UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DSIC DEPARTAMENT DE SISTEMES INFORMÀTICS I COMPUTACIÓ

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Dpto. de Sistemas Informáticos y Computación

Diseño de estrategias de Machine Learning para la gestión de humedales

Trabajo Fin de Máster

Máster Universitario en Inteligencia Artificial, Reconocimiento de Formas e Imagen Digital

AUTOR/A: Casino Sánchez, Virginia

Tutor/a: Tavares de Araujo Cesariny Calafate, Carlos Miguel

Cotutor/a: Cecilia Canales, José María

CURSO ACADÉMICO: 2023/2024

Agradecimientos

Este trabajo ha sido parcialmente financiado por el proyecto de I+D TED2021-130890B-C22, de MCIN/AEI/10.13039/501100011033, y por NextGenerationEU/PRTR de la Unión Europea.

Quiero expresar mi gratitud a la Fundación ValgrAI por su gran trabajo y por la ayuda de estudio que me otorgaron al inicio del curso.

Agradezco a mis tutores, Carlos Tavares y Chema Cecilia, por su constante apoyo y orientación durante este proceso. Su experiencia y consejos han sido de gran ayuda y su compromiso con mi desarrollo académico ha sido muy valioso.

También agradezco a todo el profesorado que me ha enriquecido con sus conocimientos a lo largo de este año.

Por último, pero no menos importante, quiero agradecer a mi familia y seres queridos. Su apoyo incondicional, ánimo y comprensión han sido esenciales para mí en cada etapa de este trabajo. Gracias por su paciencia y el ánimo constante que me han brindado, permitiéndome alcanzar mis metas y superar los desafíos.

Resum

Este Treball fi de màster (TFM) s'enfoca en el disseny i implementació de models de Machine Learning (ML) per a la caracterització/modelització, en este cas concretament la temperatura de l'aigua, de la Llacuna de La Mata (Torrevieja, Alacant). Esta llacuna representa un espai natural protegit de gran interés turístic i mediambiental, i que està subjecte a diverses pressions antropogèniques. El desenvolupament d'este model és fonamental per a generar coneixement dels comportaments d'este entorn natural a curt, mitjà i llarg termini. Els objectius del TFM se centren, d'una banda, a conéixer, amb suficient antelació, els canvis abruptes en la temperatura de l'aigua. Per a esta part, amb l'ús de models de regressió, l'enfocament del projecte va començar amb una perspectiva univariant, veient-se posteriorment la necessitat de canviar a un plantejament multivariant que incloga variables exògenes com la temperatura ambiental i el nivell de l'aigua, ja que el que es perseguix és trobar el model més adequat per a identificar els canvis abruptes de la temperatura, en un horitzó no gaire llunyà, la qual cosa constituïx el nucli central de la primera part del treball. D'altra banda, i en relació amb el segon objectiu, s'ha plantejat estudiar l'evolució de la variabilitat de la temperatura al llarg del temps, creant una reconstrucció històrica d'esta, a partir de dades de la temperatura ambient des de 1950 fins hui. D'esta manera, es constatarà com la variació de la temperatura, i més concretament els canvis abruptes d'esta, esdevinguts durant els últims anys, afecten el canvi de comportament de les principals espècies d'aus que la visiten.

Paraules clau: Aprenentatge automàtic, temperatura de l'aigua, aiguamolls, predicció, reconstrucció històrica

Resumen

Este Trabajo Fin de Máster (TFM) se enfoca en el diseño e implementación de modelos de Machine Learning (ML) para la caracterización/modelización, en este caso concretamente la temperatura del agua, de la Laguna de La Mata (Torrevieja, Alicante). Esta laguna representa un espacio natural protegido de gran interés turístico y medioambiental, y que está sujeto a diversas presiones antropogénicas. El desarrollo de este modelo es fundamental para generar conocimiento de los comportamientos de este entorno natural a corto, medio y largo plazo. Los objetivos del TFM se centran, por un lado, en conocer, con suficiente antelación, los cambios abruptos en la temperatura del agua. Para esta parte, con el uso de modelos de regresión, el enfoque del proyecto comenzó con una perspectiva univariante, viéndose posteriormente la necesidad de cambiar a un planteamiento multivariante que incluya variables exógenas como la temperatura ambiental y el nivel del agua, ya que lo que se persigue es encontrar el modelo más adecuado para identificar los cambios abruptos de la temperatura, en un horizonte no muy lejano, lo que constituye el núcleo central de la primera parte del trabajo. Por otro lado, y en relación con el segundo objetivo, se ha planteado estudiar la evolución de la variabilidad de la temperatura a lo largo del tiempo, creando una reconstrucción histórica de la misma, a partir de datos de la temperatura ambiente desde 1950 hasta hoy. De esta manera, se constatará cómo la variación de la temperatura, y más concretamente los cambios abruptos de la misma, acontecidos durante los últimos años, afectan al cambio de comportamiento de las principales especies de aves que la visitan.

Palabras clave: Aprendizaje automático, temperatura del agua, humedales, predicción, reconstrucción histórica

Abstract

This Master's Thesis (TFM) focuses on the design and implementation of Machine Learning (ML) models for the characterization/modelization, in this case specifically the water temperature, of the Laguna de La Mata (Torrevieja, Alicante). This is a protected natural area of great tourist and environmental interest that is subject to various anthropogenic pressures. The development of this model is essential to generate knowledge of the behavior of this natural environment in the short, medium, and long term. The objectives of the TFM focus, on the one hand, on predicting, sufficiently in advance, the abrupt changes in water temperature. For this part, with the use of regression models, the project approach began with a univariate perspective, seeing the need to change to a multivariate approach that includes exogenous variables such as ambient temperature and water level, since the aim is to find the most capable model for identifying abrupt changes in temperature, in a not too distant horizon, which is the core of the first part of the work. On the other hand, and in relation to the second objective, it has been proposed to study the evolution of temperature variability over time, creating a historical reconstruction of the same, based on ambient temperature data from 1950 to the present. In this way, it will be shown how temperature variation, and more specifically the abrupt changes in temperature that have occurred in recent years, related to the behavioral changes of the main species of birds that visit the area.

Key words: Machine learning, water temperature, wetlands, prediction, historical reconstruction

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1
Introducción

Los Entornos Naturales Protegidos (ENPs) son zonas de gran importancia ambiental que requieren una gestión sostenible para garantizar su conservación a corto y largo plazo [39]. Estos espacios no solo ofrecen beneficios directos e indirectos a la sociedad, como señala [65], sino que también desempeñan funciones clave en tres ámbitos principales: i) como reguladores del clima a escala local y mundial, ii) como lugares de recreo y ocio para la comunidad, y iii) como elementos vitales para impulsar las actividades económicas [12]. Contribuyen ampliamente al bienestar humano, influyendo en aspectos como la salud, la libertad, la seguridad y las relaciones sociales, además de servir como fuentes de recursos naturales [32]. Sin embargo, Sarmento y Berger [69] advierten que la salud y el valor socioambiental de estos espacios están en riesgo debido a la presión humana, como la actividad turística, la contaminación y la sobreexplotación de los recursos naturales.

La conservación y gestión eficaz de estas zonas enfrenta retos como la limitación de recursos y la falta de datos a largo plazo, esenciales para el seguimiento de sus funciones ecológicas. La evaluación continua de las interacciones entre los sistemas ecológicos y sociales es crucial para la sostenibilidad de los ENPs y el mantenimiento de las actividades económicas en su interior, como indica [35]. Las tecnologías de la información se presentan como herramientas clave para la gestión del desarrollo industrial en estos espacios, destacando [37] su importancia para el seguimiento del ecoturismo y sus impactos, la promoción de la biodiversidad, y los principios del turismo sostenible.

La regulación de la Unión Europea en el Reglamento de Ejecución 2023/138 [18] proporciona un marco legal para el uso de datos ambientales de alto valor [30], facilitando la generación de conocimiento a través de servicios inteligentes basados en Machine Learning (ML). Este avance tecnológico sustenta el potencial para mejorar la eficiencia operativa de los ENPs, optimizando la asignación de recursos y prediciendo cambios ecológicos con mayor precisión. El uso del ML puede conducir a procesos de toma de decisiones más informados, integrando grandes cantidades de datos procedentes de diversas fuentes, como imágenes de satélite, sensores sobre el terreno, y datos ecológicos históricos [34].

Sin embargo, la recogida de datos de alta calidad de los ENPs, especialmente las situadas en regiones remotas o ecológicamente sensibles, presenta varios retos que pueden limitar (o incluso impedir) la eficacia de las aplicaciones de ML. Los ENPs suelen estar en lugares remotos que a menudo carecen de la infraestructura necesaria para soportar dispositivos de recogida de datos de alta tecnología, como fuentes de energía fiables o redes de comunicación seguras, lo que complica el despliegue y mantenimiento de estos sistemas. Además, la obtención de los permisos necesarios para instalar sensores u otros equipos de vigilancia en los ENPs puede ser un proceso complejo a nivel administrativo,

en el que intervienen múltiples partes interesadas, y una normativa estricta destinada a proteger la integridad ecológica de estas zonas.

Este trabajo se centra en la generación de datos de alta calidad del Parque Natural de las Lagunas de la Mata y Torrevieja, un notable ENP situado en Torrevieja, Alicante, en el sureste de España. Este ENP comprende dos lagunas principales: la Laguna de Torrevieja, conocida por sus actividades de extracción de sal, y La Mata, que se salvaguarda principalmente por su rica biodiversidad natural. La Mata destaca especialmente por sus importantes poblaciones de aves, que incluyen hasta 2.000 flamencos reproductores y 3.000 zampullines cuellinegros, lo que pone de manifiesto su importancia como lugar de cría y santuario de estas especies. La presencia de otras aves como cigüeñuelas, tarros blancos, aguiluchos cenizos, avoceta común, chorlitejo patinegro, charrán común, charrancito común y zarapito real subraya aún más el valor ecológico de la zona. La laguna también alberga una fauna única, como la gamba de salmuera, un crustáceo que prospera en ambientes muy salinos, lo que la convierte en un hábitat crítico para especies adaptadas a condiciones extremas. La flora que rodea las lagunas es igualmente diversa y está adaptada a las condiciones salinas. Mientras que el interior de las lagunas alberga una vegetación mínima, la periferia y las zonas menos salinas sustentan una variedad de plantas tolerantes a la sal como Arthrocnemum macrostachyum, algunas especies de Juncus, y géneros como Suaeda, Salicornia y Salsola. Destaca la presencia de la orquídea silvestre Orchis collina, que posee la mayor población de la Comunidad Valenciana y está clasificada como especie vulnerable. La zona sur de la laguna de la Mata presenta una vegetación típicamente mediterránea que incluye coscoja, pino carrasco, tomillo, y un bosque de repoblación con diversas especies de pinos y eucaliptos.

La generación de datos dentro de las lagunas de La Mata y Torrevieja se lleva a cabo mediante dos metodologías principales para garantizar tanto la precisión como la exhaustividad. En primer lugar, se han implementado varios sistemas de infraestructura *in situ* para capturar datos medioambientales de alta resolución y en tiempo real. Esto incluye el despliegue de sistemas básicos de boyas, colocadas estratégicamente para medir la temperatura del agua a dos profundidades y niveles de agua diferentes. Además, se ha establecido una estación meteorológica para controlar continuamente las condiciones atmosféricas. Estas instalaciones se complementan con el acceso a datos históricos proporcionados por la AEMET (Agencia Estatal de Meteorología), que ofrece un rico archivo de condiciones meteorológicas pasadas, crucial para los estudios longitudinales. En segundo lugar, el enfoque propio incorpora el análisis de varios modelos de ML para mejorar la predicción a largo plazo, y la reconstrucción de series temporales diarias de temperatura del agua, que se remontan a 1947. Estos modelos están diseñados para predecir cambios repentinos en la temperatura del agua mediante el aprendizaje a partir del amplio conjunto de datos que combina tanto los datos en tiempo real recogidos con los equipos de campo propios, como los registros históricos. Este enfoque dual permite una capacidad de respuesta inmediata, informada por datos en vivo, y permite una visión predictiva de las condiciones futuras basada en tendencias históricas, facilitando así una gestión más proactiva de los esfuerzos ecológicos y de conservación del parque natural.

1.1 Motivación

Las principales contribuciones del documento son:

- **Conjunto de datos integral**: Proporcionar un conjunto de datos sólido que integra vigilancia ambiental en tiempo real con amplios datos históricos, ofreciendo una visión sin precedentes de las condiciones ecológicas actuales e históricas de las la-

gunas. Este recurso es valioso para investigadores y conservacionistas interesados en cambios ambientales a largo plazo y sus repercusiones en la biodiversidad local.

- **Aplicación pionera de ML**: Ser pioneros en la aplicación de técnicas de ML para predecir y reconstruir series temporales de temperatura del agua en un ENP. Esta aplicación mejora la comprensión de la dinámica ecológica y las capacidades de predicción, permitiendo decisiones informadas para gestionar la salud ecológica de las lagunas.

- **Metodología de recogida de datos**: La metodología de recogida de datos, que utiliza infraestructuras *in situ* y modelos analíticos avanzados, establece un precedente para la vigilancia ambiental en otros ENPs. Este enfoque integrado garantiza la precisión y fiabilidad de los datos y ofrece un modelo escalable para otros ecosistemas sensibles y remotos en todo el mundo.

- **Contribución a la ciencia y la conservación**: La combinación de datos ecológicos detallados, aplicaciones tecnológicas innovadoras y estrategias de seguimiento escalables contribuye significativamente a la ciencia y la conservación del medio ambiente. Además, mejora la comprensión de nichos ecológicos específicos y ofrece herramientas y metodologías aplicables globalmente para preservar y proteger entornos naturales en una era de rápidos cambios ecológicos.

1.2 Objetivos

Los objetivos de este trabajo son:

- **Modelado de la temperatura del agua a largo plazo**: Desarrollar modelos de regresión capaces de predecir los valores de la temperatura del agua en un horizonte temporal lejano.

- **Identificación y modelado de cambios abruptos en la temperatura del agua**: Desarrollar modelos de regresión capaces de predecir cambios repentinos en la temperatura del agua.

- **Estudio de la evolución histórica de la temperatura**: Crear una reconstrucción histórica de la temperatura del agua desde 1947 hasta 2007 utilizando datos de temperatura ambiental. Este objetivo busca entender cómo las variaciones y cambios abruptos en la temperatura han afectado el comportamiento de las especies de aves o fauna que visitan y residen en el área.

1.3 Estructura de la memoria

La estructura de esta memoria es la siguiente:

- **Introducción**: El documento comienza con una introducción que presenta la motivación, los objetivos del estudio y una visión general de la estructura del trabajo.

- **Estado del arte**: Esta sección investiga aplicaciones de IoT (Internet de las cosas) y técnicas de ML en entornos naturales inteligentes (ENIs), revisando trabajos anteriores y destacando innovaciones y limitaciones.

- **Series temporales**: Se explica el concepto y las características de las series temporales, detallando los métodos y modelos para su análisis, así como los desafíos en su análisis.
- **Descripción del problema**: Se detalla la naturaleza de los datos recopilados, su preprocesamiento y análisis descriptivo. Se describen los modelos utilizados, como autorregresivos, clásicos y redes neuronales artificiales (RNAs), y las métricas de evaluación.
- **Modelado predictivo a futuro**: Esta sección aborda la aplicación de modelos para predecir la temperatura del agua, describiendo conjuntos de datos, experimentos y resultados.
- **Reconstrucción histórica**: Se describe la reconstrucción histórica de la temperatura del agua desde 1947 hasta 2007, detallando el experimento y sus resultados, y discutiendo su impacto en la biodiversidad.
- **Conclusiones**: Se resumen los principales hallazgos encontrados en el trabajo, destacando la innovación en técnicas de ML, la importancia de enfoques multivariantes, y la relevancia de datos históricos. Además, se recogen distintas propuestas para un trabajo futuro.
- **Apéndice A - Entorno de ejecución**: En este anexo, se describe el entorno técnico utilizado para realizar los análisis y experimentos, incluyendo herramientas y configuración del sistema.

CAPÍTULO 2
Estado del arte

Esta sección se enfoca en los estudios que tratan sobre los ENIs desde diversas perspectivas: enfoques basados en IoT, enfoques de ML y específicamente la predicción de la temperatura del agua. Al final, se presenta un resumen de las distintas contribuciones analizadas mediante una tabla que muestra los diferentes modelos de ML empleados.

2.1 Soluciones IoT transdisciplinares para las ENIs

La incorporación del IoT en la gestión de ENIs ha demostrado ser una herramienta valiosa para mejorar la monitorización, gestión y conservación de estos entornos. Diversos estudios y aplicaciones ponen de manifiesto la relevancia y el impacto positivo del IoT en estos contextos. La integración del IoT en ENIs aporta un valor significativo al mejorar la gestión y monitorización de estos entornos. Según Norouzi et al. [43], el IoT permite la recopilación y el análisis de datos en tiempo real a través de sensores distribuidos, lo que facilita la toma de decisiones informadas y oportunas para la conservación y el uso sostenible de los recursos naturales. Estas tecnologías optimizan la eficiencia en ámbitos como la gestión del agua, la vigilancia de la calidad del aire, y la conservación de la biodiversidad, lo que permite responder con rapidez a los cambios medioambientales y mejorar la investigación científica. El IoT también promueve la conectividad y el acceso a la información, facilitando la colaboración entre diversas partes interesadas, desde investigadores hasta gestores medioambientales y el público en general.

