Trabajo Fin de Grado
Ingeniería de las Tecnologías Industriales

Técnicas de procesamiento del lenguaje natural para analizar artículos sobre Machine Learning y la gestión de redes de agua

Autor: María Granados Santos

Tutor: Alicia Robles Velasco

Dpto. Organización Industrial y Gestión de Empresas II
Escuela Técnica Superior de Ingeniería

Sevilla, 2022

Trabajo Fin de Grado
Ingeniería de las Tecnologías Industriales

Técnicas de procesamiento del lenguaje natural para analizar artículos sobre Machine Learning y la gestión de redes de agua

Autor:

María Granados Santos

Tutor:

Alicia Robles Velasco

Colaborador docente invitado

Dpto. de Organización Industrial y Gestión de Empresas II

Escuela Técnica Superior de Ingeniería

Universidad de Sevilla

Sevilla, 2022

Trabajo Fin de Grado: Técnicas de procesamiento del lenguaje natural para analizar artículos sobre Machine Learning y la gestión de redes de agua

Autor: María Granados Santos
Tutor: Alicia Robles Velasco

El tribunal nombrado para juzgar el Proyecto arriba indicado, compuesto por los siguientes miembros:

Presidente:

Vocales:

Secretario:

Acuerdan otorgarle la calificación de:

Sevilla, 2022

El Secretario del Tribunal

A mi abuelo

Agradecimientos

A mis padres y a mi hermana, por apoyarme siempre y demostrarme que no existe la suerte, sino que todo con lo que se sueña se consigue mediante esfuerzo y trabajo.

A mis amigos por la ayuda que siempre me han prestado y todas las sonrisas que me han arrancado. En especial a Ale y Nacho por apoyarme en todo momento y sacar lo mejor de mí.

Y a Alicia, por ir más allá de lo académico, y ser guía en esta vocación que compartimos.

Resumen

Las redes de distribución de agua representan uno de los mayores avances de la humanidad; por su impacto directo en la salud de la población, en las capacidades de producción de bienes y servicios y por el bienestar que conllevan. No obstante, actualmente, y pese a las tecnologías disponibles, este tipo de redes tienen en España pérdidas superiores a un 15%, con el impacto económico y medioambiental que esto supone.

Dada la importancia de estas pérdidas, se ha planteado un análisis de las investigaciones disponibles sobre las redes de abastecimiento y saneamiento entre la producción científica mundial, con el objetivo de entender sus principales condicionantes. Como la producción científica sobre las redes de agua es enormemente extensa, se ha decidido preparar una herramienta basada en técnicas de procesamiento del lenguaje natural que permita facilitar el estudio de los artículos científicos disponibles.

En este proyecto, se han diseñado y desarrollado algoritmos que permiten, de forma automatizada, valorar la importancia que los artículos dan a ciertos conceptos, como los factores empleados en cada trabajo (material, presión, diámetro, etcétera) y el uso que hacen los artículos de distintas técnicas de Machine Learning. Adicionalmente, se ha preparado un sistema de evaluación de los resultados obtenidos, de cara a facilitar las conclusiones de dicho análisis. Este sistema permite estudiar la correlación entre los diferentes conceptos definidos e identificar cuáles son los artículos que más utilizan dichos conceptos (o lista de los mismos).

Abstract

Water distribution networks are considered one of the greatest advances of humankind, due to its direct impact on the health of the population, on the production capacities of goods and services and for the well-being that they entail. However, presently, and despite the available technologies, this type of network in Spain has losses of over 15%, with the economic and environmental impact that this entails.

Due to the importance of these losses, an analysis of the available research on supply and sanitation networks among world scientific production has been proposed, with the aim of understanding its main conditioning factors. As the number of scientific papers on water networks is enormous, it has been decided to design a tool based on natural language processing techniques that will facilitate the study of available scientific articles.

In this project, algorithms have been designed and developed, so they allow, in an automated way, to assess the importance that each article assigns to certain terms, such as the characteristics of the elements considered in each research (material, pressure, diameter, etc.) and the use those articles have done of different Machine Learning techniques. Additionally, in order to facilitate the conclusions of the analysis, an evaluation system has been prepared for the achieved results. This system makes it possible to study the correlation between different defined terms and to identify which articles use more often those terms (or a list of them).

Índice

Índice de ilustraciones

Índice de tablas

Índice de gráficos

1 INTRODUCCIÓN

La comunidad científica internacional suele utilizar, como mecanismo para documentar, divulgar y dar a conocer sus trabajos de investigación, artículos en medios científicos acreditados. En estos medios, los investigadores o grupos de investigadores difunden sus investigaciones, finalizadas o en curso, y actualizan los avances que se producen en las mismas [1].

Para ilustrar la importancia que tiene la publicación científica en el mundo académico, sirva a título de ejemplo que el más conocido ranking de universidades internacionales, el Academic Ranking of World Universities (ARWU), conocido habitualmente como el ranking de Shanghái, incluye como factor evaluable de cada universidad la producción científica medida en los siguientes términos:

- Artículos publicados en prestigiosas revistas como Nature y Science.
- El número de artículos indexados en Science Citation Index Expanded (SCIE) y en Social Science Citation Index (SSCI)
- El número de investigadores más citados seleccionados por Thomson Scientific

En total, la producción científica medida de esta forma tiene un peso del 60% en la posición de una universidad en el ranking [2].

Para tener una idea del tamaño de la producción científica a nivel mundial, se recoge seguidamente la evolución del número de artículos en Scopus en el periodo 2015-2019. Como se puede ver en la *Ilustración 1-1. Gráfico de la evolución del indicador Output: número de artículos publicados (2015-2019)* [3], existe una diferencia tremenda entre Europa Occidental y el resto. Pero lo que es más significativo es que el volumen total de producción científica crece de forma sostenida.

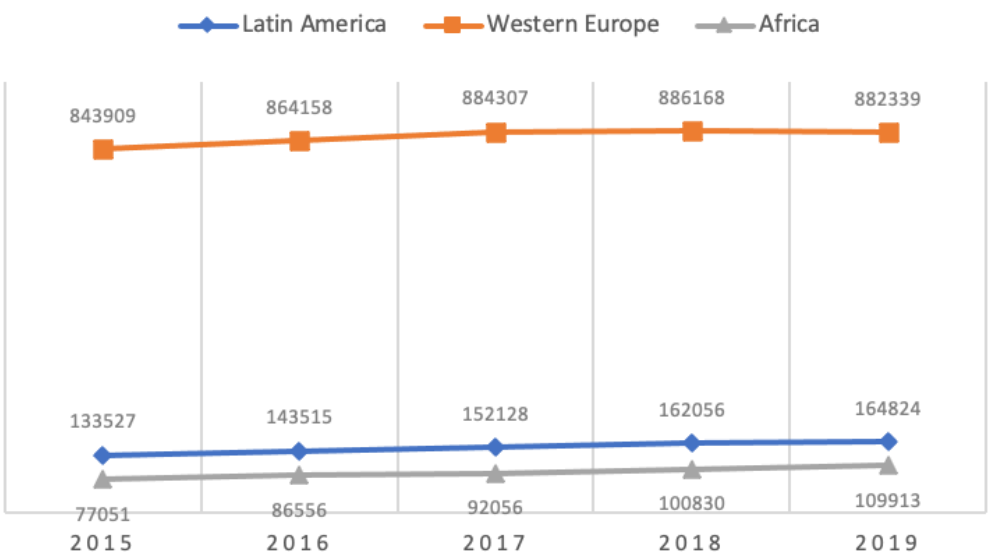

Ilustración 1-1. Gráfico de la evolución del indicador Output: número de artículos publicados (2015-2019) [3]

Seguidamente, y para dar una idea de la evolución se presenta en *Ilustración 1-2. Gráfico de la evolución de la relación output/impacto por países con mayor producción científica (Scopus 2007-2019)* [4] cómo evoluciona el volumen de producción científica de cada país (eje X) y su impacto en términos de citas (eje Y).

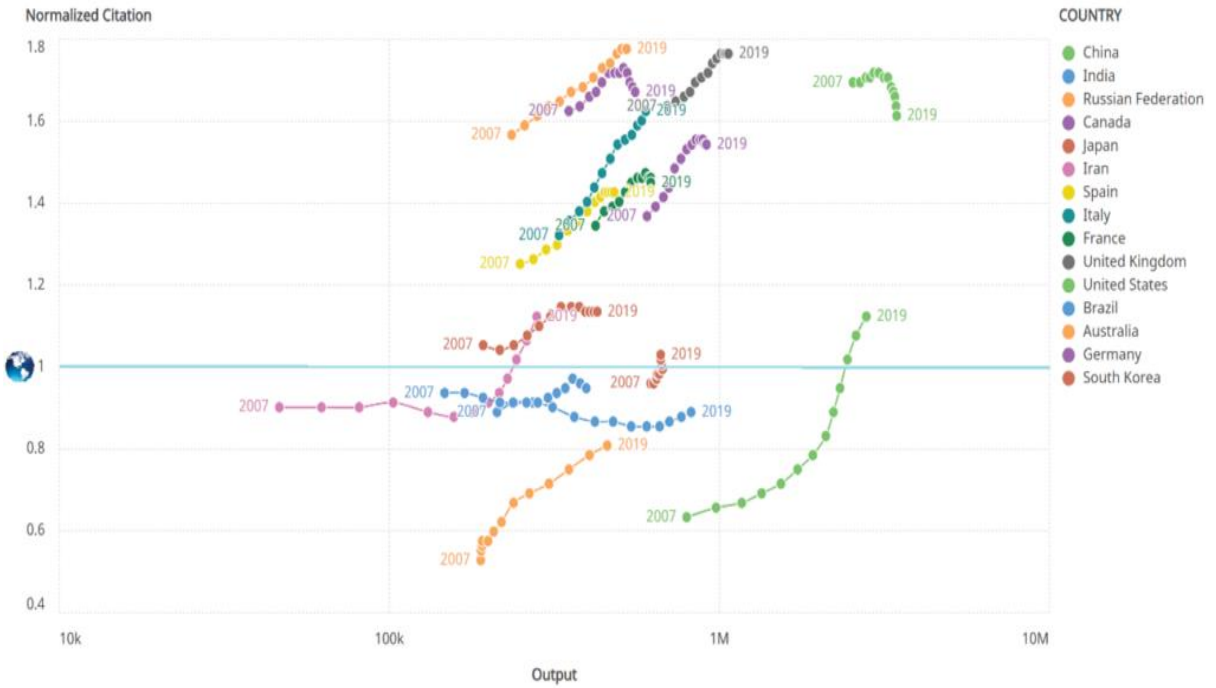

Ilustración 1-2. Gráfico de la evolución de la relación output/impacto por países con mayor producción científica (Scopus 2007-2019) [4]

Como se aprecia en dicha ilustración los mayores productores de ciencia (en términos de publicaciones) siempre estuvieron por encima de la media mundial y además casi todos han mejorado en su posición relativa.

España, representada por el color amarillo en el gráfico, ha aumentado considerablemente su producción científica en el periodo de estudio. Por otro lado, se observa que la producción científica de China ha disminuido de forma significativa desde 2010, aunque sigue siendo de las más voluminosas en términos absolutos.

La información descrita anteriormente se ha incluido a efectos de ilustrar la importancia de los artículos en revistas como divulgadores y medidores de la capacidad de producción científica de un país, institución o grupo de investigación.

Por otra parte, la distribución de agua es un aspecto clave para garantizar la calidad de vida de los ciudadanos, la producción de bienes y servicios (agrícolas, ganaderos, industriales, etcétera) y la calidad del medio ambiente.

Por lo que respecta a la eficiencia de estas redes, hay que decir que en España las pérdidas de las redes de distribución de suministro a nivel nacional suponen un 15.4% del total de agua suministrada en 2018, lo equivalente a 653 hm^3. Esto equivale a 737.890.000 €, y a esta cifra habría que sumar las pérdidas correspondientes al saneamiento.

Se trata por tanto de un tema relevante, en términos económicos y medioambientales. Como consecuencia de ellos, existe una numerosa producción científica sobre los problemas en las redes de distribución de agua, sus causas, sus posibles soluciones, y la cuantificación de estas. En particular, muchos de estos análisis se han enfocado en los últimos años a la aplicación de técnicas de aprendizaje supervisado a los modelos que estudian las pérdidas, roturas e ineficiencias en las redes de la gestión del agua.

El caso de estudio de este trabajo es la producción científica sobre Machine Learning en el sector del agua.

Tradicionalmente, la búsqueda de conceptos en artículos científicos se venía haciendo mediante el acceso a bases de datos previamente indexadas mediante palabras claves o términos concretos. El objetivo de este trabajo es aplicar técnicas de procesamiento del lenguaje natural a las búsquedas en artículos científicos, que, aunque en este caso se han circunscrito a las redes de distribución de agua, supone una metodología que podría aplicarse a cualquier tipo de búsqueda.

Así, por ejemplo, las técnicas utilizadas permiten valorar la exactitud de las palabras claves de los artículos frente al texto completo, la frecuencia de estas palabras claves en el texto completo, las posibles correlaciones entre términos, o las búsquedas simultáneas por varios conceptos.

La idea es poder hacer búsquedas más naturales, con una simplicidad que se asemeje al lenguaje natural, abandonando la rigidez inherente a tener que buscar por términos exactos.

Pero, además, se busca establecer una forma de presentar los resultados que permita al investigador orientar rápidamente dónde debe centrar los esfuerzos de estudio dentro de la producción científica disponible, sea porque ya está previamente analizado o por tratarse de un campo no explorado aún.

1.1 Estructura del trabajo

Este trabajo se ha dividido en cinco capítulos principalmente. A lo largo del segundo capítulo se estudia la producción científica del aprendizaje automático en la distribución de agua. Posteriormente se define el procesamiento del lenguaje natural y la implementación de una de sus aplicaciones para el desarrollo del programa. En el cuarto capítulo se determina el caso de estudio, incluyendo la búsqueda de artículos y palabras.

Finalmente, se ha incluido el análisis de resultados, es decir, se han representado, categorizado y ordenado los resultados más relevantes que se pueden obtener con el programa desarrollado. Además, se han realizado conclusiones sobre estos resultados y se ha sugerido futuras líneas de investigación basadas en este trabajo.

2 LA PRODUCCIÓN CIENTÍFICA SOBRE MACHINE LEARNING EN EL SECTOR DEL AGUA

En este capítulo se desarrollará la relación entre la gestión del agua y el *Machine Learning*, en español Aprendizaje Automático. Para ello, se ha dividido la sección en tres partes principales: la distribución de agua, *Machine Learning* y la relación entre ellos.

La primera subsección incluye la introducción a la distribución de agua, la definición de los elementos necesarios para la correcta gestión del agua y los fallos que se pueden producir en las redes de distribución, con los factores que los ocasionan.

Posteriormente se estudia el contexto de este término y la clasificación del aprendizaje. También se describe el procedimiento para la implantación del aprendizaje automático supervisado, resaltando las etapas más importantes: el procesamiento y análisis de los datos, la elección del modelo y el análisis de resultados.

Para finalizar, la tercera subsección justifica el vínculo entre los modelos de *Machine Learning* y la distribución de agua a través de pérdidas económicas y medioambientales.

2.1 La distribución de agua

La Asamblea General de Naciones Unidas aprobó el 28 de julio de 2010 el reconocimiento del derecho esencial al agua potable y el saneamiento de esta misma [5]. Para cumplir con este derecho es necesario garantizar el ciclo integral del agua (*Ilustración 2-1. Ciclo integral del agua*); recorrido del agua desde la captación en la naturaleza hasta su disposición en todos los hogares. La primera etapa de este proceso es la captación del agua requerida del ecosistema, como embalses, aguas superficiales, etcétera. Posteriormente se realiza el tratamiento del agua, para poder consumirla de forma adecuada, y el almacenamiento en depósitos urbanos. La cuarta etapa es la distribución de agua potable que se divide en dos escalas de redes, la red en alta y la red en baja. La diferencia entre estas redes es que la primera se encarga de garantizar el agua en la entrada de las poblaciones y la segunda de distribuirla hasta cada domicilio.

Por otra parte, es necesario recoger el agua desechada de los domicilios para poder transportarla hasta las plantas de depuración. A esta etapa se le denomina saneamiento.

Para finalizar el ciclo, se realiza la depuración en las Estaciones Depuradoras de Aguas Residuales, que consiste en el tratamiento necesario para retornar el agua a la naturaleza con el menor porcentaje de contaminantes [6].

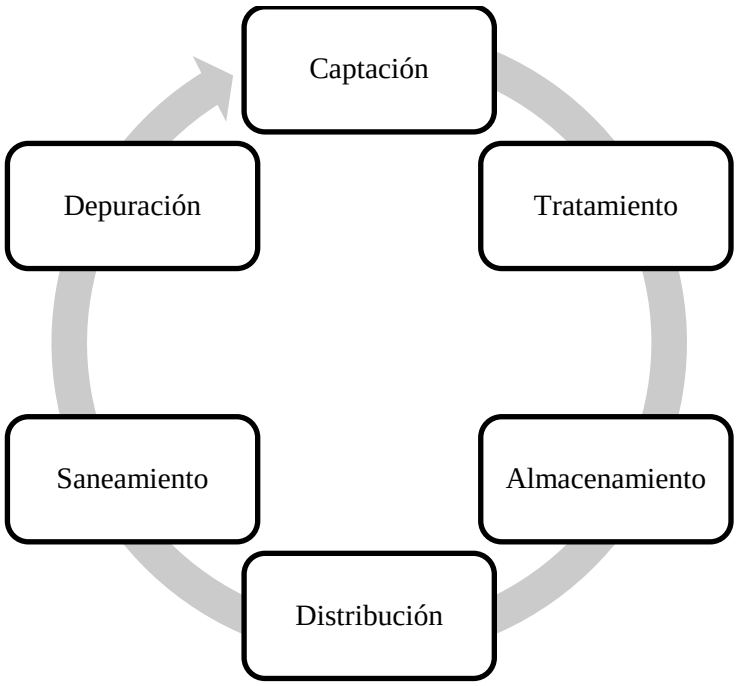

Ilustración 2-1. Ciclo integral del agua

La red de abastecimiento se compone de la captación, el tratamiento, el almacenamiento y la distribución. Las de saneamiento por su parte incluyen tanto el saneamiento en sí, como la depuración.

Las redes de distribución de agua también se dividen según su función en los siguientes tipos de redes:

- La red de transporte se encarga de trasladar el agua desde las plantas de tratamientos, los depósitos de regulación o las estaciones de bombeo hasta la denominada red arterial. Está prohibido que el usuario esté en contacto con el agua transportada por esta red.
- La red secundaria transporta el agua desde la red arterial hasta las acometidas para los suministros (unión de la red secundaria con la instalación que se pretende abastecer), bocas de riego y tomas contra incendios.
- La red de abastecimiento conecta los diferentes sectores de las zonas de abastecimiento. Desde esta red, no se pueden realizar acometidas.

2.1.1 Elementos de la red de la distribución de agua

La infraestructura de la red de la distribución de agua está compuesta por numerosos elementos, incluyendo las tuberías. Este proyecto se enfoca específicamente en el estudio de las tuberías, pero para comprender los fallos de éstas, es necesario definir los otros elementos de la red.

Se definen los diferentes elementos de la red según la Norma UNE-EN 805 2000 [7]:

- Tubos: conductos de secciones circulares, generalmente de eje recto. Para el abastecimiento del agua, se suelen emplear tubos rígidos, cuya capacidad de carga está limitada por la rotura, sin deformación significativa de la sección. No obstante, en algunas ocasiones se recurren a los tubos donde la deformación es la limitación de la capacidad de cargas, denominados tubos flexibles.
- Tuberías: conjuntos de tubos aislados, correctamente unidos, cuyas fusiones preservan la calidad esencial del agua, evitando su pérdida y contaminación. Para prevenir el deterioro mecánico, el ataque químico y la corrosión en el interior y el exterior de las tuberías, se suelen emplear revestimientos interiores y exteriores, así como otros elementos de protección.
- Juntas o uniones: elementos que permiten la conexión de dos componentes. La unión flexible permite una desviación angular significativa, durante la instalación y en momentos posteriores; también es posible un ligero desplazamiento diferencial entre ejes. Si esa desviación se produce solo en el periodo de instalación, se estará trabajando con uniones ajustables. La junta rígida no tolera la desviación angular significativa en ningún momento.
- Válvulas: componente que permite iniciar, cortar o regular el caudal y la presión mediante piezas móviles. Existen diferentes tipos de válvulas según su función, como las válvulas de aislamiento, de regulación y antirretorno.
- Depósitos de regulación: estructura encargada de almacenar agua. La relación directa entre los depósitos y los consumidores, hacen de obligado cumplimiento la garantía del agua.
- Estaciones de bombeo: conjunto de bombas instaladas para asegurar los caudales y presiones necesarias para distribuir el agua. El bombeo principal suele emplearse para transportar el caudal a los depósitos de las redes, y se encuentran en la toma de agua o en las salidas de las plantas de tratamiento de agua, en caso de ser necesario el saneamiento de esta misma. Para garantizar el caudal a las diferentes redes de abastecimiento se utilizan bombas intermediarias. Por último, se ha de mencionar las bombas de reimpulsión, instaladas en línea con las tuberías y no antes de un depósito.

2.1.2 Fallos en las redes de aguas

Los deterioros de tuberías se pueden clasificar, según el profesor Ignacio Javier Acosta, en dos categorías principalmente [8].

La primera se denomina fallos formales, y se producen debido a las acciones físicas en los tramos de tuberías. El efecto de esta rotura es muy poco notorio en comparación con las demás, pero son las más graves dado que se producen inmediatamente, sin posibilidad de reaccionar.

El segundo grupo de la clasificación se denomina fallos sustanciales, producidos por acciones químicas de sustancias que se encuentran en las tuberías. Las más frecuentes son la corrosión y la agregación o depósito interno de materias extrañas.

Seguidamente se detallarán los diferentes fallos que se pueden producir tanto en las redes de abastecimiento como en las redes de saneamiento. Para una mejor comprensión, se ha *i*ncluido la *Ilustración 2-2. Esquema sobre las patologías de fallo de tuberías en redes de abastecimiento* e *Ilustración 2-3. Esquema sobre las patologías de fallo de tuberías en redes de saneamiento.*

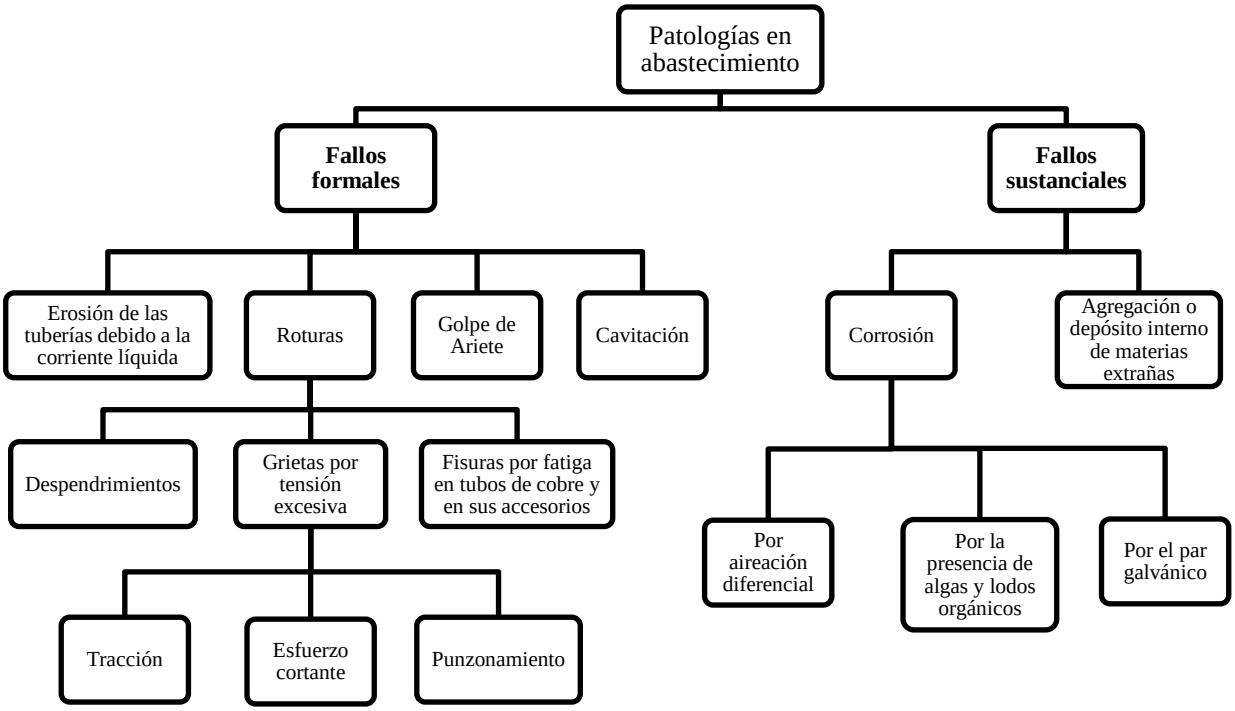

Ilustración 2-2. Esquema sobre las patologías de fallo de tuberías en redes de abastecimiento

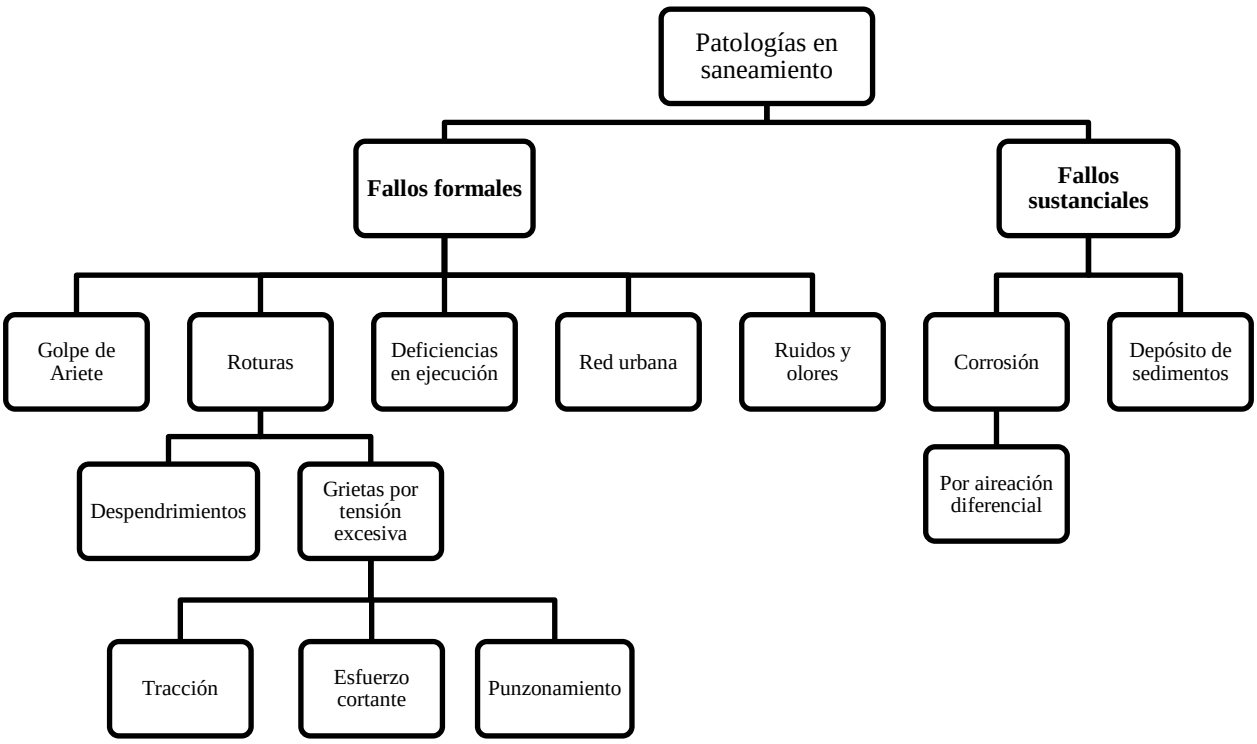

Ilustración 2-3. Esquema sobre las patologías de fallo de tuberías en redes de saneamiento

En la instalación de tuberías de redes de abastecimiento se encuentra la distinción entre deterioros debido a la erosión de las tuberías a causa de la corriente líquida, la rotura, el golpe de ariete y la cavitación.

Entre los fallos formales se encuentra la erosión de la tubería causada por la corriente líquida. Consiste en el arrastre de material debido a la alta velocidad y a las turbulencias del líquido en el interior de las tuberías. Los efectos se producen en los tramos posteriores a una curva y en los estrechamientos o ensanchamientos de los tubos, provocando la posibilidad de rotura. En estos tramos, se producen altas velocidades y turbulencias en el agua que circula por el interior de la tubería. Debido a ello, la capa protectora formada en el área afectada es arrastrada progresivamente, exponiendo el material primitivo a una nueva oxidación, provocando la estrechez de la pared de la tubería que conlleva a la erosión mecánica de la superficie de esta misma.

Las grietas por tensión excesiva y las fisuras por fatiga en tubos de cobre y en sus accesorios son los causales de las roturas. Los fallos de grietas por tensión pueden estar provocadas por tracción, esfuerzo cortante o punzonamiento. En cambio, las fisuras por fatiga en tubos de cobre se producen a causa del constante movimiento de la tubería, formando grietas por fatiga del material. La oxidación o el par galvánico con el metal de la tubería provoca corrosión de los anclajes, originando fallos en la sujeción de las tuberías exteriores. Este fenómeno, denominado desprendimiento, es una de las causas de rotura de tuberías.

