ESTUDIO SOBRE EL PAPEL DE LA DIGITALIZACIÓN EN LA GESTIÓN DE LAS INFRAESTRUCTURAS HÍDRICAS DE LA COMUNITAT VALENCIANA

Càtedra de Transformació del Model Econòmic
Economia Circular en el Sector de l'Aigua

Xarxa Càtedres de Transformació del Model Econòmic

GENERALITAT VALENCIANA
Conselleria d'Hisenda i Model Econòmic

VNIVERSITAT DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Universitat d'Alacant
Universidad de Alicante

UNIVERSITAT JAUME I

UNIVERSITAS Miguel Hernández

Contenido

1. INTRODUCCIÓN 3
2. LA DIGITALIZACIÓN EN LA GESTIÓN DE LOS RECURSOS NATURALES 4
3. EL SECTOR DEL CICLO URBANO DEL AGUA 5
2.1. Captación, tratamiento y distribución de agua. 7
2.2. Alcantarillado y depuración de aguas residuales. 8
4. GESTIÓN DE DATOS 9
5. MACHINE LEARNING: DEL DATO A LA INFORMACIÓN 11
6. FASES DE UN MODELO DE MACHINE LEARNING 13
7. ALGORITMOS DE CLASIFICACIÓN 15
- Regresión lineal 15
- Regresión logística 16
- Árboles de decisión 16
- Random Forest 17
- XGBoost 17
- Gradient Boosting 18
- Redes neuronales 18
- K-Means 19
8. SOFTWARE USADO EN EL CASO DE EJEMPLO. 19
9. CASO DE EJEMPLO: PREDICCIÓN DE AVERÍAS EN BOMBAS SOPLANTES 19
8.1 OBJETIVO Y JUSTIFICACIÓN 21
8.2 EXPLORACIÓN Y SELECCIÓN. 22
8.3 MODELADO DE DATOS: DISCRETIZACIÓN 22
8.4 MODELO ML: ÁRBOLES DE DECISIÓN 24
10. RESULTADOS 29
11. CONCLUSIONES 30

ILUSTRACIONES

Ilustración 1 Proceso ETL *10*
Ilustración 2 Tipos de aprendizaje automático *11*
Ilustración 3 Fases construcción de un modelo Machine Learning *14*
Ilustración 4 Ejemplo árbol de decisión (plot) *28*

CÓDIGOS DE DESARROLLO

Código de desarrollo R 1: Paquetes necesarios ____ *21*
Código de desarrollo R 2: Carga y lectura de datos ____ *22*
Código de desarrollo R 3: Cambio a factores (categóricas) y discretización del target (fallo). ____ *24*
Código de desarrollo R 4: Matriz de confusión y métricas de acierto, precisión, cobertura ____ *26*
Código de desarrollo R 5: Creación de variable aleatoria, data frames e identificación de variables ____ *26*
Código de desarrollo R 6: Modelo y gráfico árbol de decisión ____ *27*
Código de desarrollo R 7: Evaluación del modelo ____ *27*

TABLAS

Tabla 1 Factores y discretización *22*

1. INTRODUCCIÓN

La transformación digital está cambiando nuestro actual modo de vida, la forma de aprender, trabajar e incluso relacionarnos. Las numerosas ventajas que ofrece la digitalización ocupan un gran número de aspectos que pueden ser agrupados en económicos, sociales y ambientales. Por ejemplo, a nivel económico, las empresas han conseguido una mayor flexibilidad tanto de los productos y servicios como el modo de concebir el trabajo. La digitalización facilita a las empresas un control mucho mayor de los procesos de fabricación y posterior distribución, la deslocalización del trabajo, la rapidez, el acceso a grandes volúmenes de información y, en definitiva, la posibilidad de adaptarse de manera acelerada a un entorno globalizado en constante cambio.

Todos estos cambios tienen una elevada repercusión en el aspecto social, la forma de relacionarnos y consumir servicios ha experimentado un profundo cambio. A su vez, nuestros gustos o intereses generan ingentes volúmenes de datos que alimentan a los sistemas de información, su almacenamiento y procesamiento da lugar a nuevos servicios más innovadores, más eficientes y personalizados. Todo esto es posible gracias al uso de metodologías que son capaces de, además de gestionar grandes volúmenes de datos, evaluar patrones, analizar tendencias e ir mejorando las predicciones. Los resultados permiten, tanto a empresas como instituciones, anticipar cambios en las necesidades del mercado, disminuir la incertidumbre en las decisiones a adoptar y aumentar la eficiencia de uso de los recursos, entre otros beneficios.

Sin embargo, los aspectos ambientales tales como el control y monitorización de la contaminación atmosférica, el uso de los recursos naturales, la monitorización de residuos generados o la cuantificación de los impactos ambientales generados a nivel local o global avanzan a menor velocidad. Todas las economías se enfrentan a cambios en los patrones climáticos que interrumpen gravemente la vida de las personas y la estructura de la sociedad. El reto del cambio climático y la degradación del medio ambiente son cuestiones donde el potencial de la digitalización puede ayudar a comprender y actuar de manera más eficaz y eficiente. El presente trabajo revisa el marco metodológico para la implementación de proyectos de digitalización en el sector del agua urbana, propone numerosos ejemplos prácticos y explica, con mayor detenimiento, los

resultados de la implementación de un modelo de "*machine learning" en la gestión de activos e infraestructuras del ciclo urbano del agua.*

2. LA DIGITALIZACIÓN EN LA GESTIÓN DE LOS RECURSOS NATURALES

El sector de los recursos naturales siempre ha sido un elemento muy importante en la economía mundial. La gestión de los recursos naturales requiere metodologías innovadoras que permitan monitorizar y entender el ciclo de vida de los productos, bienes y servicios, así como identificar las oportunidades de reconversión de los residuos generados en materias primas para otros procesos industriales. El desarrollo de las tecnologías conduce al avance del uso racional de los recursos naturales y la protección del medio ambiente. Esto es posible a la gran cantidad de información relativa a la calidad del aire, agua, usos del suelo, flora y fauna, o aspectos relacionados con el uso de la energía y la generación de ruido, desechos y emisiones (Kalymbek et al., 2021).

Son numerosos los sectores que implementan en sus procesos soluciones tecnológicas, por ejemplo, en el sector agrícola se usan sensores para monitorizar las variables que afectan a la producción, optimizando de este modo la obtención de alimentos (Lioutas, Charatsari, & De Rosa, 2021), aumentar la precisión de la aplicación de fertilizantes, pesticidas y herbicidas (Carolan, 2016), determinar fechas óptimas de siembre de cultivos (López & Corrales, 2016), o para ayudar a identificar y eliminar malezas (Lottes et al., 2017; Fennimore, 2017)).

En el sector pesquero los sistemas satelitales basados en sensores ópticos proporcionan numerosos parámetros físicos del entorno acuático y localización de embarcaciones pesqueras, los drones submarinos determinan las mejores áreas para la pesca, así como otros dispositivos analizan el volumen y las características biofísicas de los peces capturados generando una elevada riqueza de información en la cadena de suministro (Mnatsakanyan, A. G., & Kharin, A. G. (2021).

Del mismo modo, en el sector ganadero, los datos recopilados por sensores (como cámaras, micrófonos, acelerómetros, analizadores de gases…) sobre animales o sobre su entorno, junto con técnicas analíticas avanzadas, proporcionan herramientas eficientes para monitorear animales para mejorar su bienestar y optimizar el uso de recursos, como

alimentación, agua, tierra (Pezzuolo et al., 2021). Incluso algunas tareas son realizadas por robots, como por ejemplo el ordeño del ganado (Driessen & Heutinck, 2015) o los dispensadores automáticos de alimentación (Chiumenti et al., 2018).

En este contexto la Unión Europea identifica como Tecnologías Habilitadoras (*Key Enabling Technologies*) a aquellas tecnologías intensivas en conocimiento que se caracterizan por un alto grado de I+D (Evangelista, R., et al., 2018). Éstas se presentan como inductoras de innovación en múltiples sectores con un fuerte potencial para provocar grandes cambios sociales, económicos y ambientales.

La estrategia digital de la UE pretende que esta transformación y adaptación digital funcione tanto para las personas y como para las empresas antes del 2050, contribuyendo, al mismo tiempo, al objetivo de conseguir una Europa climáticamente neutra. Las tecnologías digitales ofrecen toda una serie de beneficios tangibles no solamente económicos, también sociales y ambientales, aumentando aún más las ventajas de su implementación. Esto es debido a que ofrecen un gran abanico de soluciones que a su vez permiten reducir los impactos ambientales generados y reasignar los recursos de una manera más eficiente. Por este motivo, algunos sectores como el de la energía, los residuos o el agua presentan amplias posibilidades en la integración de estas herramientas digitales en sus procesos.