El estudio de Nundloll et al. [45] destaca cómo la implantación de una infraestructura de IoT en Conwy, al norte de Gales, puede superar retos como la conectividad limitada y las condiciones ambientales adversas. La integración de sensores de suelo, meteorológicos y rastreadores de ganado con análisis en la nube proporciona datos en tiempo real que permiten un seguimiento continuo y una gestión adaptativa. Esta infraestructura no solo mejora la comprensión de los ecosistemas, sino que también facilita la toma de decisiones sostenibles, lo que subraya la importancia de IoT para optimizar y proteger los entornos naturales.

En este sentido, [11] destaca la importancia del IoT en la creación de ENIs a través de un sistema de monitorización del agua. Este sistema, compuesto por sensores de nivel de agua, turbidez, pH y oxígeno, junto con módulos de control y comunicación, permite la monitorización dinámica y el ajuste de los niveles y la calidad del agua en tiempo real. La implementación de tecnologías IoT de bajo coste y bajo consumo energético, junto con algoritmos de inteligencia artificial (IA) para analizar datos históricos y previsiones meteorológicas, mejora significativamente la precisión en la medición y conservación de los recursos hídricos. Este enfoque no solo reduce los costes y mejora la precisión, sino

que también optimiza el uso de los recursos hídricos, contribuyendo a la sostenibilidad y la resiliencia de las ciudades inteligentes frente a fenómenos meteorológicos extremos.

Además, la implantación de redes de sensores IoT en ENIs, como se detalla en el caso del Parque Natural de Golija (Serbia), demuestra la importancia de esta tecnología para la gestión de catástrofes naturales, concretamente los incendios forestales. Según Novković et al. [44] la red de sensores IoT permite la recopilación continua de datos meteorológicos y medioambientales locales, mejorando significativamente la evaluación en tiempo real del riesgo de incendios. Métodos como el análisis multicriterio y el uso de SIG combinados con sensores IoT proporcionan una zonificación precisa de la susceptibilidad a los incendios, facilitando la implantación de sistemas de alerta temprana y estrategias de mitigación eficaces. Estos avances no solo optimizan la gestión del riesgo de incendios, sino que también contribuyen a la protección de zonas naturales sensibles y a la preservación de la biodiversidad.

El uso de la tecnología IoT también ha demostrado su eficacia en la restauración ecológica de las cuencas hidrográficas, ofreciendo una poderosa herramienta para la gestión y la protección del medio ambiente. El trabajo de Wang et al. [79] documenta en su estudio de la cuenca del río Jianghuai cómo la implantación de sensores inteligentes y algoritmos de optimización ha permitido comprender y gestionar mejor los recursos naturales. Los datos obtenidos mediante IoT facilitan la creación de bases de datos estructuradas para la supervisión en tiempo real, la alerta precisa, y la utilización científica de los recursos hídricos. La restauración ecológica, con el apoyo del IoT, mejora la capacidad de respuesta a los cambios ambientales, promueve la sostenibilidad, y mejora la conservación de la biodiversidad, demostrando ser una estrategia eficaz y eficiente en la recuperación de entornos naturales dañados.

Por último, la integración de las modernas tecnologías de la comunicación, en concreto el IoT, en ENIs, es crucial para la conservación de los recursos naturales en el sector agrícola. Vankayala et al. [78] describen cómo el IoT permite implantar sistemas de riego eficientes mediante el uso de sensores inalámbricos que controlan en tiempo real la humedad del suelo, la calidad del aire, y otras condiciones ambientales críticas. Esto facilita la optimización del uso del agua, reduciendo tanto el despilfarro por exceso de riego como los efectos negativos del riego insuficiente. Al automatizar y centralizar la gestión de estos datos, los agricultores pueden tomar decisiones informadas que no sólo mejoran la productividad agrícola, sino que también conservan los recursos hídricos esenciales para la sostenibilidad a largo plazo del ecosistema .

En conclusión, los estudios revisados demuestran que la tecnología IoT tiene un impacto significativo en la gestión de ENIs. Al proporcionar datos en tiempo real y facilitar la toma de decisiones basadas en información precisa, estas tecnologías no solo optimizan la conservación de los recursos naturales, sino que también mejoran la capacidad de respuesta ante desastres naturales y promueven la sostenibilidad a largo plazo. La implantación del IoT en distintos contextos, desde la gestión del agua hasta la agricultura, pone de relieve su versatilidad y su potencial para transformar la gestión medioambiental en el futuro.

2.2 Aprendizaje automático aplicado a ENIs

El ML ha demostrado un valor significativo en los ENIs, especialmente en la biología de la fauna salvaje y la gestión de los recursos naturales. Las aplicaciones de ML, como los modelos de distribución de especies y el reconocimiento de patrones, permiten una mayor precisión en la predicción y el análisis de datos ecológicos complejos y desorganizados. Esto no sólo optimiza la gestión y conservación de las especies, sino que facilita la

toma de decisiones basadas en datos concretos y actualizados, mejorando así la eficacia de las estrategias de conservación. Además, la capacidad del ML para procesar grandes volúmenes de datos de sensores y teledetección ayuda a monitorizar y gestionar ecosistemas a gran escala, proporcionando una poderosa herramienta para abordar los retos medioambientales contemporáneos [28].

En el contexto de la cartografía operativa de la vegetación natural potencial (VNP) a escala mundial, el estudio de Hengl et al. [27] evaluó la eficacia de varios algoritmos de ML, como las RNA, los bosques aleatorios, y los modelos de regresión. Utilizando un conjunto diverso de datos medioambientales y registros de presencia de especies, los bosques aleatorios demostraron ofrecer el mejor rendimiento predictivo. Esto subraya la relevancia del ML para mejorar la precisión y el detalle espacial en la modelización de la NPP, crucial para evaluar el potencial del suelo y concienciar sobre la degradación de la tierra en ENIs. Este enfoque permite la creación de mapas detallados y reproducibles, facilitando una mejor planificación y gestión medioambiental.

El uso de la IA en la planificación inteligente para la restauración ambiental de los ecosistemas terrestres ha demostrado ser crucial para abordar la degradación del suelo y preservar los recursos naturales. Un marco que incorpora modelos de ML permite la creación dinámica de Configuraciones de Recuperación Biológica (BRC) para evaluar y mitigar los impactos del crecimiento urbano en los ecosistemas. En el área metropolitana de Changsha Zhuzhou Xiangtan (CZX), la aplicación de redes bayesianas y del modelo Least Collective Resilience (LCR) ha permitido identificar y clasificar las fuentes ecológicas y los corredores medioambientales. Los resultados indican que la IA mejora la planificación medioambiental integrando factores topográficos, climáticos y socioeconómicos, optimizando la conservación ecológica, y equilibrando el desarrollo urbano con la sostenibilidad medioambiental [81].

El uso de técnicas de ML en ENIs, como las áreas marinas protegidas, también ha demostrado ser esencial para predecir y fomentar comportamientos proambientales entre los visitantes. Un estudio sobre el comportamiento ecológico de los turistas en Chipre utilizó modelos de análisis cualitativo comparativo de conjuntos difusos (fsQCA) y de sistema de inferencia neurodifusa adaptativo (ANFIS) para identificar patrones de comportamiento sostenible. Se puso de manifiesto que el ML no sólo permite predecir con precisión los comportamientos medioambientales, sino que también ayuda a diseñar experiencias turísticas que promuevan un compromiso medioambiental duradero, proporcionando herramientas valiosas para aplicar estrategias de conservación eficaces en recursos ecológicos sensibles [64].

Un estudio sobre la gestión de la sostenibilidad medioambiental en 19 naciones asiáticas entre 1990 y 2020 analizó cómo influyen la digitalización y el uso de los recursos naturales en las emisiones de CO_2 utilizando algoritmos adaptativos como las RNAs. Los resultados mostraron que la interacción entre las tecnologías de la información y la comunicación (TICs) y los recursos naturales puede moderar eficazmente los efectos negativos de la explotación de los recursos. Factores como la urbanización, el uso de la energía, y la complejidad económica son clave para predecir las emisiones de CO2. El estudio recomienda que los gobiernos den prioridad a la gestión sostenible de los recursos naturales mediante el uso de las TICs, fomentando la eficiencia energética y el consumo de energías renovables para alcanzar los objetivos de desarrollo sostenible relacionados con la energía y el clima [63].

En conjunto, estos estudios ilustran cómo el ML y la IA están transformando la gestión y conservación de ENIs. La capacidad de estas tecnologías para mejorar la precisión de los modelos ecológicos, facilitar la planificación medioambiental y fomentar comportamientos favorables al medio ambiente proporciona herramientas avanzadas para abor-

dar los retos medioambientales contemporáneos y mejorar la sostenibilidad a largo plazo. La integración de estos enfoques no sólo optimiza la conservación ecológica, sino que también equilibra el desarrollo urbano con la sostenibilidad medioambiental, proporcionando una referencia teórica y práctica para futuras estrategias de gestión y conservación del medio ambiente.

2.3 Predicción de la temperatura del agua

Conocer la temperatura futura del agua es crucial para la gestión y conservación de los ecosistemas acuáticos. La temperatura del agua influye en la calidad del hábitat, la salud de las especies, y los procesos ecológicos fundamentales. Una predicción precisa de la temperatura del agua permite aplicar estrategias de mitigación y adaptación al cambio climático, proteger la biodiversidad, y optimizar la gestión de los recursos hídricos. También ayuda a anticipar y responder a fenómenos extremos, mejorando la resiliencia de los ecosistemas y de las comunidades humanas que dependen de ellos.

El control y la predicción de la temperatura del agua son fundamentales para el éxito de las piscifactorías y la salud de los ecosistemas acuáticos, ya que la temperatura influye directamente en la fisiología y el comportamiento de los peces. En la producción europea de anguila, por ejemplo, se utilizaron regresiones múltiples y modelos ARIMA (Auto-Regressive Integrated Moving Average) calibrados con datos históricos de temperatura. Estos modelos demostraron una elevada precisión, con una varianza explicada superior al 95 %, y errores de predicción inferiores a 1ºC, lo que facilita la gestión operativa y la aplicación de medidas preventivas, mejorando la eficiencia de la producción acuícola [24].

Los algoritmos de ML han demostrado su eficacia en la predicción de la temperatura del agua gracias a su capacidad para modelizar relaciones no lineales complejas. En un estudio aplicado al río Tunga-Bhadra, en la India, se utilizaron técnicas como la regresión de cresta, K-vecinos más próximos, bosques aleatorios y regresión de vectores de soporte, combinadas con el análisis de sensibilidad global de Sobol y el filtro conjunto de Kalman. Los resultados indicaron que la temperatura máxima del aire es el predictor más sensible, siendo la regresión de vectores de soporte el modelo más robusto a escala mensual, mejorando significativamente las predicciones en sistemas fluviales tropicales [62].

En otro estudio, Abdi et al. [1] desarrollaron un modelo de red neuronal profunda multicapa para predecir la temperatura del agua de los ríos a partir de datos meteorológicos. Este modelo, validado con datos del río Los Ángeles, en el sur de California, incluía variables como la temperatura del aire, la humedad relativa, la presión barométrica, y la velocidad del viento. Los resultados mostraron que los modelos DNN (Deep Neural Network) multicapa superaron a los modelos de regresión lineal simple, reduciendo los errores absolutos medios en un 20 %, y mejorando el coeficiente de determinación en un 26 %.

Del mismo modo, Ikram et al. [31] evaluaron métodos mejorados de aprendizaje profundo para la predicción de la temperatura del agua en el río Bailong, China, utilizando redes neuronales convolucionales y de memoria a corto y largo plazo. Estos modelos se optimizaron utilizando algoritmos novedosos como el algoritmo de búsqueda reptiliana y el optimizador de media ponderada vectorial. La combinación de LSTM con el optimizador INFO (media ponderada de vectores optimizador), y el uso de variables como la temperatura del aire, el caudal, las precipitaciones, los sedimentos y el día del año, proporcionaron las predicciones más precisas, reduciendo los errores de predicción, y aumentando la eficacia en la estimación de la temperatura del agua.

El régimen térmico de los ríos, que afecta a la calidad del agua y a la distribución de las especies, está influido por factores atmosféricos, topográficos, de caudal y de cauce. La variabilidad de la temperatura del agua puede ser natural o el resultado de actividades humanas como la deforestación y el cambio climático. Los modelos para predecir la temperatura del agua se clasifican en modelos de regresión, estocásticos y deterministas. Los modelos deterministas utilizan un enfoque de balance energético, mientras que los demás se basan en las relaciones entre la temperatura del aire y la del agua. La comprensión de estos procesos es esencial para la gestión de la pesca y la evaluación del impacto ambiental, contribuyendo a la protección eficaz del hábitat acuático [9].

En resumen, la investigación sobre la predicción de la temperatura del agua ha avanzado considerablemente gracias a diversas técnicas de modelización y ML. Estos avances son cruciales para la gestión eficiente de las piscifactorías y la preservación de los ecosistemas acuáticos, ya que permiten tomar decisiones con conocimiento de causa y aplicar medidas preventivas eficaces.

En comparación con los estudios revisados, este trabajo difiere principalmente en su enfoque específico en la laguna de La Mata (Torrevieja, Alicante, España), un ENP de alto valor turístico y ambiental, frecuentemente sometido a diversas presiones antropogénicas. Mientras que los artículos revisados se centran en la predicción de la temperatura del agua en ríos o sistemas acuícolas mediante modelos y técnicas de ML, este trabajo aborda la modelización de la temperatura del agua en un entorno lagunar específico. Predecir la temperatura del agua en un río y en una laguna no es lo mismo debido a diferencias en el flujo del agua, la profundidad, el intercambio con el medio ambiente, las fuentes de aporte de agua y la ecología circundante. Los ríos tienen un flujo constante que mezcla continuamente el agua, creando una distribución más homogénea de la temperatura, mientras que las lagunas tienen aguas más estancadas, lo que permite una mayor influencia de la radiación solar y la temperatura del aire, resultando en una estratificación térmica. Además, los ríos suelen ser más superficiales y responder más rápidamente a los cambios de temperatura, mientras que las lagunas, con mayor volumen y profundidad, muestran cambios térmicos más lentos y graduales. La vegetación y la ecología también juegan roles distintos, afectando la temperatura en cada tipo de cuerpo de agua de manera única. Además, este trabajo no sólo pretende predecir la temperatura del agua en una laguna a largo plazo, considerando posibles cambios abruptos en sus valores, sino que también se propone reconstruir históricamente la variabilidad de la temperatura del agua desde 1947 hasta 2007. Esta reconstrucción histórica busca analizar cómo los cambios de temperatura han afectado el comportamiento de la fauna en la laguna o el calentamiento global, un aspecto que no se aborda en los estudios revisados, los cuales se centran en modelos de predicción a corto plazo, sin considerar la dimensión histórica ni el impacto sobre la fauna específica del entorno estudiado ni el tratamiento de cambios bruscos de los valores de la temperatura.

Para concluir esta sección, la Tabla 2.1 presenta una recopilación de las técnicas y modelos utilizados en los artículos revisados, organizados según los problemas que abordan.

Tabla 2.1: Resumen de los distintos modelos utilizados en los trabajos relacionados.

Artículo	Tema	Problema	Métodos
Yin et al. [81]	ENI + ML	Restauración ambiental en ecosistemas terrestres	BRC Redes Bayesianas Teoría de circuitos
Rai et al. [63]		Estudio adopción TIC mejora gestión recursos	AdaBoost Regressor Gradient Boosted Trees Random Forest Regressor ANN
Huettmann [28]		Predicción cambios distribución especies	Árboles de regresión Random Forest Técnicas de boosting
Rezapouraghdam et al. [64]		Predicción comportamiento ambientales en turismo y hospitalidad	Modelos ML ANFIS
Hengl et al. [27]		Mejorar precisión y detalles de los mapas de PNV globales	Redes neuronales Random Forest Gradient Boosting K-Nearest Neighbors Cubist
Abdi et al. [1]	Predicción temperatura del agua	Temperatura agua en ríos	Linear Regression DNN
Caissie [9]			Modelos de regresión Modelos estocásticos Modelos deterministas
Rajesh et al. [62]			Ridge Regression K-Nearest Neighbors Random Forest Support Vector Regression
Ikram et al. [31]			CNN LSTM RSA INFO
Gutiérrez-Estrada et al.[24]		Acuicultura	Modelo de regresión múltiple Modelo suavizado exponencial ARIMA

CAPÍTULO 3
Series temporales

Este capítulo se dedica al estudio y análisis de las series temporales, secuencias de datos obtenidos y ordenados cronológicamente a intervalos regulares o irregulares, que serán objeto de trabajo en este documento. Este tipo de datos es fundamental en numerosos ámbitos, incluyendo la economía, la meteorología, la ingeniería y las finanzas, por su capacidad para revelar patrones, tendencias y ciclos que de otro modo podrían pasar desapercibidos. El capítulo desglosará los componentes básicos de las series temporales, explorará las principales técnicas de modelado y discutirá la importancia y los desafíos inherentes a la interpretación de estos complejos conjuntos de datos [7, 41, 76].