El pulso de Zhukowsi o el golpe de ariete se origina por cualquier modificación de velocidad de circulación del agua que atraviesa por la tubería, llegando a producir sobrepresión en el interior de dicho elemento.

Otra de los fallos formales es la cavitación, también denominada aspiración en vacío. Cuando el agua pasa a

gran velocidad por una arista afilada, puede darse el caso de que las moléculas pasen a estado de vapor una vez alcanzada la presión de vapor, formando cavidades.

La rotura, el golpe de ariete, la deficiencia en ejecución, la red urbana y los ruidos y los olores son los fallos más comunes de las tuberías de la red de saneamiento del agua.

Los desprendimientos causados en conductores exteriores, como los bajantes y canalones, se deben al fallo de sujeciones, ya sea la falta de anclaje, la corrosión o la aparición de par galvánico entre el hierro y el zinc de la tubería. Las grietas por tensión excesiva están provocadas por tracción o esfuerzo cortante, ambas justificadas por la sujeción excesivamente rígida, y por punzonamiento, causante por la falta de protección de la tubería instalada bajo pavimento. Todo lo recogido anteriormente puede causar la rotura de la infraestructura de saneamiento.

El golpe de ariete es más impactante en las redes de saneamiento que en las de abastecimiento, debido a la inestabilidad de la velocidad del fluido, que genera impactos en el momento de descarga.

Las acumulaciones de sedimentos en las arquetas también provocan fallos puesto que congestionan la red. La diferencia positiva entre la cota de servicio de la red urbana y la del desagüe, hacen que esta última se llene cuando entra en carga la red urbana, levantando la tapa de registro de la arqueta (tapa que se utiliza para proteger la arqueta).

En lo que respecta a los fallos sustanciales, la corrosión puede producirse debido a la inmersión, la aireación diferencial, al par galvánico y a la presencia de algas y lodos orgánicos. La primera causa mencionada es la más habitual, y es provocada por el continuo contacto entre la tubería y el agua potable, que, dado su elevado contenido de oxígeno, facilita la corrosión al disolver los posibles depósitos de carbonato de calcio (calcáreos) que se aprecian en las paredes internas de las tuberías.

En numerosas ocasiones se pueden encontrar defectos de montajes que permiten un aumento de acumulación de oxígeno, provocando la corrosión por aireación diferencial. La lesión por par galvánico consiste en la corrosión debido al contacto entre diferentes metales en presencia de humedad, que provoca una electrólisis, afectando al estado de las tuberías. La disposición de algas y lodos orgánicos en el agua también provocan la corrosión del metal.

La lesión de depósito interno de materias extrañas se define como la variación de la característica del agua debido a la presencia de otros materiales en el interior de las tuberías, por ejemplo, la modificación del color del agua.

Los fallos sustanciales son más habituales en la red de abastecimiento que en la red de saneamiento dado que la infraestructura de esta última se realiza en policloruro de vinilo o polietileno, elementos que evitan la corrosión. Los elementos como los sumideros o cazoletas que están hechos de acero o fundición de este mismo, sí que podrían llegar a oxidarse a lo largo del tiempo. La otra lesión sustancial más común de la infraestructura de saneamiento son los depósitos de sedimentos, provocados por las arquetas o tuberías cuya pendiente no es la suficiente o la unión entre los diferentes elementos no es continua o tiene algún defecto, que

conlleva a la retención de sedimentos de materias sólidas.

2.1.2.1 Factores de fallos en tuberías

Gracias a la tecnología actual y al aumento de las investigaciones sobre la distribución de agua, se ha podido determinar los factores que provocan el deterioro de las tuberías y así poder desarrollar modelos estadísticos para prevenir la rotura de las tuberías. Estos factores están clasificados en tres grupos: físicos, operacionales y ambientales [9].

Factores físicos

Estos factores son los producidos por las características intrínsecas de las tuberías. Existen numerosos factores físicos, por lo que es imposible nombrar todos. A continuación, se detallan los más importantes.

Material de la tubería

La fundición de grafito esferoidal (hierro dúctil), el cemento aluminoso, el cloruro de polivinilo y el polietileno son los materiales más utilizados para la fabricación de tuberías. La elección de un material u otro depende de las consideraciones técnicas (disponibilidad del material, coste, experiencia, etcétera), consideraciones externas del terreno (climatología, corrosividad del suelo, etcétera) y las consideraciones del agua que se transporta como la velocidad, la temperatura y la presión entre otras.

El hierro dúctil (también denominado fundición de grafito esferoidal o fundición nodular) se introdujo como sustituto de la fundición gris al ser más resistente a los impactos y a la oxidación y poseer mayor capacidad para soportar altas presiones y amortiguar los golpes de ariete. Las tuberías de fundición de grafito esferoidal que atraviesan terrenos con diferentes condiciones naturales provocan par galvánico y electrólisis que conllevan al deterioro en la red de agua. En las tuberías con estas características, se incluyen revestimientos tanto interiores como exteriores para protegerlas y así poder evitar los fallos. Otra desventaja de este material, en comparación a los metales, es el elevado coste para fabricar tuberías cuyo diámetro deba ser inferior a 140 mm [10].

Entre 1906 y 1913 se empezó a utilizar el cemento aluminoso para la fabricación de tuberías en Italia, donde surgió la idea. El auge del uso de este material para la distribución de agua surge entre 1940 y 1960. Las ventajas del cemento aluminoso eran el bajo coste tanto del material como de la operación y la resistencia a la corrosión. La rigidez, la poca flexibilidad y la baja resistencia al movimiento del suelo eran los principales inconvenientes de este material. En la década de 1970, gran parte de la población estadounidense sufrió problemas de salud, acuñados al agua distribuida por las tuberías de cemento aluminoso; por lo que se dejó de utilizar este material. No obstante, cabe destacar que actualmente entre un 16 y 18% de las tuberías de distribución de agua en Estados Unidos y Canadá son de este cemento [11].

Debido a la detención del uso del cemento aluminoso, en 1970 se empezó a utilizar el cloruro de polivinilo, con el inconveniente de que las tuberías resultaron ser de baja calidad y tenacidad debido al proceso de fabricación del material. Al final de esta década, aumentaron el uso del material para las tuberías puesto que mejoraron el proceso de fabricación del cloruro de polivinilo. Las ventajas que proporciona este material son el bajo coste de fabricación, la resistencia a la corrosión y la facilidad de ensamblaje. El polietileno es muy similar al cloruro de polivinilo, aunque destaca por poseer una elevada resistencia a la presión y una alta duración. Es un material que se empezó a utilizar en 1980 y que actualmente es el más empleado para las redes de distribución de agua [9]. Estos dos plásticos no se pueden emplear para transportar agua cuya temperatura sea superior a 80 ºC, puesto que se deforman las tuberías [12].

Revestimientos

Como se ha mencionado con anterioridad, los revestimientos se utilizan para proteger las tuberías y están relacionados con el material con el que se fabrican las tuberías. Debido a ello, los revestimientos pueden modificar el rendimiento de las tuberías, pues es muy probable que una mala elección del material de protección conlleve a la corrosión de la infraestructura, pudiendo así provocar el fallo de esta misma.

Edad

En los periodos inmediatos a la instalación de tuberías se producen numerosos fallos. Conforme el tiempo avanza, la probabilidad del deterioro disminuye hasta que la edad de la tubería es considerada notable, en ese momento vuelven a producirse fallos en las tuberías. Por todo esto, el factor de la edad debe incluir tanto la vida de la tubería como el año de instalación [9] .

Diámetro

La elección del material de una tubería depende de su diámetro, puesto que, como se ha mencionado anteriormente, el hierro dúctil no se utiliza para tuberías con un radio inferior a una dimensión específica. La delimitación del uso del material debido al diámetro no es exclusiva para el hierro dúctil, si no que ocurre lo mismo para los demás, por lo que el diámetro se podría no tener en cuenta como un factor independiente del material de la tubería.

Si en una tubería aumenta la velocidad del líquido, se produce una reducción de la presión del agua, ya que se necesita menor fuerza para impulsar a las partículas. La diferencia de velocidades y presiones pueden producir fallos en las tuberías como el golpe de ariete y la cavitación. El diámetro es inversamente proporcional a la velocidad, por lo que es un factor fundamental para la valoración del deterioro de las tuberías.

Otros

Los defectos de fabricación, la longitud, la manipulación y el almacenamiento de las tuberías son otros factores causantes de fallos de la red. Por ejemplo, a la hora de desplazar las tuberías, existe una posibilidad de realizarla de forma incorrecta, llevando a cabo la producción de abolladuras o grietas.

Factores operacionales

Es necesario determinar para qué se utilizan las tuberías y cómo se gestionan, puesto que en estos pasos son donde se encuentran los factores operacionales. La recogida de datos de estos factores suele ser muy inferior a la de los factores ambientales y físicos, pero aun así se ha demostrado que tienen una gran importancia en el deterioro de las tuberías.

Presión interna de las tuberías

La diferencia de presiones en el interior de una tubería puede provocar fallo por fatiga (debido a la tensión de carga cíclica), acelerando la propagación de grietas. Así mismo, la sobrepresión en tuberías también provoca deterioros, como efecto del golpe de ariete [9].

Velocidad del agua

El uso de bombas y válvulas, para el correcto funcionamiento del transporte de agua, pueden llegar a ser los causantes de las diferencias de velocidades en las tuberías, llegando a producir el golpe de ariete.

Fallos anteriores

Según los datos históricos, el 22% de los fallos ocurren en el mismo metro donde anteriormente se produjeron deterioros, por lo que se debe considerar los deterioros anteriores como un factor operacional [9].

Factores ambientales

Por último, se ha de mencionar los factores ambientales, ya que, dependiendo del lugar de la instalación de las tuberías, las condiciones climatológicas pueden afectar al deterioro de estas. A continuación, se demuestra la relación entre las temperaturas y el movimiento del suelo con los fallos de tuberías, aunque existen otros factores como el tráfico y la densidad de la población por área que también puedan tener impacto.

Bajas temperaturas

Se ha demostrado que se producen mayores fallos en las tuberías en los meses de invierno. Cuando la

temperatura es inferior a 0ºC, el agua que se transporta se solidifica, incrementando su volumen. Este incremento puede provocar un aumento de presión, que puede generar fallos en las tuberías.

Altas temperaturas y movilidad del suelo

Los fallos de tuberías en los meses de verano se producen debido a los esfuerzos de flexión en las tuberías, causado por las diferencias de temperatura entre el suelo y el agua transportada.

Las propiedades del suelo, como la textura, la estructura y la porosidad, están relacionadas con la humedad que se produce durante el transporte y el almacenamiento del agua. El movimiento del suelo, producido debido a la humedad, conlleva a la degradación química de las infraestructuras de gestión del agua [9].

2.2 Machine Learning

El *Machine Learning* (ML), en español aprendizaje automático es considerado una rama de la Inteligencia Artificial (IA), por lo que, para poder profundizar en este término, es importante remontarse a los inicios de la inteligencia artificial.

En 1918, la empresa alemana Scherbius & Ritter patentó Enigma, una máquina de rotores que se utilizaba para cifrar y descifrar mensajes. Fue utilizado con fines militares durante la Segunda Guerra Mundial, y la oposición hizo uso de Bombe, una máquina desarrollada por el británico Alan Turing, para descifrar las transmisiones alemanas [13].

Después de la Segunda Guerra Mundial, en 1950, Turing se cuestionó si las máquinas podían pensar. Fue el primer científico en cuestionar esto, aunque no acuñó el término inteligencia artificial que se conoce en la actualidad.

Seis años más tarde, en una conferencia en Dartmouth, John McCarthy definió la inteligencia artificial como "la ciencia e ingenio de hacer máquinas inteligentes, especialmente programas de computación inteligentes" [14].

En 1959, se empleó el término *machine learning* como una disciplina de la inteligencia artificial que dota a los ordenadores de algoritmos generales para que sean capaces de identificar patrones y elaborar predicciones sin necesidad de programarlos [15].

Es muy común confundir *machine learning* con *Big Data*, rama que estudia los sistemas encargados del procesamiento y análisis de datos. Se diferencian en que el *big data* solo facilita la consulta de la información y en cambio el *machine learning* busca patrones que resuman la conexión de los datos [16].

El aprendizaje automático se clasifica según el conjunto de datos obtenidos para analizar el problema. Los datos etiquetados son aquellos en los que se dispone y distinguen la variable de salida. El aprendizaje

automático supervisado se aplica si el conjunto de datos está etiquetado. En caso contrario se utilizarán métodos de aprendizaje automático no supervisado, encargados de encontrar una clasificación coherente para esos datos [17]. El aprendizaje semi-supervisado engloba tanto los datos etiquetados como los no etiquetados. Por último, se ha de mencionar el aprendizaje por refuerzo, utiliza datos etiquetados, pero no consigue una solución correcta al instante, sino que con el paso del tiempo es cuando se determina si la respuesta es acertada o errónea [18].

Para la implementación del aprendizaje supervisado es necesario realizar el procedimiento iterativo basado en 4 etapas:

1. Definir el objetivo: es la primera etapa que se debe realizar y consiste en determinar los objetivos que se quieren obtener. Para ello, es necesario responder a las preguntas referentes a qué y cómo se quiere hacer y si son suficientes los datos que se tienen para resolver el problema
2. Procesamiento y análisis de exploración de datos: consiste en clasificar y preparar los datos que se tienen para que en la siguiente etapa se pueda elegir el modelo más adecuado.
3. Construcción del modelo: se determina el algoritmo que se desea implementar según los datos que se tienen.
4. Evaluación del modelo y análisis de errores y correcciones: finalmente se evalúa las soluciones del modelo y se realizan procedimientos para ver la eficiencia del algoritmo

2.2.1 Procesamiento y análisis de exploración de datos

El análisis exploratorio de datos es una herramienta que se utiliza para manipular los datos originales de cara a obtener las respuestas que se requieren. Para ello, se analiza e investiga los conjuntos de datos y se resumen sus principales características. Normalmente, en esta etapa se recurre a métodos de visualización de datos para así obtener un mejor rendimiento [19].

Esta etapa está formada a su vez por distintos procesos:

- Evaluación de los datos
- Limpieza de datos
- Tratamiento de datos
- Reducción de datos

2.2.1.1 Evaluación de los datos

Es muy importante determinar qué datos tenemos, pues en función de su clasificación se utilizarán unos modelos u otros. Además, es necesario prestar atención a los datos que no presentan toda la información

requerida o a valores atípicos del conjunto de datos recopilados.

Missing value

Missing value, en español datos incompletos, son datos que no están registrados para ciertas variables. Los datos incompletos son problemáticos ya que, si el conjunto de datos analizados es pequeño, puede sesgar los resultados.

Los datos incompletos suelen ser comunes en los problemas de investigación de ciencias sociales, puesto que lo normal es recopilar información a través de encuestas. Este procedimiento consta en realizar preguntas a los diferentes individuos y dependiendo de su respuesta, realizar otras preguntas. Debido a ello, este método recoge respuestas incompletas, puesto que las preguntas varían en función de las respuestas anteriores.

Algunas de las causas de la pérdida de datos son: la entrada incompleta de datos, el mal funcionamiento del equipo y la pérdida de archivos.

Existen numerosas soluciones para solventar este problema [20] :

- Eliminación: si el conjunto de datos que se va a analizar es muy grande, y la aportación de los datos incompletos es muy pequeña, se podrán eliminar estos valores. Esta alternativa no es la recomendable, puesto que a veces se desecha información valiosa.
- Imputación: se utiliza para reemplazar el valor faltante por un valor con sentido. Esa solución puede realizarse a través de la media o de la mediana de la variable. También es posible utilizar modelos de *machine learning* para predecir ese valor.
- Aceptación: utilizar modelos de *machine learning* que puedan manejar datos incompletos.

Outliers

Los valores atípicos, en inglés *outliers* o *atypical*, son valores extremos que difieren de la mayoría de los puntos de valores del conjunto de datos. Al igual que los datos incompletos, pueden provocar una modificación de los resultados a obtener, por lo que es importante resolver este problema. Para ello, se recurre a la aceptación de datos o a la eliminación, siendo esta alternativa poco útil puesto que en ocasiones no se puede diferenciar qué puntos son extremos y cuales están relacionados.[21].

2.2.1.2 Limpieza de datos

Este paso consiste en eliminar datos que no aporten información relevante al modelo. Por ejemplo, si se está analizando archivos de texto, las referencias de un artículo no suelen aportar información. En este proceso, también se realizan los procedimientos para resolver los problemas de datos incompletos o datos atípicos [22].

La eliminación de valores puede derivar en la insuficiencia de datos para obtener una solución óptima del problema, por lo que sería necesario recoger más datos.

2.2.1.3 Tratamiento de datos

El tratamiento de datos se realiza para obtener una solución óptima. Este procedimiento toma importancia tras la primera iteración del procedimiento de la implantación de *machine learning*, puesto que se utiliza para mejorar el modelo de aprendizaje automático. A continuación, se destacan los dos métodos de tratamiento de datos más comunes.

Sampling

Los datos desequilibrados son aquellos en los que hay un sesgo severo en la distribución de clases (mayoritarias y minoritarias). Este sesgo provoca errores en los algoritmos de aprendizaje supervisado, puesto que ignora por completo la clase minoritaria. Las predicciones más importantes suelen producirse sobre la clase minoritaria por lo que se deben modificar los datos no equilibrados. Para evitar la ignorancia de los datos de clase minoritaria se recurre al método *Sampling*, en español muestreo. Este procedimiento consiste en crear una nueva versión del conjunto de datos de entrenamiento, modificando la distribución de clase inicial. Esta nueva distribución se puede obtener eliminando aleatoriamente ejemplos en la clase mayoritaria, *oversampling* (sobremuestreo en español), o añadiendo ejemplos duplicados aleatoriamente en la clase minoritaria, *undersampling* (submuestreo en español) [23].

Crossvalidation

Para explicar este método, denominado validación cruzada en español, es fundamental clasificar los datos en función del uso de ellos en el modelo. Si se implementa el algoritmo usando esos datos u observaciones se está ante un conjunto de entrenamiento (*training set* en inglés) y si no se usan en la creación del modelo, se denominan conjunto de validación y conjunto de *test* (en inglés *validation set* and *test set*) [24]. La validación cruzada se emplea para estimar la habilidad de un modelo de aprendizaje automático en un conjunto de validación, es decir, utiliza una muestra limitada para estimar cómo se espera que funcione el modelo cuando se usa para hacer predicciones sobre datos que no se utilizaron para implementar el algoritmo [25].

El procedimiento de este método es simple:

1. Dividir el conjunto de datos en grupos
2. Para cada grupo
    2.1. Definir un grupo como conjunto de datos de validación
    2.2. Definir los grupos restantes como conjuntos de datos de entrenamiento

2.3. Ajustar el modelo según los datos de entrenamiento y evaluarlo con los datos de validación

2.4. Guardar la puntuación de evaluación y descartar el modelo

3. Resumir la eficiencia del modelo usando las puntuaciones de cada evaluación.

A través del uso de la validación cruzada se obtiene un resultado menos sesgado que con otros procedimientos, como la simple división de los conjuntos de datos.

2.2.1.4 Reducción de datos

Un gran conjunto de datos dificulta el análisis del problema, por lo que es recomendable realizar la reducción de datos.

En ocasiones el conjunto de una misma variable en diferentes datos abarca un rango muy amplio, por lo que es recomendable el uso de escala, para así garantizar una correcta asociación de los pesos a las variables en los modelos de *machine learning*.

2.2.2 Elección del modelo

Los modelos del aprendizaje supervisado se pueden dividir en dos categorías en función de la naturaleza de la variable de salida a predecir: los algoritmos de clasificación y de regresión. El primer grupo engloba los algoritmos que se suelen utilizar para estudios cuyo objetivo es obtener como resultado una etiqueta discreta o específica, es decir, la respuesta del problema es un conjunto finito de resultados posibles. Para la obtención de una cantidad como respuesta, se utilizan los modelos de regresión, ya que facilitan la predicción de variables continuas [26].

Existen numerosos algoritmos de aprendizaje supervisado, por lo que a continuación se detallarán, de forma abreviada, los modelos más recurrentes.

2.2.2.1 Logistic regression

Este modelo, al que nos referimos en español como regresión logística, es un algoritmo para problemas de clasificación y no de regresión, aunque su nombre (*logistic regression*) pueda llevar a la confusión. Se suele utilizar cuando la relación entre los datos no es muy compleja y cuando se tratan un gran número de datos. Se puede representar según la *Ecuación 2-1*.

$$y = \sigma(z) = \sigma(wx) = \sigma\left(\sum_i w_i \cdot x_i\right)$$

Ecuación 2-1

El valor de entrada es $\vec{x}$, un vector de valores reales. $\vec{w}$ representa el vector real de pesos y se obtiene de un conjunto de etiquetas de muestra de entrenamiento, que es como se le llama a la muestra de datos.

σ es la función logística representada por la siguiente fórmula, *Ecuación 2-2*.

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Ecuación 2-2

Puede tomar valores entre 0 (valor mínimo) y 1 (valor máximo). El resultado de sigma se interpreta como la probabilidad de clasificación. En caso de estudiar problemas binarios, si los valores de la función logística dan como resultado un número superior a 0.5, la muestra se clasificará en el grupo de clase 1 y en caso contrario al de 0 [27].

2.2.2.2 Decision tree

Los árboles de decisión, en inglés *decision tree*, son modelos que sirven tanto para los algoritmos de clasificación, si la variable es cualitativa (formada por un conjunto de valores), como para los modelos de regresión, cuando las variables son continuas.

Se basa en la creación de un diagrama de flujo y de la realización continua de preguntas cuyas respuestas solo pueden ser sí o no. Para inicializar el modelo, se utiliza el denominado nodo de raíz, la primera cuestión que se realiza. Cada vez que se realiza una pregunta, se agrega un nodo a la estructura del árbol y se clasifican los datos en diferentes clases dependiendo de si la respuesta es afirmativa o negativa. El nodo hoja se utiliza para detener el procedimiento y obtener una solución coherente. Para mayor compresión, se puede observar la *Ilustración 2-4. Ejemplo de árbol de decisión. Fuente: elaboración propia*.

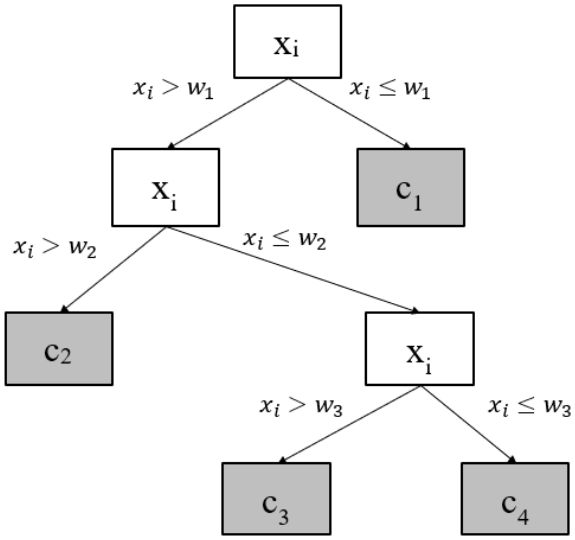

Ilustración 2-4. Ejemplo de árbol de decisión. Fuente: elaboración propia

Los nodos de hojas no siempre incluyen el conjunto de datos pertenecientes a una sola clase, es por esto por lo que la solución del algoritmo es la clase más común entre todos los puntos de los datos. Este inconveniente se define como tendencia de sobre-ajustar, *overfit* en inglés [28].

2.2.2.3 Random forest

Es una técnica que consiste en combinar un conjunto de árboles de decisión con la misma distribución, pero cada árbol depende de una muestra aleatoria de los datos, *bagging*. El resultado es mejor que el que se puede obtener con un solo árbol de decisión, e incluso se puede afirmar que el error de generalización disminuye a la hora de aumentar el número de árboles [29].

Cabe destacar que este modelo se utiliza tanto para la clasificación como para la regresión, ya que está basado en los árboles de decisión que se emplean para ambos casos, como se puede ver en la *Ilustración 2-5. Ejemplo del algoritmo random forest. Fuente: elaboración propia.*

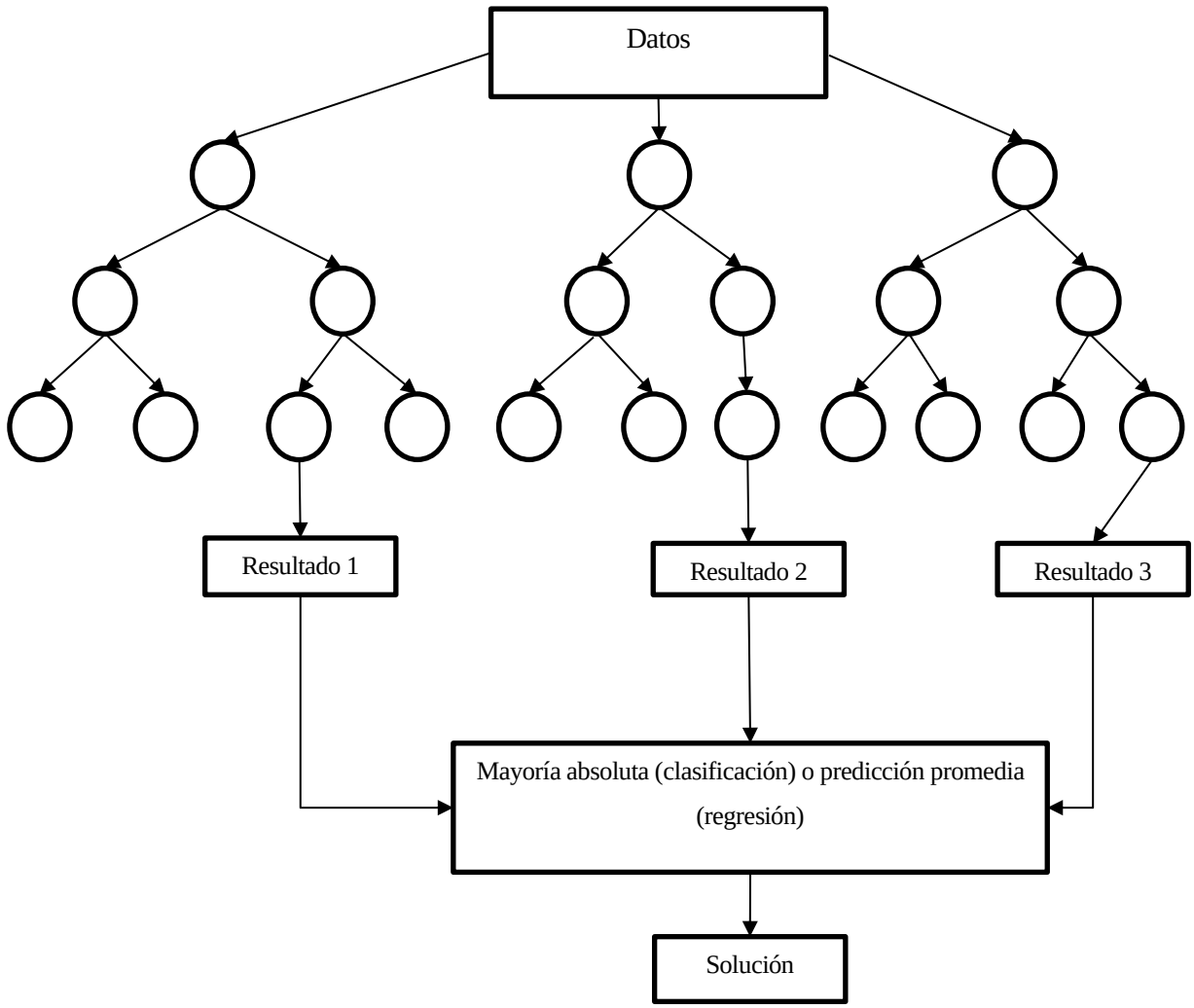

Ilustración 2-5. Ejemplo del algoritmo random forest. Fuente: elaboración propia

2.2.2.4 Artificial neural network

El modelo que se va a definir a continuación no solo responde al nombre de *artificial neural network*, sino que también se puede denominar como *simulated neural network* o como *neural network*. Su apelativo (redes neuronales en español) hace referencia a las neuronas cerebrales, puesto que imitan el comportamiento de estas.

Las redes neuronales se inicializan con una capa de entrada, conjuntos de nodos de iniciación, y finaliza con la capa de salida, nodos de salida. La entrada y la salida están unidas por capas ocultas, formadas también por diferentes nodos. Para transportar datos de una capa a otra, la salida de cualquier nodo individual tiene que ser un número superior al umbral que se especifique (este umbral depende de la investigación que se esté realizando). En el momento en el que se cumple, ese nodo se activa y así se consigue enviar los datos a la siguiente capa [30]. Esta explicación se entiende mejor observando *Ilustración 2-6. Estructura básica de una*

red neuronal.

Al igual que en los árboles de decisión, las redes neuronales se utilizan tanto para problemas de clasificación como de regresión.