3. EL SECTOR DEL CICLO URBANO DEL AGUA

El agua es un bien esencial tanto para los ciudadanos como para la actividad económica de cualquier sector, razón por la cual es necesario garantizar un suministro de agua seguro, previsible y sobre todo de calidad. La creciente demanda de agua, unida a una reducción de recursos hídricos, están provocando un mayor interés por parte de los estados miembros en concebir el agua como un bien económico escaso de creciente valor que necesita de nuevas fuentes de abastecimiento que aseguren la continuidad y sostenibilidad del ciclo hidrológico.

La escasez hídrica actual tiene consecuencias globales sobre la población, el medio ambiente y la economía. Esta situación obliga a las administraciones a tomar medidas dirigidas a fomentar no solo la conservación sino a potencial la reutilización y la economía circular. Es el caso de la Agenda 2030 para el Desarrollo Sostenible de Naciones Unidas,

la cual comprende una colección de 17 objetivos de desarrollo sostenible (ODS) que buscan el desarrollo social y económico de los territorios y la conservación del medio ambiente, donde destaca el ODS 6 sobre la gestión del agua y el saneamiento.

El ciclo urbano del agua ofrece grandes oportunidades para la implementación de la digitalización a lo largo de todas sus etapas. Un adecuado proceso digital de los flujos de información permite un mayor control de la logística a todos los niveles, desde la predicción de la demanda de agua hasta el control de residuos generados con tal asegurar su reutilización dentro de los canales apropiados (de Sousa Jabbour et al., 2018). Para garantizar el suministro de agua se requiere de una serie de infraestructuras que permitan almacenar, potabilizar, distribuir y depurar el agua. Al conjunto de procesos que transcurren desde la captación y suministro hasta su depuración y vertido se denomina ciclo urbano del agua.

En este sentido, es importante subrayar que, una vez el agua es usada, y tras la aplicación de los tratamientos de depuración, somos capaces de obtener múltiples recursos tales como fertilizantes, fangos y biogás (Richard et al., 2020; Coroamă and Mattern, 2019). Actualmente, la tecnología permite usar esta agua, previamente regenerada, y aprovecharla para otros usos, disminuyendo la presión sobre la fuente convencional.

Por todo ello, el ciclo urbano del agua representa numerosas oportunidades para la digitalización en cada una de las fases que lo integran. La elevada cantidad de indicadores técnicos y económicos que se generan desde la captación hasta la distribución y depuración proporcionan valiosa información que puede ser usada para optimizar los procesos tanto de potabilización como de depuración y distribución, generar escenarios de consumos, disminuir costes energéticos y alargar la vida útil de los activos e infraestructuras.

En los últimos años, las tecnologías de inteligencia artificial se han aplicado cada vez más para traducir los datos en conocimiento procesable para mejorar el funcionamiento de las infraestructuras del ciclo urbano del agua y respaldar la toma de decisiones en materia de operación y mantenimiento (Al Aani et al., 2019; Corominas et al.,

2018 ; Haimi et al., 2013 ; Li et al., 2021). A continuación, se resumen algunos ejemplos de digitalización en las distintas fases del ciclo urbano del agua.

2.1. Captación, tratamiento y distribución de agua.

El proceso de captación y tratamiento de agua permite generar numerosos datos que son captados a través de sensores, tales como caudales, energía, reactivos y mantenimientos requeridos, entre otros. La digitalización permite monitorizar su distribución hasta los hogares, industrias y comercios. Su constante monitorización facilita a los operadores prever los picos de demanda, adecuando de este modo los caudales tratados y maximizando los tiempos de disponibilidad. En este sentido, facilita la detección de posibles fugas en la red, así como posibles desviaciones o anomalías en los consumos de los usuarios, mejorando la eficiencia hídrica del sistema lo que repercute en un menor consumo energético de la red.

Otro ejemplo aplicado en la etapa de potabilización y distribución del agua es el relativo a la instalación de contadores inteligentes y la sectorización de la red de distribución (Moy de Vitry et al., 2019). En materia de tele lectura, los contadores inteligentes permiten a los usuarios consultar el consumo propio de agua en tiempo real a través de las aplicaciones alojadas en la nube. El software ofrece un sistema de alarma por fugas interiores ante un incremento súbito del consumo habitual, detecta posibles ocupaciones en segundas residencias o situaciones de riesgo para personas vulnerables que viven solas, en el caso de detectarse un parón en el consumo.

Los contadores permiten una facturación exacta del servicio, eliminando de este modo los consumos estimados por los hogares, industria o comercios, dotando de mayor transparencia y calidad la prestación del servicio. A su vez, los datos generados están conectados a la plataforma para el abastecimiento (ETAPs), que recopila y analiza datos diarios sobre el relacionados con la cantidad de agua suministrada o los caudales y presiones en la red. Las pérdidas de agua potable en la red de distribución es uno de los mayores problemas para los operadores, implican elevadas pérdidas económicas además de ser una fuente de ineficiencia.

Esto es debido, entre otros aspectos, al elevado consumo energético necesario para su impulsión. La sectorización permite establecer un balance acotado por zonas geográficas, barrios y edificios, identifican y cuantifican las posibles pérdidas de agua potable y, en consecuencia, permite a los operadores actuar en áreas muy delimitadas. En este caso, se combinan herramientas digitales dedicadas a monitorizar los caudales de entrada y salida y la geolocalización de las redes. El posterior análisis de los datos permite establecer alarmas en tiempo real.

2.2. Alcantarillado y depuración de aguas residuales.

En esta fase del ciclo urbano del agua las herramientas digitales ofrecen un gran número de posibilidades, por ejemplo, controlar los caudales de aguas residuales vertidos al alcantarillado, ya sea con el fin de detectar grandes fluctuaciones del volumen generado como picos de carga orgánica. Esta información resulta de elevada utilidad para las Estaciones de tratamiento de aguas residuales (EDARs), los datos recogidos permiten planificar las tareas de operación, ajustando los reactivos y la energía necesaria con el fin de garantizar el correcto funcionamiento del proceso.

La monitorización de los tratamientos en las EDARs permite automatizar los procesos con el fin de mejorar la eficiencia en estas infraestructuras. Los beneficios de la digitalización repercuten de manera directa disminuyendo los costes del proceso a la par que maximizan los beneficios obtenidos mediante el control de los residuos con tal de generar recursos para otros usos; fangos para la agricultura, fertilizantes y energía, entre otros.

Desde el punto de vista operativo de las instalaciones, la aplicación del sistema de computación en la nube permite integrar volúmenes de datos generados a partir de los sistemas de monitorización y medición para la calidad del agua, el consumo de energía y el uso de reactivos. El sistema de computación en la nube procesa datos y visualiza la operación del proceso de depuración de las aguas residuales, el consumo de energía y el análisis de costes. De este modo, con ayuda de la digitalización, es posible generar un gemelo digital del proceso completo. Los gemelos digitales son generados para simular escenarios tales como; variaciones en los caudales, precipitaciones y cargas orgánicas, entro otros aspectos.

Esta simulación permite al operador proyectar distintas alternativas en el entorno real, favoreciendo de este modo la disminución de los costes y una mayor eficiencia energética a la par que minimiza riesgos de vertido. Otro aspecto, en el que la digitalización presenta numerosas ventajas es en el de la gestión de los activos físicos de las instalaciones de forma centralizada. Su monitorización constante y la aplicación de algoritmos permite reducir el riesgo de avería y los costes asociados al ciclo de vida a partir de datos históricos, así como una mejor planificación de las tareas de mantenimiento más efectivas y el establecimiento de un plan de renovación e inversión en nuevos equipos e infraestructuras.

4. GESTIÓN DE DATOS

La monitorización permite generar una gran cantidad de datos, por ejemplo, la tele lectura de contadores en grandes ciudades puede abarcar millones de unidades conectadas en red que emiten los consumos diarios de los usuarios, estos consumos pueden enviar información en intervalos de pocos minutos, por lo que la cantidad de información diaria de todo el parque de contadores es muy elevada. Esta ingente cantidad de datos requiere de la aplicación de metodologías de análisis con tal de convertir el dato en información útil para operadores, administraciones y ciudadanos. Su correcta gestión también implica el mantenimiento histórico y almacenaje de toda esta información previamente ordenada y estructurada.

De este modo la información almacenada y correctamente actualizada podrá ser utilizada para generar un beneficio. En este sentido, la capacidad de almacenar, agregar y combinar datos y luego utilizar los resultados para realizar análisis profundos se ha vuelto cada vez más accesible a medida que el almacenamiento digital y la computación en la nube siguen reduciendo los costes y otras barreras tecnológicas (Brill, 2007). Además, la capacidad de generar, comunicar, compartir y acceder a los datos se ha visto revolucionada por el creciente número de dispositivos y sensores que ahora están conectados por redes digitales.

Otra de las razones de la importancia de almacenar la información previamente estructurada y ordenada es la posibilidad de relacionar distintas bases de datos. La

combinación y fusión de datos de distinta tipología genera nuevas vías de información. Por ejemplo, volviendo al caso de los tele contadores, los datos relativos a los consumos de agua en los hogares pueden relacionarse con las temperaturas y las precipitaciones. En caso de existir una correlación directa entre temperatura y consumo de agua la empresa suministradora puede prever el posible incremento en los consumos, anticipando los volúmenes necesarios en lo que respecta al proceso de captación y potabilización. Además, la empresa de abastecimiento podrá incorporar las predicciones meteorológicas para las zonas de suministro con tal de proyectar escenarios a corto y medio plazo.