3.1 Descripción

Una serie temporal es una secuencia de observaciones de una variable, típicamente cuantitativas, recolectadas y registradas a intervalos de tiempo sucesivos y ordenadas cronológicamente. Estos intervalos pueden ser regulares, como diarios, mensuales o anuales, o irregulares, dependiendo del fenómeno observado. El análisis de series temporales es crucial para identificar patrones, tendencias y comportamientos cíclicos y estacionales. Esto no solo proporciona una comprensión profunda del fenómeno estudiado, sino que también facilita la predicción y gestión de eventos futuros.

Una serie temporal es una colección de observaciones x_t ordenadas cronológicamente, donde t representa el tiempo en momentos específicos, como se muestra en la notación:

$$\{x_t\}, t \in T$$

Esta notación indica que los datos se recogen en los tiempos especificados por el conjunto (T). El tiempo puede ser de dos tipos:

- **Discreto**: Las observaciones se realizan en momentos específicos y pueden no ser equidistantes. Ejemplos incluyen registros diarios de temperatura o mediciones trimestrales del rendimiento económico.
- **Continuo**: Las observaciones se capturan de manera ininterrumpida a lo largo del tiempo. Este tipo de serie temporal es común en datos médicos como la monitorización continua de la presión sanguínea, donde métodos como la interpolación son necesarios para analizar los datos entre los registros capturados.

3.2 Características principales

Las series temporales poseen varias características clave esenciales para su análisis y modelado. Estas incluyen:

- **Tendencia (Trend)**: Es el movimiento de largo plazo que muestra la dirección general en la que se desplazan los datos. Este puede ser ascendente, descendente o constante. Las tendencias pueden adoptar formas lineales, como un aumento constante en las ventas a lo largo de los años, o no lineales, como el crecimiento exponencial en la base de usuarios de una plataforma digital. Estas tendencias se identifican y modelan a través de técnicas de suavización como medias móviles o métodos de descomposición.
- **Estacionariedad (Stationarity)**: Una serie es estacionaria si sus propiedades estadísticas, como la media, la varianza y la autocovarianza, permanecen constantes a lo largo del tiempo. La estacionariedad es esencial ya que muchos modelos de predicción suponen esta característica para garantizar estimaciones confiables. Métodos como la diferenciación se utilizan para transformar series no estacionarias en estacionarias, permitiendo un análisis más efectivo.
- **Estacionalidad (Seasonality)**: Refiere a los patrones o fluctuaciones que se repiten en intervalos regulares debido a influencias estacionales. La estacionalidad se modela incluyendo términos específicos en los modelos de series temporales para capturar estas variaciones periódicas.
- **Autocorrelación (Autocorrelation)**: Mide cómo las observaciones en una serie temporal se relacionan con sus valores pasados. Una alta autocorrelación indica una dependencia significativa entre observaciones consecutivas, lo cual es fundamental para construir modelos predictivos. Herramientas como la función de autocorrelación (ACF) y la función de autocorrelación parcial (PACF) son empleadas para investigar esta característica.
- **Ciclos (Cycles)**: Son oscilaciones irregulares influenciadas por dinámicas económicas, políticas o ambientales. A diferencia de la estacionalidad, los ciclos no presentan una periodicidad fija y pueden durar varios años, lo que los hace más desafiantes para predecir y modelar con precisión.
- **Ruido (Noise)**: Representa las variaciones erráticas en los datos que no se pueden explicar por la tendencia, la estacionalidad o los ciclos. El ruido puede deberse a errores de medición o factores aleatorios y externos. Es crucial identificar y filtrar el ruido para mejorar la precisión de los modelos de series temporales.
- **Nivel (Level)**: Es el valor promedio que asume la serie en un momento dado. Cambios en el nivel pueden influir significativamente en la apariencia de la serie temporal, modificando la percepción de tendencia y otros componentes.
- **Volatilidad (Volatility)**: Indica la intensidad de las variaciones en la serie temporal y es especialmente relevante en el contexto financiero para medir el riesgo y la incertidumbre en los mercados.

Estas características son fundamentales para comprender la estructura subyacente de una serie temporal y para desarrollar modelos predictivos precisos. El análisis de series temporales a menudo comienza con la identificación y cuantificación de estas características, proporcionando una base sólida para el modelado y la predicción. La correcta

descomposición y análisis de estas características permite mejorar la precisión y la interpretabilidad de los modelos de series temporales.

3.2.1. Tipos en base a la estacionariedad

Al analizar series temporales, una de las características fundamentales a considerar es la estacionariedad. Con base en esta propiedad, se pueden observar dos tipos principales de series temporales: estacionarias y no estacionarias. Esta clasificación es esencial ya que determina los métodos de análisis y predicción más adecuados, dado que cada tipo presenta características y desafíos específicos.

- **Estacionarias**: Una serie temporal se considera estacionaria si sus propiedades estadísticas, como la media y la varianza, son constantes a lo largo del tiempo. Las series estacionarias no presentan tendencias ni patrones estacionales fuertes, lo que facilita su modelado y predicción utilizando métodos estadísticos tradicionales. La estacionariedad es una condición deseable porque muchos métodos de análisis y predicción asumen que las propiedades del proceso generador de datos no cambian con el tiempo. Las series estacionarias son más fáciles de analizar debido a la simplicidad y predictibilidad de sus características estadísticas.

- **No estacionarias**: Las series no estacionarias presentan propiedades que varían con el tiempo, como tendencias (aumento o disminución) y estacionalidades (fluctuaciones periódicas). La no estacionariedad implica que las estadísticas descriptivas de la serie, como la media y la varianza, cambian con el tiempo, complicando su análisis y predicción. Para tratar con series no estacionarias, a menudo se aplican transformaciones como la diferenciación, que convierte una serie no estacionaria en una estacionaria al eliminar las tendencias a largo plazo. Un ejemplo común de serie no estacionaria es el crecimiento económico, que puede mostrar una tendencia ascendente a largo plazo debido a factores como la innovación tecnológica y el aumento de la productividad.

3.3 Métodos y modelos

El análisis de series temporales emplea una variedad de métodos y modelos para comprender y predecir los datos. A continuación, se describen algunos de los métodos y modelos más destacados:

3.3.1. Suavizado exponencial

El suavizado exponencial es una técnica que utiliza promedios ponderados de observaciones pasadas para realizar pronósticos. Este método tiene varias variantes:

- **Suavizado exponencial simple**: Adecuado para series sin tendencia ni estacionalidad.

- **Suavizado exponencial doble (Holt)**: Incorpora la tendencia además del nivel.

- **Suavizado exponencial triple (Holt-Winters)**: Añade un componente estacional al nivel y la tendencia.

Estos métodos son particularmente efectivos para pronósticos a corto plazo cuando se espera que los datos recientes sean más representativos del comportamiento futuro de la serie.

3.3.2. Descomposición de series temporales

La descomposición de series temporales separa una serie en componentes de tendencia, estacionalidad y ruido. Esta técnica permite analizar y modelar cada componente por separado, facilitando la identificación de patrones subyacentes y mejorando la precisión de los pronósticos. La descomposición puede realizarse mediante métodos clásicos o mediante la descomposición STL (Seasonal and Trend decomposition using Loess), que ofrece mayor flexibilidad al usar técnicas de suavizado local.

3.3.3. Pruebas de estacionariedad

Para asegurar que los modelos aplicados sean adecuados, es crucial realizar pruebas de estacionariedad, como la prueba Dickey-Fuller. Esta prueba ayuda a determinar si una serie temporal es estacionaria, lo cual es un requisito fundamental para la aplicación de ciertos modelos estadísticos.

3.3.4. Modelos autorregresivos y de media móvil

Entre los modelos más prominentes se encuentran:

- **Modelos autorregresivos (AR)**: Basados en la idea de que los valores presentes de la serie pueden explicarse por sus valores pasados.
- **Modelos de media móvil (MA)**: Modelan la serie temporal como la media de los términos de error observados recientemente. Estos modelos pueden combinarse para formar los modelos ARIMA (AutoRegressive Integrated Moving Average), que integran términos autorregresivos, de media móvil y de diferenciación para manejar la no estacionariedad en los datos. Los modelos ARIMA se pueden extender para incluir componentes estacionales, dando lugar a los modelos SARIMA (Seasonal ARIMA), útiles para datos con patrones estacionales significativos.

3.3.5. Modelos ARCH y GARCH

En el análisis financiero, los modelos ARCH (Autoregressive Conditional Heteroskedasticity) y GARCH (Generalized ARCH) son especialmente relevantes. Estos modelos se utilizan para prever la volatilidad de las series temporales, con los modelos GARCH extendiendo los modelos ARCH para capturar la volatilidad que cambia con el tiempo. Son particularmente populares para analizar y predecir la volatilidad en los datos financieros.

3.3.6. Modelos de Espacio de Estados y Filtros de Kalman

Los modelos de espacio de estados y los filtros de Kalman son herramientas poderosas para modelar y predecir series temporales en presencia de ruido y para manejar datos incompletos. Estos modelos permiten una representación flexible de la serie temporal y son especialmente útiles en situaciones donde los datos pueden ser ruidosos o incompletos.

3.4 Importancia del análisis de series temporales

El análisis de series temporales es crucial en diversas disciplinas debido a las siguientes razones:

- **Predicción y pronóstico**: Los modelos de series temporales permiten prever valores o eventos futuros basándose en datos históricos, lo cual es vital para la toma de decisiones en negocios, economía, finanzas, climatología y gestión de operaciones.
- **Detección de patrones**: Identifica tendencias y estacionalidades, facilitando la planificación estratégica y operativa, así como la interpretación de fenómenos temporales.
- **Control y monitoreo**: Es crucial en ingeniería y procesos industriales para el monitoreo continuo y el control de calidad, asegurando la eficiencia y detectando anomalías.
- **Análisis de impacto**: Evalúa el efecto de eventos específicos en las series temporales, proporcionando información valiosa para el desarrollo de políticas y estrategias.
- **Estrategias de negocio y operacionales**: Ayuda a las empresas a desarrollar estrategias alineadas con las tendencias identificadas, permitiendo anticipar cambios y adaptarse proactivamente.
- **Mejora del entendimiento de datos**: Facilita la comprensión de las dinámicas de los datos a lo largo del tiempo, ofreciendo una visión clara de los factores que influyen en los sistemas o fenómenos observados.

3.5 Desafíos en el análisis de series temporales

El análisis de series temporales enfrenta una variedad de desafíos que complican tanto el modelado como la interpretación de los datos. A continuación, se detallan los desafíos más comunes y las técnicas empleadas para abordarlos:

- **Datos faltantes**:
  - *Problema*: La ausencia de datos en ciertos puntos puede distorsionar el análisis y reducir la precisión de los modelos.
  - *Solución*: Se utilizan métodos de imputación como la imputación por media, regresión y vecinos más cercanos para reemplazar los valores faltantes con estimaciones basadas en otros datos disponibles.
- **Estacionalidad compleja**:
  - *Problema*: Algunas series temporales presentan patrones estacionales múltiples o irregulares, lo que dificulta su modelado con técnicas estándar.
  - *Solución*: Ajustar los modelos para considerar estos patrones puede implicar la descomposición de la serie en sus componentes estacionales. Los modelos SARIMA (ARIMA estacional) son una extensión que incorpora componentes estacionales para manejar series con patrones significativos.
- **Cambios estructurales y rupturas**:

  - *Problema*: Las series temporales pueden experimentar cambios abruptos debido a eventos externos, lo cual requiere modelos que se adapten dinámicamente a estos cambios.
  - *Solución*: Detectar y ajustarse a estos cambios es crucial. Los métodos pueden incluir la identificación de puntos de cambio y la recalibración de los modelos.

- **Estacionariedad**:
  - *Problema*: Muchas técnicas de modelado suponen que los datos son estacionarios, es decir, que sus propiedades estadísticas no cambian a lo largo del tiempo.
  - *Solución*: Convertir una serie no estacionaria en estacionaria puede requerir transformaciones como la diferenciación o el escalamiento para estabilizar la media y la varianza.

- **Dependencia temporal y autocorrelación**:
  - *Problema*: Los datos de series temporales son secuenciales, violando la suposición de independencia comúnmente requerida en otros análisis estadísticos. La autocorrelación puede llevar a estimaciones ineficientes de los parámetros del modelo.
  - *Solución*: Utilizar métodos específicos que consideren la autocorrelación, como modelos AR y MA.

- **Multicolinealidad**:
  - *Problema*: La multicolinealidad entre variables independientes puede complicar la estimación precisa de los parámetros y reducir la interpretabilidad de los modelos.
  - *Solución*: Identificar y manejar la multicolinealidad mediante técnicas como la regularización o la eliminación de variables altamente correlacionadas.

- **Ruido y errores de medición**:
  - *Problema*: Las series temporales pueden estar sujetas a errores de medición, introduciendo ruido en los análisis.
  - *Solución*: Filtrar y suavizar los datos para reducir el ruido, manteniendo la integridad de las predicciones y análisis.

- **Elección del modelo**:
  - *Problema*: Determinar el modelo más adecuado puede ser un desafío, especialmente cuando se consideran factores como estacionalidad, tendencia y ciclos.
  - *Solución*: A menudo es necesario combinar diferentes modelos para capturar todas las dinámicas de la serie, como los modelos híbridos que combinan ARIMA con modelos de ML.

- **Escalabilidad y rendimiento computacional**:
  - *Problema*: Modelar grandes conjuntos de datos de series temporales, especialmente a alta frecuencia o durante largos períodos, puede ser computacionalmente intensivo.
  - *Solución*: Utilizar métodos optimizados y algoritmos eficientes para el análisis y la predicción, aprovechando tecnologías de procesamiento paralelo y distribuidas.

Estos desafíos hacen que el análisis de series temporales sea un campo avanzado y dinámico dentro de la estadística y el ML, requiriendo enfoques y herramientas especializadas para manejar adecuadamente la complejidad de los datos temporales.

3.6 Resumen

En resumen, la exploración de series temporales ofrece perspectivas fundamentales sobre la naturaleza dinámica de los datos a lo largo del tiempo. Al desglosar y analizar tendencias, ciclos, estacionalidad y otros patrones, se obtiene una comprensión más profunda de los fenómenos estudiados, lo que es crucial para la toma de decisiones informadas y la predicción de eventos futuros. Los desafíos asociados con el análisis de series temporales, como la estacionariedad y la presencia de ruido, requieren el uso de técnicas avanzadas y modelos sofisticados, cuya correcta aplicación puede significativamente mejorar la precisión de los pronósticos y el entendimiento de las dinámicas subyacentes.

CAPÍTULO 4
Descripción del problema

En este capítulo se presenta el problema que se abordará en el trabajo y se ofrece una descripción detallada de los datos que se utilizarán. Se explicará el proceso de preprocesamiento aplicado a estos datos, seguido de un análisis de sus características. Asimismo, se introducirán los modelos que se implementarán y las métricas que se emplearán para evaluar su desempeño, proporcionando una visión completa del enfoque metodológico adoptado en este estudio.

La Figura 4.1 muestra el Parque Natural de las Lagunas de La Mata y Torrevieja, ubicado en Alicante, en el sureste de España. Este parque, que sirve como caso de estudio para este trabajo, se encuentra en la comarca de la Vega Baja del Segura y abarca los municipios de Torrevieja, Guardamar del Segura, Los Montesinos y Rojales, con una extensión de 3.743 hectáreas. Este entorno natural singular incluye dos lagunas principales: la laguna de Torrevieja, situada a la izquierda en la Figura 4.1, y la laguna de La Mata, a la derecha, separadas por una estructura anticlinal conocida como El Chaparral. Ambas lagunas están artificialmente conectadas al mar mediante canales, lo que facilita su explotación salinera. El característico color rosado del agua de la laguna de Torrevieja se debe a la presencia de Dunaliella, una microalga que prospera en ambientes salinos. Por otro lado, la laguna de La Mata, con su tonalidad más verde, presenta condiciones menos salinas que justifican su color. En esta figura se subrayan las características ecológicas contrastantes de las dos lagunas, las cuales forman parte de la diversa biosfera de la región.

Establecido como parque natural el 10 de diciembre de 1996, en virtud del Decreto 237/1996 de la Generalitat Valenciana, la zona tiene un gran interés ecológico y paisajístico. Las lagunas forman una parte crucial de un triángulo de humedales que incluye parques naturales vecinos como El Hondo y las Salinas de Santa Pola, desempeñando un papel clave en los ciclos biológicos de numerosas especies. Estas especies utilizan el parque durante las migraciones, la nidificación y la invernada, destacando su importancia como punto caliente de biodiversidad. Desde el punto de vista ecológico, el parque se caracteriza por sus hábitats salinos, como las lagunas y las salinas, así como por ambientes terrestres influidos por su alto contenido en sal. Los diversos hábitats también incluyen zonas de bajo relieve, pinares, arroyos de agua dulce, y tierras agrícolas en las que destacan los viñedos que cultivan uvas. Este rico mosaico de ambientes sustenta una amplia gama de flora y fauna, adaptada a las condiciones salinas. Las zonas interiores de las lagunas soportan una vegetación mínima debido a la alta salinidad, mientras que las periferias soportan especies tolerantes a la sal como Arthrocnemum macrostachyum, varias especies de Juncus y géneros como Suaeda, Salicornia y Salsola. La presencia de la vulnerable orquídea silvestre Orchis collina, principalmente a lo largo del borde sur de la laguna de La Mata, añade valor ecológico a la zona. Esta zona meridional también cuenta con vegetación típicamente mediterránea, como coscoja y pino carrasco, entre otras.