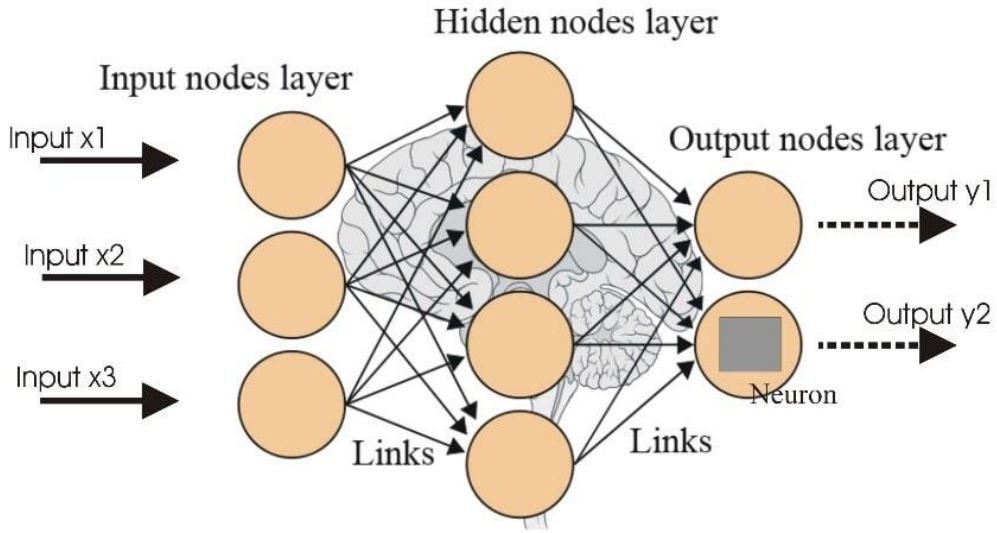

Ilustración 2-6. Estructura básica de una red neuronal [31]

2.2.2.5 Bayesian network

Bayesian network, en español red bayesiana, es un algoritmo gráfico de probabilidad utilizado mayoritariamente en el aprendizaje estadístico. Se caracteriza por su aplicación a los problemas que conllevan a conclusiones que solo se pueden construir a partir de información previa sobre ese problema. Sirve para definir la independencia condicional de las variables conocidas y compartir información y relación entre las variables desconocidas [32].

Se considera un modelo similar a las redes neuronales, diferenciándose en la estadística. Es por ello, que la clasificación y la regresión se puede realizar a través de esta red.

2.2.2.6 Support vector machine

La máquina de vectores de soporte es un algoritmo de regresión y muy similar a las redes neuronales. Se encarga de la construcción de un hiperplano que separa de forma óptima los datos en dos categorías. En estos modelos es muy frecuente recurrir a la función denominada *kernel*, cuya misión es encontrar ese hiperplano para separar los datos a través de crear una dimensión nueva [33]. En la *Ilustración* 2-7. Ejemplo de *SVM* se puede observar la solución que se obtiene utilizando este algoritmo, construyendo el hiperplano de las dos

formas, sin utilizar la función *kernel* y utilizándose.

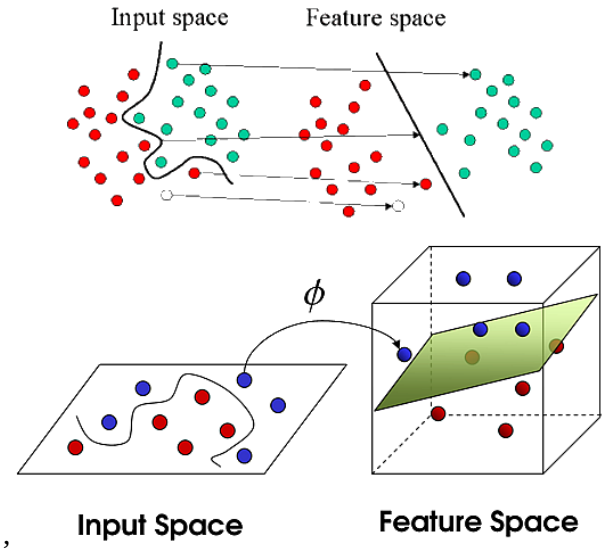

Ilustración 2-7. Ejemplo de SVM [34]

2.2.3 Análisis de errores

Para determinar la eficiencia del modelo, existen numerosos criterios de evaluación cuyo objetivo es indicar si el resultado que se ha obtenido en el modelo se aproxima a la realidad. Destacan MSE: *mean square error* (en español, error cuadrático medio) para modelos de regresión y ROC: *receiver operating characteristic* (curva característica operativa del receptor) para algoritmos de clasificación.

2.2.3.1 MSE: Error cuadrático medio

Este método se basa en la siguiente ecuación, siendo M el número de puntos, y_i: el error real, $\hat{y}_i$: el error estimado (obtenido del modelo). Si al aplicar la *Ecuación 2-3. Error cuadrático medio*, el resultado es 0, no habría error en la estimación de los resultados.

$$MSE = \frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)^2$$

Ecuación 2-3. Error cuadrático medio

El error cuadrático medio se suele emplear durante la aplicación del modelo, una vez terminado es mejor emplear el criterio de la raíz cuadrática del error cuadrático medio (RMSE), puesto que da un valor mucho más exacto. Su fórmula se puede observar en la *Ecuación 2-4. Raíz cuadrática del error cuadrático medio*.

$$RMSE = \sqrt{MSE}$$

Ecuación 2-4. Raíz cuadrática del error cuadrático medio

2.2.3.2 ROC: Curva característica operativa del receptor

Una curva ROC es un gráfico que muestra el rendimiento del modelo y para su representación se necesita la matriz de confusión. Cada columna de la matriz representa el número de predicciones de cada clase y cada fila muestra el número real de instancias de cada clase. Esta matriz proporciona la relación de las predicciones realizadas por el algoritmo con los resultados que se deberían de haber obtenido [35]. En la *Ilustración 2-8. Ejemplo de matriz de confusión* [36] se muestra un ejemplo de matriz de confusión para la clasificación en dos clases.

Ilustración 2-8. Ejemplo de matriz de confusión [36]

Los cuatro resultados se pueden clasificar en dos:

- Verdadero: el valor real coincide con la predicción obtenida del modelo. Puede darse el caso de que los valores fuesen positivos, entonces se estaría ante un verdadero positivo, o que sea negativo, que sería un verdadero negativo.
- Falso: existe una discrepancia entre el valor real y la predicción del algoritmo. En el caso de que, en la realidad el valor fuese positivo, se clasificaría como falso negativo, en caso contrario se clasificaría como falso positivo.

La curva característica operativa del receptor se representa utilizando la sensibilidad y la especificidad, métricas de la matriz de confusión. Ambas métricas indican la capacidad del estimador para discriminar los casos positivos de los casos negativos diferenciándose en la fracción de representación, puesto que la primera indica los verdaderos positivos y la segunda los verdaderos negativos, como se puede observar en la *Ecuación 2-5. Tasa de Verdaderos Positivos (Sensibilidad)* y en la *Ecuación 2-6. Tasa de Verdaderos Negativos (Especificidad)* [36].

$$TVP = \frac{VP}{VP + FN}$$

Ecuación 2-5. Tasa de Verdaderos Positivos (Sensibilidad)

$$TVN = \frac{VN}{VN + \ FP}$$

Ecuación 2-6. Tasa de Verdaderos Negativos (Especificidad)

Una vez realizada la representación gráfica, es recomendable trazar una diagonal de 45º, puesto que cuanto más se acerque la curva a esta recta, menos precisa es la prueba.

2.3 Relación de la distribución de agua y el aprendizaje automático

EurEau es la federación europea de asociaciones de abastecimientos de agua y saneamiento, representante de 29 países de Europa. En 2021 estableció que la longitud de la red de tuberías de agua potable era aproximadamente de 4.3 millones de kilómetros. Los fallos en las diferentes tuberías causaron unos valores medios de agua no facturados de aproximadamente el 25% del agua total. Si se realiza la comparación con los datos de 2017, se puede apreciar una disminución de la pérdida media en volumen debido a que numerosos países no respondieron al indicador.

En 2021, el importe global facturado del agua sin incluir el IVA fue de 108 billones de euros, un 21% mayor que en 2017. Cabe destacar que, en 2017, los datos de Alemania no se contabilizaron en el estudio puesto que no estaban disponibles [37].

En Estados Unidos, el sistema de distribución de agua está formado por 2.2 millones de millas de tuberías subterráneas, es decir, 35.41 millones de kilómetros. Se estima que se producen pérdidas diarias de alrededor de 6 billones de galones de agua, aproximadamente 3.8 billones de litros, debido a las roturas de las tuberías [38].

Por todo lo mencionado anteriormente, se está prestando gran atención a la gestión del agua, recalcando el alto porcentaje de deterioro de tuberías, que disminuyen la probabilidad de garantizar el derecho humano del agua y el PIB de cada país. Para combatir con este problema, desde 2010 se impulsó la implantación del aprendizaje automático en esta industria, es decir, se inicializó el estudio de los fallos de tuberías a través de modelos de *machine learning*.

Por ello, en este trabajo se realiza un análisis de la literatura científica más reciente sobre esta temática. Con ello se pretende destacar en qué punto está este campo en la actualidad, así como qué técnicas o procedimientos son las más recurrentes.

3 METODOLOGÍA DEL PROCESAMIENTO DEL LENGUAJE NATURAL

Este trabajo se basa en la aplicación del procesamiento del lenguaje natural. Debido a ello, en la primera subsección se define el concepto del procesamiento del lenguaje natural y se indica algunas de las numerosas aplicaciones del procesamiento del mismo.

Seguidamente, se determina qué aplicación se ha acuñado para este trabajo y la herramienta empleada para el desarrollo del programa.

Para finalizar, se han desarrollado los algoritmos principales que componen el programa.

3.1 Definición y aplicaciones

En 1960, nació el procesamiento del lenguaje natural (en inglés: *Natural Language Processing* (NLP)), un campo de conocimiento de inteligencia artificial que ayuda a las computadoras a comprender, interpretar y manipular el lenguaje humano [39]. La idea fundamental de esta tecnología es dotar a la máquina de las capacidades necesarias para conseguir comunicarse con el ser humano, es decir, la máquina tiene que ser capaz de entender, procesar y aprender el lenguaje natural. Las aplicaciones principales del procesamiento del lenguaje natural son: la recuperación y extracción de información, la minería de datos, la traducción automática, los sistemas de búsqueda de respuesta, la generación de resúmenes automáticos y el análisis de sentimientos [40].

- Recuperación y extracción de información: consiste en un buscador, donde el individuo puede introducir palabras claves y el ordenador sea capaz de encontrar esos términos en el texto seleccionado.
- Minería de datos o en inglés *Text Mining*: se utiliza para un elevado número de archivos de texto. La idea principal es conseguir que el ordenador sea capaz de determinar la información relevante, clasificar los documentos y construir una base de datos [41] .
- Traducción automática: también denominada *Machine Translation* en inglés, se aplica para traducir documentos escritos en diferentes idiomas. Se basa en el estudio del texto en el lenguaje original y la

traducción de ese texto, en lugar de la sustitución de palabra por palabra. De esta forma se consigue un resultado más óptimo, ya que un mismo término en un idioma puede tener muchos significados diferentes [42].

- Sistemas de búsquedas de respuestas: utilizados por la mayoría de las páginas web que ofrecen un servicio al cliente. Su objetivo es capacitar al sistema para que pueda responder a las preguntas más frecuentes de los usuarios.
- Generación de resúmenes automáticos: aplicación del procesamiento del lenguaje natural que genera un resumen a partir de un texto, mostrando las palabras más importantes y eliminando las que no aportan ninguna información como los artículos, las preposiciones, etcétera.
- Análisis de sentimientos: en la mayoría de las empresas se utilizan los sistemas de calificación de servicios para determinar si el cliente queda satisfecho o se puede mejorar algún aspecto del servicio prestado. Estos sistemas son ejemplos de análisis de sentimientos, puesto que se emplean para determinar los sentimiento o emociones de un individuo hacia un sistema [40].

3.2 Implementación

Como se ha mencionado en la introducción, el número de documentos de producción científica ha aumentado significativamente en los últimos años. Es por ello por lo que se ha decidido desarrollar un programa basado en el procesamiento del lenguaje natural, para así poder recoger la información de numerosos documentos y poder realizar un análisis entre los diferentes artículos sin necesidad de ayuda humana.

Python es un lenguaje de alto nivel porque contiene diferentes estructuras de datos implícitas como listas, diccionarios y tuplas. Fue desarrollado en 1991 por el holandés Guido van Rossum para manejar las excepciones del lenguaje de programación ABC e interactuar con el sistema operativo Amoeba [43].

Las numerosas librerías de Python, también denominadas bibliotecas, y la simplicidad a la hora de programar hacen que este software sea el más utilizado para proyectos de análisis de datos [44].

Por todo lo nombrado anteriormente, se ha determinado que Python es el mejor software para poder desarrollar el programa deseado.

El sistema desarrollado se emplea para determinar la frecuencia de aparición de cada palabra que componen los diferentes documentos. En la *Ilustración 3-1. Esquema ilustrativo del programa*, se observa como la entrada del programa son los documentos que se desean analizar y la salida son dos tablas. La mayoría de los artículos incluye un apartado denominado palabras claves, en inglés "*keywords*", formado por las palabras que permiten clasificar el artículo. Por ello, una de las tablas hace referencia a dichas palabras claves y la otra tabla

a los textos completos de los artículos.

En cada tabla las filas son los diferentes documentos, las columnas las palabras que aparecen en al menos un documento y las celdas indican el número de veces que aparece esa palabra en ese artículo.

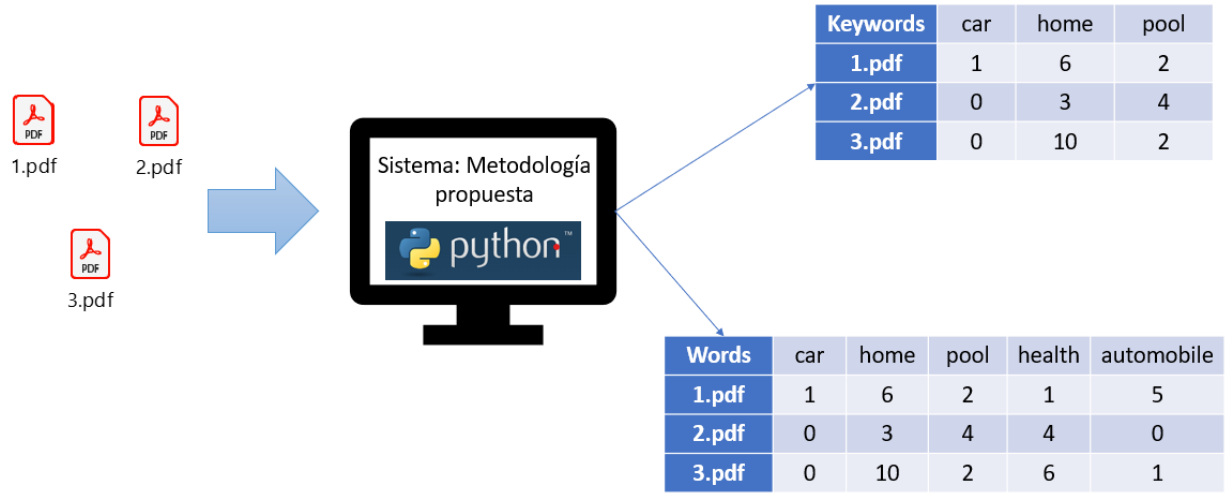

Ilustración 3-1. Esquema ilustrativo del programa

El programa consta de dos etapas o estructuras principales: la recogida de datos y la unificación de datos. En cada una se han utilizado distintas librerías y herramientas existentes en el lenguaje de programación Python, tal y como se muestra a continuación.

3.2.1 Recogida de datos

Los documentos de producción científica constan de un gran número de palabras, imágenes y ecuaciones entre otros, por lo que el primer paso es recoger la información relevante. Para ello, a la hora de obtener los datos a través del software, se han eliminado los caracteres no alfabéticos, puesto que no aportan información.

Las librerías utilizadas en este paso han sido:

- PyMuPDF: se ha empleado para convertir archivos de extensión "pdf" a archivos de texto, dado que el software es mucho más eficiente tratando esta extensión.
- NLTK: librería que emplea "tokenize", metodología para dividir un texto en palabras o términos individuales.

- Pattern: empleada para transformar una palabra plural en singular.
- Geny, Spacy: son dos librerías que junto a NLTK, permiten eliminar "*stopwords*" (en español palabras vacías), palabras que no aportan información, como los determinantes y los pronombres entre otros.

3.2.2 Unificación de datos

Tras tener toda la información de cada documento recogida en variables es necesario unificar esos datos. El nexo de la unión deseada es el conjunto de todas las diferentes palabras que conforman los artículos que se quieren estudiar.

Como se puede deducir, la recogida de las diferentes palabras es muy elevada, por lo que para poder analizar los datos de manera eficiente se ha utilizado la librería Pandas, caracterizada por ser rápida y eficiente para la manipulación de datos.

A través de esta biblioteca se crea una estructura sencilla, denominada "*DataFrame*" (en español cuadro de datos), similar a una matriz, donde las columnas son las diferentes palabras y las filas son los documentos estudiados. En cada componente, se encuentra el número de aparición de esa palabra en ese documento.

Esta estructura facilita la búsqueda de información en los artículos y la obtención de visualizaciones óptima de los resultados del estudio que se desee realizar a través de dos librerías principalmente: Matplotlib y Seaborn. Estas dos bibliotecas se han empleado para la generación de gráficas.

3.3 Algoritmos propuestos

Para realizar el proyecto, se han definido algoritmos que incluyen numerosas funciones definidas por el usuario y otras exportadas a través de los módulos y librerías de Python, mencionado anteriormente. Para una mayor comprensión, se ha incluido la *Ilustración 3-2. Estructura del programa*.

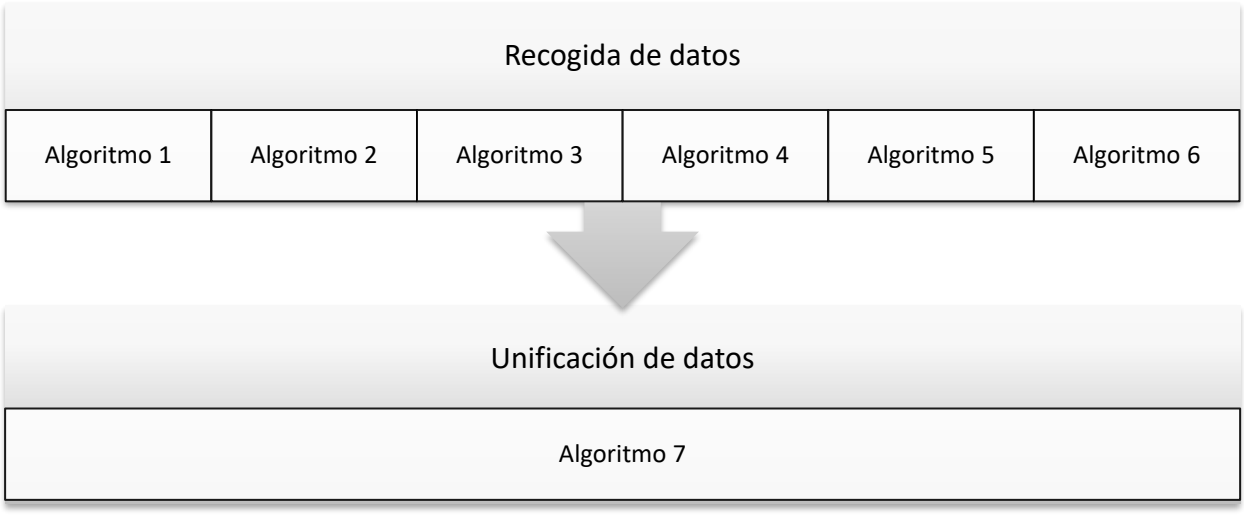

Ilustración 3-2. Estructura del programa

La *Tabla 3-1. Definición de variables* se ha incluido para determinar las variables que se van a emplear, en ella se puede observar los nombres, los tipos y las definiciones de cada variable. Cabe destacar que, al inicio del programa, la única variable definida es "documentos", las demás están vacías.

Nombre	Tipo	Definición
documentos	lista	nombres de los n documentos
w	diccionario	formado por n listas que incluyen las palabras del texto completo de cada documento
kw	diccionario	formado por n listas que incluyen las palabras claves de cada documento
documento	cadena	nombre del documento con formato pdf
documento_txt	cadena	nombre del documento con formato txt
words	lista	palabras del texto completo de un documento
keywords	lista	palabras claves de un documento
frec_w	dataframe	frecuencia de aparición de las palabras del texto completo de cada documento
frec_kw	dataframe	frecuencia de aparición de las palabras claves de cada documento

Tabla 3-1. Definición de variables

El *Algoritmo 3-1. Transformación de datos* es el algoritmo inicial del programa formado por numerosas funciones, también denominadas algoritmos. Se encarga de guardar en diferentes variables, denominadas en Python: diccionarios, toda la información que se quiere estudiar de los distintos documentos. El primer paso que realiza el algoritmo es comprobar que el archivo se encuentre en la misma ruta del programa. Una vez se ha afirmado que el documento está en el directorio correcto, a través de la función denominada os.path.isfile, se realizan los diferentes algoritmos: *Algoritmo 3-2. Convertir el documento pdf en txt*, *Algoritmo 3-3. Eliminar caracteres no alfabéticos del documento*, *Algoritmo 3-4. Guardar las palabras relevantes del*

documento, Algoritmo 3-5. Obtener las palabras claves del documento y *Algoritmo 3-6. Eliminar de words las palabras que forman el apartado de referencias en el documento*. En caso de que el archivo no se encuentre, se mostrará por pantalla qué documento no está guardado en la carpeta correcta y se finalizará el programa.

Algoritmo 1 Transformación de datos

Entrada: documentos= ["1.*pdf*", ..., "102.*pdf*"]
  n=len(*documentos*)
  i= 1
  mientras $i \leq n$ **hacer**
    si (documentos[i]) *se encuentra en la ruta del programa* **entonces**
      Convertir el documento pdf en txt (documentos[i])
      Eliminar caracteres no alfabéticos del (documentos[i])
      w[documentos[i]]= *Guardar las palabras relevantes del* (documentos[i])
      kw[documentos[i]]= *Obtener las palabras claves del documento*(w[documentos[i]])
      Eliminar de (w[documentos[i]]) *las palabras que forman el apartado de las referencias en el documento*
    si no
      print("*No se ha encontrado documentos*(*i*)")
      EXIT
    fin si
    i=i+1
  fin mientras
Salida: w, kw

Algoritmo 3-1. Transformación de datos

Python genera muchos problemas a la hora de guardar información de archivos "pdf" es por ello por lo que se ha diseñado el *Algoritmo 3-2. Convertir el documento pdf en txt*. El procedimiento del algoritmo es crear un archivo de texto en blanco, y copiar cada página del archivo "pdf" al texto.

Algoritmo 2 Convertir el documento pdf en txt

Entrada: documento
  Abrir (documento)
  documento_txt= *Archivo txt vacío con nombre del* (documento)
  para *cada* pagina *del* documento **hacer**
    Copiar la pagina *de* (documento) *en* (documento_txt)
  fin para
  Cerrar (documento_txt)
  Cerrar (documento)

Algoritmo 3-2. Convertir el documento pdf en txt

Los documentos suelen incluir ecuaciones, números y otros elementos que no son necesarios para este estudio. El *Algoritmo 3-3. Eliminar caracteres no alfabéticos del documento* se encarga de eliminar enlaces, números, símbolos y múltiples espacios vacíos. Cabe destacar que antes de suprimir los caracteres no alfabéticos, se han

sustituido las palabras compuestas (formadas por guiones) por palabras simples, como por ejemplo "*over-sampling*" en "*oversampling*".

Algoritmo 3 Eliminar caracteres no alfabéticos del documento

Entrada: documento_txt
  Abrir (documento_txt) *en modo lectura*
  texto= *Guardar todo el escrito del* (documento_txt)
  Cerrar (documento_txt)
  Abrir (documento_txt) *en modo escritura*
  para *cada* línea *del* texto **hacer**
    Eliminar enlaces de (línea)
    Eliminar todos los caracteres numéricos de (línea)
    Sustituir palabras compuestas de (línea) *por palabras simples*
    Eliminar todos los caracteres no alfanuméricos de (línea)
    Eliminar múltiples espacios vacíos de (línea)
  fin para
  Cerrar (documento_txt)

Algoritmo 3-3. Eliminar caracteres no alfabéticos del documento

El siguiente algoritmo, *Algoritmo 3-4. Guardar las palabras relevantes del documento*, se ha desarrollado para obtener la información relevante de los documentos. En la variable texto se almacena todo el texto del documento como una cadena de caracteres. Posteriormente, se almacenan todas las palabras del texto en la lista, denominada "w". Como más adelante será necesario determinar la frecuencia de aparición de cada palabra, se han convertido todas las letras mayúsculas en minúsculas y las palabras plurales en singulares, para así poder determinar de forma eficiente el número de aparición de cada palabra. Otro aspecto que se ha de mencionar es la eliminación de las palabras como los artículos, preposiciones y verbos simples (*stopwords*); puesto que no aportan información para la investigación.

Algoritmo 4 Guardar las palabras relevantes del documento

Entrada: documento_txt
  Abrir (documento_txt) *en modo lectura*
  texto= *Guardar todo el escrito del* (documento_txt)
  Cerrar (documento_txt)
  words= *palabaras del* (texto)
  Remplazar todas las letras mayúsculas por minúsculas de (words)
  Convertir todas las palabras plurales en singulares de (words)
  Eliminar de (words) *todas las palabras irrelevantes*
Salida: words

Algoritmo 3-4. Guardar las palabras relevantes del documento

Se ha mostrado interés en la diferenciación de las palabras claves y del texto completo, por lo que se ha desarrollado el *Algoritmo 3-5. Obtener las palabras claves del documento*, encargado de obtener las palabras claves. Se ha copiado el diccionario "w" en otra lista denominada "k" y de esta última se han eliminado las palabras que aparecen antes de la palabra "keyword" y las palabras posteriores a la palabra que ocupa la posición 21, puesto que se ha estimado que como máximo se incluyen 21 palabras en el apartado de palabras claves del documento. Además, como la mayoría de los artículos siguen la misma estructura, después del apartado de palabra claves, aparece la sección denominada en inglés "introduction" o "highlight"; por lo que se ha incluido un comando para determinar si la palabra "introduction" o "highlight" aparece en la lista "k" y si es así, eliminar esta palabra y las palabras posteriores a esta.

Algoritmo 5 Obtener las palabras claves del documento

Entrada: words
  keywords=words
  si *la palabra "keyword" o 'keywor" aparece en* (keywords) **entonces**
    Eliminar de (keywords) *todas las palabras hasta la primera aparicion de la palabra "keyword"*
    si len(*keywords*) > 20 **entonces**
      Eliminar de (keywords) *todas las palabras posteriores a la palabra cuya posición es la* 21
    fin si
    si *la palabra "introduction" o "highlight" aparece en* (keywords) **entonces**
      Eliminar de (keywords) *todas las palabras posteriores a la primera aparición de la palabra "introduction" o "highlight"*
    fin si
  si no
    Eliminar de (keywords) *todas las palabras posteriores a la palabra cuya posición es la* 21
  fin si
Salida: keywords

Algoritmo 3-5. Obtener las palabras claves del documento

Al final de los artículos, se suele redactar el apartado de referencias, en inglés "references", que incluyen todas las citas a las que hace referencia el autor del documento. Estos datos son innecesarios para el estudio, por lo que el *Algoritmo 3-6. Eliminar de words las palabras que forman el apartado de referencias en el documento* se encarga de eliminar las palabras posteriores a ese apartado. Para ello, se determina la máxima posición de la palabra "reference" en la lista "w" y se elimina esta palabra y las posteriores.

Algoritmo 6 Eliminar de words las palabras que forman el apartado de referencias en el documento

Entrada: words
  si *la palabra "reference" aparece en* (words) **entonces**
    Eliminar de (words) *todas las palabras posteriores a la última aparición de la palabra "reference"*
  fin si
Salida: words

Algoritmo 3-6. Eliminar de words las palabras que forman el apartado de referencias en el documento

Todo lo mencionado respecto a la recogida de información de las palabras del texto completo y las palabras claves, se recoge en los diccionarios "w" y "kw". Cada diccionario está formado por n listas que recogen las palabras de cada caso.

Seguidamente, se incluye el segundo algoritmo principal del programa que, como el primero, incluye otras funciones que se describirán más adelante. El *Algoritmo 3-7. Crear dataframe* tiene como objetivo crear dos cuadros de datos, donde el primero representa la frecuencia de aparición de las diferentes palabras del texto completo del documento ("frec_w"). En cambio, el segundo ("frec_kw") solo muestra la frecuencia de las palabras claves, por lo que este cuadro de texto presenta una dimensión menor al primero. Dado que para los dos "dataframe" los pasos a seguir son los mismos, diferenciándose en los diccionarios seleccionados, se va a explicar solo uno.

Para "frec_w", es necesario definir "index" y "columns" (en español filas y columnas) del cuadro de datos. Para determinar el nombre de las columnas, se ha creado una lista de palabras denominadas "total_w", donde se encuentre las diferentes palabras que componen los n documentos. El nombre de las filas corresponde a los diferentes nombres de los artículos. Para poder incluir en cada celda la frecuencia de aparición de esa palabra en ese documento es necesario crear previamente un cuadro de datos auxiliar, cuyas columnas son los nombres de los documentos, y las celdas están rellenas con las palabras de las listas de palabras de cada documento.

Los "dataframe" tienen que estar formados por listas con las mismas dimensiones, por lo que previamente se ha definido una función para igualar la dimensión de las listas. Para ello, se han recorrido las diferentes listas correspondientes a los documentos para determinar la dimensión máxima. En los casos en los que la dimensión de alguna lista sea menor que el máximo, se ha incluido tantos "0" como elementos necesarios para tener la misma longitud. Una vez realizado este paso, se incluyen en el "dataframe" auxiliar, todas las palabras correspondientes a cada columna.

Se recorre este cuadro auxiliar y se determina el número de aparición de cada palabra en cada documento.