Con tal de generar información de calidad, los datos en origen deben ser correctamente procesados, ello implica distintas tareas, desde la limpieza de la base de datos con tal de detectar outlayers o posibles valores erróneos hasta la selección de la información que requerirá la empresa.

Estos procesos suelen componer un flujo de trabajo intensivo y constituyen una parte integral de la fase posterior de las arquitecturas de los almacenes de datos, donde se lleva a cabo la recogida, extracción, limpieza, transformación y transporte de los datos. Para hacer frente a este flujo de trabajo y para facilitar y gestionar los procesos operativos del almacén de datos ya existen en el mercado herramientas especializada disponibles, bajo el título general de herramientas de extracción-transformación-carga (ETL).

Ilustración 1 Proceso ETL

Los datos proceden de las fuentes tales como ordenadores y cualquier otro periférico que pueda almacenar datos. A continuación, se procesan mediante una serie de rutinas de transformación que homogeneizarán su formato, paralelamente se comprueba su

calidad e integridad y, por último, una vez que los datos están en el formato de destino, se cargan en el almacén de datos.

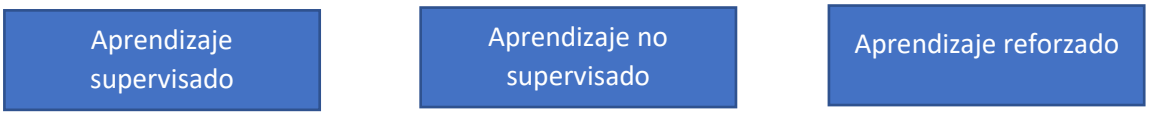
Estos datos se pueden clasificarse principalmente en las siguientes formas:

- Gestión de datos relacionados: Almacena datos en filas y columnas conformando tablas conectadas por claves comunes.
- Gestión de datos jerárquicos: Se basa en un esquema de árbol para organizar los datos. Puede representarlos en tablas, las filas se componen por cada registro y las columnas por cada atributo. Se relacionan por medio de correspondencias.
- Gestión de datos en red: Es una estructura con relaciones más complejas, pues cada registro puede rastrearse desde distintos caminos.
- Gestión de datos dirigidos a objetos: Puede almacenar cualquier tipo de dato gráfico (imagen, sonido y texto).

5. MACHINE LEARNING: DEL DATO A LA INFORMACIÓN

El objetivo del aprendizaje automático es proporcionar alguna conclusión o predicción a partir del conjunto de datos disponible utilizando un modelo. Las técnicas de aprendizaje automático (ML) permiten usar estos datos con el fin de ayudar a identificar y clasificar aspectos que pueden resultar de interés. El objetivo es analizar la existencia de posibles patrones que pueden estar relacionados con un objetivo establecido (Chen et al., 2020). Principalmente se pueden clasificar en aprendizaje supervisado, aprendizaje no supervisado y aprendizaje reforzado.

Ilustración 2 Tipologías de aprendizaje automático



El aprendizaje supervisado tiene un enfoque reactivo debido a que el evento o consecuencia que se desea analizar es conocido de antemano (Gupta et al., 2018; Adesanwo et al., 2017). Por lo tanto, el principal objetivo es construir un modelo de aprendizaje a partir de datos de entrenamiento previamente etiquetados. Las observaciones etiquetadas permiten hacer predicciones futuras junto con el margen de

error o tolerancia que el modelo genera. Un mayor número de observaciones permitirá mejorar los resultados del modelo, disminuyendo el margen de error asociado (Rojas, E. 2018). En este caso, el aprendizaje supervisado puede devolver variables discretas (mediante tarea de clasificación) o valores continuos (regresión).

En el primer de los casos, variables discretas, se usa para predecir las etiquetas de clase categórica de nuevas instancias basadas en observaciones pasadas, por ejemplo, si deseamos conocer si una bomba de impulsión de agua fallará o no, daremos el valor 1 a Fallo y el valor 0 al no Fallo. Estas etiquetas vienen definidas por una serie de variables explicativas asociadas a las bombas que estamos evaluando, éstas pueden ser las horas de operación, la edad del equipo, el tipo de agua residual, la existencia de cavitaciones etc.… de modo que, a partir de los datos de entrenamiento, el sistema ofrecerá una respuesta de 0 o 1 a partir de la similitud de las nuevas bombas evaluadas con observaciones pasadas.

Esta tarea de clasificación permite usar más de etiquetas de clase, por lo que no tiene que ser necesariamente una clasificación binaria. Un ejemplo del uso de más etiquetas podría ser el relacionado con los consumos de agua, pudiendo clasificar el consumo mensual en intervalos que oscilen entre 1 y 10 y que éstos dependan de características relacionadas con las viviendas, superficie, zona residencial, número de baños, número promedio de personas en la vivienda etc…

Por otro lado, el aprendizaje supervisado puede predecir resultados continuos, es lo que conocemos con análisis de regresión. En este caso el objetivo es encontrar una relación entre las variables explicativas que permita predecir un resultado continuo. Para seleccionar la mejor combinación de variables capaces de explicar el modelo se usan técnicas estadísticas. En primer lugar, las variables técnicas incluidas en la regresión deben ser significativas con un nivel de confianza del 95%, de igual modo cabe considerar la mejor bondad de ajuste R^2, así como el menor error estándar en las estimaciones (Gujarathi, 2022).

Además, el modelo estimado debe ser validado, para ello debe cumplir distintas hipótesis. En primer lugar, las variables utilizadas deben ser independientes entre sí, los problemas de multicolinealidad entre los regresores ofrecen resultados sesgados. Para

ello se usa el factor de inflación de la varianza (Kleinbaum et al., 1988). Además, para garantizar la robustez del modelo, se analiza el comportamiento de los residuos a través del test de normalidad, una elevada dispersión generaría mayores errores en la proyección del cálculo. Para finalizar con la validación se aplica el test de homocedasticidad (Breus and Pagan, 1979) y autocorrelación de los residuos (Durbin & Watson, 1951).

Algunos trabajos aplican regresiones al sector urbano del agua, por ejemplo, Hernandez-Chover, V. et al., (2019) modelizan el coste del mantenimiento a través del tiempo. Para lograrlo emplean una serie de variables explicativas relacionadas con la edad de la planta, la carga contaminante y los habitantes equivalentes tratados. Por otro lado, Castellet-Viciano, Ll. et al., (2018) obtienen un modelo que explica los consumos energéticos en las plantas de tratamiento de aguas residuales. Para lograrlo diferencia el tipo de tecnología usada y el tamaño (según habitantes equivalentes) de las plantas de tratamiento de aguas residuales.

El aprendizaje no supervisado no cuenta con datos que definan un objetivo concreto, por lo que clasifican principalmente los datos a partir de patrones que permitan separar conjuntos similares. Esta técnica explora la estructura de los datos para extraer información significativa o similitudes entre los datos analizados que nos ayuden a comprender el patrón analizado. Para ello, el análisis recurre a la generación de grupos (clústeres), cada grupo viene definido por un grupo de observaciones que comparten cierto grado de semejanza (Bishop, C., 2007).

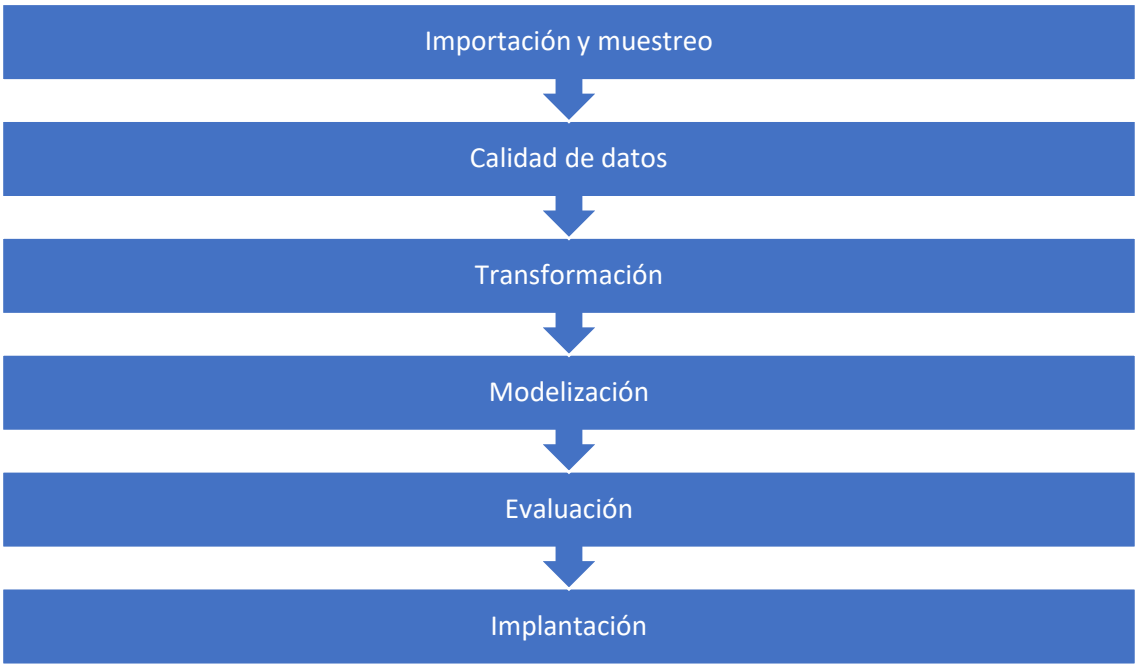
Por último, el aprendizaje reforzado consiste en desarrollar un sistema que mejore su rendimiento, para ello se usa una señal de recompensa. De modo que, a través de la interacción con el entorno, un agente puede utilizar el aprendizaje reforzado para aprender una serie de acciones que maximicen esta recompensa (López, J., López, B., & Díaz, V., 2004).