Figura 4.1: Vista aérea del Parque Natural de las lagunas de La Mata y Torrevieja. Imagen del repositorio de Google Earth.

Además, el parque es conocido por su avifauna, que incluye importantes poblaciones de flamenco común y zampullín cuellinegro. Otras especies aviares notables son la cigüeñuela, el tarro blanco, el aguilucho cenizo y la avoceta común. La existencia de la gamba Artemia salina, que requiere altos niveles de salinidad, subraya las características ecológicas únicas de las lagunas.

Este trabajo se centra en la digitalización de la laguna de La Mata, que forma parte del Parque Natural de las Lagunas de La Mata y Torrevieja, una zona en la que no se realizan actividades económicas. Entre las infraestructuras IoT desarrolladas para este estudio, destacan las siguientes:

- **Boya para medir la temperatura del agua**: Se desplegó un sistema de boyas para medir la temperatura del agua a dos profundidades diferentes. Dado que las lagunas son relativamente poco profundas, con una profundidad máxima de 1'5 metros, este sistema proporciona datos críticos para el seguimiento de las variaciones térmicas. Se compone de una robusta estructura esférica hueca de 800 x 1.610 mm, equipada con un ancla y una cadena para un amarre estable. Los componentes electrónicos de la boya incluyen una CubeCell - AB01 Dev-Board (V2), responsable de la adquisición y procesamiento de datos. Está alimentada por una batería de polímero de litio EEMB (3'7 V, 510 mAh, 532248) complementada por un panel solar redondo de 60 mm capaz de generar 0'28 W a 5 V, lo que garantiza un funcionamiento continuo. También cuenta con dos unidades de DS18B20, sensores digitales de sonda de temperatura de 5 m de longitud fabricados en acero inoxidable, diseñados para ser impermeables y duraderos en condiciones acuáticas. Por último, una carcasa impresa en 3D sirve para alojar y proteger los componentes electrónicos del medio acuático.
- **Mareógrafo para medir el nivel del agua**: Se instaló un mareógrafo para medir continuamente el nivel del agua en la laguna. Es importante señalar que el nivel del agua fluctúa debido principalmente a las transferencias de agua gestionadas entre las lagunas por la empresa salinera, y a las precipitaciones. Este sensor está gestionado por una CubeCell - AB01 Dev-Board (V2), que gestiona la recogida de datos y

el control de los sensores. Además, se alimenta con el mismo modelo de batería de polímero de litio EEMB que la boya (3'7 V, 510 mAh, 532248) y está equipada con un panel solar más pequeño (53x30 mm, 30 mA, 5 V) adaptado a sus necesidades energéticas. Un sensor ultrasónico de distancia MaxSonar MB7360-200 proporciona mediciones de los cambios del nivel del agua. La carcasa se basa en una caja de policarbonato con clasificación IP65 de 110x70x30 mm, que garantiza la durabilidad y la protección frente a las condiciones ambientales. Por último, este sistema se conecta a una estructura de soporte de madera hecha a medida, empleada para fijar el sensor en una posición óptima para una medición precisa del nivel.

- **Estación meteorológica Davis Vantage Pro2 con cable**: Esta estación meteorológica mide diversas variables meteorológicas a intervalos subhorarios, incluyendo: velocidad y dirección del viento, temperatura y humedad interior y exterior, sensación térmica y temperatura del punto de rocío, precipitaciones diarias, mensuales y anuales actuales y acumuladas, intensidad de las precipitaciones, presión atmosférica actual y tendencia, previsión meteorológica, fase lunar y horas de puesta y salida del sol.

Además, se construyó una LoRaWAN gateway utilizando la RAK7391 CM4 gateway, alimentada por un módulo de computación Raspberry Pi 4. La RAK7391 sirve como placa portadora para el módulo de computación Raspberry Pi 4 (CM4), y cuenta con varias ranuras para módulos de radio y WisBlock. Esta placa admite varias interfaces para satisfacer las diversas necesidades de los desarrolladores, como HDMI, Ethernet de 1 GB (con soporte PoE opcional), Ethernet de 2'5 GB, USB2 y USB3, mPCIe, CSI, DSI, M.2 y WisBlock.

El RAK7391 es versátil y puede utilizarse como pasarela, dispositivo de borde, u ordenador de uso general (véase la Figura 4.2). Admite hasta cuatro módulos LoRaWAN independientes, lo que permite crear un producto de LoRaWAN gateway multicanal/-multibanda. Las opciones flexibles de alimentación incluyen terminal DC, terminal Phoenix y PoE opcional. La placa cuenta con una interfaz de ventilador para la disipación del calor de la CPU, monitorización de la fuente de alimentación, y ultracondensadores para suministrar energía durante los cortes, garantizando la estabilidad del sistema. Este LoRaWAN gateway se desplegó centrándose en el centro de interpretación del Parque Natural de La Mata-Torrevieja, para dar soporte a los exhaustivos esfuerzos de recopilación y análisis de datos necesarios para una monitorización y gestión medioambiental eficaces.

Además, la Agencia Estatal de Meteorología (AEMET) proporcionó un amplio conjunto de datos históricos procedentes de una estación meteorológica situada estratégicamente dentro del parque natural. En funcionamiento desde 1947, este conjunto de datos incluye mediciones diarias esenciales para el análisis medioambiental a largo plazo. Las variables registradas incluyen temperatura, humedad, presión atmosférica, velocidad del viento, dirección del viento, precipitación total, intensidad de la precipitación, precipitación en la última hora, punto de rocío y sensación térmica. Este rico conjunto de datos permitió una evaluación exhaustiva de las tendencias y variaciones climáticas, proporcionando un contexto inestimable para los esfuerzos de seguimiento medioambiental dentro de las lagunas de La Mata y Torrevieja. Estos datos climáticos longitudinales son cruciales para correlacionar los cambios ecológicos con las condiciones meteorológicas a lo largo de casi un siglo, ofreciendo información sobre el impacto de los cambios climáticos en el ecosistema local.

A continuación, se explican en detalle los datos que se utilizarán en este estudio, los cuales se dividen en dos conjuntos: los datos recolectados por los dispositivos IoT, y los datos de AEMET. Estos conjuntos se han tratado por separado debido a sus diferentes

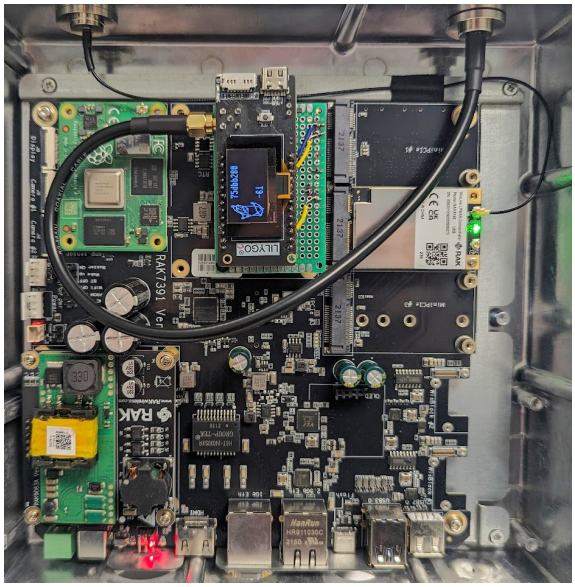

Figura 4.2: LoRaWAN Gateway basada en RAK.

periodos de datación. Los datos de los dispositivos IoT abarcan menos de un año, específicamente el año anterior, mientras que los datos de AEMET cubren aproximadamente 60 años, desde 1947.

4.1 Datos IoT

Dada la infraestructura IoT descrita anteriormente, se recopilaron tres conjuntos de datos, que son los distintos parámetros con los que se van a trabajar: temperatura del agua en el fondo y en la superficie, temperatura ambiente y nivel del agua de la laguna.

En la Tabla 4.1, se presenta un resumen general de los datos recopilados inicialmente. Esta incluye las fechas de inicio y fin de las mediciones, las unidades de medida, y los totales de registros recogidos, junto con la cantidad de filas que presentan valores faltantes o fechas duplicadas. A pesar de que el volumen total de datos no es extenso, es importante señalar que hay una cantidad significativa de registros incompletos y duplicados. Esto último reduce efectivamente la cantidad de datos útiles para análisis posteriores, y subraya la necesidad de revisar los procesos de recolección y manejo de datos para mejorar la integridad y precisión de la información ambiental recopilada.

Tabla 4.1: Resumen general de los datos de estudio.

Parámetro	Fecha		Medida	Total valores	Valores faltantes	Fechas duplicadas
	Inicio	Fin				
Tª fondo	06-01-2023	14-09-2023	ºC	7.843	110	1.516
Tª superficie			ºC		112	
Tª ambiente	26-09-2022		ºC	2.755	160	36
Nivel agua			cm	2.646	86	34

Además, se debe considerar que la frecuencia con la que se han tomado los datos varía significativamente entre los diferentes parámetros y periodos. Existen también intervalos de tiempo durante los cuales no se registraron datos, interrumpiendo la continuidad y potencialmente comprometiendo la calidad del análisis. Este patrón indica que existen

periodos de tiempo en los que no se realizó monitoreo, a pesar de las fechas de inicio y finalización registradas para cada parámetro.

La Tabla 4.2 muestra el proceso de monitorización de los parámetros, detallando las fechas específicas de inicio y finalización, así como las frecuencias de medición empleadas. Para los parámetros de temperatura de fondo y de superficie, las mediciones comenzaron el 06-01-2023, y hasta el 17-02-2023 se realizaron cada 30 minutos. Posteriormente, del 26-04-2023 al 04-06-2023, las mediciones se efectuaron cada hora. A continuación, del 16-06-2023 al 11-07-2023, la frecuencia aumentó a cada veinte minutos y, desde el 11-07-2023 hasta el 14-09-2023, se volvió a una frecuencia horaria. Por su parte, la temperatura ambiente se registró desde el 26-09-2022 al 11-07-2023 cada seis horas, y luego del 11-07-2023 al 14-09-2023 cada hora. El nivel de agua se controló desde el 26-09-2022 hasta el 16-06-2023 de forma diaria, y a partir del 16-06-2023 hasta el 14-09-2023, las mediciones se realizaron cada hora. Esta variabilidad en las frecuencias de medición, y los cambios en los periodos de registro, son cruciales para la interpretación y análisis de los datos ambientales recopilados.

Tabla 4.2: Datación y frecuencia de la monitorización de datos.

Parámetro	**Fecha**		**Frecuencia medición**
	Inicio	**Fin**	
Tª fondo Tª superficie	06-01-2023	17-02-2023	30 min
	26-04-2023	04-06-2023	1 hora
	16-06-2023	11-07-2023	20 min
	11-07-2023	14-09-2023	1 hora
Tª ambiente	26-09-2022	11-07-2023	6 horas
	11-07-2023	14-09-2023	1 hora
Nivel agua	26-09-2022	16-06-2023	1 día
	16-06-2023	14-09-2023	1 hora

4.1.1. Análisis descriptivo

En la Tabla 4.3 se muestra el análisis descriptivo estadístico de las variables. Con respecto a las medias, indican diferencias potenciales en escalas de medición o en las condiciones de medición entre las variables. Notablemente, las temperaturas de superficie y de fondo son comparativamente similares, mientras que la temperatura ambiente es ligeramente inferior.

La variabilidad en los datos, representada por la desviación estándar, es más pronunciada en la temperatura de superficie, seguida por la temperatura de fondo y el nivel, lo que podría reflejar fluctuaciones significativas o errores en la medición. Los valores extremos, especialmente el mínimo sorprendentemente bajo de -42’33 en la temperatura de superficie, sugieren errores o valores atípicos que requieren de revisión.

Los cuartiles indican una distribución asimétrica, con terceros cuartiles significativamente altos en comparación con las medias para las temperaturas, lo que puede señalar una concentración de valores altos. Esta tendencia es menos pronunciada para la temperatura ambiente y el nivel.

La presencia de valores extremos, y la distribución de los datos, subrayan la importancia de investigar estas anomalías para comprender mejor sus causas, que pueden incluir errores de medición o fallos instrumentales.

Tabla 4.3: Análisis descriptivo estadístico.

	Tª fondo	Tª superficie	Tª ambiente	Nivel
Total	6.271	6.215	2.559	2.344
Media	25′10	25′77	23′32	40′36
Desviación estándar	8′66	9′48	6′24	7′54
Mínimo	9′06	-42′33	4′28	-6′00
Primer cuartil	15′25	15′03	19′42	35′00
Segundo cuartil	29′50	30′31	25′60	40′00
Tercer cuartil	32′00	32′59	27′88	46′00
Máximo	42′08	41′14	35′28	56′20

Se han realizado histogramas para cada parámetro con el fin de observar la distribución de las frecuencias, y entender mejor la forma general y las tendencias de los datos recolectados. En la Figura 4.3 se puede sacar la siguiente información:

- **Temperatura del agua en el fondo**: Muestra una distribución bimodal de temperaturas en el fondo, con dos picos alrededor de los 15°C y 30°C. Esto sugiere que hay dos rangos comunes de temperatura en el fondo, lo que podría indicar dos tipos diferentes de ambientes o condiciones estacionales.
- **Temperatura del agua en la superficie**: También revela una distribución bimodal con picos alrededor de -20°C y 20°C. La presencia de temperaturas negativas tan extremas podría relacionarse con condiciones invernales o de áreas geográficas muy frías, mientras que el otro pico sugiere temperaturas típicas en condiciones más templadas o en verano.
- **Temperatura ambiente**: Tiene una distribución que tiende a ser normal centrada alrededor de 20°C, indicando que la mayoría de las mediciones de temperatura ambiente se agrupan en este rango, lo que puede ser típico de un clima templado.
- **Nivel del agua**: La distribución es multimodal, con tres picos prominentes alrededor de 10, 30, y 45. Esto podría indicar diferentes estados o niveles comunes de un cuerpo de agua, posiblemente relacionados con temporadas de lluvias o ciclos de sequía.

También se han llevado a cabo boxplots que se pueden observar en la Figura 4.4 con el objetivo de analizar la dispersión de los datos cuantitativos mediante la visualización de la mediana, los cuartiles, y detectar la presencia de valores atípicos. Se concluye lo que se indica a continuación:

- **Temperatura del agua en el fondo**: Es bastante estable, con una mediana cercana a los 30ºC. No presenta valores atípicos, lo que indica poca variabilidad en las mediciones.
- **Temperatura del agua en la superficie**: Varía más que en el fondo, con una mediana aproximadamente de 10ºC. Existen algunos valores atípico muy bajos, cercanos -40ºC, lo que en principio podría ser un error de medición.
- **Temperatura ambiente**: Muestra una mediana de aproximadamente 25ºC, pero se observa una colección de valores atípicos en la parte baja, cerca de los 5ºC o menos. Esto podría indicar varias ocasiones en las que las temperaturas descendieron significativamente, posiblemente debido a condiciones climáticas nocturnas o eventos

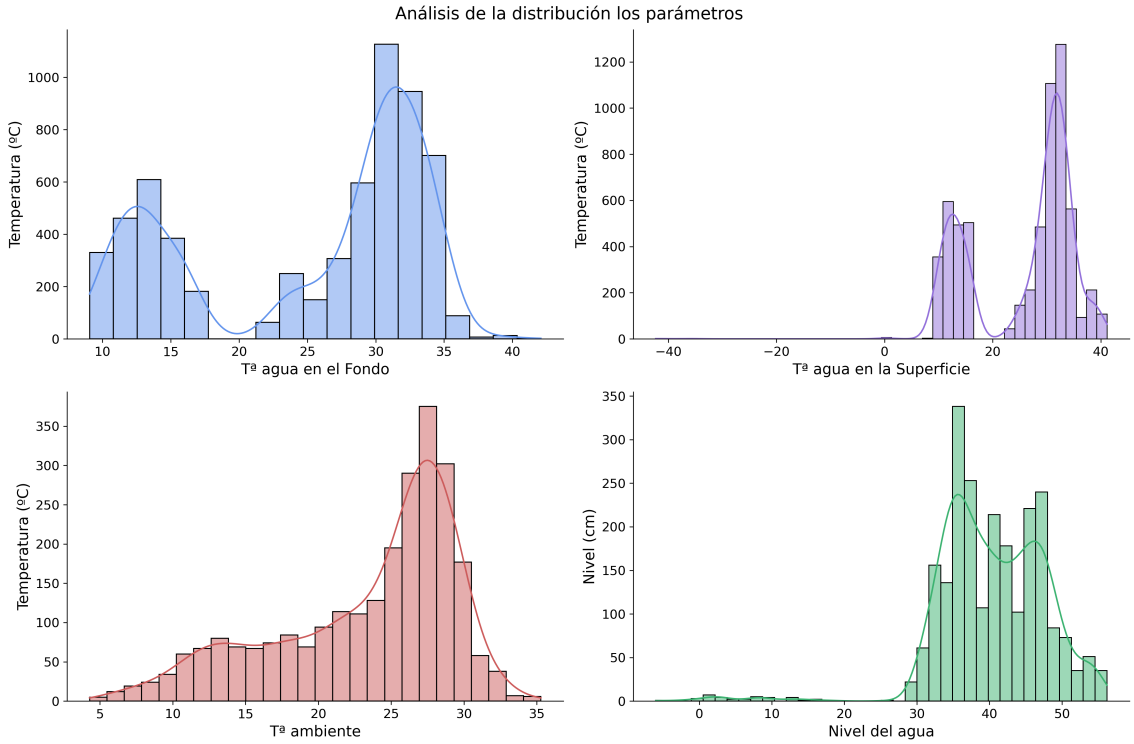

Figura 4.3: Distribución en frecuencia de los valores de los principales parámetros monitorizados.

climáticos inusuales. Esto sugiere que, aunque las temperaturas suelen ser estables, hay períodos de frío notable que podrían tener implicaciones para la vida vegetal y animal, así como para las actividades humanas en la región.