Algoritmo 7 Crear dataframe

Entrada: documentos, kw, w
  total_w= *Lista de todas las palabras sin repetir de* (words)
  total_ks= *Lista de todas las palabras sin repetir de* (kw)
  frec_w = *dataframe donde index* = (total_w), *columns* = (documentos) *y cada celda es la frecuencia de aparición de la palabra de* index *en ese documento*
  frec_kw = *dataframe donde index* = (total_kw), *columns* = (documentos) *y cada celda es la frecuencia de aparición de la palabra de* index *en ese documento*
  Trasponer(frec_w)
  Trasponer(frec_kw)
Salida: frec_w, frec_kw

Algoritmo 3-7. Crear dataframe

Tras realizar estos dos algoritmos principales, solo queda representar la información que se desea mostrar. Para

ello, se han desarrollado funciones que se incluyen en el anexo.

4 CASO DE ESTUDIO

El caso de estudio de este trabajo es la producción científica sobre el aprendizaje automático en el sector del agua. Se trata de recopilar documentos que incluyan el estudio de aplicaciones de aprendizaje supervisado a la distribución de agua. Para ello se han localizado 102 artículos científicos que cubren estas temáticas.

Posteriormente, se han inventariado los términos que se quieren analizar en el cuerpo de los documentos. Esos términos están relacionados con el *machine learning* y los factores de fallos en las tuberías. Esta búsqueda de términos en los diferentes documentos, sin la necesidad de abrir los 102 artículos, tiene numerosas finalidades, como determinar qué modelos son los más empleados o la relación entre ellos, y los factores para la distribución de agua, entre otros.

4.1 Búsqueda de artículos

Para la obtención de los artículos que se enfocan en la aplicación de modelos de *machine learning* en la predicción de fallos o roturas en las tuberías que conforman las redes de distribución de agua, se ha utilizado Scopus, una base de datos de referencias bibliográficas, artículos y revistas científicas. Esta herramienta permite realizar búsquedas con diferentes filtros.

En este caso se ha realizado la búsqueda, en inglés, para documentos que contengan las palabras *machine learning* o *machine-learning*, *water supply* o *water distribution* y *pipe fail* o *pipe break**, como se puede observar en la *Ilustración 4-1. Búsqueda realizada*.

```
("machine learning" OR "machine-learning") AND ("water supply" OR "water distribution") AND
("pipe fail*" OR " pipe break*") AND ( LIMIT-TO ( LANGUAGE,"English" ) )
```

Ilustración 4-1. Búsqueda realizada

La importancia del idioma es debido al alto número de artículos científicos redactados en inglés. El segundo filtro se utiliza para acortar la búsqueda a artículos de modelos de *machine learning* que contengan las palabras de suministro o distribución de agua. Por último, se han introducido las palabras *pipe fail* o *pipe break* que hacen referencia a fallo o rotura de tuberías. Los asteriscos que se encuentran justo después de la palabra *fail* y *break* sirven para encontrar artículos que contengan diferentes tiempos verbales del verbo *fail* o *break*. Por ejemplo, en la búsqueda realizada se encontrará documentos que contengan la palabra *failed* en lugar de *fail*, en español sería equivalente a roto y romper.

La idea inicial era encontrar los artículos que incluyesen las palabras mencionadas anteriormente en algunos de los siguientes apartados: títulos, resúmenes y palabras claves. Tras haber realizado esa búsqueda se encontraron aproximadamente 30 artículos y puesto que, no todos los artículos se pueden descargar sin coste, se obtenía un número muy limitado de artículos. Este número reducido de documentos se debe a numerosas causas como son una estructura diferente a la habitual de los documentos y la exclusión de las palabras "*machine learning"* en los diferentes apartados puesto que es más común mencionar el modelo de aprendizaje automático. Debido a ello, se ha buscado todos los documentos que incluyeran los diferentes términos.

A la hora de realizar esa búsqueda, se encontraron 281 documentos que cumplían esas características. Se ha de tener en cuenta la limitación de la descarga no gratuita de los artículos, por lo que no ha sido posible realizar el programa para ese número de archivos. La búsqueda engloba las publicaciones desde el año 1997, pero dado que no se disponía de la misma tecnología que en la última década, se ha trabajado con los artículos desde el 2012 hasta el 30 de abril de 2022, fecha de realización de la búsqueda.

A continuación, se muestra en la *Tabla 4-1. Artículos de modelos del aprendizaje artificial en problemas de rotura de tuberías*, los títulos de los artículos, cómo se han nombrado para el programa y el año de publicación.

Año	Nombre documento	Número
2013	Dynamic soft sensors for detecting factors affecting turbidity in drinking water [45]	1
2013	Entropy-Based Sensor Placement Optimization for Waterloss Detection in Water Distribution Networks [46]	2
2014	Forecasting hurricane-induced power outage durations [47]	3
2014	Water pipe condition assessment: A hierarchical beta process approach for sparse incident data [48]	4
2015	Water distribution networks [49]	5
2016	Advances in Knowledge Discovery and Data Mining [50]	6
2016	Oriented review to potential simulator for faults modeling in diesel engine [51]	7
2017	A prioritization method for replacement of water mains using rank aggregation [52]	8
2017	Evaluation of chlorine decay models under transient conditions in a water distribution system [53]	9
2017	State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains [54]	10

Año	Nombre documento	Número
2017	Stochastic data mining tools for pipe blockage failure prediction [55]	11
2018	Application of K-nearest neighbours method for water pipes failure frequency assessment [56]	12
2018	Artificial neural networks: Applications in the drinking water sector [57]	13
2018	K-nearest neighbours method as a tool for failure rate prediction [58]	14
2018	Neural network using the Levenberg–Marquardt algorithm for optimal real-time operation of water distribution systems [59]	15
2018	Pipe failure modelling for water distribution networks using boosted decision trees [60]	16
2018	Pipeline failure prediction in water distribution networks using weather conditions as explanatory factors [61]	17
2018	Proposed probabilistic models of pipe failure in water distribution system [62]	18
2018	Scaling-laws of flow entropy with topological metrics of water distribution networks [63]	19
2018	Sewer condition prediction and analysis of explanatory factors [64]	20
2019	A brief review of random forests for water scientists and practitioners and their recent history in water resources [65]	21
2019	An observatory framework for metropolitan change: Understanding urban social-ecological-technical systems in texas and beyond [66]	22
2019	Effects of weather conditions on drinking water distribution pipe failures in the Netherlands [67]	23
2019	Extreme learning machine model for water network management [68]	24
2019	Is bigger always better? A controversial journey to the center of machine learning design, with uses and misuses of big data for predicting water meter failures [69]	25
2019	Multitask learning for sparse failure prediction [70]	26
2019	Prediction and sensitivity analysis of bubble dissolution time in 3D selective laser sintering using ensemble decision trees [71]	27
2019	Review of the quantitative resilience methods in water distribution networks [72]	28
2019	Software for on-line testing of pipeline modeling methods [73]	29
2019	The analysis of water supply operating conditions systems by means of empirical exponents [74]	30
2020	A decision tree approach to the risk evaluation of urbanwater distribution network pipes [75]	31
2020	An Approach to Predict the Failure of Water Mains Under Climatic Variations [76]	32
2020	Bayesian network-based methodology for selecting a cost-effective sewer asset management model [77]	33
2020	Causal reasoning application in smart farming and ethics: A systematic review [78]	34
2020	Combining recorded failures and expert opinion in the development of ANN pipe failure prediction models [79]	35
2020	Comparison of statistical and machine learning models for pipe failure modeling in water distribution networks [80]	36
2020	Deep learning approach for diabetes prediction using PIMA Indian dataset [81]	37
2020	Development of a binary model for evaluating water distribution systems by a pressure driven analysis (PDA) approach [82]	38
2020	Geoadditive quantile regression model for sewer pipes deterioration using boosting optimization algorithm [83]	39
2020	Long-Term Water Pipe Condition Assessment: A Semiparametric Model Using Gaussian Process and Survival Analysis [84]	40
2020	Pipe fault prediction for water transmission mains [85]	41

Año	Nombre documento	Número
2020	Predicting failures in electronic water taps in rural sub-Saharan African communities: An LSTM-based approach [86]	42
2020	Risk analysis of chemical plant explosion accidents based on bayesian network [87]	43
2020	Survival analysis of water distribution network under intermittent water supply conditions [88]	44
2020	Twenty years of asset management research for Dutch drinking water utilities [89]	45
2020	Urban water management: A pragmatic approach [90]	46
2020	Water network-failure data assessment [91]	47
2021	A genetic algorithm-based support vector machine to estimate the transverse mixing coefficient in streams [33]	48
2021	Adaptive Metaheuristic Scheme for Generalized Multiple Abnormality Detection in a Reservoir Pipeline Valve System [92]	49
2021	Anomaly detection in dam behaviour with machine learning classification models [93]	50
2021	Artificial neural networks to forecast failures in water supply pipes [94]	51
2021	Estimation of a logistic regression model by a genetic algorithm to predict pipe failures in sewer networks [95]	52
2021	Extreme wave height detection based on the meteorological data, using hybrid NOF-ELM method [96]	53
2021	Failure detection methods for pipeline networks: From acoustic sensing to cyber-physical systems†[97]	54
2021	Improved ensemble-learning algorithm for predictive maintenance in the manufacturing process [98]	55
2021	Improving nonconformity responsibility decisions: a semi-automated model based on CRISP-DM [99]	56
2021	Incorporation of covid-19-inspired behaviour into agent-based modelling for water distribution systems' contamination responses [100]	57
2021	Integrated Planning of Operational Maintenance Programs for Water and Gas Distribution Networks [101]	58
2021	Iterative classifier optimizer-based pace regression and random forest hybrid models for suspended sediment load prediction [102]	59
2021	Knowledge management and operational capacity in water utilities, a balance between human resources and digital maturity—the case of ags [103]	60
2021	Long-Term Pipeline Failure Prediction Using Nonparametric Survival Analysis [104]	61
2021	Machine Learning and AI for Water Utilities: Junk or Jewel? Triumph or Trash? [105]	62
2021	Microplastics in combined sewer overflows: An experimental study [106]	63
2021	Modeling Pipe Break Data Using Survival Analysis with Machine Learning Imputation Methods [107]	64
2021	Modelling the impact of water temperature, pipe, and hydraulic conditions on water quality in water distribution networks [108]	65
2021	Multi-task learning by hierarchical Dirichlet mixture model for sparse failure prediction [109]	66
2021	Performance evaluation of machine learning algorithms for seismic retrofit cost estimation using structural parameters [110]	67
2021	Pipe Break Rate Assessment While Considering Physical and Operational Factors: A Methodology based on Global Positioning System and Data-Driven Techniques [111]	68
2021	Pipeline in-line inspection method, instrumentation and data management [112]	69

Año	Nombre documento	Número
2021	Predicting pipeline corrosion in heterogeneous soils using numerical modelling and artificial neural networks [113]	70
2021	Predictive analytics for water main breaks using spatiotemporal data [114]	71
2021	Regression models utilization to the underground temperature determination at coal energy conversion [115]	72
2021	Rehabilitation of an industrial water main using multicriteria decision analysis [116]	73
2021	Sensitivity analysis for performance evaluation of a real water distribution system by a pressure driven analysis approach and artificial intelligence method [117]	74
2021	Smart technologies for sustainable water management: An urban analysis [118]	75
2021	System-level prognostics and health management: A graph convolutional network–based framework [119]	76
2021	Teaching-learning-based optimization of neural networks for water supply pipe condition prediction [120]	77
2021	The Value of Machine Learning Main Break Prediction [121]	78
2021	The varying threshold values of logistic regression and linear discriminant for classifying fraudulent firm [122]	79
2021	Trends and applications of machine learning in water supply networks management [123]	80
2021	Water Connection Bursting and Leaks Prediction Using Machine Learning [124]	81
2021	Water Leakage Detection Using Neural Networks [125]	82
2021	Water pipe failure prediction using AutoML [126]	83
2021	Water poverty assessment based on the random forest algorithm: application to Gansu, Northwest China [127]	84
2021	Water treatment and artificial intelligence techniques: a systematic literature review research [128]	85
2021	Why are You Here? Modeling Illicit Massage Business Location Characteristics with Machine Learning [129]	86
2022	A machine learning based credit card fraud detection using the GA algorithm for feature selection [130]	87
2022	A Risk-Based Approach in Rehabilitation of Water Distribution Networks [131]	88
2022	Contribution of Internet of things in water supply chain management: A bibliometric and content analysis [132]	89
2022	Current Trends in Fluid Research in the Era of Artificial Intelligence: A Review [133]	90
2022	Data-driven approach to predict the sequence of component failures: a framework and a case study on a process industry [134]	91
2022	Decision with Uncertain Information: An Application for Leakage Detection in Water Pipelines [135]	92
2022	Development of Methods for Diagnosing the Operating Conditions of Water Supply Networks over the Last Two Decades [136]	93
2022	Evaluation of sewer network resilience index under the perspective of ground collapse prevention [137]	94
2022	Failure Prediction of Municipal Water Pipes Using Machine Learning Algorithms [138]	95
2022	Hybrid models for suspended sediment prediction: optimized random forest and multi-layer perceptron through genetic algorithm and stochastic gradient descent methods [139]	96
2022	Machine learning applied on the district heating and cooling sector: a review [140]	97
2022	Optimization of water resources utilization by GA–PSO in the Pinshuo open pit combined mining area, China [141]	98

Año	Nombre documento	Número
2022	Real-Time burst detection based on multiple features of pressure data [142]	99
2022	Runoff Simulation Under Future Climate Change Conditions: Performance Comparison of Data-Mining Algorithms and Conceptual Models [143]	100
2022	The challenges of predicting pipe failures in clean water networks: A view from current practice [144]	101
2022	Water quality prediction based on Naïve Bayes algorithm [145]	102

Tabla 4-1. Artículos de modelos del aprendizaje artificial en problemas de rotura de tuberías

4.2 Búsqueda de palabras

Una de las aplicaciones del programa desarrollado es la búsqueda de la frecuencia de una palabra en los documentos analizados. Esto es muy interesante para determinar el número de artículos que se centran en el estudio de algún modelo de *machine learning* o de los factores de la distribución de agua.

De cara a simplificar el mecanismo de búsqueda, se ha optado por asignar a cada elemento, objeto del análisis, una palabra de búsqueda que los representa. Así, por ejemplo, para el modelo *random forest* se ha asignado la palabra de búsqueda "*forest*". Como todos los artículos están en inglés, la búsqueda se ha realizado en ese idioma.

Seguidamente se detallará la relación completa de elementos con sus palabras de búsqueda y la traducción al español.

4.2.1 Machine learning

4.2.1.1 Modelos de machine learning

La mayoría de los artículos recopilados aplican modelos de aprendizaje automático, por ello es interesante analizar cuáles son los más utilizados (aparecen en las *keywords*, apartado del documento donde se encuentran las palabras claves) y los más recurrentes en general (analizando el texto completo). Para ello, se ha creado una lista denominada modelos, donde aparecen las palabras de los algoritmos más empleados según lo mencionado anteriormente. Para realizar la búsqueda se emplea un término representativo del modelo, en la *Tabla 4-2. Modelos de machine learning* se ilustra los términos buscados.

Palabra	Model	Modelo
neural	artifical neural network	redes neuronales
forest	random forest	bosques aleatorios
tree	decision tree	árbol de decisión
vector	support vector machine	máquinas de vectores de soporte
regression	regression	regresión
classification	classification	clasificación

Tabla 4-2. Modelos de machine learning

4.2.1.2 Procesamiento y análisis de exploración de los datos

En base al artículo científico "Trends and Applications of Machine Learning in Water Supply Networks Management" [123] en el cual se analizan 13 artículos sobre el uso de *machine learning* para la predicción de rotura de tuberías en redes de agua, se analizan los términos que se muestran en la *Tabla 4-3. Procesamiento y análisis de exploración de los datos*, relacionados con el procesamiento de los datos y el entrenamiento y la validación de los modelos.

Palabra	Concept	Concepto
missing	missing value	datos incompletos
outlier	outlier	datos atípicos
sampling	sampling	muestreo
undersampling	undersampling	submuestreo
oversampling	oversampling	sobremuestreo
crossvalidation	crossvalidation	validación cruzada
mse	mean square error	error cuadrático medio
rmse	root mean square error	raíz cuadrática del error cuadrático medio
roc	receiver operating characteristic	curva característica operativa del receptor
confusion	confusion matrix	matriz de confusión

Tabla 4-3. Procesamiento y análisis de exploración de los datos

4.2.2 Redes de distribución de agua

Con relación a las redes de distribución de agua, el estudio se enfoca en analizar qué variables son las más empleadas, y, por tanto, las más registradas en las bases de datos de las empresas. Para este caso, no ha sido necesario simplificar el elemento de búsqueda en una palabra, puesto que la mayoría de los factores se definen con un solo término. En la *Tabla 4-4. Lista de factores de fallos de la distribución de agua* se muestran las palabras buscadas.

Palabra	Traducción
material	material
diameter	diámetro
age	edad
installation	instalación
pressure	presión
protection	protección
traffic	tráfico
temperature	temperatura
freezing	congelación
corrosion	corrosión
corrosivity	corrosividad

Tabla 4-4. Lista de factores de fallos de la distribución de agua

5 ANÁLISIS DE RESULTADOS

A lo largo de este capítulo se recogen los principales resultados obtenidos. En la primera subsección, se realizará un breve esquema de las palabras más repetidas en las palabras claves y en los textos completos de los 102 documentos.

Posteriormente, se realiza el estudio de la frecuencia de los términos descritos en el capítulo anterior en las palabras claves y los textos completos de todos los artículos. Además, se han incluido diagramas de dispersión para determinar la correlación entre algunos de los términos más importantes.

5.1 Resumen

A la hora de realizar búsquedas en internet de artículos que traten el tema elegido, se ha de tener en cuenta que, en numerosas ocasiones, se puede encontrar documentos que mencionen alguna de las palabras buscadas, pero el tema del artículo no sea el que se está trabajando. Es por ello por lo que se han analizado las palabras más frecuentes, tanto en las palabras claves como en el texto completo, de los 102 documentos. Este análisis se ha llevado a cabo con una nube de palabras, una representación gráfica donde las palabras más empleadas están escritas con un tamaño de letra mucho mayor que las palabras que apenas aparecen.

Como se puede observar en la imagen de la derecha de la *Ilustración 5-1. Nube de palabras de las palabras claves o keywords (izquierda) y de los textos completos (derecha)* en las palabras claves, algunos de los términos más recurrentes son: "*water*", "*network*", "*learning*" y "*machine*". Las dos primeras palabras son agua y red, y se puede deducir que las dos últimas corresponden a "*machine learning*", aprendizaje automático en español.

En la parte derecha de la ilustración, se muestra que los términos más repetidos son "*model*", "*water*", "*datum*", "*pipe*" y "failure". En español, los términos son modelo, agua, dato, tubería y fallo. Se puede apreciar que en esta ilustración aparecen mayor número de palabras debido a que se está estudiando un texto completo y no solo las palabras claves como ocurre en el otro caso.

Aun así, se puede observar que la mayoría de las palabras más frecuentes de *keywords* aparecen también en el texto completo, dando por conclusión que efectivamente las palabras claves son resúmenes correspondientes

de los artículos completos.

Ilustración 5-1. Nube de palabras de las palabras claves o keywords (izquierda) y de los textos completos (derecha)

5.2 Lista de palabras

Dentro del alcance del trabajo, se han recopilado los resultados de analizar los 102 artículos para las listas de las palabras, previamente definidas.

El análisis de los modelos se ha ejecutado de dos maneras distintas, aunque complementarias.

Por una parte, se han analizado las apariciones de las palabras de la lista modelos entre las palabras claves del artículo (*keywords*). Por otra parte, se ha realizado el mismo estudio sobre el artículo completo. De esta manera y para cada término de la lista, se han obtenido dos resultados:

- Frecuencia de aparición de cada palabra de la lista en las palabras claves del artículo.
- Frecuencia de aparición de cada palabra en todo el texto del artículo.

Para la lista de palabras referentes a la evaluación de datos, el tratamiento de datos, el análisis de errores y la distribución de agua, se ha recurrido al estudio completo del texto, pues apenas aparecen en las palabras claves.

5.2.1 Modelos de machine learning

De cara a facilitar el análisis de los resultados, se ha preparado una tabla resumen, *Tabla 5-1. Frecuencia de aparición de la lista de modelos*, que recoge la frecuencia de aparición de cada una de las palabras de la lista, tanto entre las palabras claves de los artículos como en el texto completo del artículo.

Frecuencia de aparición	**Palabras claves**			**Texto completo**				
	0	1	2	0	1-10	11-20	21-50	>50
logistic	99	3		75	21	4	2	
tree	98	4		41	38	12	7	4
forest	91	10	1	58	34	5	4	1
neural	91	10	1	35	49	12	6	
bayesian	98	4		55	42	2	2	1
vector	101	1		38	57	5	2	
regression	90	11	1	34	40	14	10	4
classification	101	1		59	31	9	3	

Tabla 5-1. Frecuencia de aparición de la lista de modelos

Las principales conclusiones que se pueden extraer son:

- Las palabras más repetidas de los modelos son "*neural*", "*vector*" y "*tree*", aunque hay que tomarlo con precaución puesto que la palabra "*vector*" puede emplearse a la hora de definir los datos de un modelo.
- Por palabras claves, la conclusión sería que los modelos más utilizados son los que contienen las palabras "*neural*" y "*forest*". Este resultado difiere con el texto completo, ya que en el modelo "*random forest*" está basado en árboles de decisiones, "*tree decision*", y es por ello por lo que en los artículos donde se utilicen bosques aleatorios se menciona recursivamente el término *"tree"*
- En general, se observa que hay una desproporción significativa entre la frecuencia de "*regression*" y "*classification*". Aunque, hay que tomar este dato con precaución debido a que el modelo de regresión logística incluye la palabra regresión y, sin embargo, es un modelo de clasificación.

Seguidamente se recogen los resultados en detalle para cada uno de los términos de la lista modelos. El resultado del estudio se representa en formato gráfico. En el eje de abscisas se indica el máximo del número de veces que se repite la palabra en un artículo (frecuencia) y el de ordenadas el número de artículos. Se debe tener en cuenta que el rango de los ejes difiere en los distintos histogramas.

Así, por ejemplo, para la palabra "*logistic*", representado en el *Gráfico 5-1. Histogramas de la palabra "logistic" en las palabras claves o keywords (izquierda) y los textos completos(derecha)*, referido al modelo "*logistic regression*", el resultado es el siguiente:

- El gráfico de la izquierda, indica que la palabra "*logistic*" no aparece entre las palabras claves de 99 artículos, pero si se manifiesta una vez en 3 documentos.
- El gráfico de la derecha indica las veces que la palabra "*logistic*" aparece en el texto completo del número de artículos que se indica en el eje de ordenadas. Resalta el valor de que solo en 75 documentos no se incluye la palabra "*logistic*"; es decir, 27 de los artículos analizados incluyen este término ya sea porque usan el modelo de regresión logística o bien porque lo mencionan en algún momento.

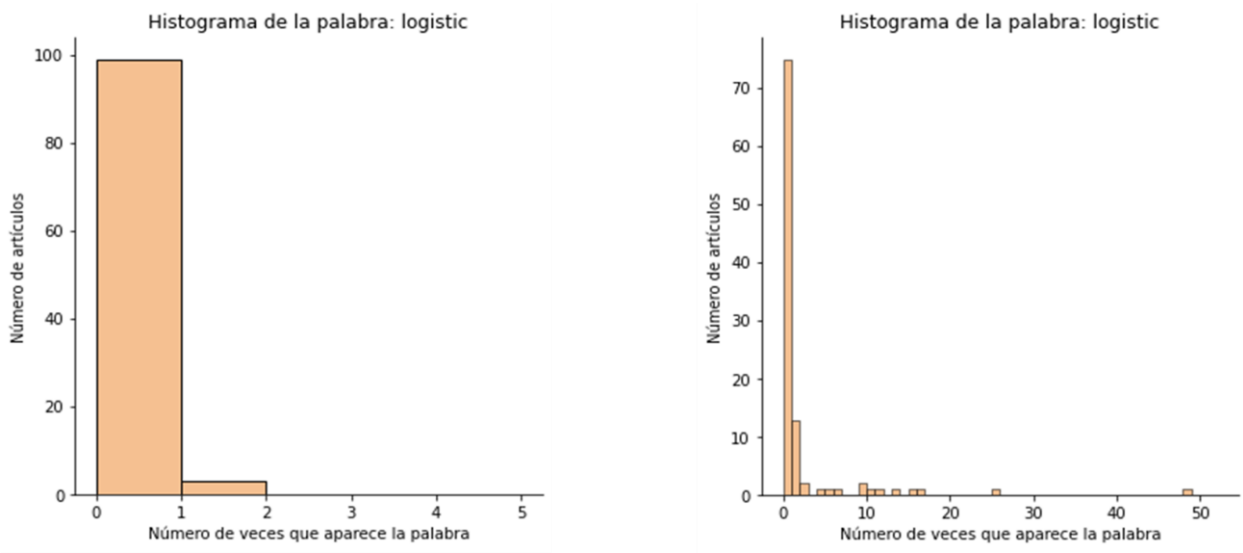

Gráfico 5-1. Histogramas de la palabra "logistic" en las palabras claves o keywords (izquierda) y los textos completos(derecha)

A continuación, se incluyen los demás histogramas para la lista de modelos, con la misma distribución, a la izquierda el diagrama para las palabras claves y a la derecha la frecuencia de la aparición de las palabras en el texto completo del artículo: *Gráfico 5-2. Histogramas de la palabra "tree" en las palabras claves o keywords (izquierda) y los textos completos (derecha)*, *Gráfico 5-3. Histogramas de la palabra "forest" en las palabras claves o keywords (izquierda) y los textos completos (derecha)*, *Gráfico 5-4. Histogramas de la palabra "neural" en las palabras claves o keywords (izquierda) y los textos completos (derecha)*, *Gráfico 5-5. Histogramas de la palabra "bayesian" en las palabras claves o keywords (izquierda) y los textos completos (derecha)* y *Gráfico 5-6. Histogramas de la palabra "vector" en las palabras clave o keywords (izquierda) y los textos completos (derecha).*

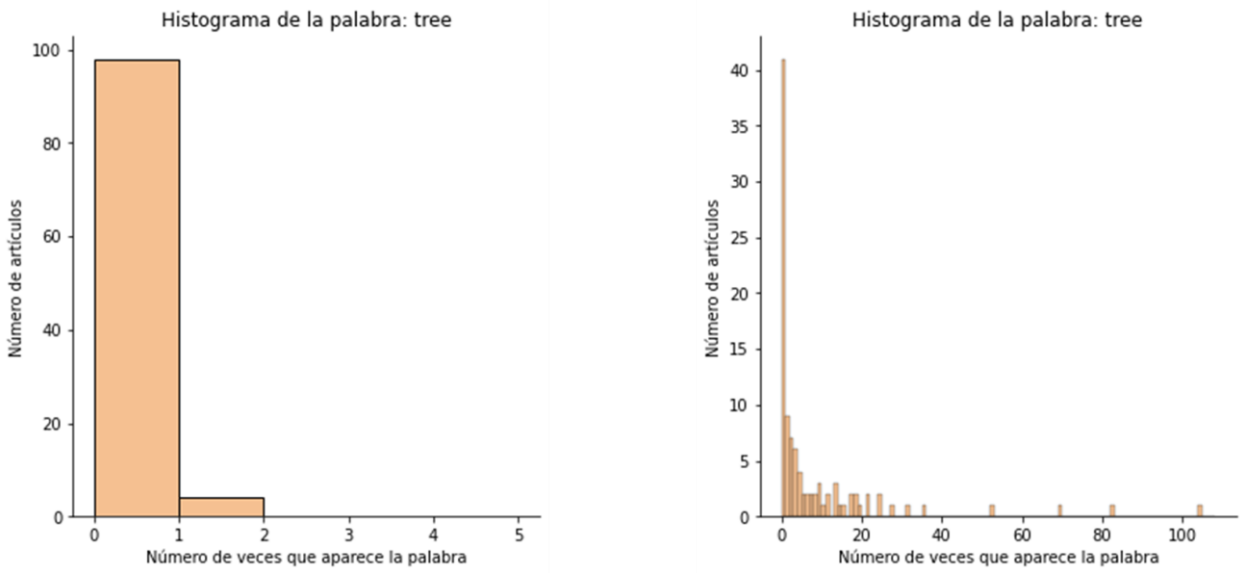

Gráfico 5-2. Histogramas de la palabra "tree" en las palabras claves o keywords (izquierda) y los textos completos (derecha)

Del histograma anterior, hay que resaltar que "*tree*" aparece más de 80 veces en dos artículos ("A brief review of random forests for water scientists and practitioners and their recent history in water resources" [65] y "Failure detection methods for pipeline networks: From acoustic sensing to cyber-physical systems†" [97]), aunque hay que tener en cuenta que en numerosas ocasiones "*forest*" incluye también el término "*tree*". Para identificar cuáles son esos documentos se ha utilizado el algoritmo explicado en el apartado 5.3.