6. FASES DE UN MODELO DE MACHINE LEARNING

Generar un modelo de Machine Learning implica un proceso, en puntos anteriores se menciona la importancia de recolectar y grabar datos para poder ser analizados con

posterioridad. La calidad de los datos resulta de elevada importancia, la siguiente ilustración (3) resume los distintos procesos necesarios.

Ilustración 3 Fases construcción de un modelo Machine Learning



1. Importación y muestreo: Los datos se pueden importar de fuentes tal como un sitio web, utilizando una API o una base de datos. Este paso es uno de los más complicados y requiere un tiempo determinado. A continuación, es importante realizar un primer muestreo de la tipología de los valores que incluye, numéricos, ordinales, dicotómicos.

2. Calidad de datos: A partir de un primer muestreo, es necesario asegurar la calidad de los datos para garantizar un rendimiento óptimo del algoritmo. Con este objetivo se lleva a cabo una exploración general y el cálculo de los estadísticos básicos, los nulos y ceros que contiene, posibles valores atípicos y un análisis de coherencia.

3. Transformación de datos: En este punto, se debe depurar valores, discretizar los datos, generar intervalos y seleccionar cuales son los más importantes para el objetivo que hemos diseñado. Seleccionar las variables target y las predictoras del modelo.

4. Modelización: los algoritmos de aprendizaje se alimentan con los datos que se procesaron en las etapas anteriores. La idea es que los algoritmos pueden extraer información útil de los datos iniciales. Paso previo a aplicar un método de clasificación, es la partición del conjunto de datos en dos conjuntos de datos más pequeños que serán

utilizadas con los siguientes fines: entrenamiento y test. El subconjunto de datos de entrenamiento es utilizado para estimar los parámetros del modelo y el subconjunto de datos de test se emplea para comprobar el comportamiento del modelo estimado.

5. Evaluación el algoritmo. Se realizan las pruebas de la información que genera el conocimiento del entrenamiento previo que se obtuvo a través del algoritmo. Se realiza una evaluación sobre la precisión del algoritmo en sus predicciones y, si no está satisfecho con su rendimiento, debe volver a la etapa anterior y continuar entrenando el algoritmo cambiando algunos parámetros hasta que se logre un rendimiento aceptable. Para comparar los distintos modelos generamos una unidad para medir el rendimiento. Esta unidad de medida utiliza la precisión en la clasificación. Un método para evaluar clasificador alternativo a la métrica expuesta es la curva ROC (Receiver Operating Characteristic). La curva ROC es una representación gráfica del rendimiento del clasificador que muestra la distribución de las fracciones de verdaderos positivos y de falsos positivos. La fracción de verdaderos positivos se conoce como sensibilidad, sería la probabilidad de clasificar correctamente a un individuo cuyo estado real sea definido como positivo.

6. Implantación. Finalmente, tras haber seleccionado el modelo óptimo, se adapta el código con tal de asegurar la sincronización con la fuente de datos. Esta etapa garantizará la continua alimentación y aprendizaje del modelo a partir de las nuevas observaciones que se vayan generando.

7. ALGORITMOS DE CLASIFICACIÓN

Existen muchos algoritmos de Machine Learning, desde los más básicos a otros más complejos. Podemos destacar algunos como los siguientes:

- Regresión lineal**:** La regresión lineal es una técnica de modelado estadístico que se emplea para describir una variable de respuesta continua como una función de una o varias variables predictoras. Las técnicas de regresión lineal permiten crear un modelo lineal. Este modelo describe la relación entre una variable dependiente *Y* (también conocida como la respuesta) como una función de una o varias variables independientes *X* (denominadas predictores). Dependiendo del número de variables predictoras las regresiones pueden ser simples o múltiples.

En el primer de los casos únicamente se utilizará un predictor y, en el segundo de los casos se utilizarán múltiples predictores para predecir la respuesta. Las regresiones permiten generar predicciones, comparar ajustes lineales, representar valores residuales y evaluar la bondad de ajuste.

- Regresión logística**:** La regresión logística es similar a la regresión lineal, está adaptado para modelos en los que la variable dependiente es dicotómica. Los coeficientes de regresión logística pueden utilizarse para estimar la razón de probabilidad de cada variable independiente del modelo. Normalmente, la predicción tiene un número finito de resultados, como un sí o un no.
- Árboles de decisión**:** Los árboles de decisión generalmente se construyen de forma recursiva, siguiendo un enfoque de arriba hacia abajo **(**Sakthivel, N. et al., 2010). El acrónimo TDIDT, que significa Inducción descendente en árboles de decisión, se refiere a este tipo de algoritmo. Un árbol estándar inducido con C5.0 (o posiblemente ID3 o C4.5) consiste en un número de ramas, una raíz, un número de nodos y un número de hojas (Tan, P. et al., 2011). Una rama es una cadena de nodos desde la raíz hasta una hoja; y cada nodo implica un atributo. La aparición de un atributo en un árbol proporciona la información sobre la importancia del atributo asociado. El procedimiento para formar el árbol de decisión y explotar el mismo se caracteriza por lo siguiente:
    1. El conjunto de características estadísticas extraídas de los estudios constituye la entrada al algoritmo; La salida es el árbol de decisión.
    2. El árbol de decisión tiene nodos de hoja, que representan etiquetas de clase, y otros nodos asociados con las clases (nivel de magnitud en este caso) que se están analizando.
    3. Las ramas del árbol representan cada valor posible del nodo de parámetro del cual se originan.
    4. El árbol de decisión se puede usar para expresar la información estructural presente en los datos comenzando en la raíz del árbol (el nodo más alto) y moviéndose a través de una rama hasta un nodo de hoja.
    5. El nivel de contribución de cada parámetro individual viene dado por una medida estadística dentro del paréntesis en el árbol de decisión. El primer

número entre paréntesis indica el número de puntos de datos que se pueden clasificar usando ese conjunto de parámetros. Los parámetros que aparecen en los nodos del árbol de decisión están en orden descendente de importancia.

6. En cada nodo de decisión en el árbol de decisión, se puede seleccionar el parámetro más útil para la clasificación utilizando los criterios de estimación apropiados. El criterio utilizado para identificar el mejor parámetro invoca el concepto de entropía y la ganancia de información que se analiza en detalle en las siguientes subsecciones. El algoritmo del árbol de decisión (C4.5) tiene dos fases: construcción y poda. La fase de construcción también se conoce como la "fase de crecimiento".

- Random Forest**:** Este método es de tipo “ensamblador”, está formado de un grupo de modelos predictivos que permiten alcanzar una mejor precisión y estabilidad del modelo. Estos proveen una mejora significativa a los modelos de árboles de decisión (Rani, T. et al., 2019). Un árbol de decisión también sufre de los problemas de sesgo y varianza. Es decir, cuánto en promedio son los valores proyectados diferentes de los valores reales (sesgo) y cuan diferentes serán las predicciones de un modelo en un mismo punto si muestras diferentes se tomarán de la misma población (varianza). De manera que se generan múltiples árboles y cada árbol da una clasificación (vota por una clase), el resultado es la clase con mayor número de votos en todo el bosque (forest).
- XGBoost**:** Para ajustar un dataset de entrenamiento utilizando XGBoost, se realiza una predicción inicial. Los residuales se calculan en función del valor predicho y de los valores observados (Chen and Gestrin, 2016). Se crea un árbol de decisión con los residuales utilizando una puntuación de similitud de los residuales. Se calcula la similitud de los datos de una hoja, así como la ganancia de similitud de la división posterior. Se comparan las ganancias para determinar una entidad y un umbral para un nodo. El valor de salida de cada hoja también se calcula mediante los residuales. Para la clasificación, los valores se calculan generalmente utilizando el registro de momios y probabilidades. La salida del árbol se convierte en el nuevo residual para el dataset, que se utiliza para construir otro árbol. Este

proceso se repite hasta que los residuales dejan de reducirse, o bien el número de veces especificado. Cada árbol subsiguiente aprende a partir de los árboles anteriores y no tiene asignado el mismo peso, a diferencia de cómo funciona Bosque aleatorio.