- **Nivel del agua**: Es relativamente constante, con una mediana cerca de los 40 cm. Los múltiples valores atípicos bajos sugieren fluctuaciones en el nivel del agua que podrían ser causadas por la evaporación, extracción de agua, o cambios en las precipitaciones.

Para finalizar con el análisis descriptivo se realiza una gráfica de dispersión de los datos a lo largo del tiempo (ver Figura 4.5). Unas primeras conclusiones que se pueden obtener son:

- **Estacionalidad**: Las temperaturas de la superficie y ambiente parecen exhibir una estacionalidad clara, con picos durante los meses de verano (de junio a agosto, aproximadamente) y valores más bajos durante los meses de invierno (de diciembre a febrero, aproximadamente). Esto es típico en climas templados donde las temperaturas fluctúan con las estaciones. Más adelante, tras todo el preprocesamiento, se observará con más detalle esta característica.
- **Consistencia de temperatura del fondo**: La temperatura en el fondo muestra poca variación a lo largo del tiempo, manteniéndose relativamente constante en comparación con las otras dos medidas. Esto es típico en ambientes acuáticos como lagos o mares, donde la temperatura en profundidades más grandes tiende a ser más estable.
- **Picos anómalos**: Hay ciertos puntos, especialmente en la temperatura ambiente, donde se observan picos muy altos o bajos que podrían representar eventos anómalos o errores en la medición. Sería importante verificar estos datos para asegurar su precisión.

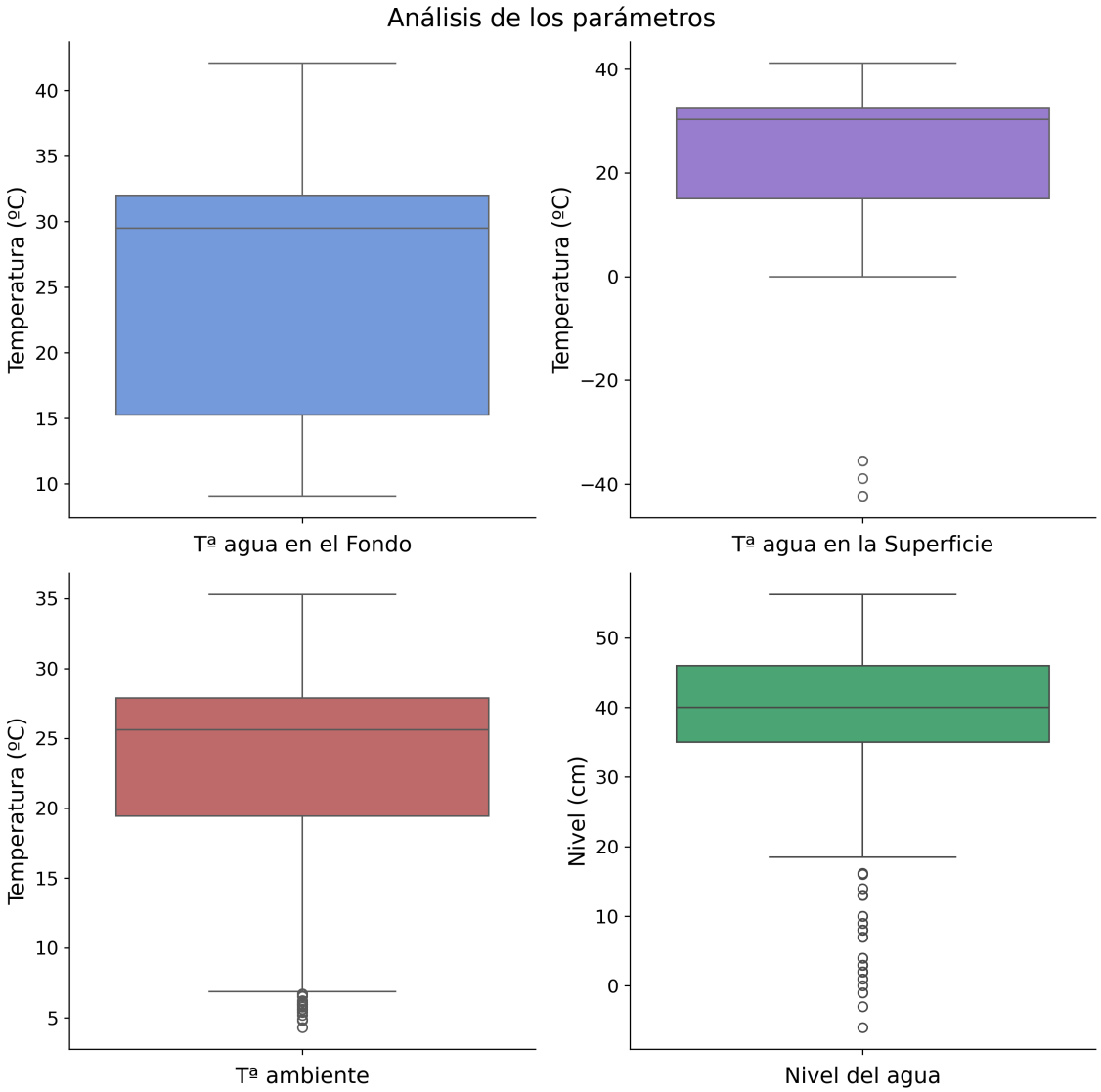

Figura 4.4: Boxplots relativos a los valores de los principales parámetros a estudiar.

- **Relación entre temperaturas de superficie y ambiente**: Se puede observar que las tendencias en las temperaturas de superficie y ambiente son bastante similares, lo que sugiere una fuerte correlación entre estas dos medidas. Es probable que el calor del ambiente sea transferido a la superficie del agua, o viceversa, dependiendo de las condiciones climáticas.

En resumen, los datos gráficos analizados revelan patrones importantes sobre las condiciones ambientales y las dinámicas climáticas en un entorno específico. A continuación se presenta una síntesis integrada de las conclusiones derivadas de estos análisis:

- **Patrones estacionales claros**: Las temperaturas de superficie y ambiente exhiben fluctuaciones estacionales, alcanzando picos en verano y mínimos en invierno. Esta estacionalidad marcada sugiere que ambos niveles térmicos están influenciados por cambios climáticos estacionales, lo que es crucial para comprender las dinámicas de un cuerpo de agua a lo largo del año.
- **Estabilidad en el fondo y ambiente**: A diferencia de la superficie, la temperatura en el fondo muestra menor variabilidad, indicativo de un ambiente más constante y

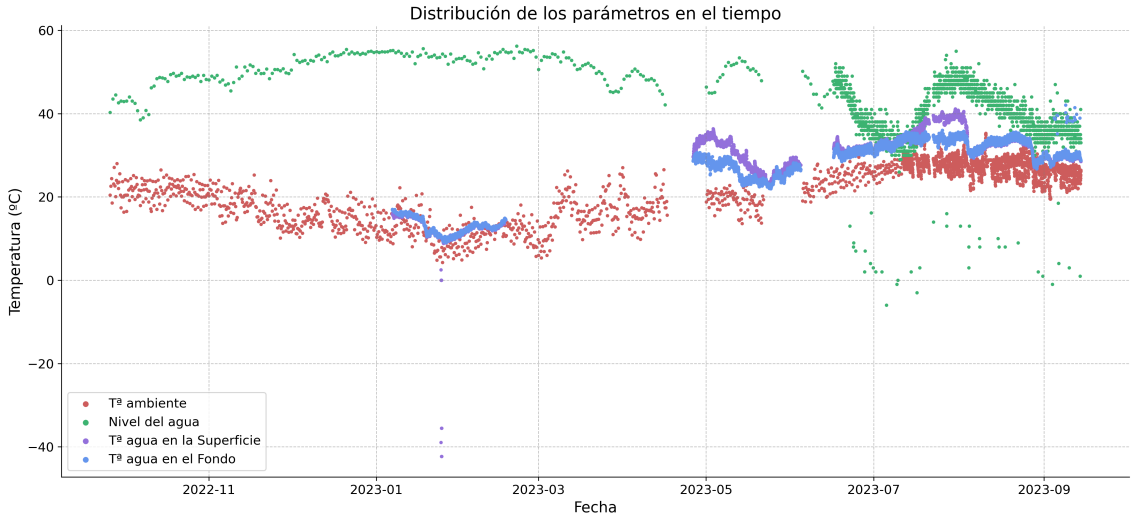

Figura 4.5: Dispersión de los valores en el tiempo.

estable. Esto es vital para especies y procesos ecológicos que requieren condiciones térmicas constantes. Similarmente, la temperatura ambiente tiende a ser estable, salvo por algunos valores atípicos que indican episodios esporádicos de frío extremo.

- **Interconexión entre superficie y ambiente**: La similitud en los patrones de temperatura entre la superficie y el ambiente sugiere un fuerte vínculo y un intercambio directo de calor, lo cual es fundamental para entender la interacción entre el aire y el agua.

- **Anomalías y variabilidad extrema**: Se observa una variabilidad significativa y valores extremos, especialmente en las temperaturas de superficie, reflejando la exposición a condiciones atmosféricas extremas y la influencia de factores microclimáticos. Estos hallazgos subrayan la necesidad de verificar la precisión de los datos, y explorar las causas subyacentes de estas anomalías.

4.1.2. Preprocesamiento

El proceso de limpieza de datos comenzó con la conversión de valores de cadena a numéricos para facilitar su análisis, seguido de la estandarización de todos los formatos de fecha para asegurar consistencia. Para las fechas duplicadas, se calculó el valor medio. En cuanto a los valores atípicos, no se hicieron cambios en la temperatura en el fondo por falta de valores atípicos. Para la temperatura en la superficie y el nivel del agua, los pocos valores atípicos identificados que estaban por debajo de un umbral preestablecido se reemplazaron por valores nulos. Por último, en la temperatura ambiente, se definieron los límites de los valores atípicos utilizando el rango intercuartílico y aquellos que estuvieran fuera de este se reemplazaron como en los otros parámetros por valores nulos. Una vez limpiados y tratados, se unifican todos los parámetros en un solo conjunto de datos (ver Figura 4.6).

En la siguiente fase del proceso es necesario unificar la frecuencia temporal y abordar los valores faltantes o las lagunas en los datos temporales. Si se observa la Figura 4.6, se ve que la disponibilidad de datos mejora significativamente a partir de mediados de junio de 2023. Adicionalmente, según se indica en la Tabla 4.2, a partir del 11 de julio de 2023, la frecuencia temporal se homogeneiza para todos los datos.

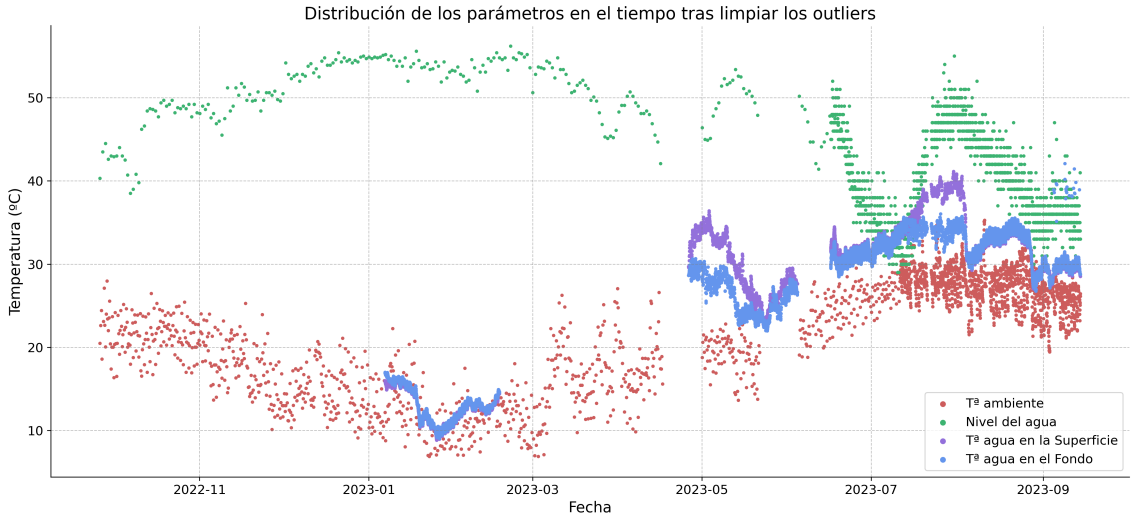

Figura 4.6: Datos tras la primera parte del preprocesamiento.

Aunque considerar únicamente los datos a partir del 11 de julio resulta en una muestra pequeña, se propone comenzar el análisis desde el 17 de junio de 2023. A pesar de que será necesario estimar algunos valores faltantes, se ha decidido utilizar los datos desde esta fecha por su consistencia temporal, y porque presentan el menor número de valores faltantes. Optar por este período minimiza la necesidad de grandes ajustes en los datos previos, lo que reduciría la precisión por la introducción de errores potenciales. No obstante, procede insistir en que algunas estimaciones seguirán siendo necesarias para evitar un tamaño de muestra demasiado reducido. Por consiguiente, el conjunto de datos a considerar será desde el 17 de junio de 2023 hasta el 14 de septiembre de 2023.

Con los datos organizados como se contempla en la Figura 4.7, se divisa que aún persisten algunos valores faltantes, aunque en menor cantidad que antes. Adicionalmente, hacia el final del conjunto de datos, es notable la presencia de picos en la temperatura del agua en el fondo. Estos no son representativos de valores reales, dado que los sensores se contaminaron en ocasiones, afectando la precisión de las mediciones.

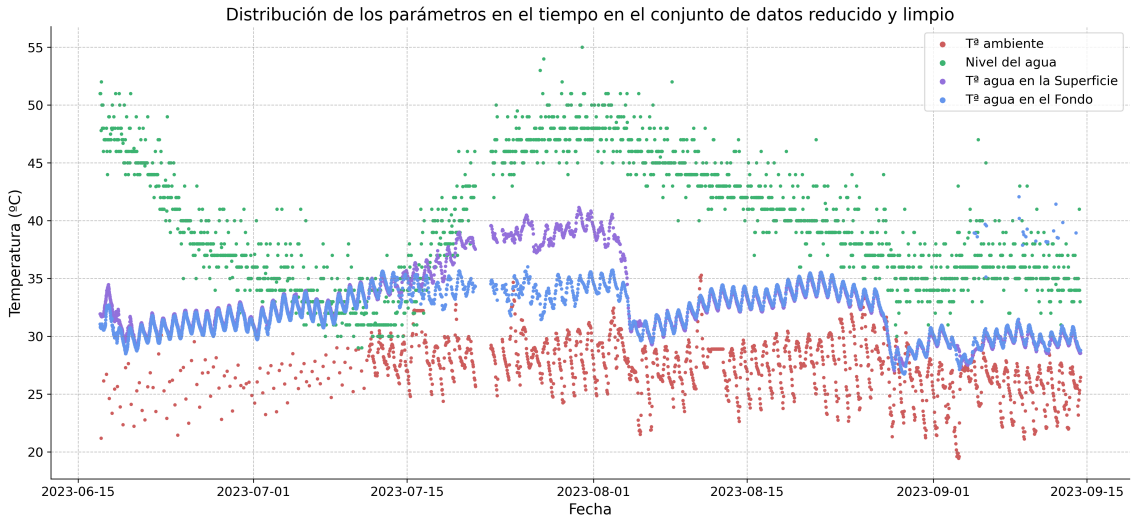

Figura 4.7: Reducción de los datos de estudio.

En este caso, se adoptará un enfoque diferente para el tratamiento de valores atípicos. Si se detecta un aumento considerable en la temperatura, específicamente un incremento

de al menos dos grados en menos de una hora, se sustituirá dicho pico por el valor de temperatura anterior.

Además, considerando que el objetivo principal de este estudio es la predicción de futuros valores de la temperatura del agua, se ha decidido unificar los datos de temperatura tanto del fondo como de la superficie. Se realiza mediante el cálculo del promedio de ambos, dado que, en ciertos segmentos, estos valores son prácticamente idénticos.

Antes de tratar los valores faltantes se va a unificar la frecuencia temporal de los valores ya que, desde el 17 de junio hasta el 11 de julio, es diferente (4.2). Por ello, como a partir del 11 de julio la frecuencia es horaria, se decide emplear esta porque hacerlo cada 20 minutos resultaría en que faltarían muchos datos y si, fuera diaria, el conjunto de datos se vería más reducido incluso. La manera de unificar valores en caso de que la frecuencia sea cada 20 minutos, para tener valores cada hora, se hace mediante la media de los diferentes valores para cada hora.