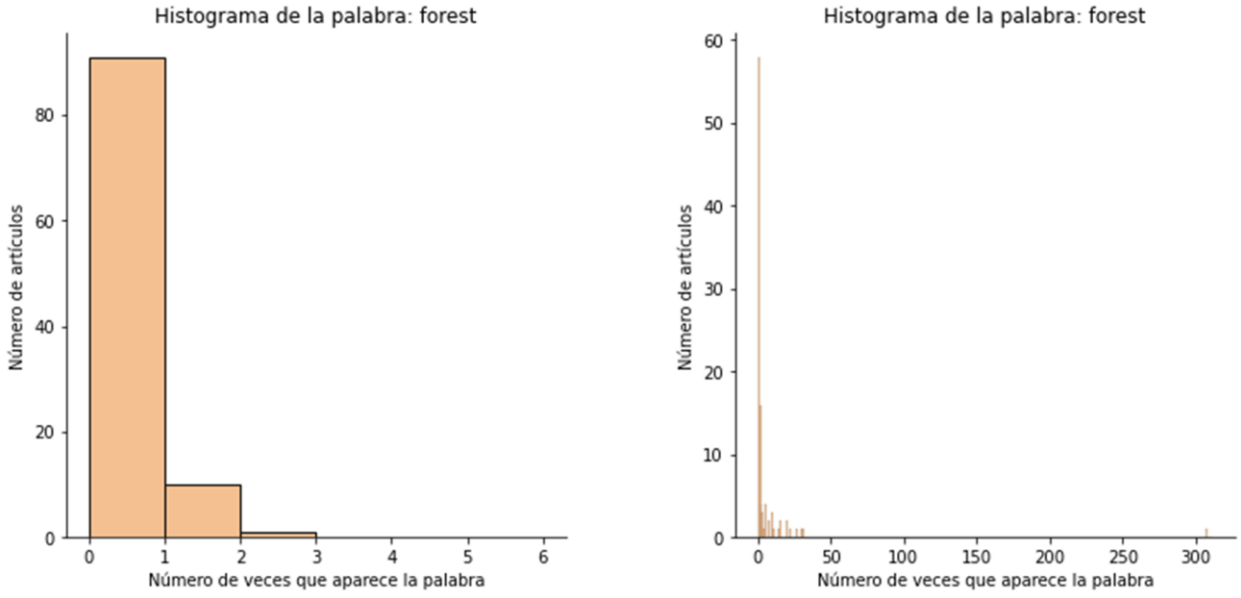

Gráfico 5-3. Histogramas de la palabra "forest" en las palabras claves o keywords (izquierda) y los textos completos (derecha)

En el *Gráfico 5-3. Histogramas de la palabra "forest" en las palabras claves o keywords (izquierda) y los textos completos (derecha)* hay que destacar que hay un artículo ("A brief review of random forests for water scientists and practitioners and their recent history in water resources" [65]) en el que "*forest*" aparece mencionado más de 300 veces.

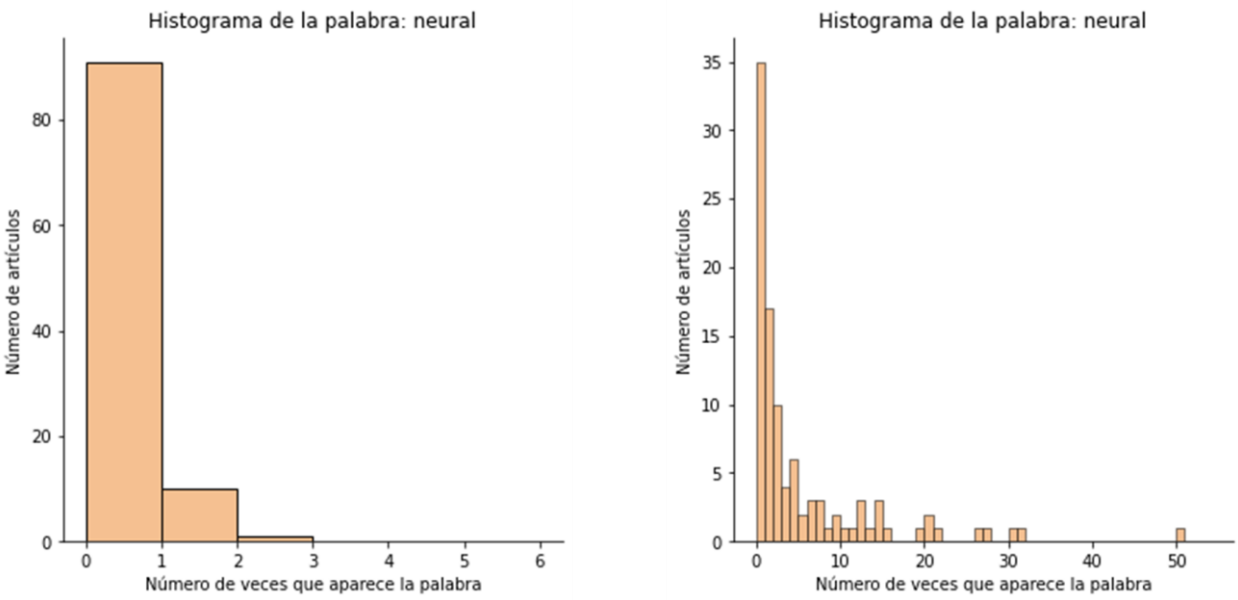

Gráfico 5-4. Histogramas de la palabra "neural" en las palabras claves o keywords (izquierda) y los textos completos (derecha)

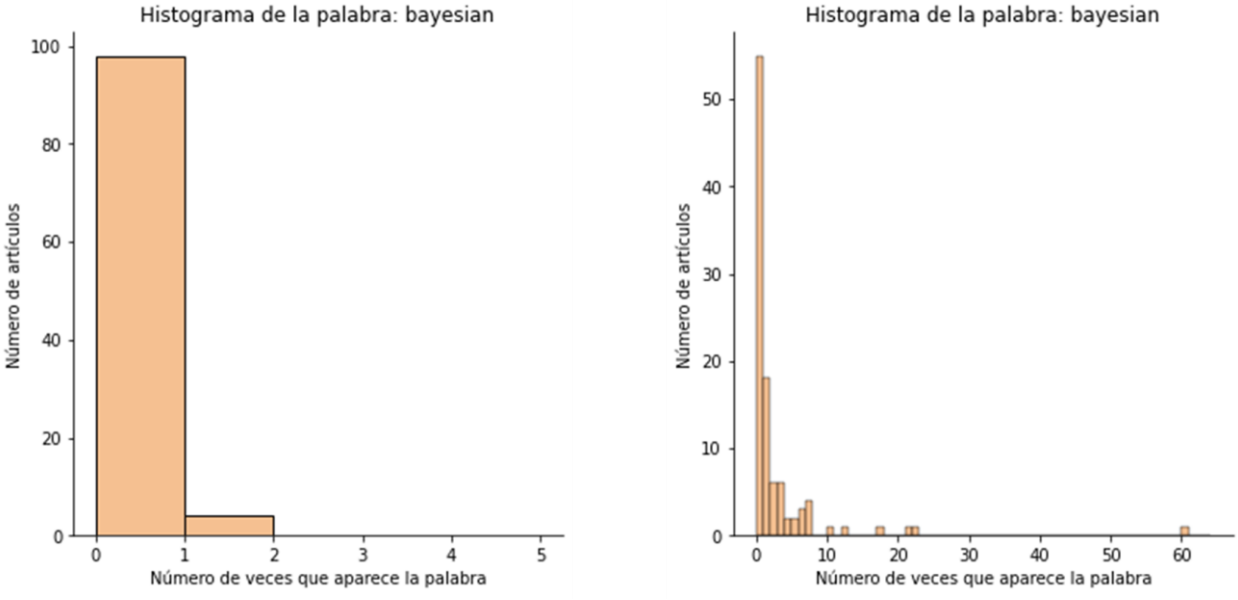

Gráfico 5-5. Histogramas de la palabra "bayesian" en las palabras claves o keywords (izquierda) y los textos completos (derecha)

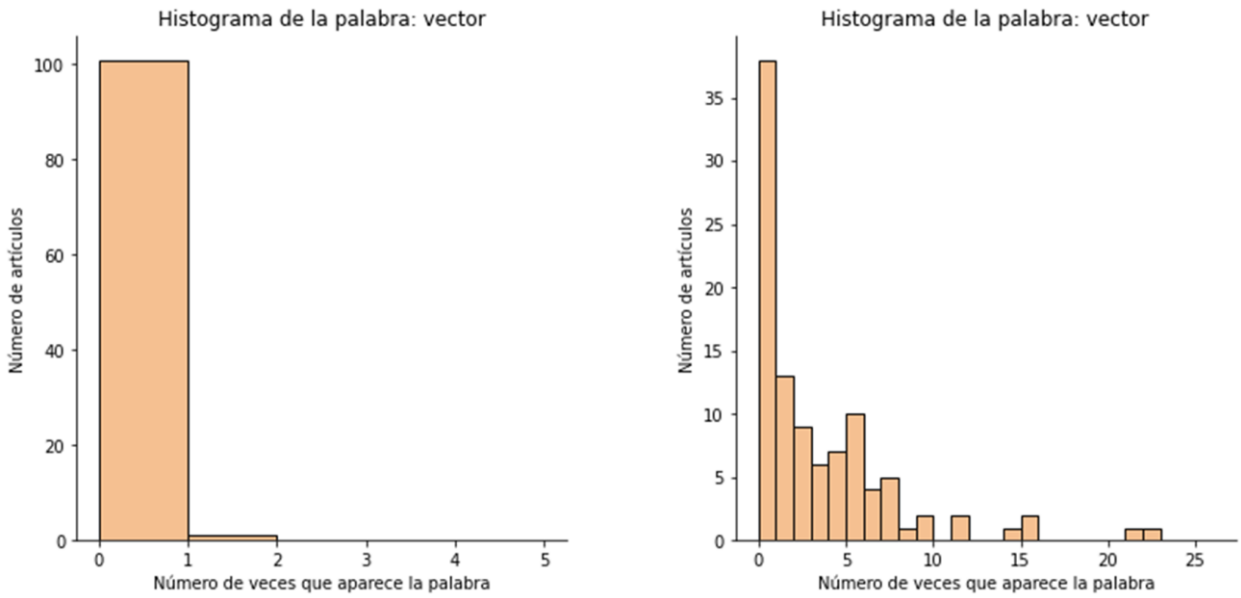

Gráfico 5-6. Histogramas de la palabra "vector" en las palabras clave o keywords (izquierda) y los textos completos (derecha)

Aunque se indicó anteriormente, resaltar que "*vector*" es el término que presenta menos apariciones. Obsérvese que su escala de "número de veces que aparece la palabra" es con diferencia la más pequeña. Incluso en las palabras clave, sólo aparece en una ocasión (gráfico de la izquierda), mientras que hay dos artículos que lo mencionan más de 25 veces.

Como se estudió en la subsección de *machine learning*, los modelos se pueden clasificar según el conjunto de datos, por lo que también se han incluido los histogramas para esta clasificación: *Gráfico 5-7. Histogramas de la palabra "regression" en las palabras clave o keywords (izquierda) y los textos completos (derecha)* y *Gráfico 5-8. Histogramas de la palabra "classification" en las palabras claves o keywords (izquierda) y los textos completos (derecha)*.

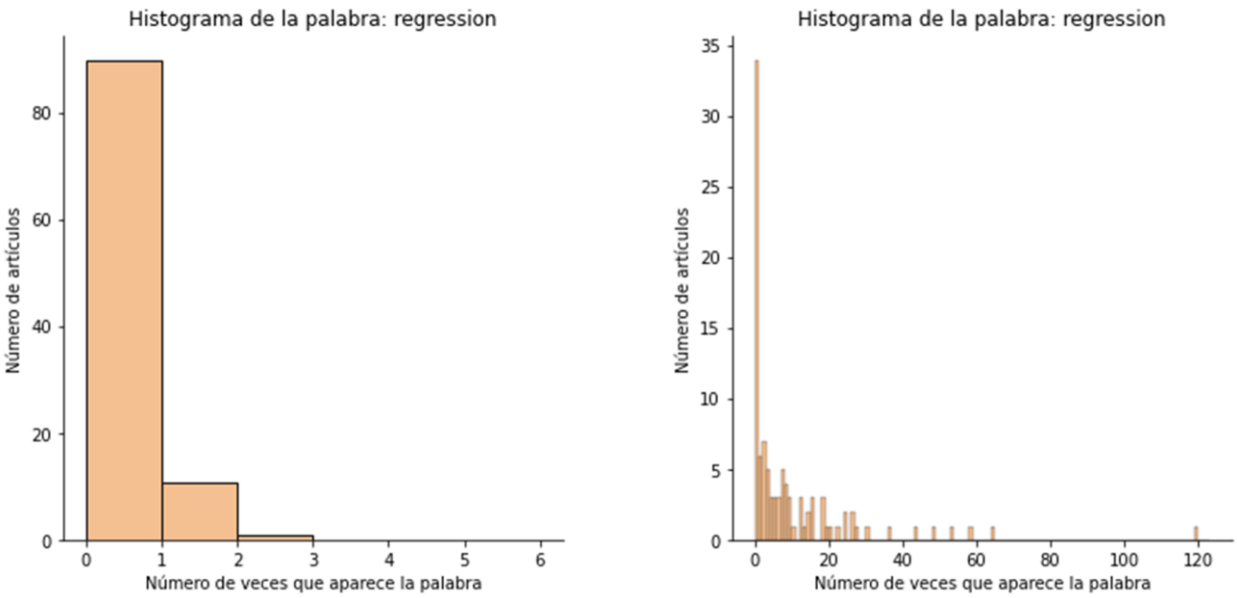

Gráfico 5-7. Histogramas de la palabra "regression" en las palabras clave o keywords (izquierda) y los textos completos (derecha)

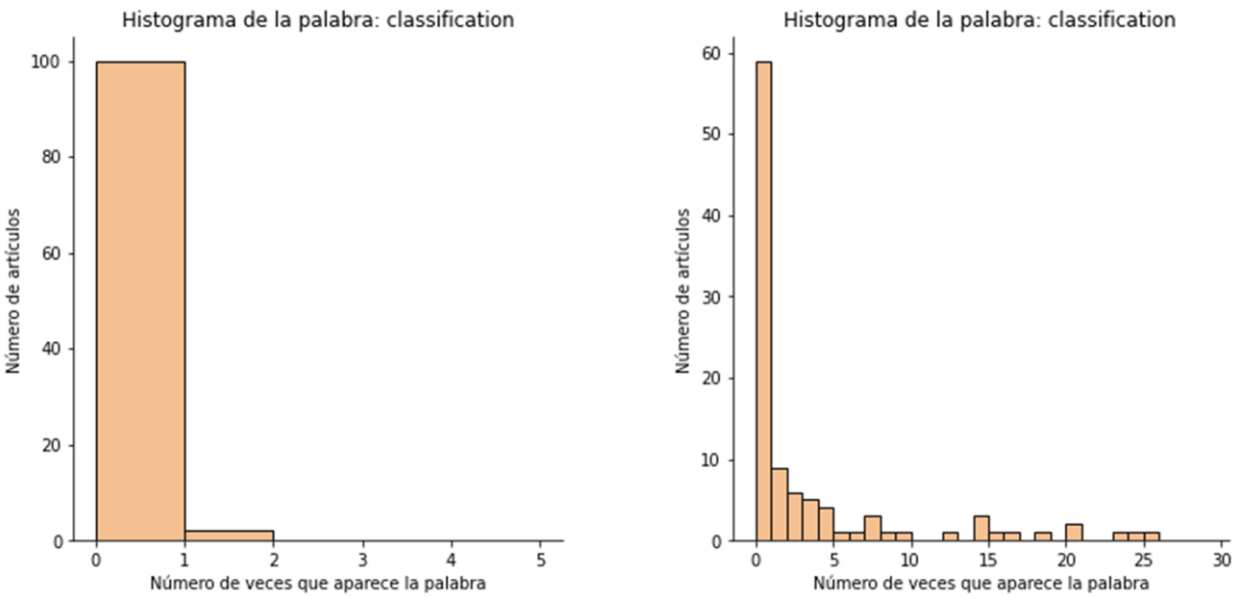

Gráfico 5-8. Histogramas de la palabra "classification" en las palabras claves o keywords (izquierda) y los textos completos (derecha)

5.2.2 Evaluación de datos

En relación a la evaluación de datos, según las frecuencias de aparición en los textos completos, se muestra en la *Tabla 5-2. Frecuencia de aparición de los términos de la evaluación de datos* y el *Gráfico 5-9. Histogramas de las palabras "missing" y "outlier" de la evaluación de datos* podemos obtener dos conclusiones

básicamente:

- Los datos incompletos tienen una frecuencia de aparición mucho mayor que los datos atípicos (un 33% frente a un 22%).
- Es necesario considerar que el término "*missing*" puede aparecer debido a ser una palabra muy utilizada en otros contextos, lo cual puede falsear la conclusión descrita en el punto anterior.

Por ello, y en relación a la evaluación de datos se entiende que no se puede extraer una conclusión precisa con respecto a ninguna de las dos tipologías.

Frecuencia de aparición	**Texto completo**				
	0	1-5	6-10	11-15	>15
missing	69	28	4		1
outlier	80	20	1		1

Tabla 5-2. Frecuencia de aparición de los términos de la evaluación de datos

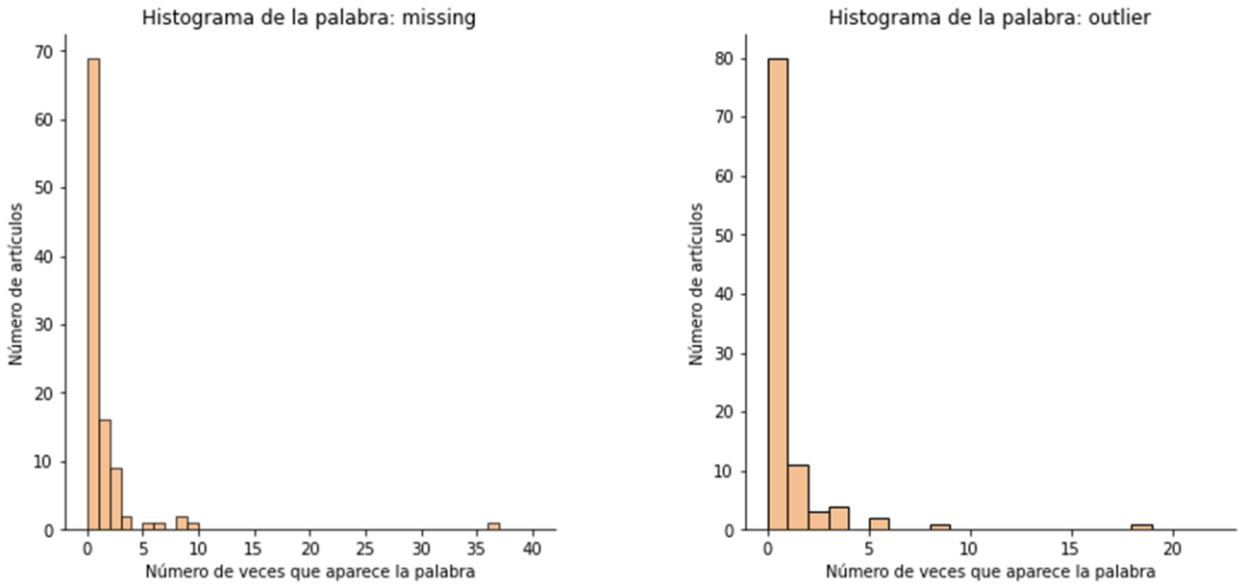

Gráfico 5-9. Histogramas de las palabras "missing" y "outlier" de la evaluación de datos

5.2.3 Tratamiento de datos

En relación con las tipologías de tratamiento de datos, hay claramente mayor aparición de las técnicas de *sampling* y *crossvalidation* (39% y 26%) frente al resto (inferiores al 10%). Más aún cuando las otras dos son subtipos de *sampling*. Esta información se resume en la *Tabla 5-3. Frecuencia de aparición de los términos del tratamiento de datos* y en los gráficos: *Gráfico 5-10. Histogramas de las palabras "sampling" y*

"crossvalidation" del tratamiento de datos y *Gráfico 5-11. Histogramas de las palabras "oversampling" y "undersampling" del tratamiento de datos.*

Frecuencia de aparición	**Texto completo**				
	0	1-5	6-10	11-15	>15
sampling	62	34	2	1	3
oversampling	93	7		1	1
undersampling	94	5	2		1
crossvalidation	75	25	1	1	

Tabla 5-3. Frecuencia de aparición de los términos del tratamiento de datos

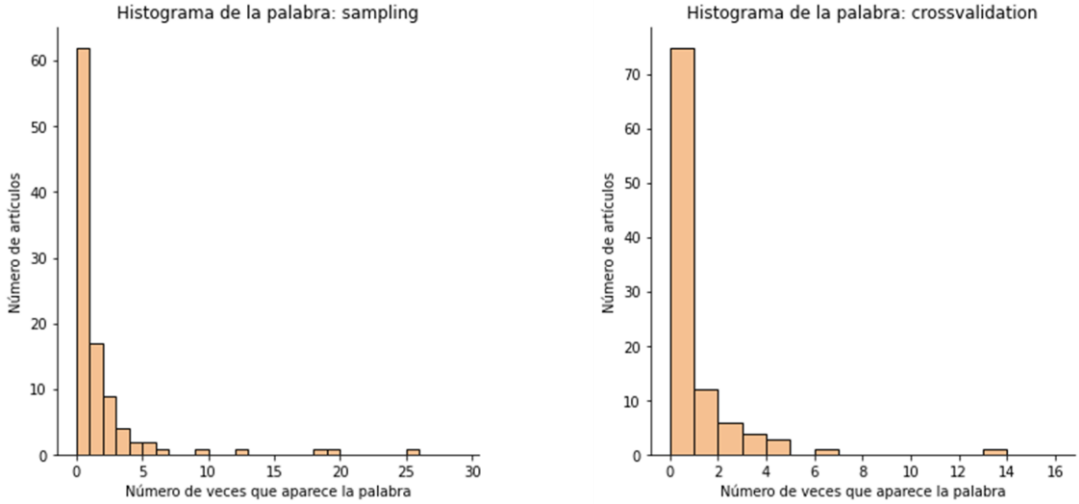

Gráfico 5-10. Histogramas de las palabras "sampling" y "crossvalidation" del tratamiento de datos

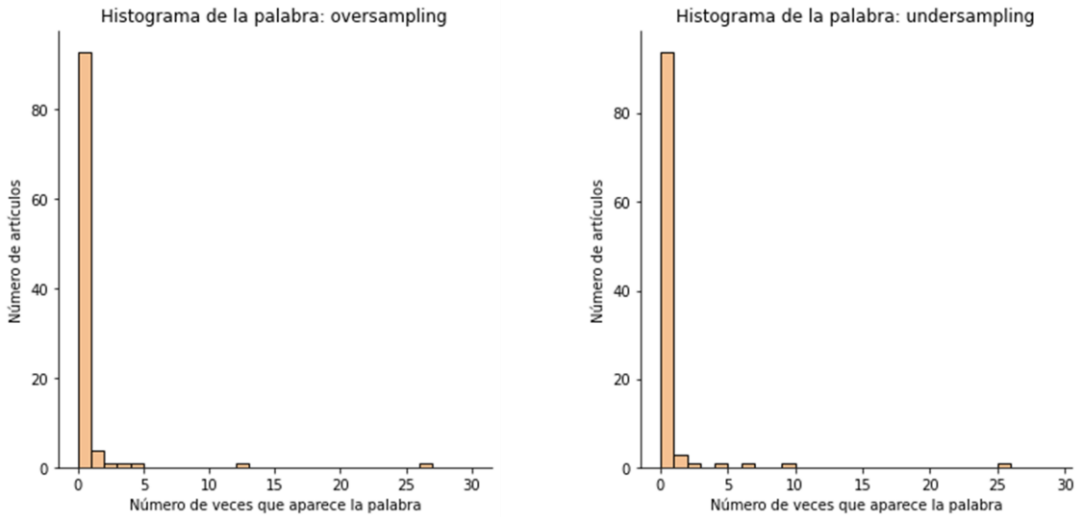

Gráfico 5-11. Histogramas de las palabras "oversampling" y "undersampling" del tratamiento de datos

5.2.4 Análisis de errores

Con respecto a los textos completos, son los métodos: matriz de confusión y RMSE los que más se mencionan. RMSE se utiliza más que RSE por ser este mucho más preciso. En cuanto a ROC, hay que tener en cuenta que se basa a su vez en el método de matriz de confusión, por lo que es muy probable que cuando se menciona ROC se menciona "*confusion*" a su vez. Estas conclusiones se resumen en la *Tabla 5-4. Frecuencia de aparición del análisis de errores* y en los diferentes gráficos: *Gráfico 5-12. Histogramas de las palabras "mse" y "rmse" del análisis de errores* y *Gráfico 5-13. Histogramas de las palabras "roc" y "confusion" del análisis de errores.*

Frecuencia de aparición	**Texto completo**				
	0	1-10	11-20	21-50	>50
mse	88	12	1	1	
rmse	80	13	6	2	1
roc	83	14	5		
confusion	79	19	2	2	

Tabla 5-4. Frecuencia de aparición del análisis de errores

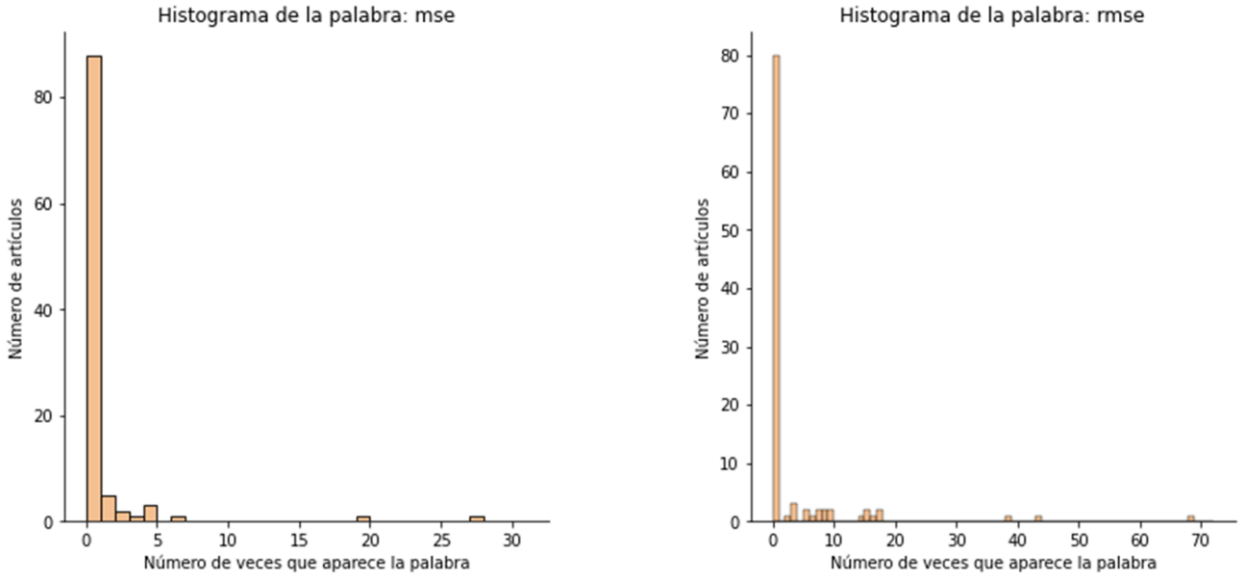

Gráfico 5-12. Histogramas de las palabras "mse" y "rmse" del análisis de errores

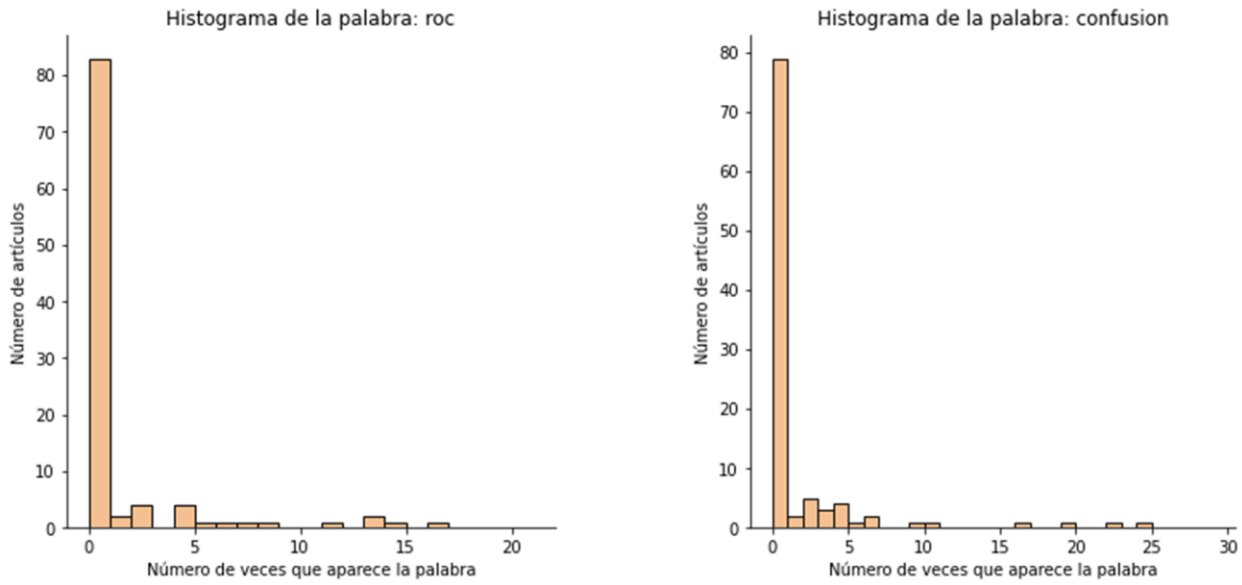

Gráfico 5-13. Histogramas de las palabras "roc" y "confusion" del análisis de errores

5.2.5 Distribución de agua

De las gráficas de frecuencia de aparición de los términos de las redes de distribución de agua en el texto completo (*Gráfico 5-14. Histogramas de las palabras "corrosivity" y "freezing"*, *Gráfico 5-15. Histogramas de las palabras "traffic" y "protection"*, *Gráfico 5-16. Histogramas de las palabras "corrosion" y "installation"*, *Gráfico 5-17. Histogramas de las palabras "temperature" y "diameter"*, *Gráfico 5-18. Histogramas de las palabras "age" y "pressure"* y *Gráfico 5-19. Histograma de la palabra "material"*) podemos deducir fácilmente que hay cinco factores claros de impacto en la distribución; material, presión, edad, diámetro y temperatura, como se muestra en la *Tabla 5-5. Frecuencia de aparición de los elementos de la distribución de agua*.