- Gradient Boosting**:** Un modelo Gradient Boosting está formado por un conjunto de árboles de decisión individuales, entrenados de forma secuencial, de forma que cada nuevo árbol trata de mejorar los errores de los árboles anteriores. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo. La flexibilidad de este algoritmo ha hecho posible aplicar boosting a multitud de problemas (regresión, clasificación múltiple...) convirtiéndolo en uno de los métodos de machine learning de mayor éxito. Si bien existen varias adaptaciones, la idea general de todas ellas es la misma: entrenar modelos de forma secuencial, de forma que cada modelo ajusta los residuos (errores) de los modelos anteriores.
- Redes neuronales**:** Una red neuronal es un método de la inteligencia artificial que enseña a las computadoras a procesar datos de una manera que está inspirada en la forma en que lo hace el cerebro humano. Se trata de un tipo de proceso de machine learning llamado aprendizaje profundo, que utiliza los nodos o las neuronas interconectados en una estructura de capas que se parece al cerebro humano. Crea un sistema adaptable que las computadoras utilizan para aprender de sus errores y mejorar continuamente. De esta forma, las redes neuronales artificiales intentan resolver problemas complicados, como la realización de resúmenes de documentos o el reconocimiento de rostros, con mayor precisión.
  Su arquitectura se basa en una serie de neuronas interconectadas en tres capas;
  Capa de entrada: La información del mundo exterior entra en la red neuronal artificial desde la capa de entrada. Los nodos de entrada procesan los datos, los analizan o los clasifican y los pasan a la siguiente capa.
  Capa oculta: Las capas ocultas toman su entrada de la capa de entrada o de otras capas ocultas. Las redes neuronales artificiales pueden tener una gran cantidad de capas ocultas. Cada capa oculta analiza la salida de la capa anterior, la procesa aún más y la pasa a la siguiente capa.

Capa de salida: La capa de salida proporciona el resultado final de todo el procesamiento de datos que realiza la red neuronal artificial. Puede tener uno o varios nodos. Por ejemplo, si tenemos un problema de clasificación binaria (sí/no), la capa de salida tendrá un nodo de salida que dará como resultado 1 o 0.

Es usada para distintas aplicaciones como: Visión artificial, reconocimiento de voz procesamiento de lenguaje natural o motores de recomendación.

- K-Means**:** Es un algoritmo de aprendizaje no supervisado que se utiliza para resolver los problemas de agrupación en el aprendizaje automático o la ciencia de datos. Para procesar los datos de aprendizaje, el algoritmo K-means en la minería de datos comienza con un primer grupo de centroides seleccionados al azar, que se utilizan como puntos de inicio para cada clúster, y luego realiza cálculos iterativos (repetitivos) para optimizar las posiciones de los centroides.

  Se detiene la creación y optimización de clústeres cuando:

  Los centroides se han estabilizado - no hay cambios en sus valores porque la agrupación ha sido exitosa.

  Se ha alcanzado el número definido de iteraciones.

8. SOFTWARE USADO EN EL CASO DE EJEMPLO.

Con el objetivo de facilitar la replicabilidad del siguiente ejemplo se insertan algunos códigos (R) en las siguientes secciones. R es un lenguaje y un entorno para la computación estadística y los gráficos. R está disponible como Software Libre bajo los términos de la Licencia Pública General GNU de la Fundación del Software Libre en forma de código fuente. Se compila y ejecuta en una amplia variedad de plataformas UNIX y sistemas similares (incluyendo FreeBSD y Linux), Windows y MacOS.

9. CASO DE EJEMPLO: PREDICCIÓN DE AVERÍAS EN BOMBAS SOPLANTES

En estos últimos años han sido numerosos los avances orientados a optimizar la gestión de los activos. Un número creciente de estudios muestra que la falta de mantenimiento conduce al deterioro progresivo de los equipos y, en última instancia, a fallas (Souris, 1992). Con un mayor envejecimiento, aumenta el riesgo de eventos que interrumpen las operaciones de la planta y pueden conducir a fallas, mientras que la

viabilidad de mantener la infraestructura disminuye (Yerri et al., 2017). A medida que los equipos envejecen, los costos de reparación y mantenimiento aumentan (Younis & Knight, 2010) y el consumo de energía aumenta (Rojas & Zhelev, 2012). Varios autores (Castellet-Viciano, et. Al., 2018; Rojas y Zhelev, 2012; Terrazas, Vázquez, Briones, Lázaro y Rodríguez, 2010; Zhang et al., 2015) han subrayado el problema de que las EDAR envejecen y aumentan el consumo de energía de los equipos, lo que dificulta la eficiencia.

El deterioro de la infraestructura no solo aumenta los costos de mantenimiento, sino que también implica preocupaciones de impacto ambiental y social (Kong y Frangopol, 2005). En este sentido, conocer el estado de los activos permite optimizar las distintas tareas de mantenimiento para prolongar la vida útil de los equipos y establecer reemplazos o renovaciones parciales con tal de asegurar un correcto funcionamiento. Si bien, existen numerosos factores que pueden afectar a la condición del activo, muchos de ellos están relacionados con las propias tareas dedicadas a mantener los activos en óptimas condiciones.

La proliferación de tecnologías de detección y monitorización ha permitido generar una gran cantidad de datos de los procesos productivos. Los datos, una vez analizados y procesados pueden aportar conocimientos valiosos sobre los procesos de fabricación, producción y los equipos. Las técnicas de aprendizaje automático (ML) permiten usar estos datos con el fin de ayudar a identificar y clasificar aspectos que pueden resultar de interés. El objetivo es analizar la existencia de posibles patrones que pueden estar relacionados con un objetivo establecido. Estos enfoques analíticos han permitido comprender con mayor profundidad las relaciones existentes entre los factores de diseño, de operación y mantenimiento de los equipos y permitiendo, en consecuencia, la reducción de los costes dedicados a los mantenimientos, la reducción de averías en los equipos, en las paradas de reparación, inventarios, repuestos y el aumento de la vida útil de los activos mejorando la seguridad y sostenibilidad económica de los operadores (Carvalho et al., 2019, Sakthivel et al., 2010).

Entre todos los activos electromecánicos de las EDAR, las bombas soplantes son ampliamente utilizadas en las plantas de tratamiento de aguas residuales. Su objetivo es proporcionar un flujo de aire continuo al reactor, facilitando así la eliminación de la materia orgánica y los nutrientes que contiene el agua residual. Para planificar su

sustitución y anticiparse a posibles averías, es importante evaluar su estado o condición. Identificar las diferentes causas que aceleran el deterioro de estos equipos y producen averías es de gran importancia para los operadores, ya que pueden anticipar posibles roturas o averías. Entre los distintos aspectos podemos encontrar las horas de operación acumulada por el equipo, la existencia de vibraciones, si equipa sistemas automatizados tales como sistemas de parada, la mayor o menor frecuencia de los mantenimientos preventivos o aspectos técnicos como la presión y el caudal.

Los paquetes necesarios para manipular datos en R, generar gráficos, crear modelos, evaluarlos y crear los árboles de decisión son los siguientes (Código 1);

Código de desarrollo R 1: Instalación de paquetes necesarios

```
paquetes <- c('data.table',
              'dplyr',
              'tidyr',
              'ggplot2',
              'randomForest',
              'ROCR',
              'purrr',
              'smbinning',
              'rpart',
              'rpart.plot'
)
```

8.1 OBJETIVO Y JUSTIFICACIÓN

El objetivo del estudio es calcular la probabilidad de que una bomba se averíe con la consiguiente parada del equipo. Para ello se identifican aquellos factores que mayores implicaciones pueden tener el desgaste del equipo y la consecuente rotura. La muestra la componen un total de 400 bombas soplantes a lo largo de los últimos 10 años. El campo objetivo viene definido por una respuesta dicotómica (0) o (1), donde 0 significa que el equipo no ha sufrido avería y 1 que sí ha sufrido alguna avería. Con tal de conseguir una muestra balanceada se revisan los datos, el 43% de los equipos evaluados han sufrido alguna avería.

8.2 EXPLORACIÓN Y SELECCIÓN.

Siguiendo las fases comentadas en la sección 5, se carga la base de datos en el software, se seleccionan las variables que formaran parte del estudio y se obtiene un primer resumen (código 2).

- Las horas acumuladas de operación.
- Vibraciones.
- Mantenimiento: El historial de los mantenimientos preventivos realizados.
- Parámetros técnicos de operación.
- Sistemas de automatización y control.
- Presencia de arena en aguas residuales (corrosión).
- Existencia de cavitaciones.