Por último, en esta fase de preprocesamiento se aborda el manejo de los valores faltantes. Para este propósito, se han aplicado diversos métodos y herramientas, incluyendo la interpolación lineal y temporal, la interpolación con spline cúbico, y la FFT (Transformada Rápida de Fourier). Sin embargo, estos enfoques no lograron resultados satisfactorios, lo que llevó a optar por una corrección manual de los datos. El conjunto de datos resultante se presenta en la Figura 4.8

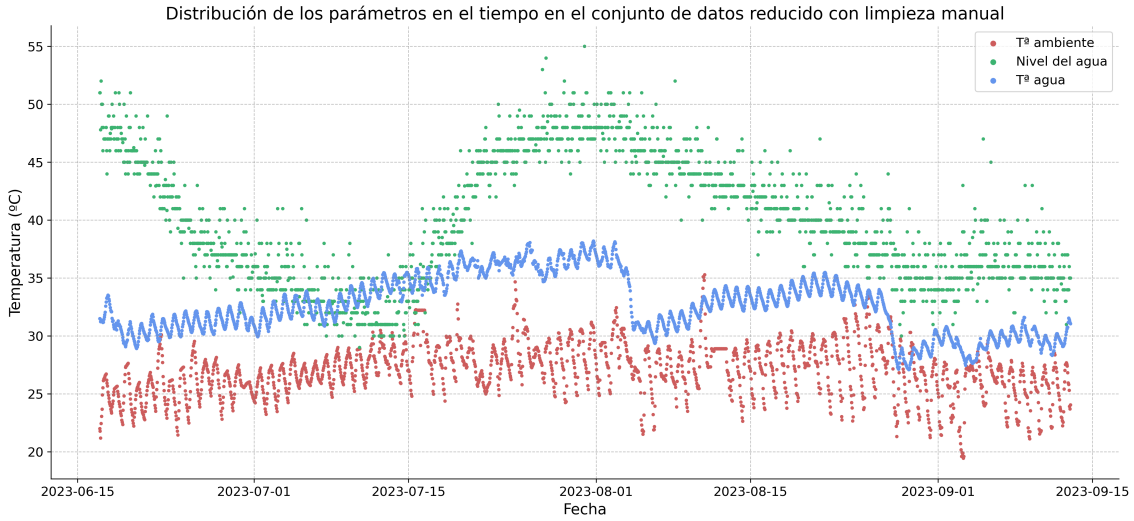

Figura 4.8: Datos para el estudio procesados.

Cabe destacar que el periodo del cual se tiene datos es solo para la época estival, por lo tanto, el análisis solo se puede realizar para la temporada de verano, ya que solo se dispone de datos del 16 de junio al 11 de septiembre, abarcando un total de 90 días. Por ende, con estos datos no es posible desarrollar un modelo climático anual generalizado, dado que las temperaturas invernales difieren significativamente de las estivales. Esto limitará la capacidad para predecir las variaciones climáticas que podrían ocurrir de septiembre a junio.

4.2 Datos AEMET

Respecto a dichos datos, procede señalar que, inicialmente, no se disponía de dichos datos en la primera etapa correspondiente al modelado predictivo futuro, por lo que se utilizan únicamente para la reconstrucción histórica. Además, es importante mencionar

que los datos han sido previamente procesados, asegurando así la ausencia de valores faltantes y valores atípicos.

El resumen general de estos datos se puede ver en la Tabla 4.4. Se trata de un histórico de la temperatura ambiente habida en Torrevieja desde el 1 de enero de 1947 hasta el 30 de noviembre de 2007, con una frecuencia de medición diaria y un total de 22.185 registros. Abarcan un total de casi 61 años, proporcionando una visión amplia y detallada de las condiciones climáticas a lo largo de varias décadas.

Tabla 4.4: Resumen general de los datos de AEMET.

Fecha		Medida	Total	Frecuencia
Inicio	Fin		valores	medición
01-01-1947	30-11-2007	ºC	22. 185	1 día

4.2.1. Análisis descriptivo

En la Tabla 4.5 se describe el análisis descriptivo estadístico de los datos de AEMET. La media es de 17'92°C, y la mediana es de 17'5°C. Esto sugiere que la distribución de las temperaturas es relativamente simétrica alrededor de su centro, ya que media y mediana son muy próximas. Con respecto a la desviación estándar, un valor de 5'68°C indica una variabilidad moderada en las temperaturas, lo que significa que, aunque la mayoría de las temperaturas están cerca de la media, también hay un rango considerable de temperaturas más frías o cálidas. En cuanto a los valores mínimos y máximos, las temperaturas varían desde -1'5°C hasta 31'8°C, lo que muestra un rango amplio de condiciones, desde heladas hasta días muy cálidos. Por último, más del 75 % de las temperaturas están por debajo de los 23°C, y más del 25 % están por debajo de los 13°C, indicando que las temperaturas frías son bastante comunes en este ambiente.

Tabla 4.5: Análisis descriptivo estadístico de la temperatura ambiente según datos AEMET.

	Histórico ambiente
Total	22.185
Media	17'92
Desviación estándar	5'68
Mínimo	-1'50
Primer cuartil	13'00
Segundo cuartil	17'50
Tercer cuartil	23'00
Máximo	31'80

La Figura 4.9 muestra el histograma y el diagrama de cajas y bigotes para estos datos. El histograma sugiere que la distribución de las temperaturas puede ser aproximadamente normal, aunque con alguna desviación, como se observa en los extremos más fríos y más cálidos. Respecto al boxplot, este confirma que no hay valores atípicos extremos en el conjunto de datos.

La variación a lo largo del tiempo muestra que, aunque hay fluctuaciones en las temperaturas, la mayoría de las mediciones se concentran entre 10°C y 30°C. Es notable que no hay una tendencia clara de aumento o disminución significativa a lo largo de las décadas, aunque sí se observan variaciones estacionales o anuales (ver Figura 4.10).

El datos de ambiente de AEMET experimenta una amplia gama de temperaturas a lo largo del año, desde condiciones de helada hasta días muy calurosos, indicativo de un

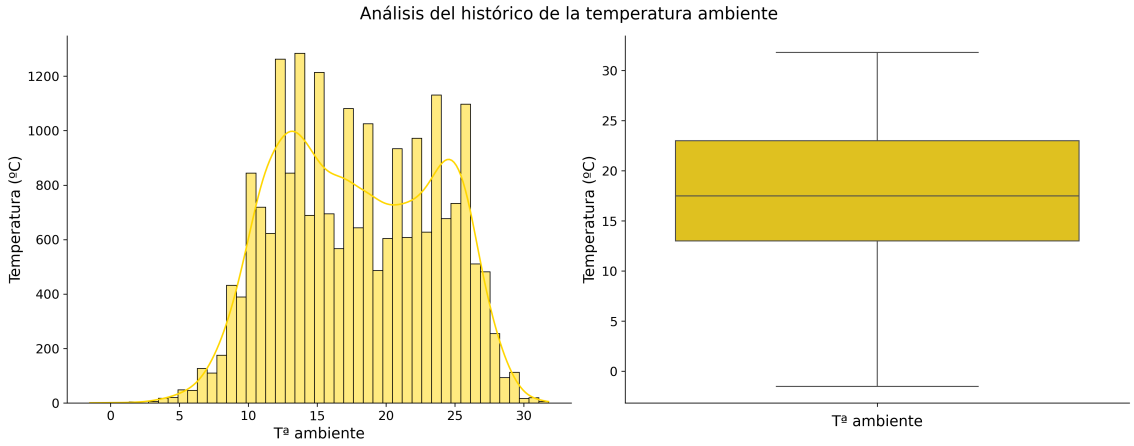

Figura 4.9: Distribución y dispersión de frecuencia de los datos de AEMET.

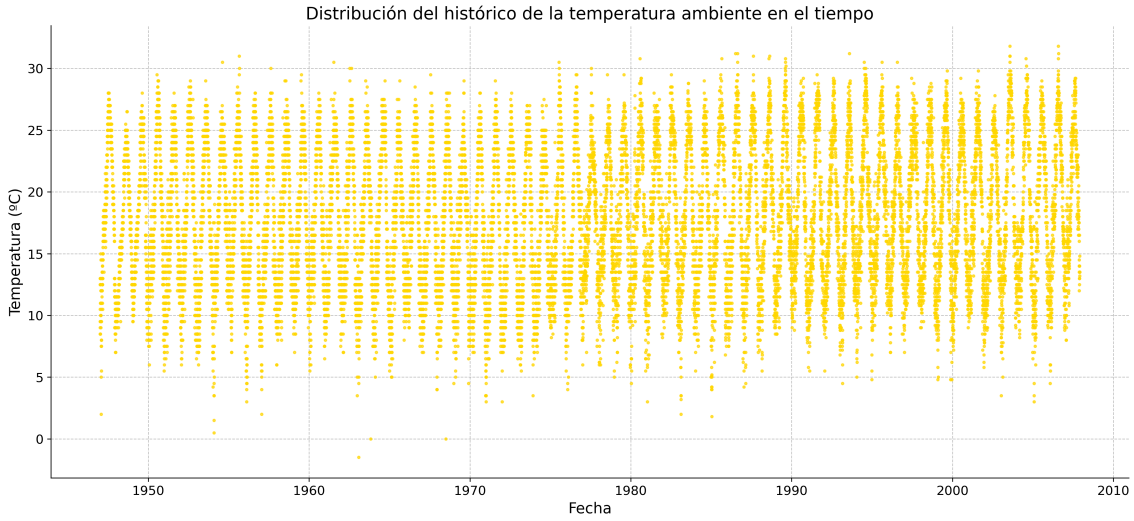

Figura 4.10: Dispersión datos AEMET en el tiempo.

clima continental, donde las diferencias de temperatura entre estaciones son marcadas, y la consistencia en la distribución de temperaturas a lo largo de las décadas sugiere que no han habido cambios drásticos en el clima de esta región, al menos en términos de temperaturas extremas.

4.3 Análisis de series temporales

Para conocer si las distintas series son estacionarias se ha aplicado el Augmented Dickey-Fuller [58]. Es una técnica estadística utilizada para determinar si una serie temporal es estacionaria, es decir, si la media, varianza y autocorrelación de la serie permanecen constantes a lo largo del tiempo. Esta prueba es una extensión de la prueba de Dickey-Fuller, diseñada para incluir una mayor cantidad de términos autorregresivos y diferencias de rezagos, mejorando así la capacidad de la prueba para manejar series temporales con estructuras de autocorrelación más complejas [68].

La temperatura del agua y el nivel del agua no son estacionarias, y la temperatura ambiente y el histórico de temperatura ambiente sí lo son. La temperatura del agua no es estacionaria principalmente debido a los cambios bruscos a principios y finales de agosto.

Se ha empleado el método *seasonal_decompose* de la librería *statsmodels* [72] para cada uno de los parámetros. Esta función se utiliza para descomponer una serie temporal en sus componentes básicos, lo que ayuda a entender mejor la estructura subyacente de los datos, y facilita el modelado y la predicción. Los componentes en los que típicamente se descompone una serie temporal son: tendencia, estacionalidad, y residuo. Se concluye lo siguiente para los datos de estudio:

- **Temperatura del agua**:
  - *Serie original*: Se pueden ver fluctuaciones diarias con un patrón general donde la temperatura tiende a disminuir ligeramente hacia finales de agosto.
  - *Tendencia*: Se nota un pico inicial en julio, seguido de una disminución gradual en la tendencia, lo que podría indicar una caída en la temperatura media a medida que avanza el verano hacia el otoño.
  - *Estacionalidad*: Muestra un patrón muy regular y repetitivo. Esto sugiere una variabilidad estacional típica, que probablemente refleja las diferencias entre las temperaturas diurnas y nocturnas.
  - *Residuos*: Son bastante bajos y consistentes, indicando que el modelo captura bien la mayoría de las variaciones en los datos de temperatura.
- **Temperatura del ambiente**:
  - *Serie original*: Parece ser más volátil que la serie de temperatura anterior, con fluctuaciones más agudas y frecuentes.
  - *Tendencia*: También presenta altibajos, con picos notables a mediados de julio y una caída a finales de agosto, aunque con un patrón menos claro que en la gráfica de temperatura del agua.
  - *Estacionalidad*: Similar a la de temperatura del agua, con patrones claros y consistentes que se repiten, indicativo de variaciones periódicas (probablemente diarias).
  - *Residuos*: Son más dispersos y fluctuantes que en la gráfica de temperatura del agua, aunque todavía relativamente bajos y estables. Esto sugiere que el modelo ajustado captura bien la mayoría de las variaciones, aunque con algo más de error residual que el modelo de temperatura.
- **Nivel del agua**:
  - *Serie original*: Aumenta inicialmente, alcanza un máximo en mediados de julio, y luego muestra una tendencia descendente con algunas fluctuaciones hacia el final.
  - *Tendencia*: Incremento general hasta mediados de julio, seguido por un descenso gradual. Esta tendencia suave sugiere un cambio significativo en los niveles durante el periodo estudiado.
  - *Estacionalidad*: Variaciones cíclicas consistentes en el nivel, lo que podría corresponder a ciclos diarios o a factores periódicos influenciados por condiciones ambientales externas.
  - *Residuos*: Son bastante bajos y consistentes, indicando que el modelo de descomposición capta adecuadamente las características principales de los datos.
- **Histórico ambiente**:

- *Serie original*: Parece ser bastante volátil a lo largo de todo el período, con altas y bajas frecuentes, lo que podría indicar variaciones estacionales o respuestas a eventos ambientales específicos.
- *Tendencia*: Aunque hay fluctuaciones, no hay un cambio notable en la tendencia a lo largo de las décadas, lo que indica que los valores promedio de estos datos ambientales no han cambiado drásticamente durante este tiempo.
- *Estacionalidad*: Es constante y casi invariable la componente estacional. Esta uniformidad sugiere un patrón repetitivo fuerte y estable en los datos a lo largo del año, probablemente vinculado a cambios estacionales o ciclos naturales.
- *Residuos*: Son bastante dispersos, con algunos picos notables, lo que indica que hay variaciones no capturadas completamente por el modelo de tendencia y estacionalidad.

Por lo tanto, la temperatura del agua presenta un comportamiento estacional fuerte y una tendencia general de disminución hacia el final del período estudiado. La baja variabilidad en los residuos sugiere que el modelo de descomposición es adecuado para explicar la mayoría de las fluctuaciones en la temperatura, con las variaciones diurnas y nocturnas, capturadas efectivamente en la componente estacional. La temperatura ambiente exhibe una fuerte regularidad estacional y una tendencia significativa, aunque con más variabilidad inexplicada que la temperatura, lo que podría reflejar la influencia de más factores ambientales o de medición. En cuanto al nivel del agua, los datos tienen una fuerte componente estacional y una tendencia general bien definida. Los residuos bajos indican que el modelo es adecuado, sugiriendo que los niveles son predecibles basados en patrones estacionales y de tendencia. Por último, el histórico de la temperatura ambiental muestra una gran estabilidad en la tendencia a lo largo de muchas décadas, así como una estacionalidad pronunciada y consistente. Los residuos dispersos sugieren que ocasionalmente hay factores o eventos que el modelo básico no captura completamente.

También se han realizado las gráficas de la ACF [57] y la PACF para los distintos parámetros [56].

La ACF es utilizada para medir la correlación lineal entre una serie de datos y sus valores en rezagos anteriores. Esta función es esencial para identificar la dependencia temporal dentro de una serie temporal, es decir, cuánto un valor en un tiempo dado es influenciado por sus valores previos. Además, la ACF es clave para detectar la presencia de estacionalidad en los datos, ya que patrones repetitivos a intervalos regulares en la ACF indican ciclos estacionales.

La PACF, por su parte, ofrece una medida de correlación entre una serie y sus rezagos que excluye los efectos de los rezagos intermedios. Esto proporciona una visión clara de la relación directa entre observaciones separadas por varios intervalos de tiempo, sin la interferencia de las correlaciones que involucran a esos intervalos intermedios. La PACF es fundamental para determinar el número de términos a utilizar en modelo AR, ya que ayuda a identificar cuántos rezagos pasados tienen un impacto significativo en los valores actuales de la serie, facilitando así la construcción de modelos precisos y eficientes [29].

La información que se obtiene es la siguiente:

- **Temperatura del agua**: Es probable que las mediciones estén influenciadas por patrones estacionales (como cambios estacionales en la temperatura) o por factores físicos que inducen dependencias autocorrelativas (como la capacidad del agua para retener calor). Se podría utilizar un modelo ARIMA, ajustado para capturar tanto la dependencia inmediata como la estructura de largo plazo de la serie temporal.

- **Temperatura del ambiente**: Sugieren que se podría considerar un modelo ARIMA estacional para capturar estas dinámicas. Un modelo como SARIMA podría ser particularmente útil aquí para incorporar componentes estacionales explícitos que parecen estar presentes.

- **Nivel del agua**: Un modelo AR podría ser un buen punto de partida para modelar estos datos, pero dado el lento decaimiento en la ACF, podrías considerar probar también modelos con más rezagos, o incluso un modelo ARIMA con componentes integrados si la serie no es estacionaria.

- **Histórico ambiente**: Un modelo SARIMA podría ser más apropiado. Este tipo de modelo permite especificar componentes estacionales que capturan las autocorrelaciones en distintas partes del año.