Frecuencia de aparición	**Texto completo**					**Porcentaje de artículos que mencionan los términos al menos una vez**
	0	1-10	11-20	21-50	>50	
corrosivity	96	6				6%
freezing	94	7	1			8%
traffic	85	17				17%
protection	79	21		2		23%
corrosion	77	19	1	2	3	25%
installation	60	37	5			41%
temperature	52	39	2	7	2	49%
diameter	39	43	15	4	1	62%
age	38	49	9	5	1	63%
pressure	33	42	14	8	5	68%
material	24	53	13	10	2	76%

Tabla 5-5. Frecuencia de aparición de los elementos de la distribución de agua

El material es significativamente más mencionado que los otros tres, siendo la edad y el diámetro de una frecuencia de aparición similar.

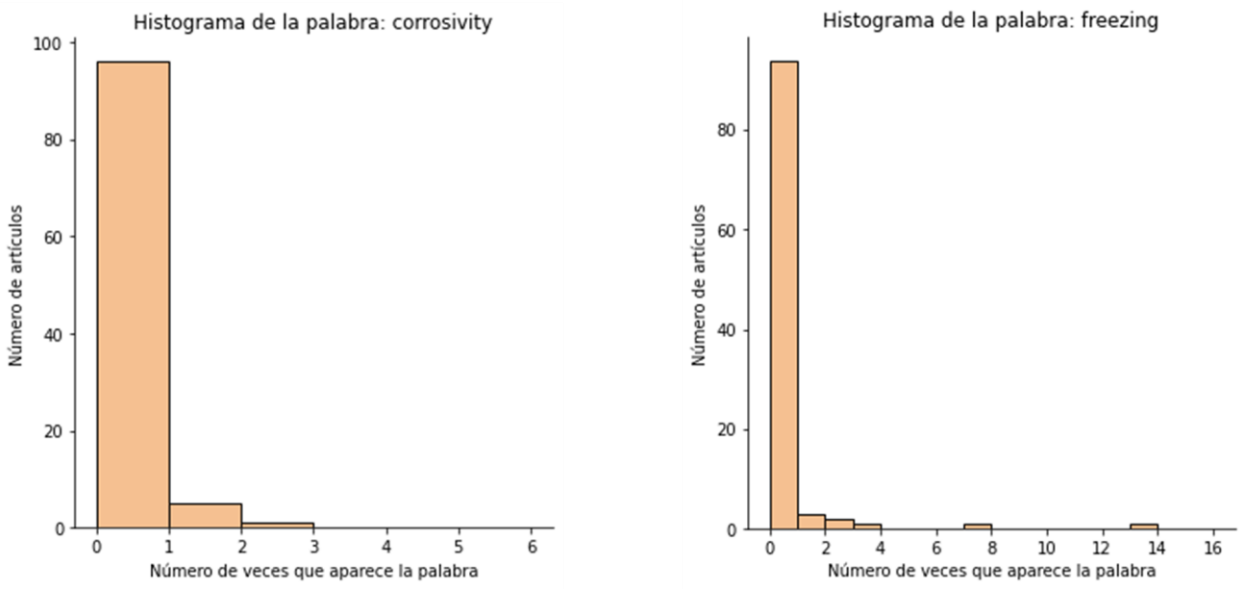

Gráfico 5-14. Histogramas de las palabras "corrosivity" y "freezing"

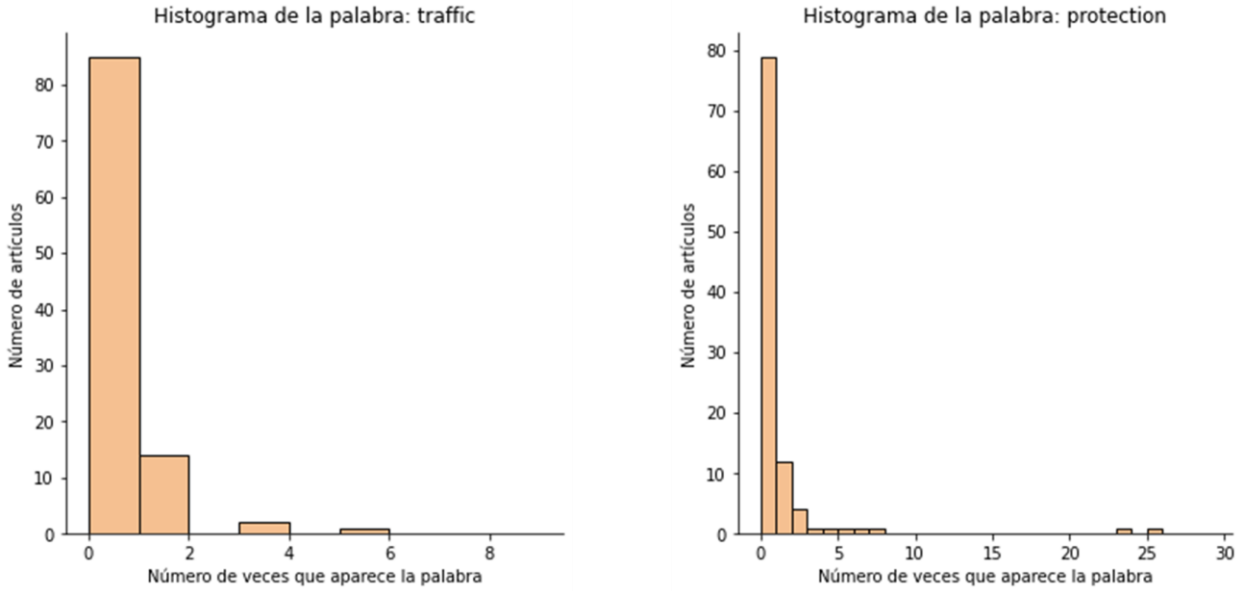

Gráfico 5-15. Histogramas de las palabras "traffic" y "protection"

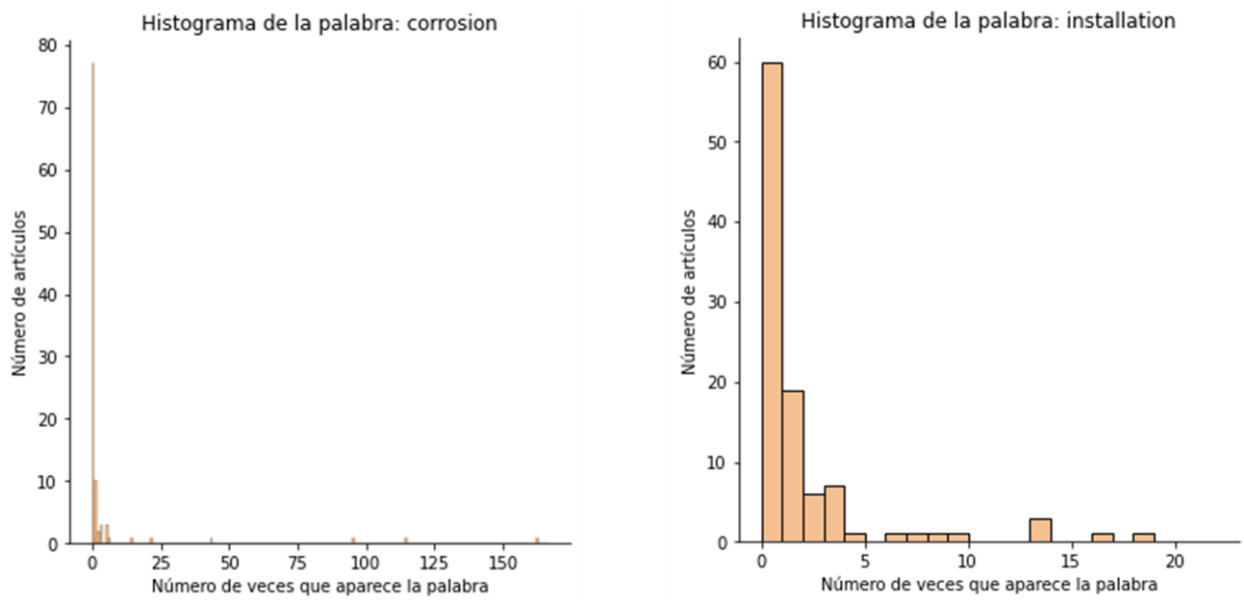

Gráfico 5-16. Histogramas de las palabras "corrosion" y "installation"

Hay que destacar cómo el término "*corrosion*" se presenta más de 80 veces en al menos tres artículos, cuando no es uno de los factores que más se estudia. No obstante, la alta frecuencia de aparición en estos tres artículos, algo que solo ocurre además con "*pressure*", indica que aunque no es de los términos más estudiados, pertenece al segundo grupo de factores a tener en cuenta.

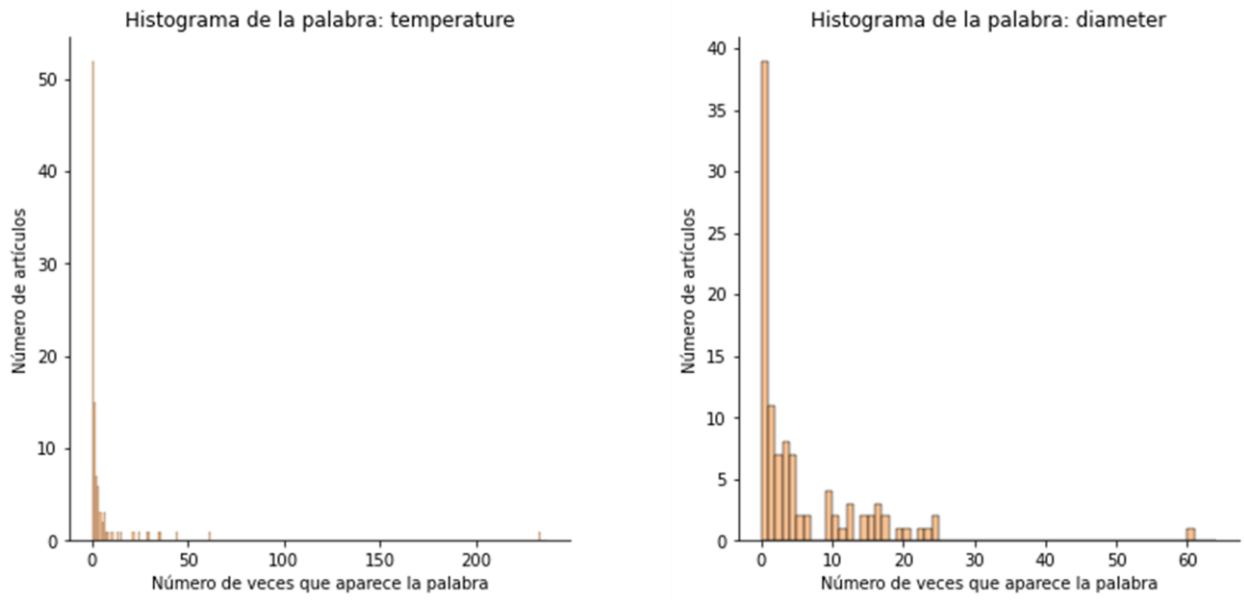

Gráfico 5-17. Histogramas de las palabras "temperature" y "diameter"

Como se aprecia en el gráfico anterior, "*temperature*" es citada más de 200 veces en un artículo ("Regression

models utilization to the underground temperature determination at coal energy conversion" [115]).

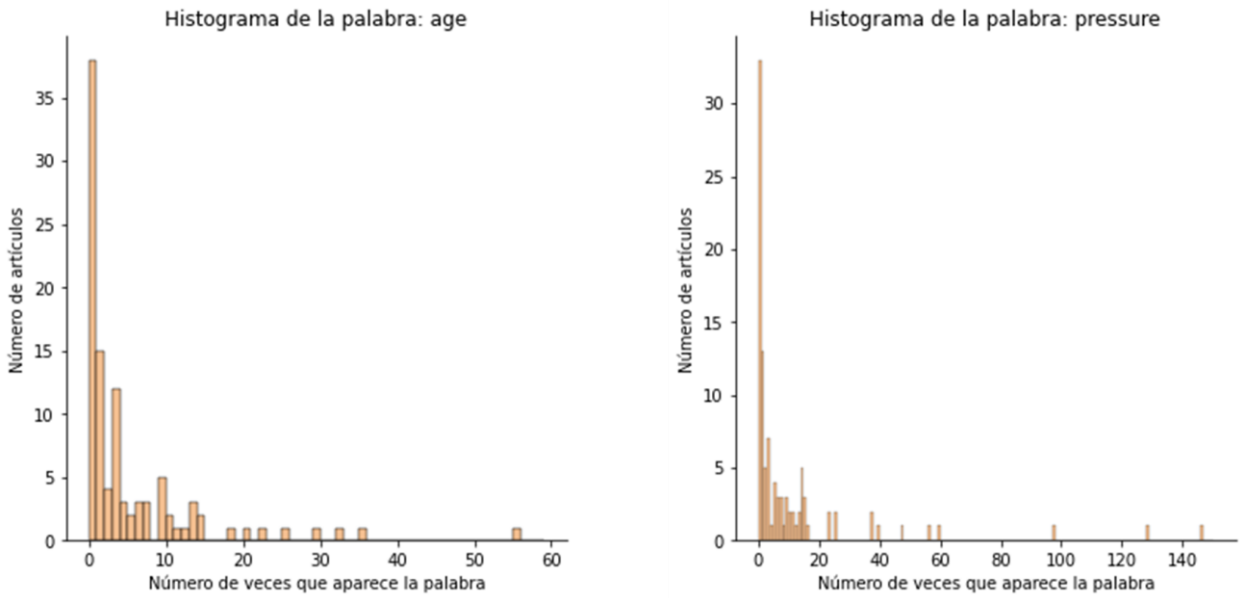

Gráfico 5-18. Histogramas de las palabras "age" y "pressure"

Como se indicó anteriormente, "*pressure*", que es de los factores más importantes por su frecuencia de aparición, es además mencionado más de 80 veces en al menos tres artículos. Obviamente, la presión es de los factores más analizados en la literatura científica.

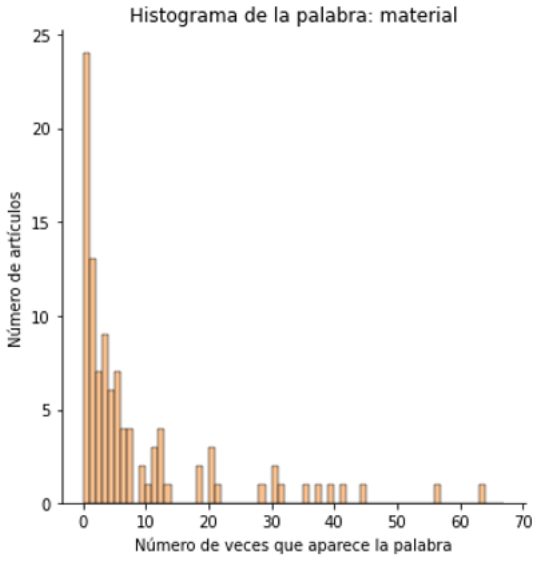

Gráfico 5-19. Histograma de la palabra "material"

5.3 Diagramas de dispersión

Para determinar las correlaciones entre los diferentes términos, se han realizado gráficos de dispersión, donde los ejes de abscisas y de ordenadas muestran la frecuencia de los dos términos que se quieren estudiar. También se ha incluido una escala, representada en distintos colores, que determinan el número de artículos que representan ese punto. Para ello, se han utilizado las frecuencias de aparición de los términos en los textos completos.

5.3.1 Sampling- classification y sampling-regression

Como puede verse comparando ambos diagramas, en el caso de "*regression"* existe un fuerte alineamiento de los puntos (con colores más intensos) hacia el eje de ordenadas, esto indica que la correlación entre "*sampling*" y "*regression*" es menor que entre "*sampling*" y "*classification*", por lo que se puede deducir que el método "*sampling*" se aplica con mayor frecuencia a los modelos de clasificación.

De hecho, se puede ver en *Gráfico 5-20. Diagramas de dispersión de "sampling" frente a "classification" y "regression"*, como los puntos de dispersión centrales del diagrama de la izquierda forman incluso una recta de 45 grados.

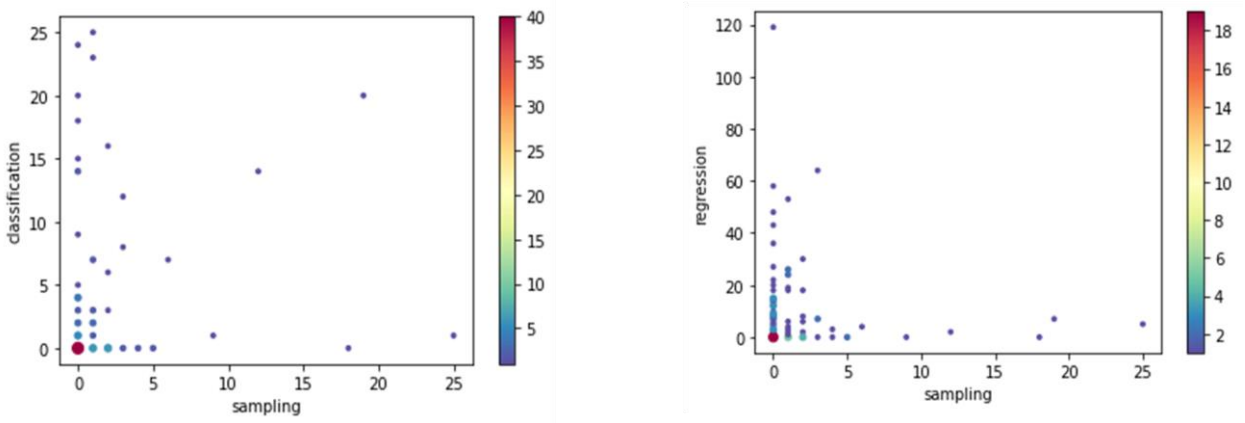

Gráfico 5-20. Diagramas de dispersión de "sampling" frente a "classification" y "regression"

5.3.2 Modelos

Existen numerosos algoritmos de aprendizaje automático que se pueden emplear para problemas de clasificación o de regresión, es por ello por lo que se ha realizado el estudio para determinar la correlación entre los diferentes algoritmos.

5.3.2.1 Tree

En este caso, se observa en *Gráfico 5-21. Diagramas de dispersión de "tree" frente a "classification" y "regression"* una mayor correlación de *regression* frente a *classification*. Para analizar los gráficos, téngase en cuenta que la escala en el caso de regression es bastante mayor (25 frente a 120).

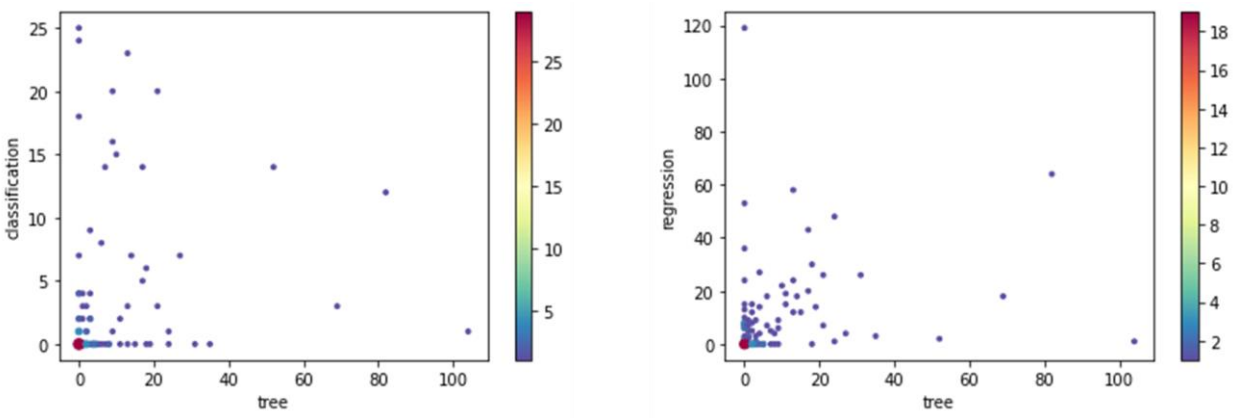

Gráfico 5-21. Diagramas de dispersión de "tree" frente a "classification" y "regression"

5.3.2.2 Forest

Los gráficos siguientes (*Gráfico 5-22. Diagramas de dispersión de "forest" frente a "classification" y "regression"*) muestran una mayor correlación entre *forest-regression* que entre *forest-classification*. A la hora de leer el gráfico, de nuevo es importante tener en cuenta que la escala de regression es mayor.

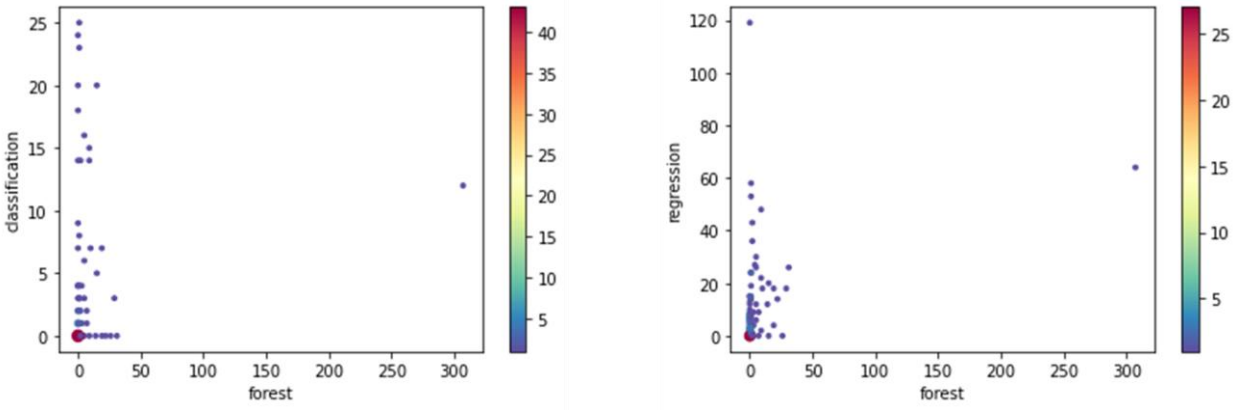

Gráfico 5-22. Diagramas de dispersión de "forest" frente a "classification" y "regression"

5.3.2.3 Neural

Con respecto a la utilización de *neural* frente a *classification* y *regression*, podemos ver cómo la mayor correlación entre ambos se presenta, según *Gráfico 5-23. Diagramas de dispersión de "neural" frente a "classification" y "regression"* ,en el caso de regression.

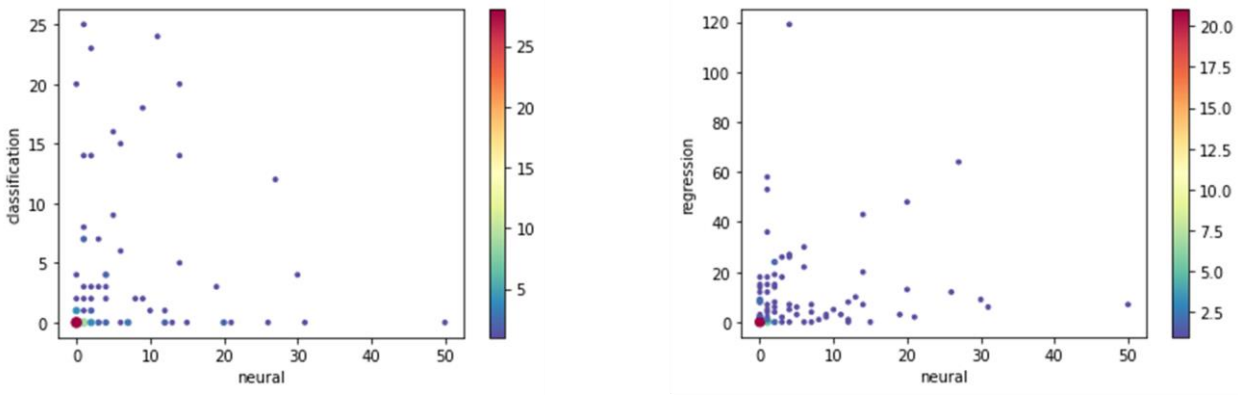

Gráfico 5-23. Diagramas de dispersión de "neural" frente a "classification" y "regression"

5.3.2.4 Bayesian

En relación a la posible correlación entre "*bayesian*" y "*classification-regression*", se puede observar en el *Gráfico 5-24. Diagramas de dispersión de "bayesian" frente a "classification" y "regression"* que en el diagrama de la izquierda no se da ninguna correlación. Sin embargo, puede apreciarse una cierta correlación entre "*bayesian*" y "*regression*", aunque no es totalmente significativa.

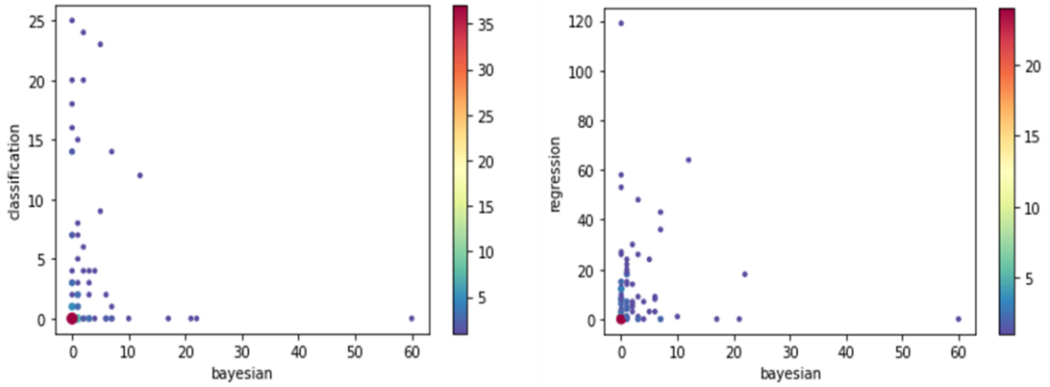

Gráfico 5-24. Diagramas de dispersión de "bayesian" frente a "classification" y "regression"

5.3.3 Factores de fallos en la red de distribución de agua

A continuación, se ha recogido en *Gráfico 5-25. Diagramas de dispersión de "material" frente a "pressure", "age", "diameter" y "temperature"*, *Gráfico 5-26. Diagramas de dispersión de "pressure" frente a "age", "diameter" y "temperature"*, *Gráfico 5-27. Diagramas de dispersión de "age" frente a "diameter" y "temperature"* y *Gráfico 5-28. Diagramas de dispersión de "diameter" frente a "temperature"* los diagramas de dispersión que correlacionan los siguientes factores entre ellos; material, presión, edad, diámetro y temperatura.

A modo de resumen, las conclusiones son las siguientes;

- El factor de temperatura no presenta correlación frente a los demás factores.
- No existe correlación entre presión y edad.
- Alta correlación entre material, edad y diámetro.
- No está claro que exista correlación entre material y presión y entre diámetro y presión.

La elección del material depende del diámetro de la tubería, lo que es coherente con los resultados obtenidos. Asimismo, la presión, por su fórmula matemática, depende del diámetro, lo cual justifica igualmente la correlación obtenida entre ambos.