Código de desarrollo R 2: Carga y lectura de datos

```
Nombre_BD <- fread('datos.csv')
as.data.frame(sort(names(df)))
str(Nombre_BD)
glimpse(Nombre_BD)
lapply(Nombre_BD,summary)
```

8.3 MODELADO DE DATOS: DISCRETIZACIÓN

La fase de discretización permite transformar variables numéricas en categóricas, de manera que facilitamos la relación de causa y consecuencia. Algunas variables presentan una salida dicotómica, por ejemplo, si trabaja en el óptimo o no, y otras requieren acotar diferentes intervalos, como es el caso de la vida útil de la bomba. Además, en este punto, se debe depurar valores, seleccionar las variables target y las predictoras del modelo.

A continuación, se discretizan las variables dando valores del 1 al 10 dependiendo de su menor o mayor influencia en el deterioro y mayor desgaste del equipo (ver tabla 1).

Tabla 1 Factores y discretización

Factores	Valor
Horas de operación	

Vida útil < 10%	10
10% > Vida útil < 25%	8
25% > Vida útil < 50%	6
50% > Vida útil < 75%	4
75% > Vida útil < Max.	3
Vida útil > Max	1
Vibraciones (*milímetros por segundo/media cuadrática*)	
< 2	10
2-4	7
4-5.5	5
5.5-9	3
>9	1
Mantenimiento (cumplimiento con recomendaciones del fabricante)	
100%	10
80%	8
60%	6
40%	4
20%	3
0%	1
Operación Óptima (caudal/presión)	
Dentro de diseño óptimo	10
Fuera de diseño óptimo	1
Automatización y sistemas de control	
Velocidad variable y sistema de parada automática.	10
Velocidad variable	5
Sin sistema	1
Presencia de arenas en aguas	
No	10
Si	1
Cavitaciones	
Leves	10
Moderadas	5
Elevadas	1

Con respecto a los mantenimientos preventivos, a partir de las recomendaciones del fabricante, se realiza un inventario de las tareas a realizar anualmente. Se dividen tareas

relacionadas con el desgaste del equipo, comprobaciones generales de los componentes y las revisiones completas (Hard-time). Se puntúan siguiendo la frecuencia recomendada por el fabricante, de modo que aquellos equipos que han realizado el total de las tareas anuales reciben la máxima puntuación. A continuación será necesario etiquetar las variables como categóricas y diferenciar la variable target (código 3).

Código de desarrollo R 3: Cambio a factores (categóricas) y discretización del target (fallo).

```
a_factores <- c('horas', 'vibraciones', 'mantenimiento', 'operacion',
'automatizacion','arenas', 'cavitaciones')
ind_larga <-names(Nombre_BD)
no_usar <- c('Fail')
"Fail"=factor("Fail, levels =c(0:1")
ind_larga<-setdiff(ind_larga,no_usar)
```

8.4 MODELO ML: ÁRBOLES DE DECISIÓN

El algoritmo del árbol de decisión se utiliza como clasificador para detectar fallos en la bomba centrífuga que impliquen parada del equipo (Sugumaran y Ramachandran, 2007; Sun et al., 2007). Este algoritmo puede realizar tanto la extracción como la clasificación de características simultáneamente. Por lo tanto, el algoritmo del árbol de decisión C4.5 se utiliza en este documento para el diagnóstico de averías y/o roturas de la bomba centrífuga.

El procedimiento de formar el árbol de decisión y evaluar la probabilidad de averías se caracteriza por lo siguiente:

1. El conjunto de características estadísticas extraídas de los estudios de vibración de bombas centrífugas monobloque forma la entrada al algoritmo; la salida es el árbol de decisión.
2. El árbol de decisión tiene nodos hoja, que representan etiquetas de clase, y otros nodos asociados con las clases (nivel de magnitud en este caso) que se analizan.
3. Las ramas del árbol representan cada valor posible del nodo de parámetro del que se originan.

4. El árbol de decisión se puede usar para expresar la información estructural presente en los datos comenzando en la raíz del árbol (el nodo más alto) y moviéndose a través de una rama hasta un nodo de hoja.
5. El nivel de contribución de cada parámetro individual viene dado por una medida estadística entre paréntesis en el árbol de decisión. El primer número entre paréntesis indica el número de puntos de datos que se pueden clasificar utilizando ese conjunto de parámetros. Los parámetros que aparecen en los nodos del árbol de decisión están en orden descendente de importancia.
6. En cada nodo de decisión del árbol de decisión, se puede seleccionar el parámetro más útil para la clasificación utilizando los criterios de estimación apropiados. El criterio utilizado para identificar el mejor parámetro invoca el concepto de entropía y ganancia de información.

A continuación, se añaden los códigos a usar en R (código 4 y 5)

Código de desarrollo R 4: Matriz de confusión y métricas de acierto, precisión, cobertura

```
confusion<-function(real,scoring,umbral){
  conf<-table(real,scoring>=umbral)
  if(ncol(conf)==2) return(conf) else return(NULL)
}
metricas<-function(matriz_conf){
  acierto <- (matriz_conf[1,1] + matriz_conf[2,2]) / sum(matriz_conf) *100
  precision <- matriz_conf[2,2] / (matriz_conf[2,2] + matriz_conf[1,2]) *100
  cobertura <- matriz_conf[2,2] / (matriz_conf[2,2] + matriz_conf[2,1]) *100
  F1 <- 2*precision*cobertura/(precision+cobertura)
  salida<-c(acierto,precision,cobertura,F1)
  return(salida)
umbrales<-function(real,scoring){
  umbrales<-
data.frame(umbral=rep(0,times=19),acierto=rep(0,times=19),precision=rep(0,times=19),co
bertura=rep(0,times=19),F1=rep(0,times=19))
  cont <- 1
  for (cada in seq(0.05,0.95,by = 0.05)){
    datos<-metricas(confusion(real,scoring,cada))
    registro<-c(cada,datos)
    umbrales[cont,]<-registro
    cont <- cont + 1
  }
  return(umbrales)
}
roc<-function(prediction){
  r<-performance(prediction,'tpr','fpr')
  plot(r)
}
auc<-function(prediction){
  a<-performance(prediction,'auc')
  return(a@y.values[[1]])
}
```

Código de desarrollo R 5: Creación de variable aleatoria, data frames e identificación de variables

```
Nombre_BD$random<-sample(0:1,size = nrow(df),replace = T,prob = c(0.3,0.7))
train<-filter(Nombre_BD,random==1)
test<-filter(Nombre_BD,random==0)
Nombre_BD$random <- NULL
independientes <- setdiff(names(Nombre_BD),c('Fail'))
target <- 'Fail'
formula <- reformulate(independientes,target)
```

Con los datos organizados e identificadas la variable dependiente y las variables explicativas, se crean las métricas de acierto, precisión y cobertura. A partir de este punto

se separan las muestras de la base de datos en grupos de entrenamiento y test. En el ejemplo se sigue el 30% para entrenamiento y 70% para el test. A continuación, se genera el modelo de árbol de decisión (código 6).

