedad de emisión de luz de agua en NTU (numérico). La variable por predecir o variable de salida contempla 0 para indicar agua potable y 1 para agua no potable.

Variable	Descripción	Tipo	Min	Max
Ph	Ph del agua	Numérico	1.43	14.00
Hardness	Capacidad del agua para precipitar jabón	Numérico	73.49	306.63
Solids	Sólidos disueltos totales en ppm.	Numérico	320.94	56351.39
Chloramines	Cantidad de cloraminas en	Numérico	1.39	13.13

	ppm.			
Sulfate	Cantidad de sulfatos disueltos en mg/L	Numérico	182.39	481.03
Conductivity	Conductividad eléctrica del agua μS/cm	Numérico	201.62	753.34
Organic carbon	Cantidad de carbono orgánico en ppm	Numérico	2.20	27.01
Trihalomethanes	Cantidad de trihalometanos en μg/L	Numérico	14.34	124.00
Turbidity	Medida de la propiedad de emisión de luz del agua en NTU	Numérico	1.45	6.49
Potability	Indica si el agua es segura para el consumo humano	Numérico	0	1

Tabla N°1 Variables del dataset

4. Resultados y discusión

Los datos que sean utilizado para diseñar el modelo se extrajeron de un dataset que cuenta con 3276 muestras donde se encontró que varias muestras contaban con valores vacíos, las variables Ph, Sulfate y Trihalomethanes contaban con 491, 781 y 162 valores vacíos respectivamente. Para realizar data cleaning en el dataset se optó por la eliminación de datos debido a la cantidad de instancias que se iban a eliminar que alcanzó el 23% del total que no era una cantidad que afecta la consistencia del dataset, luego del proceso se lograron eliminar 1265 instancias vacías.

Se realizó un análisis de las instancias resueltas luego del proceso de data cleaning y se encontró un desbalanceamiento de los datos respecto a la variable a predecir (Potability), contando con 1200 instancias con valor 0 (Potable) y 811

Revista Innovación y Software
Vol. 3, No. 2, Mes Septiembre-Febrero, 2022
ISSN: 2708-0935
Pág. 121-131
https://revistas.ulasalle.edu.pe/innosoft

instancias con valor 1 (No potable). Para resolver el desbalance se aplicó la técnica de oversampling definiendo así 2400 instancias en total.

Se empleó un árbol de decisión para determinar si el agua es potable, los datos se dividieron en dos conjuntos, el de entrenamiento y el de pruebas contando con el 80% y 20% de los datos respectivamente. Luego del entrenamiento y las pruebas se realizó una evaluación del modelo mediante la matriz confusión contando con las métricas de precisión, exactitud, exhaustividad y puntuación F1.

El resultado de la métrica de precisión es de 0.77, refiriendo que de 100 muestras de agua, el modelo logra reconocer 77 muestras como agua potable de forma correcta mientras que el resto de 23 muestras las clasifica como agua potable de forma incorrecta. En la métrica de exactitud el modelo obtuvo 0.80, es decir, el modelo logra predecir el 80% de las veces, ejemplo, de 100 muestras de agua el modelo clasifica el 80% de forma correcta.

La métrica de exhaustividad del modelo fue evaluada con 0.85, mostrando que de 100 muestras de agua potable el modelo puede reconocer de manera correcta 85 muestras, mientras que el resto se clasifican como agua no potable. Respecto a la puntuación F1 el modelo recibió un puntaje de 0.81 indicando que se reconocen el 81% de los casos positivos donde el agua es potable. El árbol resultado del modelo se muestra en la Imagen N°1.

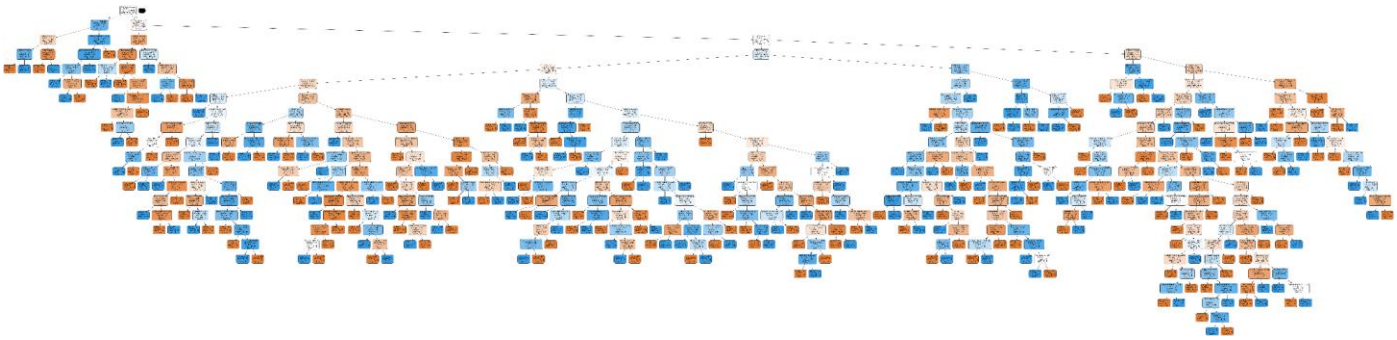

Imagen N°1 Árbol de Decisión

Conclusiones

Se logró definir un modelo capaz de predecir la potabilidad del agua, un modelo diferente al tratado en los trabajos relacionados utilizando una técnica de clasificación. El modelo fue basado en un árbol de decisión que trabajó con un dataset inicialmente con valores vacíos y desbalanceado, el modelo fue evaluado con las

métricas de precisión, exactitud, exhaustividad y puntuación F1 obteniendo un puntaje de 0.77, 0.80, 0.85 y 0.81 respectivamente.