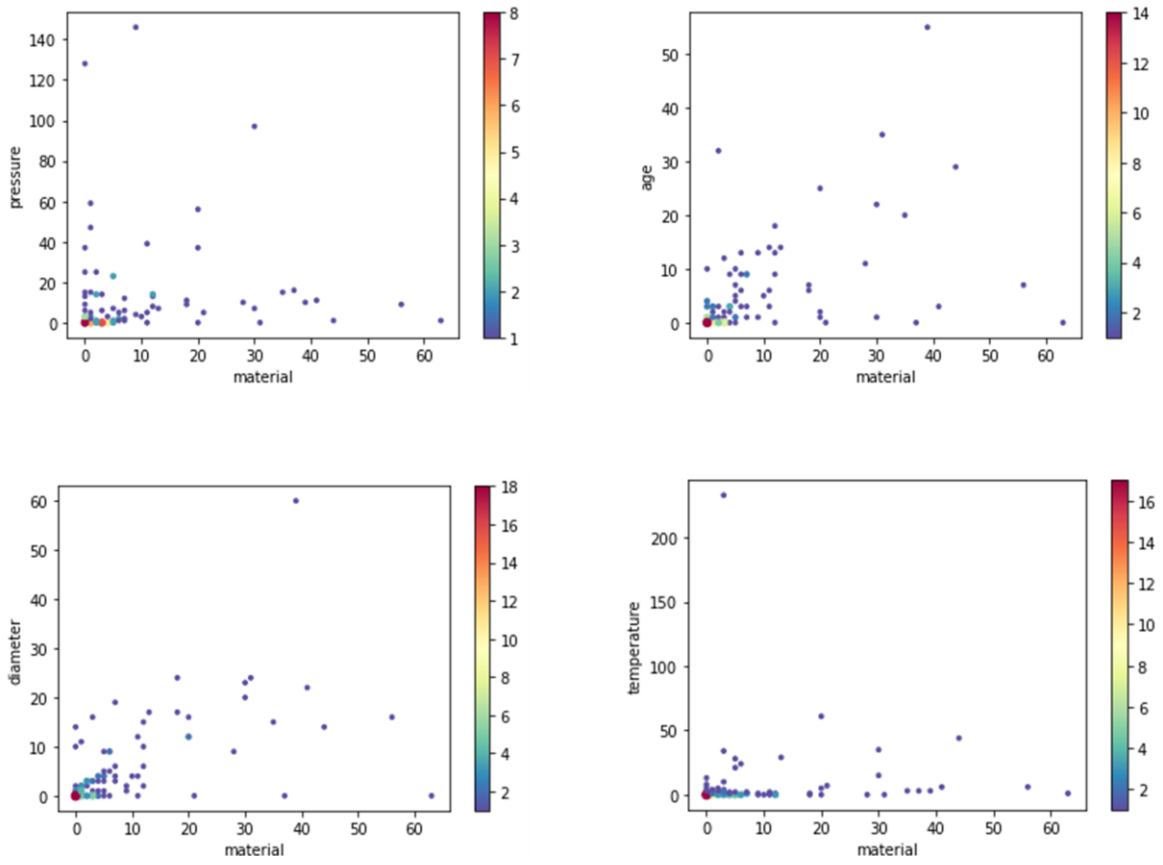

Gráfico 5-25. Diagramas de dispersión de "material" frente a "pressure", "age", "diameter" y "temperature"

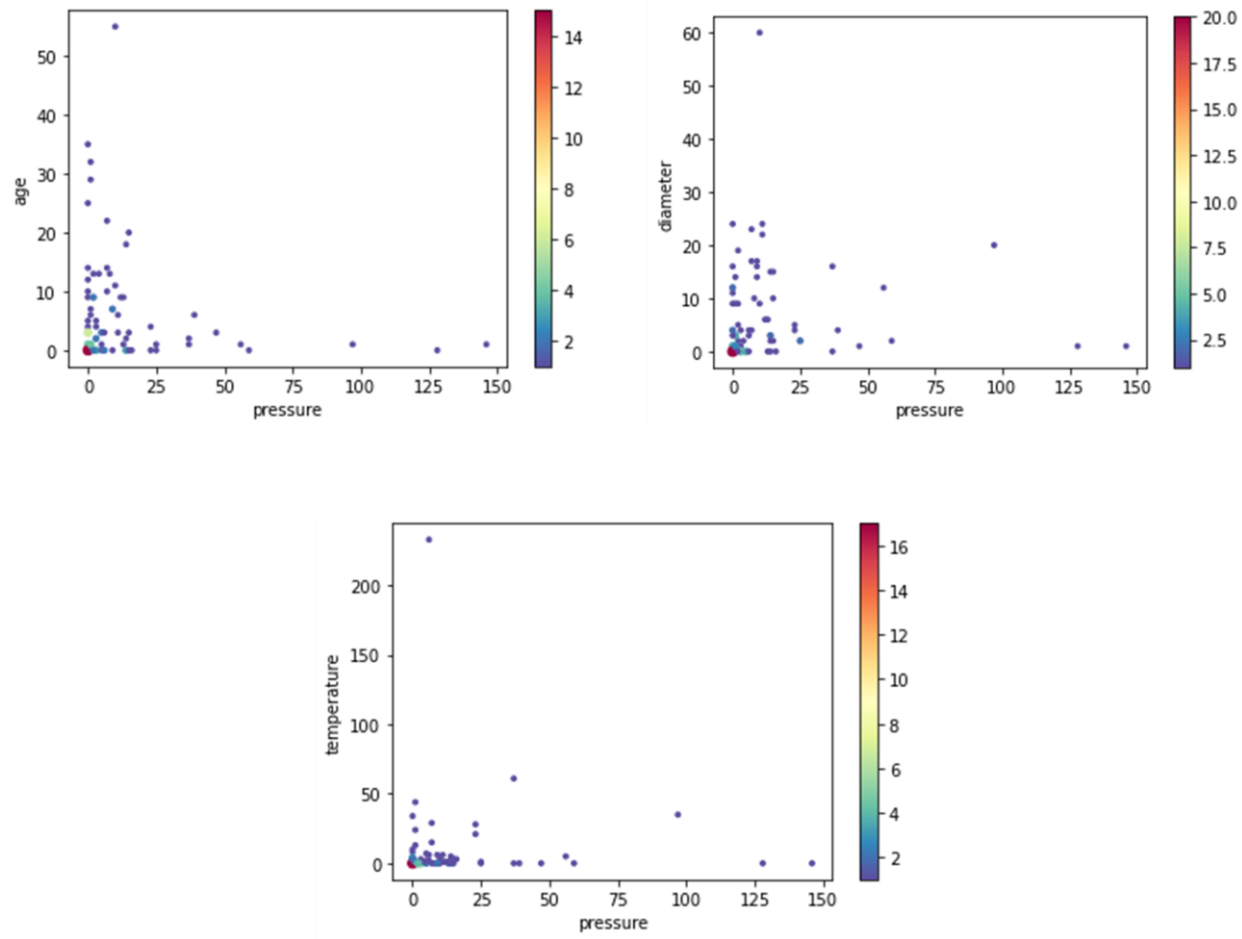

Gráfico 5-26. Diagramas de dispersión de "pressure" frente a "age", "diameter" y "temperature"

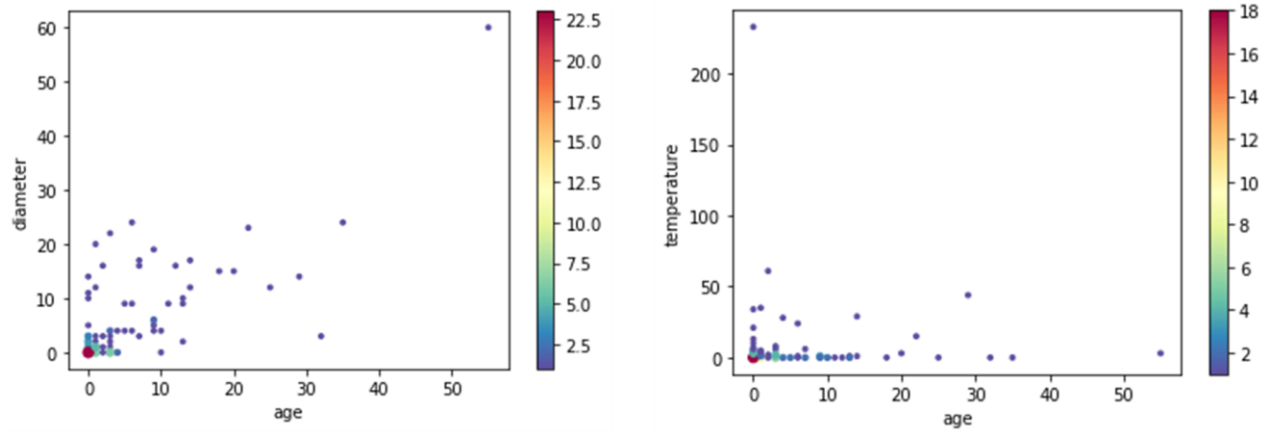

Gráfico 5-27. Diagramas de dispersión de "age" frente a "diameter" y "temperature"

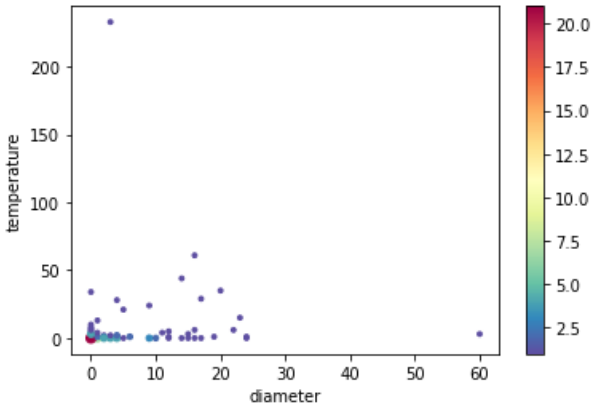

Gráfico 5-28. Diagramas de dispersión de "diameter" frente a "temperature"

5.4 Análisis de la búsqueda

Una vez determinada las correlaciones entre los diferentes términos, se ha considerado relevante determinar qué documentos son los que cumplen con una correlación alta. Esta relevancia se debe a que en numerosas ocasiones la búsqueda en internet de documentos que tratan diferentes temas no es muy eficiente. Cuando se buscan varias palabras en el navegador para encontrar un artículo que trate de esos temas, se pueden encontrar documentos que solos los mencionan pero que no sean el tema principal de los documentos.

Es por ello, por lo que se ha desarrollado un algoritmo para determinar los documentos que cumplan los criterios de búsqueda, es decir, traten de esos temas principalmente, y no se limiten a mencionarlos de forma tangencial. Con esta finalidad, se ha realizado la media ponderada de la frecuencia de aparición de los términos deseados de cada documento. Posteriormente, se ha representado los 20 documentos que cumplen que la aparición de ambos términos se realiza de forma simultánea.

A continuación, se muestra la *Tabla 5-6. Resultados resumidos del análisis de búsqueda*, donde se resumen todas las soluciones recogidas de las diferentes ilustraciones que se han obtenido: *Ilustración 5-2. Artículos que estudian los términos "classification" y "sampling"*, *Ilustración 5-3. Artículos que estudian los términos "regression" y "neural"*, *Ilustración 5-4. Artículos que estudian los términos "regression" y "tree"*, *Ilustración 5-5. Artículos que estudian los términos "regression" y "forest"*, *Ilustración 5-6. Artículos que estudian los términos "material" y "age"*, *Ilustración 5-7. Artículos que estudian los términos "material" y "diameter"* e *Ilustración 5-8. Artículos que estudian los términos "age" y "diameter"*.

Se va a utilizar el gráfico de correlación entre "*classification*" y "*sampling*" del *Gráfico 5-20. Diagramas de dispersión de "sampling" frente a "classification" y "regression"* para explicar los resultados de la tabla. En ese gráfico se puede observar que los términos aparecen más de 20 veces cada uno en un único documento. La *Ilustración 5-2. Artículos que estudian los términos "classification" y "sampling"* muestra que ese artículo es el "37.pdf", cuyo nombre real es "Deep learning approach for diabetes prediction using PIMA Indian dataset"

[81]. Los otros artículos que se representan en esta nube también tratan esos términos en menor medida.

De este ejemplo, se puede extraer la conclusión de que cuando realizamos la búsqueda de artículos, los documentos obtenidos no son siempre los deseados. En este trabajo se han tratado documentos enfocados al aprendizaje automático aplicado a la distribución de agua. Claramente del título de este artículo (en español es "Enfoque de aprendizaje profundo para la predicción de la diabetes utilizando el conjunto de datos indio PIMA") se puede deducir que el tema principal es el aprendizaje aplicado a la diabetes. Es por ello, por lo que se ha decidido realizar otra búsqueda empleando los términos '*classification'*, '*sampling'* y '*water'* con el algoritmo desarrollado para comprobar que efectivamente no estudie el tema del agua. Al llamar al algoritmo, se obtiene la *Ilustración 5-9. Artículos que estudian los términos "sampling", "classification" y "water"*, donde se puede ver que el artículo 37 no aparece por ningún lado.

Cabe resaltar que para los términos "*material*" y "*diameter*" se han encontrado tres documentos que estudian esos dos factores principalmente. Este número tiene sentido puesto que se explicó que el diámetro es un factor para la elección del material.

También se tiene que mencionar que el artículo "Comparison of statistical and machine learning models for pipe failure modeling in water distribution networks" [80], en español "Comparación de modelos estadísticos y de aprendizaje automático para el modelado de fallas de tuberías en red de distribución de agua", estudia los tres factores que presentan mayor correlación: el material, el diámetro y el año de instalación.

Términos		**Número**	**Nombre del documento**
classification	sampling	37	Deep learning approach for diabetes prediction using PIMA Indian dataset [81]
regression	neural	59	Iterative classifier optimizer-based pace regression and random forest hybrid models for suspended sediment load prediction [102]
regression	tree	21	A brief review of random forests for water scientists and practitioners and their recent history in water resources [65]
regression	forest	20	Sewer condition prediction and analysis of explanatory factors [64]
material	age	36	Comparison of statistical and machine learning models for pipe failure modeling in water distribution networks [80]
material	diameter	39	Geoadditive quantile regression model for sewer pipes deterioration using boosting optimization algorithm [83]
		54	Failure detection methods for pipeline networks: From acoustic sensing to cyber-physical systems [97]
		71	Predictive analytics for water main breaks using spatiotemporal data [114]
age	diameter	36	Comparison of statistical and machine learning models for pipe failure modeling in water distribution networks [80]

Tabla 5-6. Resultados resumidos del análisis de búsqueda

Ilustración 5-2. Artículos que estudian los términos "classification" y "sampling"

Ilustración 5-3. Artículos que estudian los términos "regression" y "neural"

Ilustración 5-4. Artículos que estudian los términos "regression" y "tree"

Ilustración 5-5. Artículos que estudian los términos "regression" y "forest"

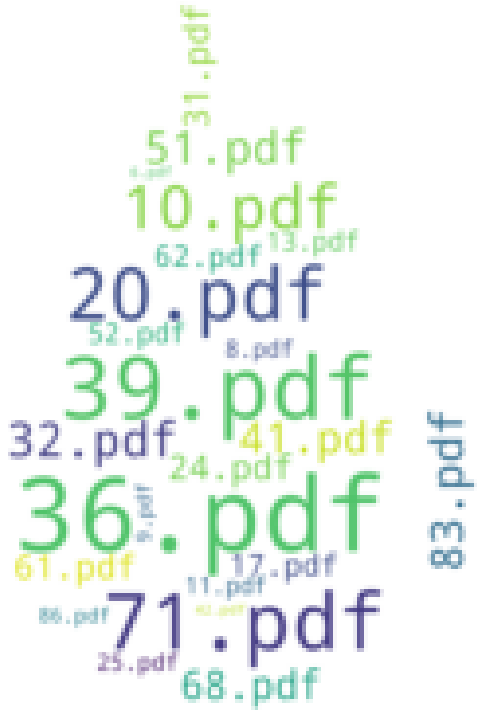

Ilustración 5-6. Artículos que estudian los términos "material" y "age"

Ilustración 5-7. Artículos que estudian los términos "material" y "diameter"

Ilustración 5-8. Artículos que estudian los términos "age" y "diameter"

Ilustración 5-9. Artículos que estudian los términos "sampling", "classification" y "water"

6 CONCLUSIONES

De cara a resumir las principales conclusiones del presente trabajo, se van a considerar desde tres puntos de vista; las que hacen referencia a la metodología desarrollada, las herramientas informáticas que se han construido para implementar la metodología, y las relativas a los problemas en la distribución de redes de agua.

En cuanto a la parte metodológica, el presente trabajo ha puesto de manifiesto que es posible, mediante la aplicación de técnica de procesamiento del lenguaje natural, analizar la producción científica disponible para establecer, ante un problema determinado, cuáles son los principales factores que hay que considerar e incluso los modelos de simulación y cálculo.

Este tipo de técnicas pueden ser de gran ayuda a los investigadores ya que permiten, de una forma rápida y estructurada, condensar la producción científica previa para obtener únicamente los enfoques más relevantes.

Por lo que respecta a las herramientas informáticas. Se han desarrollado una serie de algoritmos (en Python) que aplican esta metodología, de una forma eficiente en tiempo y consumo de recursos informáticos, amigable de utilizar para aquellos análisis que se precisen. Igualmente, estos algoritmos se han completado con un conjunto de herramientas para el análisis de sus resultados (nubes de palabras, tablas de resultados medios, histogramas de frecuencia, diagramas de dispersión, etcétera) que facilitan la obtención de conclusiones.

Finalmente, y en cuanto a los principales factores que afectan a las redes de distribución de agua, los resultados han sido muy claros, tal y como se recoge en el apartado 5.3.3. En una red de distribución, la aparición de fallos en tuberías depende principalmente de la tipología de material, del diámetro, presión, antigüedad y temperatura.

Como futuras líneas de investigación, se podrían considerar básicamente dos:

- Incluir el análisis de producción científica sobre contenido audiovisual, que con técnicas de reconocimiento de voz e imagen permitan incluir no sólo artículos, sino también conferencias.

- Utilizar técnicas de *clustering* para realizar búsquedas de argumentos complementadas con otros parámetros (fecha de publicación, institución de origen, autores, etcétera) y con resultados que incluyan documentos agrupados.

REFERENCIAS

[1] Inicio - Producción científica y datos de investigación - Biblioguías at Universidad Autónoma de Madrid n.d. https://biblioguias.uam.es/memorias (accessed June 4, 2022).

[2] Indicadores de productividad científica en ránkings universitarios: criterios y metodologías | SciELO en Perspectiva n.d. https://blog.scielo.org/es/2013/08/16/indicadores-de-productividad-cientifica-en-rankings-universitarios-criterios-y-metodologias/#.YpuMBqhBzrd (accessed June 4, 2022).

[3] Análisis de la producción científica, ¿qué tan importante es el tamaño de la institución? - SCImago n.d. https://www.scimagolab.com/analisis-de-la-produccion-cientifica-que-tan-importante-es-el-tamano-de-la-institucion/ (accessed June 4, 2022).

[4] Producción e impacto científico en el mundo - SCImago n.d. https://www.scimagolab.com/produccion-e-impacto-cientifico-en-el-mundo/ (accessed June 4, 2022).

[5] Naciones Unidas. Consejo de Derechos Humanos. 2011.

[6] Las 6 etapas del ciclo integral del agua - IDRICA n.d. https://www.idrica.com/es/blog/las-6-etapas-del-ciclo-integral-del-agua/ (accessed April 4, 2022).

[7] UNE-EN 805. 2000.

[8] Ignacio Javier Acosta García. LESIONES PROPIAS DIRECTAS. n.d.

[9] Barton NA, Farewell TS, Hallett SH, Acland TF. Improving pipe failure predictions: Factors affecting pipe failure in drinking water networks. Water Research 2019;164:114926. https://doi.org/10.1016/J.WATRES.2019.114926.

[10] fundición gris – El blog de Víctor Yepes n.d. https://victoryepes.blogs.upv.es/tag/fundicion-gris/ (accessed April 27, 2022).

[11] Hu Y, Wang DL, Cossitt K. Asbestos cement water mains: history, current state, and future planning. 2008.

[12] Tema 51 – Instalaciones de agua – elementos componentes y su funcionamiento - Oposinet n.d. https://www.oposinet.com/temario-de-tecnologia/temario-3-tecnologia/tema-51-instalaciones-de-agua-elementos-componentes-y-su-funcionamiento-2/ (accessed April 27, 2022).

[13] Alan Turing, el padre de la inteligencia artificial | Ministerio de Cultura n.d. https://www.cultura.gob.ar/alan-turing-el-padre-de-la-inteligencia-artificial-9162/ (accessed March 22, 2022).

[14] Ciencia Canaria - ¿Qué es en realidad la inteligencia artificial? n.d. https://www.cienciacanaria.es/secciones/a-fondo/796-que-es-en-realidad-la-inteligencia-artificial (accessed March 23, 2022).

[15] "Machine Learning": definición, tipos y aplicaciones prácticas - Iberdrola n.d. https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico (accessed March 28, 2022).

[16] Machine Learning, Big Data y el futuro del procesamiento de datos n.d. https://www.beetrack.com/es/blog/machine-learning-big-data-procesamiento-de-datos (accessed March 28, 2022).

[17] Aprendizaje supervisado y no supervisado - healthdataminer.com n.d. https://healthdataminer.com/data-mining/aprendizaje-supervisado-y-no-supervisado/ (accessed May 7, 2022).

[18] ¿Cómo aprende la Inteligencia Artificial? - IArtificial.net n.d. https://www.iartificial.net/como-aprende-la-inteligencia-artificial/#Aprendizaje_No_Supervisado (accessed May 8, 2022).

[19] ¿Qué es el análisis exploratorio de datos? - España | IBM n.d. https://www.ibm.com/es-es/cloud/learn/exploratory-data-analysis (accessed May 16, 2022).

[20] Las 7 Fases del Proceso de Machine Learning - IArtificial.net n.d. https://www.iartificial.net/fases-del-proceso-de-machine-learning/#Fase_4_Preparar_los_datos (accessed May 16, 2022).

[21] How to Find Outliers | 4 Ways with Examples & Explanation n.d. https://www.scribbr.com/statistics/outliers/ (accessed May 16, 2022).

[22] What Is Data Preprocessing & What Are The Steps Involved? n.d. https://monkeylearn.com/blog/data-preprocessing/ (accessed May 16, 2022).

[23] Random Oversampling and Undersampling for Imbalanced Classification n.d. https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/ (accessed May 16, 2022).

[24] Validación de modelos predictivos (machine learning): Cross-validation, OneLeaveOut, Bootstraping n.d. https://www.cienciadedatos.net/documentos/30_cross-validation_oneleaveout_bootstrap (accessed May 17, 2022).

[25] A Gentle Introduction to k-fold Cross-Validation n.d. https://machinelearningmastery.com/k-fold-cross-validation/ (accessed May 17, 2022).

[26] Algoritmos Supervisados: Clasificación vs. Regresión – Datlas INVESTIGACIONES – Blog Datlas n.d. https://blogdatlas.wordpress.com/2020/06/28/algoritmos-supervisados-clasificacion-vs-regresion-datlas-research/ (accessed May 8, 2022).

[27] Regresión Logística para Clasificación - IArtificial.net n.d. https://www.iartificial.net/regresion-logistica-para-clasificacion/ (accessed May 7, 2022).

[28] Decision Tree Classifier explained in real-life: picking a vacation destination | by Carolina Bento | Towards Data Science n.d. https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575 (accessed May 7, 2022).

[29] Leo Breiman. RANDOM FORESTS 2001.

[30] What are Neural Networks? | IBM n.d. https://www.ibm.com/cloud/learn/neural-networks (accessed May 8, 2022).

[31] Deep Learning: clasificando imágenes con redes neuronales - LIS Data Solutions n.d. https://www.lisdatasolutions.com/blog/deep-learning-clasificando-imagenes-con-redes-neuronales/ (accessed June 5, 2022).

[32] Redes Bayesianas — Matemática y Estadística — DATA SCIENCE n.d. https://datascience.eu/es/matematica-y-estadistica/redes-bayesianas/ (accessed May 8, 2022).

[33] Nezaratian H, Zahiri J, Peykani MF, Haghiabi A, Parsaie A. A genetic algorithm-based support vector machine to estimate the transverse mixing coefficient in streams. Water Quality Research Journal 2021;56:127–42. https://doi.org/10.2166/wqrj.2021.003.

[34] Minería de Datos: Máquinas de Vectores soporte | AnálisisDeDatos.net n.d. https://analisisdedatos.net/mineria/tecnicas/SVM/SVM.php (accessed May 10, 2022).

[35] ¿Qué es la matriz de confusión y cómo interpretarla? n.d. https://blogs.imf-formacion.com/blog/tecnologia/matriz-confusion-como-interpretarla-202106/ (accessed May 17, 2022).

[36] La matriz de confusión y sus métricas – Inteligencia Artificial – n.d. https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/ (accessed May 17, 2022).

[37] EurEau. Europe's Water in Figures. 2021.

[38] ASCE. A comprehensive Assessment of America's infraestructure. 2021.

[39] Vásquez AC, huerta HV, Quispe JP, Huayna AM. Procesamiento de lenguaje natural. Revista de Investigación de Sistemas e Informática 2009;6:45–54.

[40] Lenguaje Natural: Definición de la tecnología - Elipse n.d. https://elipse.ai/blog/definicion-de-la-tecnologia-lenguaje-natural/ (accessed May 11, 2022).

[41] Conceptos del Procesamiento del Lenguaje Natural - IIC n.d.

https://www.iic.uam.es/innovacion/conceptos-del-procesamiento-del-lenguaje-natural/ (accessed May 11, 2022).

[42] NLP Natural Language Processing: Introducción - DataScientest n.d. https://datascientest.com/es/nlp-natural-language-processing-introduccion (accessed May 15, 2022).

[43] Holguín C, Díaz-Ricardo Y, Antonio Becerra-García R. Ciencias Holguín, Revista trimestral, Año XX, abril-junio 2014 El lenguaje de programación Python/The programming language Python Ivet Challenger-Pérez n.d.

[44] Python o Java: ¿Cuál es mejor? Profesional Online n.d. https://www.profesionalonline.com/blog/programacion/python-o-java-cual-es-mejor/ (accessed March 23, 2022).

[45] Juntunen P, Liukkonen M, Lehtola MJ, Hiltunen Y. Dynamic soft sensors for detecting factors affecting turbidity in drinking water. Journal of Hydroinformatics 2013;15:416–26. https://doi.org/10.2166/hydro.2012.052.

[46] Christodoulou SE, Gagatsis A, Xanthos S, Kranioti S, Agathokleous A, Fragiadakis M. Entropy-Based Sensor Placement Optimization for Waterloss Detection in Water Distribution Networks. Water Resources Management 2013;27:4443–68. https://doi.org/10.1007/s11269-013-0419-8.

[47] Nateghi R, Guikema SD, Quiring SM. Forecasting hurricane-induced power outage durations. Natural Hazards 2014;74:1795–811. https://doi.org/10.1007/s11069-014-1270-9.

[48] Li Z, Zhang B, Wang Y, Chen F, Taib R, Whiffin V, et al. Water pipe condition assessment: A hierarchical beta process approach for sparse incident data. Machine Learning 2014;95:11–26. https://doi.org/10.1007/s10994-013-5386-z.

[49] Ostfeld A. Water distribution networks. Studies in Computational Intelligence 2015;565:101–24. https://doi.org/10.1007/978-3-662-44160-2_4.

[50] Bailey J, Khan L, Washio T, Dobbie G, Huang JZ, Wang R, editors. Advances in Knowledge Discovery and Data Mining. vol. 9651. Cham: Springer International Publishing; 2016. https://doi.org/10.1007/978-3-319-31753-3.

[51] Nahim HM, Younes R, Shraim H, Ouladsine M. Oriented review to potential simulator for faults modeling in diesel engine. Journal of Marine Science and Technology (Japan) 2016;21:533–51. https://doi.org/10.1007/s00773-015-0358-6.

[52] Choi GB, Kim JW, Suh JC, Jang KH, Lee JM. A prioritization method for replacement of water mains using rank aggregation. Korean Journal of Chemical Engineering 2017;34:2584–90. https://doi.org/10.1007/s11814-017-0191-1.

[53] Kim H, Kim S. Evaluation of chlorine decay models under transient conditions in a water distribution

system. Journal of Hydroinformatics 2017;19:522–37. https://doi.org/10.2166/hydro.2017.082.

[54] Wilson D, Filion Y, Moore I. State-of-the-art review of water pipe failure prediction models and applicability to large-diameter mains. Urban Water Journal 2017;14:173–84. https://doi.org/10.1080/1573062X.2015.1080848.

[55] Santos P, Amado C, Coelho ST, Leitão JP. Stochastic data mining tools for pipe blockage failure prediction. Urban Water Journal 2017;14:343–53. https://doi.org/10.1080/1573062X.2016.1148178.

[56] Kutyłowska M. Application of K-nearest neighbours method for water pipes failure frequency assessment. E3S Web of Conferences, vol. 59, EDP Sciences; 2018. https://doi.org/10.1051/e3sconf/20185900021.

[57] O'Reilly G, Bezuidenhout CC, Bezuidenhout JJ. Artificial neural networks: Applications in the drinking water sector. Water Science and Technology: Water Supply 2018;18:1869–87. https://doi.org/10.2166/ws.2018.016.

[58] Kutyłowska M. K-nearest neighbours method as a tool for failure rate prediction. Periodica Polytechnica Civil Engineering 2018;62:318–22. https://doi.org/10.3311/PPci.10045.

[59] Moura G de A, Bezerra S de TM, Gomes HP, Silva SA da. Neural network using the Levenberg–Marquardt algorithm for optimal real-time operation of water distribution systems. Urban Water Journal 2018;15:692–9. https://doi.org/10.1080/1573062X.2018.1539503.

[60] Winkler D, Haltmeier M, Kleidorfer M, Rauch W, Tscheikner-Gratl F. Pipe failure modelling for water distribution networks using boosted decision trees. Structure and Infrastructure Engineering 2018;14:1402–11. https://doi.org/10.1080/15732479.2018.1443145.

[61] Kakoudakis K, Farmani R, Butler D. Pipeline failure prediction in water distribution networks using weather conditions as explanatory factors. Journal of Hydroinformatics 2018;20:1191–200. https://doi.org/10.2166/hydro.2018.152.

[62] Pham TML, Pham HH, Do NAT, Le DH. Proposed probabilistic models of pipe failure in water distribution system. MATEC Web of Conferences, vol. 193, EDP Sciences; 2018. https://doi.org/10.1051/matecconf/201819302002.

[63] Santonastaso GF, di Nardo A, di Natale M, Giudicianni C, Greco R. Scaling-laws of flow entropy with topological metrics of water distribution networks. Entropy 2018;20. https://doi.org/10.3390/e20020095.

[64] Laakso T, Kokkonen T, Mellin I, Vahala R. Sewer condition prediction and analysis of explanatory factors. Water (Switzerland) 2018;10. https://doi.org/10.3390/w10091239.

[65] Tyralis H, Papacharalampous G, Langousis A. A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water (Switzerland) 2019;11.

https://doi.org/10.3390/w11050910.