Código de desarrollo R 6: Modelo y gráfico árbol de decisión

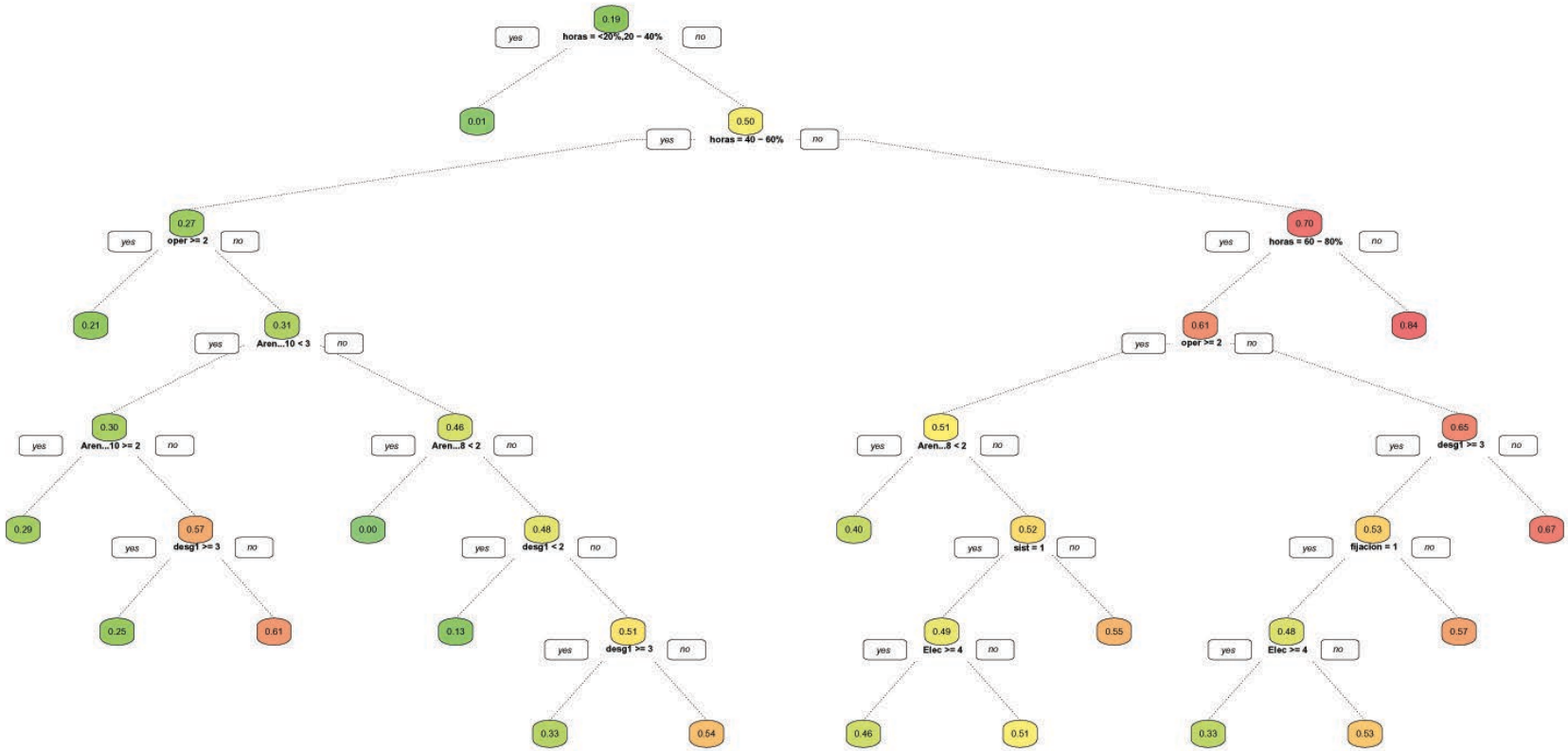
```
formula_ar <- formula
ar<-rpart(formula_ar, train, method = 'class', parms = list(
  split = "information"),
  control = rpart.control(cp = 0.00001))
printcp(ar)
plotcp(ar)
rpart.plot(ar,type=2,extra = 7, under = TRUE,under.cex = 0.7,fallen.leaves=F,gap =
0,cex=0.2,yesno = 2,box.palette = "GnYlRd",branch.lty = 3)
```

Y se evalúa el modelo mediante el cálculo de scorings, matriz de confusión y ROC (Código 8).

Código de desarrollo R 7: Evaluación del modelo

```
ar_predict<-predict(ar,test,type = 'prob')[,2]
confusion(test$Fail,ar_predict,umbral_final_ar)
ar_metricas<-filter(umb_ar,umbral==umbral_final_ar)
ar_metricas
ar_prediction<-prediction(ar_predict,test$Fail)
roc(ar_prediction)
ar_metricas<-cbind(ar_metricas,AUC=round(auc(ar_prediction),2)*100)
print(t(ar_metricas))
```

Ilustración 4 Ejemplo árbol de decisión (plot)



10. RESULTADOS

La metodología de árboles de decisión permite identificar aquellas variables (nodos) a partir de las cuales existe un cambio significativo hacia la variable objetivo (fallo). De manera que, observando la ilustración 4, el primer de los nodos que obtenemos es la variable "horas acumuladas =<20%,20 – 40%". Aquellos equipos que no cumplen con la condición, es decir, las bombas que superan el 40% de vida útil según diseño tienen una mayor probabilidad de sufrir una avería (50%). Este resultado viene explicado por el desgaste que provoca el propio funcionamiento de la bomba. Si bien, existen otros factores que pueden acelerar la fatiga y producir una avería. Siguiendo la ilustración, la siguiente variable sigue siendo el número de horas acumuladas, en este caso, sitúa la mayor probabilidad de avería para aquellas bombas soplantes que superan el 60% de vida útil. A partir de este punto, el árbol de bifurca en dos nodos, aquellas bombas que se sitúan entre el 40 y 60% de la vida útil puede sufrir una avería con una probabilidad del 27%. En cambio, aquellos equipos que están por encima de esa vida útil tienen una probabilidad de sufrir una avería mucho mayor (70%).

Siguiendo con el ejemplo, se observa que, la siguiente variable a tener en cuenta es el mantenimiento. Aquellos equipos que tienen una frecuencia de mantenimiento inferior al 20% recomendado por el fabricante tienen una probabilidad del 65% de sufrir una avería. Por otro lado, los equipos que tienen un mayor mantenimiento la probabilidad de avería disminuye hasta el 51%. El siguiente factor que puede generar mayor probabilidad de fallo es la presencia de arena en las aguas residuales. La arena tiene un efecto abrasivo que aumenta el desgaste de los componentes de la bomba. En este caso, la muestra analizada confirma que una mayor presencia de arena generaría un 52% más de probabilidades de sufrir una avería. Seguir los nodos del árbol permite entender que variables son las que mayor probabilidad acumulan y, en consecuencia, permite que el operador cambie la estrategia de mantenimiento con el objetivo de disminuir el número de averías de los equipos analizados.

La riqueza de los resultados que ofrece el árbol de decisión viene explicada por el gran número de alternativas generadas, cada alternativa está asociada a una probabilidad

de avería determinada. Este aspecto facilita al operador seleccionar aquellos escenarios donde es más interesante actuar. Por ejemplo, los escenarios que concentran mayor probabilidad de averías son los siguientes:

Escenario 1: Los equipos que superan el 80% de las horas de uso con una frecuencia de mantenimientos inferior al 30% recomendada por el fabricante y existencia elevada de cavitaciones tienen una probabilidad de avería del 67%.

Escenario 2: Los equipos que superan el 60% de horas de uso, tienen una frecuencia de mantenimiento inferior al 40% según recomendaciones del fabricante y sufren desgaste por presencia de arena en agua residual tienen una probabilidad de avería del 61%.

Escenario 3: Los equipos que superan el 80% de horas de uso, tienen una frecuencia de mantenimiento elevada y no sufren cavitaciones tienen una probabilidad de avería del 48%.

11. CONCLUSIONES

Las técnicas de aprendizaje automático (ML) permiten convertir en información una elevada cantidad de datos. El presente documento revisa las potencialidades de la digitalización en el sector de los recursos hídricos, concretamente en el ciclo urbano del agua. Explica brevemente las diferentes técnicas de aprendizaje e identifica los algoritmos más utilizados. Con tal de ser un documento práctico se focaliza en el beneficio potencial que puede generar el "*machine learning*" en la gestión de activos e infraestructuras. En concreto, se analizan la probabilidad de avería de un equipo electromecánico usado en el sector del ciclo urbano del agua.

La muestra la compone el historial de un total de 400 bombas soplantes funcionando durante 10 años, con tal de facilitar la comprensión se usa el algoritmo de los árboles de decisión. Es un modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable respuesta o dependiente, Consta de una serie de nodos que asemejan la estructura jerárquica de un árbol; raíz, ramas, nodos interno y nodos hoja.

A continuación, se analiza la base de datos con el objetivo de homogeneizar los datos e identificar posibles errores y se seleccionan las variables operacionales del historial de los equipos con tal de discretizar los valores. Como último paso, en lo que a preparación de datos se refiere, se identifica y discretiza el campo objetivo (Avería). Una vez aplicado el algoritmo, los resultados señalan que las horas de operación acumuladas son una variable importante a tener en cuenta, en concreto los equipos que superan el 80% de su vida útil presentan más probabilidad de avería. Sin embargo, equipos con menores horas de operación pueden averiarse por una menor frecuencia de mantenimientos, presencia de arena en las aguas residuales y la existencia de cavitaciones. Los resultados obtenidos permiten al operador generar un plan de acción que minimice las averías en este tipo de equipos, por ejemplo, aumentando la frecuencia de las tareas de mantenimiento pueden reducir los fallos y, en consecuencia, los costes económicos dedicados a su reparación.

Además, es importante señalar que estas metodologías son capaces de aprender por iteración, aumentando la calidad en la información que estos modelos son capaces de generar a partir de los datos. El aprendizaje de los modelos no se limita a un aspecto en concreto, la posibilidad de conectar nuevas fuentes de información y generar nuevas variables enriquece los resultados y abre nuevas vías de investigación. En conclusión, La aplicabilidad del “machine learning” en el sector de los recursos hídricos ocupa numerosas posibilidades.

REFERENCIAS

Adesanwo, M., Bello, O., Lazarus, S., & Denney, T. (2017). Smart alarming for intelligent surveillance of Electrical Submersible Pump Systems. In SPE Annual Technical Conference and Exhibition. OnePetro.

Al Aani, S.; Bonny, T.; Hasan, S.W.; Hilal, N. (2019). Can machine language and artificial intelligence revolutionize process automation for water treatment and desalination? Desalination, 458, 84–96.

Bishop, C. (2007). Pattern recognition and machine learning. New York, NY: Springer.