Por último, se procede a analizar visualmente los tres parámetros recopilados por los dispositivos IoT, excluyendo los datos de AEMET (Figura 4.11). Al comparar las temperaturas del agua y del ambiente, se observa un aumento en ambas durante los meses de verano, lo cual es típico. No obstante, la temperatura ambiente presenta picos más altos, a lo mejor debido a días especialmente cálidos o a influencias climáticas extremas. Esto sugiere que las mediciones del ambiente son más susceptibles a las condiciones diarias y a eventos meteorológicos, mientras que la temperatura del agua es más estable y menos afectada por cambios extremos diarios. En cuanto al nivel y la temperatura del agua, no se observa ninguna relación aparente. Finalmente, es importante destacar que los cambios bruscos, también conocidos como *shocks*, en los valores de la temperatura del agua, representan un desafío significativo para la predicción, como se va a ver a lo largo del trabajo.

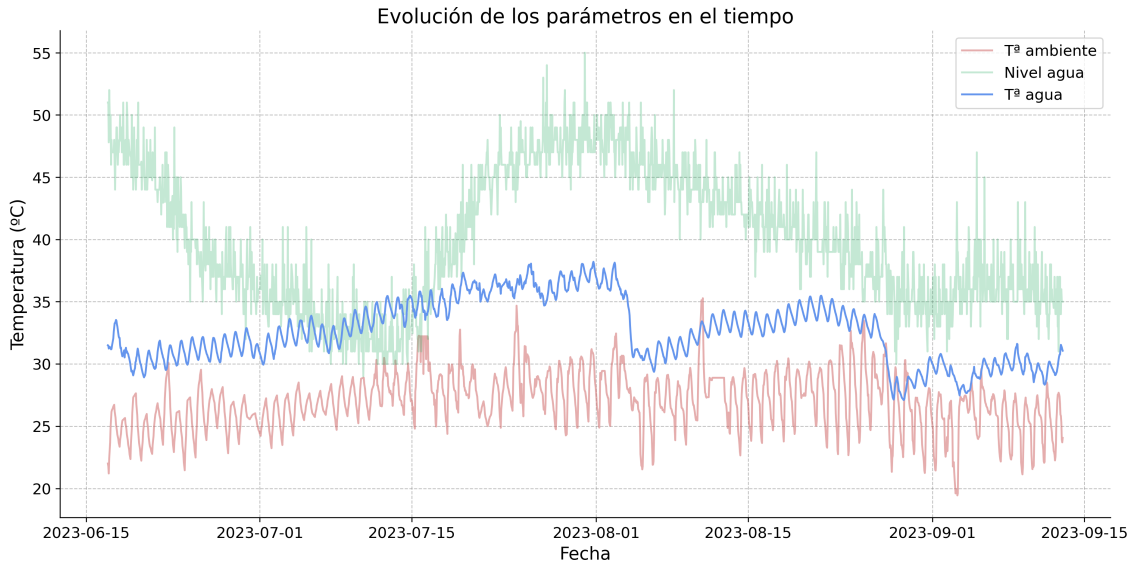

Figura 4.11: Evolución de los parámetros a lo largo del tiempo.

4.4 Datos de estudio

En resumen, los datos de estudio de los cuáles se consta para este trabajo, y una breve recopilación de la información más general de estos, se puede ver en la Tabla 4.6.

Tabla 4.6: Información datos de estudio.

Conjunto de datos	Fecha Inicio	Fin	Variables	Total filas	Frecuencia medición
IoT	17-06-2023	02-08-2023	Tª agua Tª ambiente Nivel del agua	2.113	Horaria
AEMET	01-01-1947	30-11-2007	Tª ambiente	22.185	Diaria

4.5 Modelos

En este capítulo se exploran diversos modelos predictivos utilizados para analizar y pronosticar los datos objeto de estudio en esta investigación. La selección de modelos abarca tanto técnicas estadísticas tradicionales como métodos avanzados de ML.

4.5.1. Modelos autorregresivos

A continuación, se presentan diversos modelos autorregresivos (AR) para el análisis y pronóstico de series temporales, proporcionando herramientas esenciales para comprender y predecir el comportamiento de los datos [55, 21, 46, 22, 54, 15, 42, 7].

El modelo autorregresivo AR(p) es fundamental en el análisis de series temporales estacionarias. Un aspecto clave es determinar el orden p, que se refiere al número de retrasos de la serie utilizados como predictores. La selección de p se realiza mediante criterios de información como el AIC (Criterio de Información de Akaike) o el BIC (Criterio de Información Bayesiano). Se define como:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t$$

donde:

- X_t: Valor de la serie temporal en el tiempo t.
- c: Constante o intercepto del modelo.
- ϕ_i: Coeficientes para los retrasos de la serie temporal, influencia de los valores pasados X_{t-i} en X_t.
- ϵ_t: Término de error en el tiempo t, parte no explicada por los retrasos.

Este modelo se caracteriza por:

- La función de autocorrelación (ACF) que decae exponencialmente o de forma sinusoidal.
- Un corte claro en la función de autocorrelación parcial (PACF) después del retraso p.

Modelo ARIMA (Autoregressive Integrated Moving Average)

El modelo ARIMA(p, d, q) es particularmente útil para analizar series no estacionarias que se les aplica una transformación para que los datos sean estacionarios tras d diferencias. Es adecuado para datos con tendencias y sin patrones estacionales fijos. La forma general del modelo es:

$$\nabla^D \nabla^d X_t = c + \left(\sum_{i=1}^{p} \phi_i \nabla^d X_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \right) + \left(\sum_{i=1}^{P} \Phi_i \nabla^D X_{t-i\times s} + \sum_{j=1}^{Q} \Theta_j \epsilon_{t-j\times s} \right) + \epsilon_t$$

- $\nabla^d X_t$: Resultado de diferenciar la serie d veces.
- p y q: Órdenes de los componentes autorregresivo y de medias móviles, respectivamente.
- θ_j: Coeficientes de los términos de error pasados.

Es crucial evaluar los residuos para asegurar que se comporten como ruido blanco, indicativo de un buen ajuste del modelo.

Modelo SARIMA (Seasonal ARIMA)

El modelo SARIMA(p, d, q)(P, D, Q)_s incorpora componentes estacionales a la estructura ARIMA, lo que lo hace ideal para series con patrones estacionales evidentes. Se formula como:

$$\nabla^D \nabla^d X_t = c + \left(\sum_{i=1}^{p} \phi_i \nabla^d X_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \right) + \left(\sum_{i=1}^{P} \Phi_i \nabla^D X_{t-i\times s} + \sum_{j=1}^{Q} \Theta_j \epsilon_{t-j\times s} \right) + \epsilon_t$$

- P, D, Q: Análogos estacionales de p, d y q.
- s: Periodicidad de la estacionalidad.
- Φ_i, Θ_j: Coeficientes para los componentes estacionales autorregresivos y de medias móviles.

Modelo SARIMAX (SARIMA with eXogenous variables)

El modelo SARIMAX extiende SARIMA al incorporar variables exógenas, permitiendo modelar influencias externas en la serie temporal. Se define como:

$$\nabla^D \nabla^d X_t = c + \left(\sum_{i=1}^{p} \phi_i \nabla^d X_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} \right) + \left(\sum_{i=1}^{P} \Phi_i \nabla^D X_{t-i\times s} + \sum_{j=1}^{Q} \Theta_j \epsilon_{t-j\times s} \right) + \beta Z_t + \epsilon_t$$

- Z_t: Variables exógenas que pueden influir en la serie temporal.
- β: Coeficientes asociados a cada variable exógena.

Las variables exógenas Z_t pueden incluir elementos como días festivos, acciones de marketing o cambios legislativos. La estimación de β, coeficientes para las variables exógenas, se realiza juntamente con los demás parámetros del modelo.

4.5.2. Modelos clásicos

Una forma alternativa de abordar este problema es mediante el empleo de modelos clásicos de estadística y ML, como los que se pueden ver a continuación [26, 40, 25, 4, 33, 8, 10].

Regresión lineal (LR)

La regresión lineal es un método estadístico utilizado para modelar y analizar la relación entre una variable dependiente y una o más variables independientes. El objetivo principal de este método es encontrar una línea recta que se ajuste óptimamente a los datos, minimizando las diferencias entre los valores reales y los valores que el modelo predice [80, 61]. Este proceso implica:

- Variable dependiente (Y): Es la variable que se pretende predecir o explicar.
- Variable(s) independiente(s) (X): Son las variables que se utilizan para predecir la variable dependiente. Pueden ser una o varias.

En su forma más básica, donde solo se considera una variable independiente, la ecuación de la regresión lineal simple se formula como:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

donde:

- β_0 es el término de intercepción, que representa el valor esperado de Y cuando X es 0.
- β_1 es el coeficiente de la variable independiente, indicando la pendiente de la línea de mejor ajuste.
- ϵ es el término de error, que captura todas las influencias no explicadas por la variable independiente sobre Y.

El propósito fundamental de la regresión lineal es calcular los coeficientes β_0 y β_1 que reducen al mínimo la suma de los cuadrados de los residuos, es decir, las diferencias entre los valores reales y los valores estimados por el modelo. Este procedimiento es conocido como el método de Mínimos Cuadrados Ordinarios (MCO).

La regresión lineal múltiple es una extensión de la regresión lineal simple, diseñada para incluir dos o más variables independientes. Esta variante permite modelar relaciones más complejas y ajustar modelos que consideran múltiples factores simultáneamente. La ecuación para la regresión lineal múltiple se expresa como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

donde:

- $X_1, X_2, ...X_k$ son las variables independientes.
- $\beta_0, \beta_1, ..., \beta_k$ son los coeficientes del modelo, que representan el efecto estimado de cada variable independiente sobre la variable dependiente.
- ϵ continúa siendo el término de error.

Least Absolute Shrinkage and Selection Operator (Lasso)

El método Least Absolute Shrinkage and Selection Operator (Lasso) es una técnica de regresión que integra la contracción (shrinkage) y la selección automática de variables. La contracción implica reducir el tamaño de los coeficientes estimados en un modelo de regresión [61, 5].

La idea central de Lasso es mejorar el método tradicional de regresión de mínimos cuadrados mediante la adición de una penalización equivalente a la suma de los valores absolutos de los coeficientes. Matemáticamente, el objetivo es minimizar:

$$\text{Minimizar} \quad \frac{1}{2n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

donde:

- y_i son los valores observados.
- x_{ij} son los valores de los predictores.
- β_j son los valores de los predictores.
- λ es un parámetro que controla el grado de penalización. Cuanto mayor es el valor de λ, mayor es la contracción de los coeficientes hacia cero.

La penalización impuesta por Lasso puede llevar a que algunos de los coeficientes se reduzcan exactamente a cero cuando λ es suficientemente grande. Esto permite que Lasso no solo prediga, sino que también seleccione variables, identificando las más importantes para el modelo. Esta capacidad de selección ayuda a simplificar el modelo y a evitar el sobreajuste, especialmente en situaciones con muchos predictores.

Lasso es particularmente útil en casos donde el número de predictores es significativamente mayor que el número de observaciones, o cuando existe colinealidad entre los predictores.

Decision Tree (DT)

Un árbol de decisión o Decision Tree es un tipo de modelo de aprendizaje supervisado ampliamente utilizado en estadística, minería de datos y ML. Este tipo de modelo se emplea tanto en tareas de clasificación como de regresión. En el contexto de regresión, el regresor de árbol de decisión predice un valor continuo para la variable dependiente basándose en las variables independientes [36].

Las características clave de este modelo incluyen:

- **Estructura del árbol**: Comienza con un nodo raíz, que es el punto de partida para la primera división, basada en una variable seleccionada. A partir de este nodo, se desarrollan nodos intermedios a través de divisiones subsiguientes basadas en las condiciones derivadas de las variables de entrada. Los nodos hoja constituyen los puntos terminales del árbol, y contienen los valores de predicción determinados por las condiciones de las divisiones anteriores.
- **División basada en homogeneidad**: Los nodos se dividen de manera que las observaciones dentro de cada nodo sean lo más homogéneas posible respecto a la variable objetivo. Para maximizar esta homogeneidad, se utilizan criterios estadísticos como la reducción máxima de la varianza o el error cuadrático medio.

- **Modelo no paramétrico**: A diferencia de los modelos lineales, los árboles de decisión no presuponen una relación funcional específica entre las variables de entrada y salida. Esta característica les permite capturar relaciones complejas y no lineales entre las variables, lo que ofrece una gran flexibilidad en la modelización de diferentes tipos de datos.

- **Fácil interpretación**: Los árboles de decisión son especialmente apreciados por su claridad visual y facilidad de interpretación. La estructura del árbol proporciona una representación intuitiva de cómo se toman las decisiones, facilitando la capacidad de explicación y la transparencia del proceso.

- **Propenso al sobreajuste**: Los árboles de decisión tienen tendencia a ajustarse excesivamente a los detalles específicos y al ruido de los datos de entrenamiento. Para mitigar este riesgo, se pueden aplicar técnicas como la poda del árbol, que implica restringir la profundidad del árbol y establecer límites en el número de niveles o en el número mínimo de muestras por nodo.

- **Sensibilidad a los datos**: Estos modelos pueden ser altamente sensibles a variaciones menores en los datos de entrenamiento, lo que puede llevar a diferencias significativas en la estructura del árbol. Esta sensibilidad resalta la importancia de una adecuada preparación y comprensión de los datos antes de proceder al entrenamiento del modelo.

Este conjunto de características subraya la utilidad y los desafíos asociados con los regresores de árbol de decisión, haciéndolos herramientas efectivas pero que requieren una implementación cuidadosa.

Random Forest (RF)

El Random Forest es un método de regresión y clasificación que forma parte de los algoritmos de aprendizaje supervisado en estadística y ML. Este método se fundamenta en el ensamblaje de múltiples árboles de decisión para incrementar la robustez y precisión de las predicciones [66].

Los puntos clave del Random Forest desde una perspectiva estadística incluyen:

- **Reducción de Varianza**: El Random Forest promedia las predicciones de varios árboles de decisión, disminuyendo la varianza. Esto mejora la estabilidad y la precisión de las predicciones, sin aumentar significativamente el sesgo.

- **Ley de los Grandes Números**: A medida que aumenta el número de árboles en el modelo, el promedio de sus predicciones se estabiliza más cerca del valor esperado verdadero. Esto contribuye a una mayor precisión en las predicciones globales del modelo.

- **Importancia de las Características**: Este método también permite evaluar la importancia de cada característica, basándose en su impacto en la reducción de la varianza en los nodos de los árboles. Esto proporciona percepciones profundas sobre cuáles variables son más influyentes en las predicciones.

- **Bootstrap sampling**: Cada árbol en el bosque se construye a partir de una muestra bootstrap. Esto permite que cada árbol entrene con diferentes subconjuntos de datos, lo que enriquece la diversidad del modelo y refuerza su capacidad para generalizar.

- **Error Out-of-Bag (OOB)**: Cada árbol en el bosque se construye a partir de una muestra bootstrap. Esto permite que cada árbol entrene con diferentes subconjuntos de datos, lo que enriquece la diversidad del modelo y refuerza su capacidad para generalizar.

Estas características estadísticas aseguran que el Random Forest sea una herramienta eficaz para abordar complejidades y relaciones no lineales en grandes conjuntos de datos.

Support Vector Machine Regressor (SVR)

El Support Vector Machine Regressor (SVR) es una extensión del algoritmo de Support Vector Machines (SVM) para problemas de regresión. Al igual que el SVM para clasificación, el SVR busca encontrar un hiperplano en un espacio de alta dimensión que pueda modelar la relación entre las variables de entrada y las salidas de manera que minimice el error de predicción. Sin embargo, en lugar de encontrar un hiperplano que maximice el margen de separación entre clases, el SVR busca encontrar una función que tenga al menos un margen epsilon (ϵ) de tolerancia de error con respecto a los datos de entrenamiento [19, 61].

El problema de optimización para un SVR puede formularse como:

Minimizar:

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*)$$

Sujeto a:

$$\begin{cases} y_i - (\mathbf{w}\cdot\mathbf{x}_i + b) \leq \epsilon + \xi_i \\ (\mathbf{w}\cdot\mathbf{x}_i + b) - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

Donde:

- $\mathbf{w}$ y b son los parámetros del modelo
- C es un parámetro que controla el *trade-off* entre la complejidad del modelo y la tolerancia al error. Siendo *trade-off* la necesidad de equilibrar dos características opuestas en un modelo o proceso.
- ξ_i y ξ_i^* son variables de holgura que permiten que algunos puntos estén fuera del margen ϵ.

Sus componentes principales son:

- **Función de pérdida epsilon-insensible**: La función de pérdida utilizada en SVR ignora los errores que están dentro de un margen ϵ alrededor de la predicción. Esto significa que los errores menores que ϵ no contribuyen a la función de pérdida, permitiendo una cierta flexibilidad en las predicciones.

$$L_\epsilon(y, f(x)) = \begin{cases} 0 & \text{si } |y - f(x)| \leq \epsilon \\ |y - f(x)| - \epsilon & \text{en otro caso} \end{cases}$$

- **Maximización del margen y regularización**: El SVR busca maximizar el margen (distancia entre el hiperplano y los vectores de soporte) mientras minimiza una función de regularización que controla la complejidad del modelo. Esta función de regularización suele ser la norma L2 de los coeficientes del modelo.
- **Vectores de soporte**: Los puntos de datos que están fuera del margen ϵ (aquellos con errores mayores que ϵ) se denominan vectores de soporte. Estos puntos son los únicos que afectan la solución final del modelo.