[66] Bixler RP, Lieberknecht K, Leite F, Felkner J, Oden M, Richter SM, et al. An observatory framework for metropolitan change: Understanding urban social-ecological-technical systems in texas and beyond. Sustainability (Switzerland) 2019;11. https://doi.org/10.3390/su11133611.

[67] Wols BA, Vogelaar A, Moerman A, Raterman B. Effects of weather conditions on drinking water distribution pipe failures in the Netherlands. Water Science and Technology: Water Supply 2019;19:404–16. https://doi.org/10.2166/ws.2018.085.

[68] Sattar AMA, Ertuğrul ÖF, Gharabaghi B, McBean EA, Cao J. Extreme learning machine model for water network management. Neural Computing and Applications 2019;31:157–69. https://doi.org/10.1007/s00521-017-2987-7.

[69] Roccetti M, Delnevo G, Casini L, Cappiello G. Is bigger always better? A controversial journey to the center of machine learning design, with uses and misuses of big data for predicting water meter failures. Journal of Big Data 2019;6. https://doi.org/10.1186/s40537-019-0235-y.

[70] Luo S, Chu VW, Li Z, Wang Y, Zhou J, Chen F, et al. Multitask learning for sparse failure prediction. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11439 LNAI, Springer Verlag; 2019, p. 3–14. https://doi.org/10.1007/978-3-030-16148-4_1.

[71] Ly HB, Monteiro E, Le TT, Le VM, Dal M, Regnier G, et al. Prediction and sensitivity analysis of bubble dissolution time in 3D selective laser sintering using ensemble decision trees. Materials 2019;12. https://doi.org/10.3390/ma12091544.

[72] Shuang Q, Liu HJ, Porse E. Review of the quantitative resilience methods in water distribution networks. Water (Switzerland) 2019;11. https://doi.org/10.3390/w11061189.

[73] Mikhailovskii E. Software for on-line testing of pipeline modeling methods. IOP Conference Series: Materials Science and Engineering, vol. 667, IOP Publishing Ltd; 2019. https://doi.org/10.1088/1757-899X/667/1/012064.

[74] Stańczyk J, Burszta-Adamiak E. The analysis of water supply operating conditions systems by means of empirical exponents. Water (Switzerland) 2019;11. https://doi.org/10.3390/w11122452.

[75] Yang Y, Hu Y, Zheng J. A decision tree approach to the risk evaluation of urbanwater distribution network pipes. Safety 2020;6. https://doi.org/10.3390/safety6030036.

[76] Almheiri Z, Meguid M, Zayed T. An Approach to Predict the Failure of Water Mains Under Climatic Variations. International Journal of Geosynthetics and Ground Engineering 2020;6. https://doi.org/10.1007/s40891-020-00237-8.

[77] Guzmán-Fierro J, Charry S, González I, Peña-Heredia F, Hernández N, Luna-Acosta A, et al. Bayesian

network-based methodology for selecting a cost-effective sewer asset management model. Water Science and Technology 2020;81:2422–31. https://doi.org/10.2166/wst.2020.299.

[78] Luma-Osmani S, Ismaili F, Raufi B, Zenuni X. Causal reasoning application in smart farming and ethics: A systematic review. Annals of Emerging Technologies in Computing 2020;4:10–9. https://doi.org/10.33166/AETiC.2020.04.002.

[79] Kerwin S, Garcia de Soto B, Adey B, Sampatakaki K, Heller H. Combining recorded failures and expert opinion in the development of ANN pipe failure prediction models. Sustainable and Resilient Infrastructure 2020:1–23. https://doi.org/10.1080/23789689.2020.1787033.

[80] Giraldo-González MM, Rodríguez JP. Comparison of statistical and machine learning models for pipe failure modeling in water distribution networks. Water (Switzerland) 2020;12. https://doi.org/10.3390/W12041153.

[81] Naz H, Ahuja S. Deep learning approach for diabetes prediction using PIMA Indian dataset. Journal of Diabetes and Metabolic Disorders 2020;19:391–403. https://doi.org/10.1007/s40200-020-00520-5.

[82] Morosini AF, Haghshenas SS, Haghshenas SS, Geem ZW. Development of a binary model for evaluating water distribution systems by a pressure driven analysis (PDA) approach. Applied Sciences (Switzerland) 2020;10. https://doi.org/10.3390/app10093029.

[83] Balekelayi N, Tesfamariam S. Geoadditive quantile regression model for sewer pipes deterioration using boosting optimization algorithm. Sustainability (Switzerland) 2020;12:1–24. https://doi.org/10.3390/su12208733.

[84] Weeraddana D, Hapuarachchi H, Kumarapperuma L, Khoa NLD, Cai C. Long-Term Water Pipe Condition Assessment: A Semiparametric Model Using Gaussian Process and Survival Analysis. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12085 LNAI, Springer; 2020, p. 487–99. https://doi.org/10.1007/978-3-030-47436-2_37.

[85] Gorenstein A, Kalech M, Hanusch DF, Hassid S. Pipe fault prediction for water transmission mains. Water (Switzerland) 2020;12. https://doi.org/10.3390/w12102861.

[86] Offiong NM, Wu Y, Memon FA. Predicting failures in electronic water taps in rural sub-Saharan African communities: An LSTM-based approach. Water Science and Technology 2020;82:2776–85. https://doi.org/10.2166/wst.2020.542.

[87] Zhu R, Li X, Hu X, Hu D. Risk analysis of chemical plant explosion accidents based on bayesian network. Sustainability (Switzerland) 2020;12. https://doi.org/10.3390/SU12010137.

[88] Mohammadi A, Jalili-Ghazizadeh M, Moslehi I, Yousefi-Khoshqalb E. Survival analysis of water distribution network under intermittent water supply conditions. Water Science and Technology: Water Supply 2020;20:3531–41. https://doi.org/10.2166/ws.2020.228.

[89] Beuken R, Eijkman J, Savic D, Hummelen A, Blokker M. Twenty years of asset management research for Dutch drinking water utilities. Water Science and Technology: Water Supply 2020;20:2941–50. https://doi.org/10.2166/ws.2020.179.

[90] Brunone B, Franchini M. Urban water management: A pragmatic approach. Water (Switzerland) 2020;12. https://doi.org/10.3390/w12123589.

[91] Pietrucha-Urbanik K, Tchórzewska-Cieślak B, Eid M. Water network-failure data assessment. Energies (Basel) 2020;13. https://doi.org/10.3390/en13112990.

[92] Kim S. Adaptive Metaheuristic Scheme for Generalized Multiple Abnormality Detection in a Reservoir Pipeline Valve System. Water Resources Management 2021;35:4581–600. https://doi.org/10.1007/s11269-021-02968-3.

[93] Salazar F, Conde A, Irazábal J, Vicente DJ. Anomaly detection in dam behaviour with machine learning classification models. Water (Switzerland) 2021;13. https://doi.org/10.3390/w13172387.

[94] Robles-Velasco A, Ramos-Salgado C, Muñuzuri J, Cortés P. Artificial neural networks to forecast failures in water supply pipes. Sustainability (Switzerland) 2021;13. https://doi.org/10.3390/su13158226.

[95] Robles-Velasco A, Cortés P, Muñuzuri J, Onieva L. Estimation of a logistic regression model by a genetic algorithm to predict pipe failures in sewer networks. OR Spectrum 2021;43:759–76. https://doi.org/10.1007/s00291-020-00614-9.

[96] Mahmoodi K, Nowruzi H. Extreme wave height detection based on the meteorological data, using hybrid NOF-ELM method. Ships and Offshore Structures 2021. https://doi.org/10.1080/17445302.2021.2005357.

[97] Wong B, McCann JA. Failure detection methods for pipeline networks: From acoustic sensing to cyber-physical systems†. Sensors 2021;21. https://doi.org/10.3390/s21154959.

[98] Hung YH. Improved ensemble-learning algorithm for predictive maintenance in the manufacturing process. Applied Sciences (Switzerland) 2021;11. https://doi.org/10.3390/app11156832.

[99] Ziv B, Parmet Y. Improving nonconformity responsibility decisions: a semi-automated model based on CRISP-DM. International Journal of Systems Assurance Engineering and Management 2021. https://doi.org/10.1007/s13198-021-01318-1.

[100] Kadinski L, Ostfeld A. Incorporation of covid-19-inspired behaviour into agent-based modelling for water distribution systems' contamination responses. Water (Switzerland) 2021;13. https://doi.org/10.3390/w13202863.

[101] Kerwin S, Adey BT, Asce M. Integrated Planning of Operational Maintenance Programs for Water and Gas Distribution Networks n.d. https://doi.org/10.1061/(ASCE)IS.1943.

[102] Meshram SG, Jafar M, Safari S, Khosravi K, Meshram C. Iterative classifier optimizer-based pace regression and random forest hybrid models for suspended sediment load prediction n.d. https://doi.org/10.1007/s11356-020-11335-5/Published.

[103] Feliciano JF, Arsénio AM, Cassidy J, Santos AR, Ganhão A. Knowledge management and operational capacity in water utilities, a balance between human resources and digital maturity—the case of ags. Water (Switzerland) 2021;13. https://doi.org/10.3390/w13223159.

[104] Weeraddana D, MallawaArachchi S, Warnakula T, Li Z, Wang Y. Long-Term Pipeline Failure Prediction Using Nonparametric Survival Analysis. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 12460 LNAI, Springer Science and Business Media Deutschland GmbH; 2021, p. 139–56. https://doi.org/10.1007/978-3-030-67667-4_9.

[105] Martin P, Daly C. Machine Learning and AI for Water Utilities: Junk or Jewel? Triumph or Trash? n.d.

[106] di Nunno F, Granata F, Parrino F, Gargano R, de Marinis G. Microplastics in combined sewer overflows: An experimental study. Journal of Marine Science and Engineering 2021;9. https://doi.org/10.3390/jmse9121415.

[107] Xu H, Sinha SK. Modeling Pipe Break Data Using Survival Analysis with Machine Learning Imputation Methods. Journal of Performance of Constructed Facilities 2021;35:04021071. https://doi.org/10.1061/(asce)cf.1943-5509.0001649.

[108] Mohammed H, Tornyeviadzi HM, Seidu R. Modelling the impact of water temperature, pipe, and hydraulic conditions on water quality in water distribution networks. Water Practice and Technology 2021;16:387–403. https://doi.org/10.2166/wpt.2021.002.

[109] Luo S, Chu VW, Li Z, Wang Y, Zhou J, Chen F, et al. Multi-task learning by hierarchical Dirichlet mixture model for sparse failure prediction. International Journal of Data Science and Analytics 2021;12:15–29. https://doi.org/10.1007/s41060-020-00219-z.

[110] Safaeian Hamzehkolaei N, Alizamir M. Performance evaluation of machine learning algorithms for seismic retrofit cost estimation using structural parameters. Journal of Soft Computing in Civil Engineering 2021;5:32–57. https://doi.org/10.22115/SCCE.2021.284630.1312.

[111] Amiri-Ardakani Y, Najafzadeh M. Pipe Break Rate Assessment While Considering Physical and Operational Factors: A Methodology based on Global Positioning System and Data-Driven Techniques. Water Resources Management 2021. https://doi.org/10.1007/s11269-021-02911-6.

[112] Ma Q, Tian G, Zeng Y, Li R, Song H, Wang Z, et al. Pipeline in-line inspection method, instrumentation and data management. Sensors 2021;21. https://doi.org/10.3390/s21113862.

[113] Azoor R, Deo R, Shannon B, Fu G, Ji J, Kodikara J. Predicting pipeline corrosion in heterogeneous soils using numerical modelling and artificial neural networks. Acta Geotechnica 2021.

https://doi.org/10.1007/s11440-021-01385-5.

[114] Aslani B, Mohebbi S, Axthelm H. Predictive analytics for water main breaks using spatiotemporal data. Urban Water Journal 2021;18:433–48. https://doi.org/10.1080/1573062X.2021.1893363.

[115] Durdán M, Benková M, Laciak M, Kačur J, Flegner P. Regression models utilization to the underground temperature determination at coal energy conversion. Energies (Basel) 2021;14. https://doi.org/10.3390/en14175444.

[116] Carriço N, Covas D, Almeida MDC. Rehabilitation of an industrial water main using multicriteria decision analysis. Water (Switzerland) 2021;13. https://doi.org/10.3390/w13223180.

[117] Morosini AF, Haghshenas SS, Haghshenas SS, Choi DY, Geem ZW. Sensitivity analysis for performance evaluation of a real water distribution system by a pressure driven analysis approach and artificial intelligence method. Water (Switzerland) 2021;13. https://doi.org/10.3390/w13081116.

[118] Aivazidou E, Banias G, Lampridi M, Vasileiadis G, Anagnostis A, Papageorgiou E, et al. Smart technologies for sustainable water management: An urban analysis. Sustainability (Switzerland) 2021;13. https://doi.org/10.3390/su132413940.

[119] Ruiz-Tagle Palazuelos A, Droguett EL. System-level prognostics and health management: A graph convolutional network–based framework. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability 2021;235:120–35. https://doi.org/10.1177/1748006X20935760.

[120] Elshaboury N, Abdelkader EM, Al-Sakkaf A, Alfalah G. Teaching-learning-based optimization of neural networks for water supply pipe condition prediction. Water (Switzerland) 2021;13. https://doi.org/10.3390/w13243546.

[121] Hughes DM, Fitchett J. The Value of Machine Learning Main Break Prediction. n.d.

[122] Handoyo S, Chen YP, Irianto G, Widodo A. The varying threshold values of logistic regression and linear discriminant for classifying fraudulent firm. Mathematics and Statistics 2021;9:135–43. https://doi.org/10.13189/ms.2021.090207.

[123] Robles-Velasco A, Muñuzuri J, Onieva L, Rodríguez-Palero M. Trends and applications of machine learning in water supply networks management. Journal of Industrial Engineering and Management 2021;14:45–54. https://doi.org/10.3926/jiem.3280.

[124] Gonçalves C, Gouveia N, Soares AK. Water Connection Bursting and Leaks Prediction Using Machine Learning. n.d.

[125] Shah K, Sabu S, Chaphekar V. Water Leakage Detection Using Neural Networks. Advances in Intelligent Systems and Computing, vol. 1288, Springer Science and Business Media Deutschland GmbH; 2021, p. 484–98. https://doi.org/10.1007/978-3-030-63128-4_37.

[126] Zhang C, Ye Z. Water pipe failure prediction using AutoML. Facilities 2021;39:36–49.

https://doi.org/10.1108/F-08-2019-0084.

[127] Gao X, Wang K, Lo K, Wen R, Huang X, Dang Q. Water poverty assessment based on the random forest algorithm: application to Gansu, Northwest China. Water Policy 2021;23:1388–99. https://doi.org/10.2166/wp.2021.133.

[128] Ismail W, Niknejad N, Bahari M, Hendradi R, Zaizi NJM, Zulkifli MZ. Water treatment and artificial intelligence techniques: a systematic literature review research. Environmental Science and Pollution Research 2021. https://doi.org/10.1007/s11356-021-16471-0.

[129] White A, Guikema S, Carr B. Why are You Here? Modeling Illicit Massage Business Location Characteristics with Machine Learning. Journal of Human Trafficking 2021. https://doi.org/10.1080/23322705.2021.1982238.

[130] Ileberi E, Sun Y, Wang Z. A machine learning based credit card fraud detection using the GA algorithm for feature selection. Journal of Big Data 2022;9. https://doi.org/10.1186/s40537-022-00573-8.

[131] Raspati GS, Bruaset S, Bosco C, Mushom L, Johannessen B, Ugarelli R. A Risk-Based Approach in Rehabilitation of Water Distribution Networks. International Journal of Environmental Research and Public Health 2022;19. https://doi.org/10.3390/ijerph19031594.

[132] Velani AF, Narwane VS, Gardas BB. Contribution of Internet of things in water supply chain management: A bibliometric and content analysis. Journal of Modelling in Management 2022. https://doi.org/10.1108/JM2-04-2021-0090.

[133] Sofos F, Stavrogiannis C, Exarchou-Kouveli KK, Akabua D, Charilas G, Karakasidis TE. Current Trends in Fluid Research in the Era of Artificial Intelligence: A Review. Fluids 2022;7:116. https://doi.org/10.3390/fluids7030116.

[134] Antomarioni S, Ciarapica FE, Bevilacqua M. Data-driven approach to predict the sequence of component failures: a framework and a case study on a process industry. International Journal of Quality and Reliability Management 2022. https://doi.org/10.1108/IJQRM-12-2020-0413.

[135] Meydani R, Giertz T, Leander J. Decision with Uncertain Information: An Application for Leakage Detection in Water Pipelines. Journal of Pipeline Systems Engineering and Practice 2022;13. https://doi.org/10.1061/(asce)ps.1949-1204.0000644.

[136] Stańczyk J, Burszta-Adamiak E. Development of Methods for Diagnosing the Operating Conditions of Water Supply Networks over the Last Two Decades. Water (Switzerland) 2022;14. https://doi.org/10.3390/w14050786.

[137] Zhang C, Oh J, Park K. Evaluation of sewer network resilience index under the perspective of ground collapse prevention. Water Science and Technology 2022;85:188–205. https://doi.org/10.2166/wst.2021.503.

[138] Liu W, Wang B, Song Z. Failure Prediction of Municipal Water Pipes Using Machine Learning Algorithms. Water Resources Management 2022;36:1271–85. https://doi.org/10.1007/s11269-022-03080-w.

[139] Samadianfard S, Kargar K, Shadkani S, Hashemi S, Abbaspour A, Safari MJS. Hybrid models for suspended sediment prediction: optimized random forest and multi-layer perceptron through genetic algorithm and stochastic gradient descent methods. Neural Computing and Applications 2022;34:3033–51. https://doi.org/10.1007/s00521-021-06550-1.

[140] Ntakolia C, Anagnostis A, Moustakidis S, Karcanias N. Machine learning applied on the district heating and cooling sector: a review. Energy Systems 2022;13. https://doi.org/10.1007/s12667-020-00405-9.

[141] Chen P, Yang J, Duan S, Xie X. Optimization of water resources utilization by GA–PSO in the Pinshuo open pit combined mining area, China. Environmental Earth Sciences 2022;81. https://doi.org/10.1007/s12665-022-10212-3.

[142] Zhang X, Long Z, Yao T, Zhou H, Yu T, Zhou Y. Real-Time burst detection based on multiple features of pressure data. Water Supply 2022;22:1474–91. https://doi.org/10.2166/ws.2021.337.

[143] Yoosefdoost I, Khashei-Siuki A, Tabari H, Mohammadrezapour O. Runoff Simulation Under Future Climate Change Conditions: Performance Comparison of Data-Mining Algorithms and Conceptual Models. Water Resources Management 2022;36:1191–215. https://doi.org/10.1007/s11269-022-03068-6.

[144] Barton NA, Hallett SH, Jude SR. The challenges of predicting pipe failures in clean water networks: A view from current practice. Water Supply 2022;22:527–41. https://doi.org/10.2166/ws.2021.255.

[145] Ilić M, Srdjević Z, Srdjević B. Water quality prediction based on Naïve Bayes algorithm. Water Science and Technology 2022;85:1027–39. https://doi.org/10.2166/wst.2022.006.

ANEXO

Programa principal

```
import os
import sys
from funciones import *
from representaciones import *

#Definición de variables
documentos=[]
w={}
kw={}

#Definición de la lista de nombre de documentos
for i in range(102): #102
    documentos.append(str(i+1)+".pdf")

#Recodiga de datos
for i in documentos:
    if os.path.isfile(i):
        conversion_pdf_a_txt(i)
        limpieza(i)
        w[i]=resumen(i)
        kw[i]=claves(w[i])
        w[i]=referencia(w[i])
    else:
        print(i,"no se ha encontrado")
        sys.exit()

#Obtenter palabras totales
words=palabras(w,documentos)
keywords=palabras(kw,documentos)

#Obtener la misma longitud
dic=igualar(w,documentos)
dic_key=igualar(kw,documentos)

import pandas as pd
#Usar data frame
df=pd.DataFrame(dic)
df_key=pd.DataFrame(dic_key)

#Crear el dataframe de 0 y 1
frec_w=frecuencia(words,df)
frec_kw=frecuencia(keywords,df_key)
```

```
#Invertir dataframe para la representación de resultados
frec_w_inv=frec_w.transpose()
frec_kw_inv = frec_kw.transpose()

#NUBE DE PALABRAS MÁS FRECUENTES
#Realizar la suma para la frecuencia de palabras de todos los pdf
text=frec_w.sum(axis=1).to_dict()
text_k=frec_kw.sum(axis=1).to_dict()
#Nube
frec(text_k,10)
frec(text,30)
```

Funciones

```
#Definición de funciones
import fitz
import re
import pandas as pd
import numpy as np
import string
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from gensim.parsing.preprocessing import STOPWORDS
from spacy.lang.en.stop_words import STOP_WORDS
from pattern.text.en import singularize

#Convertir el pdf en texto
def conversion_pdf_a_txt(documento):
    doc=fitz.open(documento)
    salida=open(documento+".txt","wb")
    for pagina in doc:
        texto=pagina.get_text().encode("utf-8")
        salida.write(texto)
        salida.write(b"\n----\n")
    salida.close()
    doc.close()
    return()

#Eliminar los links
def limpieza(documento):
    documento_txt=open(documento+".txt",'r',encoding='utf-8')
    alllines=documento_txt.readlines()
    documento_txt.close()
    documento_txt=open(documento+".txt",'w+',encoding='utf-8')
    for eachline in alllines:
        a=re.sub(r"http\S+",'',eachline)
        b=re.sub(r"www\S+",'',a)
        c=re.sub(r'[0-9]+','',b)
        d=re.sub(r"-",'',c)
        e=re.sub(r"\W",' ',d)
        g=re.sub("\\s+",' ',e)
        documento_txt.writelines(g)
    documento_txt.close()
    return()

#Eliminar palabras y letras sueltas
```

```
def resumen(documento):
    documento_txt=open(documento+".txt",'r',encoding='utf-8')
    text=documento_txt.read()
    documento_txt.close()
    text_tokens = word_tokenize(text)
    #Convertir mayúsculas a minúsculas
    tokens_without_sw = [w.lower() for w in text_tokens]
    #Eliminar plurales
    tokens_without_sw = [singularize(plural) for plural in tokens_without_sw]
    #Incluir en la lista de stopwords_gensim: letras_palabras(conjunto de
palabras de parada)
    letras_palabras=list(string.ascii_lowercase) +list(STOP_WORDS) +
list(stopwords.words())
    all_stopwords_gensim = STOPWORDS.union(set(letras_palabras))
    terms = [term for term in tokens_without_sw if not term in
all_stopwords_gensim]
    return(terms)

#Función que devuelve las palabras claves
def claves(terms):
    key=terms.copy()
    if 'keyword' in key:
        a_1=key.index('keyword')
        del key[0:a_1+1]
        if(len(key)>20):
            m=len(key)
            del key[21:m]
        if 'highlight' in key:
            a=key.index('highlight')
            m=len(key)
            del key[a:m]
        if 'introduction' in key:
            a=key.index('introduction')
            m=len(key)
            del key[a:m]
    elif 'word' in key:
        pos1=[p for p, v in enumerate(key) if v=='word']
        k=0
        bandera=0
        if(len(pos1)>0):
            while((k<len(pos1)) and bandera==0):
                a_2=pos1[k]
                pos2=[p for p, v in enumerate(key) if v == 'key']
                z=0
                if  len(pos2)>0:
                    while ((z<len(pos2)) and bandera==0):
                        j=pos2[z]
                        if j==(a_2-1):
                            del key[0:a_2+1]
                            bandera=1
                            if 'introduction' in key:
                                a=key.index('introduction')
                                m=len(key)
                                del key[a:m]
                            if 'highlight' in key:
                                a=key.index('highlight')
                                m=len(key)
                                del key[a:m]
```

```
                            if(len(key)>20):
                                m=len(key)
                                del key[21:m]
                        z=z+1
                else:
                    m=len(key)
                    del key[0:m]
                k=k+1
            if (bandera==0 and k>len(pos1)) :
                m=len(key)
                del key[0:m]
        else:
            m=len(key)
            del key[0:m]


    else:
        m=len(key)
        del key[0:m]
    return(key)

#Función que elimina las referencias
def referencia(terms):
    c=terms.copy()
    if 'reference' in c:
        m=len(c)
        d = np.array(c)
        result = np.where(d == 'reference')
        i=max(max(result))
        del c[i:m]
    return(c)
#Función de palabras totales
def palabras(dic_palabras,lista_nombre_documentos):
    total=[]
    for i in lista_nombre_documentos:
        #Crear un diccionario con las palabras que forman ese document
        a=dict.fromkeys(dic_palabras[i])
        #Convertir en una lista las palabras clave
        d=list(a.keys())
        total=total+d
        total=list(dict.fromkeys(total))
    total=list(set(total))
    total.sort()
    return(total)
#Función para igualar la longitud de las listas
def igualar(dic_palabras,lista_nombre_documentos):
    len_max=0
    for i in range(len(lista_nombre_documentos)):
        if(len_max<len(dic_palabras[lista_nombre_documentos[i]])):
            len_max=len(dic_palabras[lista_nombre_documentos[i]])

    for i in range(len(lista_nombre_documentos)):
        while(len(dic_palabras[lista_nombre_documentos[i]])<len_max):
            dic_palabras[lista_nombre_documentos[i]].append("0")
    return(dic_palabras)

#Función frecuencia
```

```
def frecuencia(lista_nombre_documentos,dataframe):
    #Crear un dataframe

frequency=pd.DataFrame(index=lista_nombre_documentos,columns=dataframe.column
s)
    for j in dataframe.columns:
        dicc=dataframe[j].tolist()
        for i in lista_nombre_documentos:
            counter=dicc.count(i)
            frequency.at[i,j]=counter;
    return(frequency)
```

Representaciones

```
import matplotlib.pyplot as plt
from imageio import imread
from wordcloud import WordCloud
from pattern.text.en import singularize
import matplotlib.pyplot as plot
import seaborn as sb
import numpy as np
from bokeh.plotting import figure
from bokeh.io import show

#Creación de nube de palabras
def frec(text,num):
    mk=imread("agua.jpg")
    wordcloud =
WordCloud(width=1000,height=800,background_color="white",mask=mk,max_words=nu
m).generate_from_frequencies(text)
    # crear imagen como gota de agua
    plt.imshow(wordcloud, interpolation='bilinear')
    plt.axis("off")
    plt.show()
    return()

def dibujar(palabra,df):
    #singularizar la palabra
    palabras=[]
    palabras.append(palabra)
    palabras = [singularize(plural) for plural in palabras]
    palabra = "".join(palabra)
    #Dibujar los diferentes histogramas
    intervalos=range(min(df[palabra]),max(df[palabra]+5))
    sb.displot(df[palabra],color='#F2AB6D',bins=intervalos)
    plot.ylabel('Número de artículos')
    plot.xlabel('Número de veces que aparece la palabra')
    plot.title('Histograma de la palabra: %s'%palabra)
    plot.show()
    frecuencias, bordes = np.histogram(df[palabra], bins=intervalos)
    histograma = figure(title='Histograma de la palabra: %s'%palabra,
                        x_axis_label='Número de veces que aparece la
palabra',
                        y_axis_label='Número de artículos')
    histograma.xaxis.ticker = bordes
    histograma.quad(bottom=0, top=frecuencias,
                    left=bordes[:-1], right=bordes[1:],
                    fill_color='#F2AB6D', line_color='black')
```

```
    show(histograma)
    return()

#diagramas de dispersión
def dispersion(palabra1,palabra2,df):
    test_x=df[palabra1]
    test_y=df[palabra2]
    points=list(set(zip(test_x,test_y)))
    count=[len([x for x,y in zip(test_x,test_y) if x==p[0] and y==p[1]]) for
p in points]
    plot_x=[i[0] for i in points]
    plot_y=[i[1] for i in points]
    count=np.array(count)
    plot.scatter(plot_x,plot_y,c=count,s=8*count**0.5,cmap='Spectral_r')
    plot.colorbar()
    plot.ylabel(palabra2)
    plot.xlabel(palabra1)
    plot.show()
    return()


#localizar qué documentos tienen mayor frecuencia de aparición de varias
palabras
def sol(palabra,documentos,df):
    df1=df.copy()
    if 'water' in palabra:
        num=0.005
    elif 'pipe' in palabra:
        num=0.005
    else:
        num=2/3
    for i in palabra:
        if i in df.columns:
            a=1
        else:
            a=0
    if a==1:
        n=len(palabra)
        for j in documentos:
            i=0
            while i<n:
                s=i+1
                while s<n:
                    e1=df1.at[j,palabra[i]]
                    e2=df1.at[j,palabra[s]]
                    if(e1>0 and e2>0):
                        if(e1<e2):
                            aux=e1
                            e1=e2
                            e2=aux
                        if((e2/e1)<(num)):
                            df1=df1.drop([j])
                            i=n
                            s=n
                        if((e2/e1)>=(num)):
                            s=s+1
                    else:
```

```
                        df1=df1.drop([j])
                        i=n
                        s=n
                i=i+1
        if (df1.empty is False):
            if 'water' in palabra:
                n=20
            elif 'pipe' in palabra:
                num=0.005
            for i in palabra:
                if i ==( 'water'or 'pipe'):
                    n=20
                else:
                    n=len(palabra)
                df1[i]=(1/n)*df1[i]
                df1_inv=df1.transpose()
            text=df1_inv.loc[palabra,:].transpose().sum(axis=1).to_dict()
            frec(text,20)
        else:
            print("No se encuentran documentos que traten esos temas a la
vez")
    else:
        print("No se encuentran el conjunto de palabras en ningún pdf")
    return()
```