Breusch, T.S., Pagan, A.R. (1979). A simple test for heteroscedasticity and random coefficient variation. Econometrica, pp. 1287-1294

Brill, K. (2007). The invisible crisis in the data center: the economic meltdown of Moore's law. Uptime Institute White Paper, 7, 1-8

Chiumenti, A.; da Borso, F.; Pezzuolo, A.; Sartori, L.; Chiumenti, R. (2018). Ammonia and greenhouse gas emissions from slatted dairy barn floors cleaned by robotic scrapers. Res. Agric. Eng., 64, 26–33.

Carolan, M. (2017). Publicising food: Big data, precision agriculture, and co-experimental techniques of addition. Sociol. Rural. 57, 135–154.

Carvalho, T. P., Soares, F. A., Vita, R., Francisco, R. D. P., Basto, J. P., & Alcalá, S. G. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. Computers & Industrial Engineering, 137, 106024.

Castellet-Viciano, L., Hernández-Chover, V., & Hernández-Sancho, F. (2018). Modelling the energy costs of the wastewater treatment process: The influence of the aging factor. Science of the Total Environment, 625, 363-372.

Chen, R.C.; Dewi, C.; Huang, S.W.; Caraka, R.E. (2020) Selecting critical features for data classification based on machine learning methods. Jurnal of Big Data, 2020; 7: 52

Chen Tianqi y Guestrin Carlos. (2016). XGBoost: A scalable tree boosting system. En Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Coroama, V.C.; Mattern, F. (2019). Digital rebound–why digitalization will not redeem us our environmental sins. In Proceedings of the 6th International Conference on ICT for Sustainability, Lappeenranta, Finland, Volume 2382. Available online: http://ceur-ws.org (accessed on 14 October 2022).

Corominas, L.; Garrido-Baserba, M.; Villez, K.; Olsson, G.; Cortés, U.; Poch, M. (2018).Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. Environ. Model. Software, 106, 89–103.

de Sousa Jabbour, A.B.L.; Jabbour, C.J.C.; Foropon, C.; Godinho Filho, M. (2018). When titans meet–Can industry 4.0 revolutionise the environmentally-sustainable manufacturing wave? The role of critical success factors. Technol. Forecast. Soc. Change, 132, 18–25.

Driessen, C.; Heutinck, L.F. Cows. (2015). desiring to be milked? Milking robots and the co-evolution of ethics and technology on Dutch dairy farms. Agric. Hum. Values, 32, 3–20.

Duda, R.O.; Hart, P.E.; Stork, D.G. (2001). Unsupervised Learning and Clustering. Pattern classification, Wiley

Durbin, J., Watson, G.S. (1971). Testing for serial correlation in least squares regression. Biometrika, 58 (1), pp. 1-19, 10.1093/biomet/58.1.1

Evangelista, R.; Meliciani, V.; Vezzani, A. (2018). Specialisation in key enabling technologies and regional growth in Europe. Econ. Innov. New Technology, 27, 273–289.

Fennimore, S.A. (2017). Automated Weed Control: New Technology to Solve an Old Problem in Vegetable Crops. In Proceedings of the Conference Presentation at ASA Section: Agronomic Production Systems, Tampa, FL, USA.

Gujarathi, D. M. (2022). Gujarati: Basic Econometrics. McGraw-hill.

Gupta, A., Eysenbach, B., Finn, C., & Levine, S. (2018). Unsupervised meta-learning for reinforcement learning. arXiv preprint arXiv:1806.04640.

Haimi, H.; Mulas, M.; Corona, F.; Vahala, R. (2013). Data-derived soft-sensors for biological wastewater treatment plants: An overview. Environ. Model. Softw. 47, 88–107.

Hernández-Chover, V., Castellet-Viciano, L., & Hernández-Sancho, F. (2019). Cost analysis of the facilities deterioration in wastewater treatment plants: A dynamic approach. Sustainable Cities and Society, 49, 101613.

Kalymbek, B.; Yerkinbayeva, L.; Bekisheva, S.; Saipinov, D. (2021). The Effect of Digitalization on Environmental Safety. J. Environ. Manag. Tour. 12, 1299–1306.

Kleinbaum, D., L.L. Kupper, L.L., Muller, K.E. (1988). Applied Regression Analysis and Other Multivariable Methods. PWS-Kent Publishing Co., Boston

Kong, J.S. Frangopol, D.M. (2005). Probabilistic optimization of aging structures considering maintenance and failure costs. Journal of Structural Engineering, 131 (4) (2005), pp. 600-616, 10.1061/(ASCE)0733-9445(2005)131:4(600)

Li, L.; Rong, S.; Wang, R.; Yu, S. (2021) Recent advances in artificial intelligence and machine learning for nonlinear relationship analysis and process control in drinking water treatment: A review. Chem. Eng. J., 405, 126673.

Lioutas, E. D., Charatsari, C., & De Rosa, M. (2021). Digitalization of agriculture: A way to solve the food problem or a trolley dilemma? *Technology in Society, 67*, 101744. doi:https://doi.org/10.1016/j.techsoc.2021.101744

López, J., López, B., & Díaz, V. (2004). ALGORITMO DE APRENDIZAJE POR REFUERZO CONTINUO PARAEL CONTROL DE UN SISTEMA DE SUSPENSIÓN SEMI-ACTIVA. Revista Iberoamericana de Ingeniería Mecánica, 9(2), 77-91.

Lottes, P.; Hörferlin, M.; Sander, S.; Stachniss, C. (2017). Effective vision-based classification for separating sugar beets and weeds for precision farming. J. Field Robot. 34, 1160–1178.

Mnatsakanyan, A.G.; Kharin, A.G. (2021). Digitalization in the context of solving ecosystem problems in the fishing industry. In IOP Conference Series: Earth and Environmental Science; IOP Publishing: Bristol, UK. Volume 689, p. 012008.

Moy de Vitry, M.; Schneider, M.Y.; Wani, O.F.; Manny, L.; Leitão, J.P.; Eggimann, S. (2019). Smart urban water systems: What could possibly go wrong? Environ. Res. Lett. 14, 081001.

Pezzuolo, A.; Guo, H.; Marchesini, G.; Brscic, M.; Guercini, S.; Marinello, F. (2021). Digital Technologies and Automation in Livestock. Production Systems: A Digital Footprint from Multisource Data. In Proceedings of the 2021 IEEE International Workshop on Metrology for Agriculture and Forestry (MetroAgriFor), Perugia, Italy, 3–5 November 2021; IEEE: Piscataway, NJ, USA, pp. 258–262.

Rani, T.U.; Priyanka, C.H.S.; Monica, B.S.S. (2019) A dynamic data classification techniques and tools for big data. Journal of Physics: Conference Series, 2019, 1228: 1-12.

Richard, R.; Hamilton, K.A.; Westerhoff, P.; Boyer, T.H. (2020). Tracking copper, chlorine, and occupancy in a new, multi-story, institutional green building. Environ. Sci. Water Res. Technol., 6, 1672–1680.

Rojas, E. (2018). Glosario de los seis términos básicos del Machine Learning, Retrieved from: https://www.muycomputerpro.com/2018/02/07/glosario-terminosbasicosmachine-learning.

Rojas, J., Zhelev, T. (2012). Energy efficiency optimisation of wastewater treatment: Study of ATAD. Computers and Chemical Engineering, 38, pp. 52-63, 10.1016/j.compchemeng.2011.11.016

Sakthivel, N. R., Sugumaran, V., & Babudevasenapati, S. (2010). Vibration based fault diagnosis of monoblock centrifugal pump using decision tree. *Expert Systems with Applications*, *37*(6), 4040-4049.

Souris, J.P. Diaz de Santos, (Ed.), Maintenance, source of benefits, Les Editions d'organisation (1992) (In Spanish)

Sugumaran, V., Muralidharan, V., Ramachandran, K.I. (2007). Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing. Mechanical Systems and Signal Processing, 21, pp. 930-942

Sun, W., Chen, J., Li, J., (2007). Decision tree and PCA-based fault of rotating machinery. Mechanical Systems and Signal Processing, 21 (2007), pp. 1300-1317

Tan, P.N.; Steinbach, M.; Kumar, V. (2011) Introduction to Data Mining

Terrazas E., Vázquez., A., Briones, R., Lázaro, I., Rodríguez, I. EC treatment for reuse of tissue paper wastewater: Aspects that affect energy consumption. Journal of Hazardous Materials, 181 (1-3) (2010), pp. 809-816, 10.1016/j.jhazmat.2010.05.086

Yerri, S.R., Piratla, K.R., Matthews, J.C., Yazdekhasti, S. , Cho, J., Koo. D. (2017). Empirical analysis of large diameter water main break consequences. Resources, Conservation and Recycling, 123, pp. 242-248, 10.1016/j.resconrec.2016.03.015

Younis, R., M.A. Knight. (2010). A probability model for investigating the trend of structural deterioration of wastewater pipelines. Tunnelling and Underground Space Technology, 25, pp. 670-680, 10.1016/j.tust.2010.05.007

Zhang, X., Cao, J., Li, J., Deng, S., Zhang, Y., Wu. J. (2015). Influence of sewage treatment on China' s energy consumption and economy and its performances. Renewable and Sustainable Energy Review, 49, pp. 1009-1018