El resultado obtenido no logra alcanzar el estándar de 0.80 en todas las métricas, por ello no se puede considerar un gran modelo para la predicción de la potabilidad del agua, sin embargo, se puede concluir que la técnica de árbol de decisión puede ser vista como un poderoso predictor que puede ayudar con la problemática de la potabilidad del agua.

Referencias

[1] C. C. Sánchez, "Enfermedades infecciosas relacionadas con el agua en el Perú," Revista peruana de medicina experimental y salud publica, *35*, 309-316.2018

[2] B. Serrano Pérez, R. Tendero Caballero, & M. D. Río Merino, "Parámetros indicadores del agua potable doméstica urbana, umbrales y consecuencias para la salud," 2018.

[3] F. L. Meoño, C. G. Taranco, & Y. M. Olivares, "Las aguas residuales y sus consecuencias en el Perú. Saber y hacer," 2(2), 8-25. 2015

[4] M. F. Amado Camargo, "Determinación bacteriológica de la calidad del agua de consumo humano, regadío y bebida de animales del Distrito de Majes, Provincia de Caylloma, Departamento de Arequipa, Abril-Mayo 2017", DSpace. Tesis. Arequipa, 2018. Disponible: http://repositorio.unsa.edu.pe/handle/UNSA/5890

[5] A. J. Espinosa Ramírez, "El agua, un reto para la salud pública: la calidad del agua y las oportunidades para la vigilancia en salud ambiental." UNAL. Tesis. Bogotá, 2018. Disponible: https://repositorio.unal.edu.co/handle/unal/63149

[6] C. Idrovo. "Optimización de la planta de tratamiento de Uchupucún," B.S. Tesis. Cuenca, 2010. Disponible: http://dspace.ucuenca.edu.ec/handle/123456789/2426

[7] I. D. López, A. Figueroa y J. C. Corrales, "Un mapeo sistemático sobre predicción de calidad del agua mediante técnicas de inteligencia computacional", Revista Ingenierías Universidad de Medellín, vol. 15, n.º 28, pp. 35–52, 2016. Accedido el 10 de agosto de 2022. Disponible: https://doi.org/10.22395/rium.v15n28a2

[8] L. Quiñones Huatangari, L. Ochoa Toledo, N. Kemper Valverde, O. Gamarra Torres, J. Bazán Correa y J. Delgado Soto, "Red neuronal artificial para estimar un índice de calidad de agua", Enfoque UTE, vol. 11, n.º 2, pp. 109–120, abril de 2020. Accedido el 11 de agosto de 2022. Disponible: https://doi.org/10.29019/enfoque.v11n2.633

[9] A. F. Siles, “Desarrollo de software y diseño de un sistema automatizado para monitoreo y predicción de eventos de contaminación en sistemas de distribución de agua, utilizando inteligencia artificial”. Repositorio DspaceDesarrollo de software y diseño de un sistema automatizado para monitoreo y predicción de eventos de contaminación, 1 octubre, 2019. Accedido el 16 de agosto de 2022. Disponible: http://literatura.ciidiroaxaca.ipn.mx:8080/xmlui/handle/LITER_CIIDIROAX/230

[10] A. C. Aguilar Aguilar y F. F. Obando - Díaz, "APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN DE CALIDAD DE AGUA POTABLE", Ingeniare, n.º 28, junio de 2020.Disponible: https://doi.org/10.18041/1909-2458/ingeniare.28.6215

[11] Iberdrola. "¿Qué es la Inteligencia Artificial? - Iberdrola". Iberdrola. https://www.iberdrola.com/innovacion/que-es-inteligencia-artificial.

[12] INSTITUTO DE INGENIERÍA DEL CONOCIMIENTO. "Machine Learning y Deep Learning - Expertos en IIC". 2022. Instituto de Ingeniería del Conocimiento. https://www.iic.uam.es/inteligencia-artificial/machine-learning-deep-learning/.

[13] K. Kelley. "What is Data Analysis? Types, Methods and Techniques 2022, Simplilearn". Simplilearn.com. https://www.simplilearn.com/data-analysis-methods-process-types-article#what_is_data_analysis.

[14] Lucid Software Inc.. "Qué es un diagrama de árbol de decisión". Lucidchart. https://www.lucidchart.com/pages/es/que-es-un-diagrama-de-arbol-de-decision.

[15] I. Moreno Hojas y StatPlans. "Construyendo árboles de decisión - StatDeveloper". StatDeveloper. https://www.statdeveloper.com/construyendo-arboles-de-decision/.

[16] F. Charte. "Cómo es el proceso de extraer conocimiento a partir de bases de datos - campusMVP.es". campusMVP.es. https://www.campusmvp.es/recursos/post/el-proceso-de-extraccion-de-conocimiento-a-partir-de-bases-de-datos.aspx.

[17] "Water Quality". Kaggle: Your Machine Learning and Data Science Community. 2021, [En línea] https://www.kaggle.com/datasets/adityakadiwal/water-potability.