En resumen, es un método eficaz para encontrar una función de predicción con un margen de tolerancia al error, maximizando la precisión de las predicciones mientras controla la complejidad del modelo.

K-Nearest Neighbors Regressor (KNN)

El KNN predice el valor de una variable dependiente continua basada en los valores de las k observaciones más cercanas en el espacio de características [6].

Algunas de las principales características son:

- **Distancia euclidiana**: La cercanía entre observaciones se mide usando la distancia euclidiana, definida como:

  $$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{l=1}^{p} (x_{il} - x_{jl})^2}$$

  donde $\mathbf{x}_i$ y $\mathbf{x}_j$ son vectores de características de dos observaciones y p es el número de características.
- **Selección de los vecinos (k)**: El parámetro k es crucial y representa el número de vecinos más cercanos a considerar. Elegir un valor adecuado para k es importante, ya que valores muy pequeños pueden hacer que el modelo sea demasiado sensible al ruido en los datos, mientras que valores muy grandes pueden hacer que el modelo sea demasiado general y menos preciso.

El proceso de predicción del KNN consiste en dos pasos principales:

1. **Identificación de vecinos**: Para una nueva observación $\mathbf{x}_0$, el algoritmo identifica las k observaciones en el conjunto de entrenamiento más cercanas a $\mathbf{x}_0$ basándose en la distancia euclidiana.
2. **Cálculo del valor predicho**: El valor predicho $\hat{y}_0$ para la nueva observación $\mathbf{x}_0$ es la media de los valores de la variable dependiente de los k vecinos más cercanos:

   $$\hat{y}_0 = \frac{1}{k} \sum_{i=1}^{k} y_i$$

   donde y_i son los valores de la variable dependiente de los k vecinos seleccionados.

4.5.3. Redes neuronales artificiales

Antes de explorar modelos específicos, es importante entender algunos de los fundamentos de las redes neuronales artificiales (RNAs). Inspirados en el funcionamiento del cerebro humano, estos modelos de ML están diseñados para procesar información compleja de manera eficaz [4].

Las RNAs se componen de neuronas, las unidades básicas de procesamiento, organizadas en varias capas: una capa de entrada que recibe señales, capas ocultas que procesan las señales, y una capa de salida que entrega el resultado final. Cada neurona en estas capas maneja sus entradas $x_1, x_2, \ldots, x_n$ ponderadas por pesos $w_1, w_2, \ldots, w_n$, y produce una salida calculada a través de una función de activación f, típicamente sigmoide, tanh (tangente hiperbólica), o ReLU (Rectified Linear Unit), aplicada a la suma ponderada de sus entradas más un sesgo b:

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

La introducción de la función de activación permite a la red aprender y modelar relaciones complejas y no lineales entre los datos. El proceso de aprendizaje en las redes involucra dos etapas principales: la propagación hacia adelante, donde las entradas se procesan para obtener una salida, y la propagación hacia atrás, donde se ajustan los pesos según el error de la salida, utilizando la regla de la cadena para calcular los gradientes.

Para evaluar qué tan bien la red realiza predicciones, se utiliza una función de pérdida, como el error cuadrático medio (MSE) para la regresión. Los pesos se optimizan mediante algoritmos como el descenso de gradiente estocástico (SGD) o Adam, con el objetivo de minimizar esta función de pérdida.

Además, para asegurar que la red generalice bien a nuevos datos y evitar el sobreajuste, se emplean técnicas de regularización como las regularizaciones $L1$ y $L2$, así como el dropout. Estas técnicas son esenciales para mantener la robustez del modelo cuando se expone a datos fuera del conjunto de entrenamiento.

Con estos principios básicos establecidos, se puede comprender mejor cómo los diferentes tipos de RNAs, como MLP, CNN y LSTM, aplican y adaptan estos conceptos para abordar desafíos específicos en el ML.

Multilayer Perceptron Regressor (MLP)

El MLP es una red neuronal que consta de múltiples capas, incluyendo una capa de entrada, varias capas ocultas y una capa de salida. Cada neurona en una capa se conecta con todas las neuronas de la siguiente capa, y se utilizan funciones de activación para introducir no linealidades. Los MLP son ampliamente usados para clasificación y regresión en datos tabulares donde no existen relaciones espaciales o temporales inherentes entre las características. Son particularmente eficientes para problemas con relaciones directas entre características y etiquetas, destacando por su simplicidad y facilidad de implementación [77].

Convolutional Neural Network (CNN)

Las CNNs son especialmente potentes para trabajar con datos que tienen una estructura de rejilla, como imágenes. Estas redes utilizan capas convolucionales para realizar

operaciones de convolución que extraen características locales, seguidas de capas de pooling que reducen la dimensionalidad. Las CNN son capaces de aprender jerarquías de características automáticamente, lo que las hace muy eficientes para tareas de reconocimiento y procesamiento de imágenes y video. Su capacidad para aprender características relevantes directamente de los datos reduce la necesidad de preprocesamiento intensivo [67].

Redes Neuronales Recurrentes (RNN)

Las Redes Neuronales Recurrentes (RNN) son una clase de redes neuronales diseñadas para manejar datos secuenciales y dependencias temporales. A diferencia de las redes convencionales, las RNN tienen la capacidad de mantener un estado interno o memoria que captura información sobre los inputs anteriores, lo que las hace adecuadas para tareas como el procesamiento de lenguaje natural y el reconocimiento de voz. En una RNN, las conexiones entre los nodos forman un ciclo dirigido, lo que permite que la información persista. Sin embargo, las RNN estándar suelen enfrentar dificultades con dependencias de largo plazo debido al problema de desvanecimiento o explosión de gradientes. Soluciones como las LSTM y las GRU (Unidad Recurrente Puerta) se han desarrollado para abordar estos problemas, mejorando la capacidad de la red para aprender secuencias con dependencias temporales largas y complejas. Las RNN son fundamentales en aplicaciones donde es esencial capturar dinámicas temporales para realizar predicciones precisas [60].

Long Short Term Memory (LSTM)

Las Long Short Term Memory (LSTM) son un tipo de RNN diseñada para manejar secuencias y dependencias de largo plazo. Utilizan unidades especiales que incluyen compuertas de entrada, salida y olvido para controlar el flujo de información. Esto les permite recordar información durante periodos extensos, lo que es crucial para tareas como la traducción automática, la generación de texto y la predicción de series temporales. Las LSTMs son adecuadas para modelar secuencias complejas y manejar datos secuenciales donde es importante no solo la información más reciente, sino también eventos pasados significativos [23].

4.6 Métricas

En este capítulo, se presentan las métricas seleccionadas para la evaluación de los modelos predictivos desarrollados en los experimentos de esta investigación. Las métricas escogidas son [10, 8, 33, 25, 40, 26, 2]:

- **Coeficiente de determinación (R^2)**: También conocido como el coeficiente de correlación múltiple ajustado para el número de predictores, R^2 es una medida estadística que ofrece una indicación de qué tan bien los resultados observados son replicados por el modelo. Se calcula como la proporción de la variación total de los resultados explicada por el modelo:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  - y_i: valores reales observados de la variable dependiente.
  - $\hat{y}_i$: valores predichos por el modelo.

- $\bar{y}$: media aritmética de los valores observados .
- n: número total de observaciones.

Un R^2 de 1 indica que el modelo ajusta perfectamente los datos con toda la variabilidad explicada, mientras que un valor de 0 indica que el modelo no explica la variabilidad en comparación con simplemente tomar la media de los datos.

- **Error cuadrático medio (MSE)**: El MSE es una métrica robusta que cuantifica el promedio de los errores al cuadrado, es decir, la diferencia cuadrada entre los valores estimados y los observados. Ofrece una vista global de la magnitud de los errores sin considerar su dirección:

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

El MSE es particularmente útil en contextos donde los errores grandes son inaceptables, ya que estos contribuyen de forma exponencial al valor de la métrica. Un MSE bajo es indicativo de un modelo que predice con alta precisión.

- **Error Absoluto Medio (MAE)**: El MAE proporciona una medida lineal de los errores entre los valores predichos y los reales, calculando el promedio de las diferencias absolutas:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

A diferencia del MSE, el MAE no penaliza tanto los errores grandes, lo que puede ser preferible en aplicaciones prácticas donde los errores grandes no son necesariamente críticos. Un valor bajo de MAE indica un modelo que, en promedio, se acerca mucho a los valores reales.

Estas métricas no solo permiten una evaluación cuantitativa del rendimiento del modelo, sino que también ayudan a discernir la calidad de los modelos desde una perspectiva práctica y teórica.

CAPÍTULO 5
Modelado predictivo a futuro

En este capítulo se presenta el estudio y la experimentación de un modelo predictivo a largo plazo. El objetivo es desarrollar un modelo de regresión que prediga los valores de la temperatura del agua en un horizonte lejano con la mayor precisión posible.

El horizonte es un término crucial para describir la predicción en series temporales. Se refiere al número de pasos de tiempo en el futuro para los cuales se desea realizar una predicción. En otras palabras, el horizonte de predicción es la extensión del período de tiempo que se espera predecir utilizando un modelo de series temporales.

Para esta sección, se utilizará exclusivamente el conjunto de datos recopilados por los dispositivos IoT, que incluyen la temperatura del agua, la temperatura ambiente y el nivel del agua durante la temporada estival.

5.1 Conjunto de datos de estudio

Basándose en los datos de estudio descritos en el Capítulo 4, Sección 4.1, se proponen diversos conjuntos de entrenamiento con el objetivo de abordar la predicción de la temperatura del agua desde diferentes enfoques.

5.1.1. Primer conjunto

El primer conjunto de datos tiene como objetivo estudiar una serie temporal sin muchas anomalías. Es decir, se analizarán los datos de la temperatura del agua evitando los cambios bruscos mencionados previamente, que ocurrían a principios y finales de agosto. Por lo tanto, para este conjunto de datos, se tomarán los valores desde el inicio hasta que se detecte el primer cambio brusco que se corresponde con el 2 de agosto, y un total de 1.110 filas.

En la Figura 5.1 se muestran los valores para la temperatura del agua que se van a emplear para este conjunto; la temperatura ambiente y el nivel, aunque no se muestren, también forman parte de este.

En este conjunto se pretende estudiar un horizonte cercano para comprender el comportamiento en predicciones a corto plazo, y luego ampliar el análisis a horizontes más largos. Así, se establecen los horizontes de estudio en: 1, 6, 12 horas, y 1, 2, 5 y 10 días. De esta manera, si se expresa en porcentaje la cantidad de los datos que son para entrenar y para validar, el resultado es el que se ilustra en la Tabla 5.1.

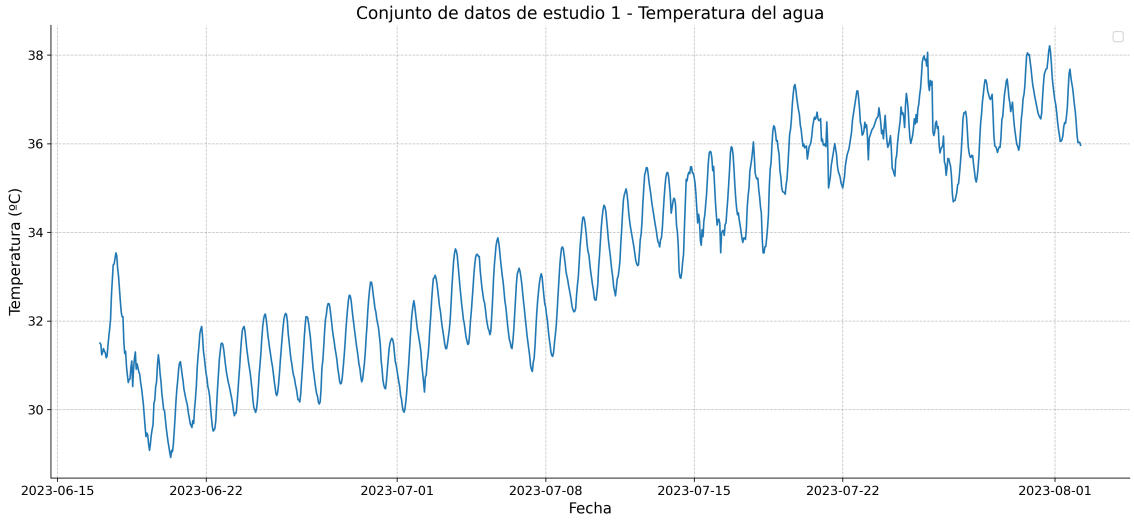

Figura 5.1: Primer conjunto de datos de entrenamiento (solo la temperatura del agua).

Tabla 5.1: Porcentajes de entrenamiento y validación para los distintos horizontes del primer conjunto de datos.

Horizonte	Entrenamiento	Validación
1 hora	99'91	0'09
6 horas	99'46	0'54
12 horas	98'92	1'08
1 día	97'84	2'16
2 días	95'68	4'32
5 días	89'19	10'81
10 días	78'38	21'62

5.1.2. Segundo conjunto

En este segundo conjunto de datos se introduce una mayor complejidad al incluir un cambio brusco. Se utilizarán los valores desde el inicio hasta un poco después del primer *shock*, que ocurre el 18 de agosto, abarcando un total de 1.500 filas.

Se pueden ver en la Figura 5.2 los valores de la temperatura de agua que se van a utilizar para el segundo conjunto.

Al igual que en el caso anterior, se pretende estudiar un horizonte cercano para comprender el comportamiento de las predicciones a corto plazo, y posteriormente ampliar el análisis a horizontes más largos estableciendo los mismos horizontes de estudio en: 1, 6, 12 horas, y 1, 2, 5 y 10 días. Los porcentajes para entrenamiento y validación usados se muestran en la Tabla 5.2.

5.1.3. Tercer conjunto

El tercer y último conjunto de datos trabajará con toda la información disponible, pero, a diferencia de los conjuntos anteriores, se definirán horizontes de predicción distintos. Dado que los cambios abruptos representan uno de los principales desafíos en las predicciones de series temporales, este conjunto se enfocará en estudiar estos *shocks*.

En particular, se plantean tres horizontes de estudio con el objetivo de analizar los *shocks* en diferentes momentos:

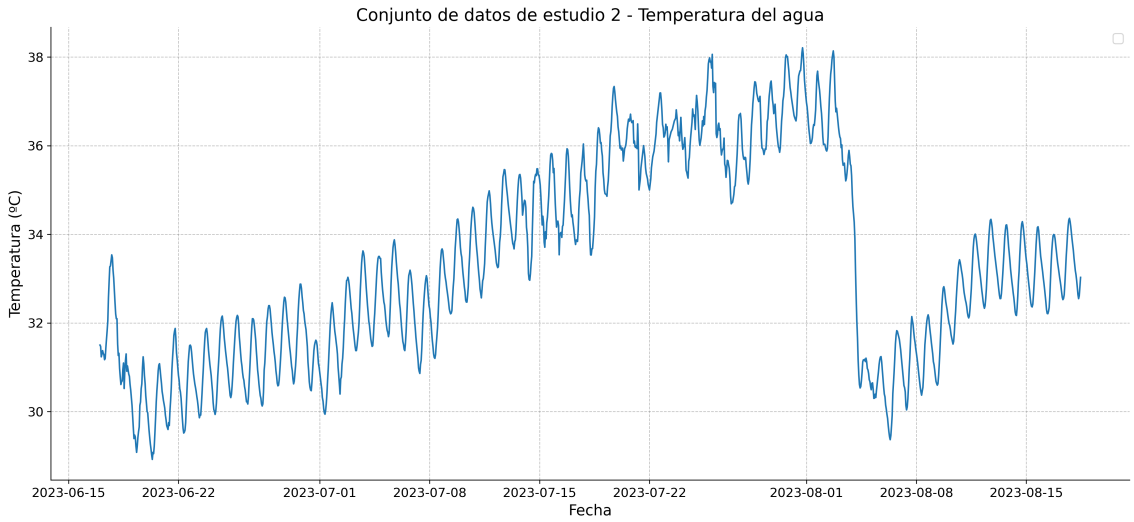

Figura 5.2: Segundo conjunto de datos de entrenamiento (solo la temperatura del agua).

Tabla 5.2: Porcentajes de entrenamiento y validación para los distintos horizontes del segundo conjunto de datos.

Horizonte	Entrenamiento	Validación
1 hora	99'93	0'07
6 horas	99'60	0'40
12 horas	99'20	0'80
1 día	98'40	1'60
2 días	96'8	3'20
5 días	92'00	8'00
10 días	84 '00	16'00

- Entrenamiento hasta justo antes del *shock*, y predicción posterior al *shock*, con un horizonte de aproximadamente 17 días (420 horas).
- Entrenamiento durante el *shock* y predicción del comportamiento durante el *shock*, con un horizonte de poco más de 16 días (400 horas).
- Entrenamiento después del *shock* y predicción para el período posterior al *shock*, con un horizonte de 5 días (120 horas).

En la Figura 5.3 se pueden ver los valores de la temperatura de agua que se van a utilizar para el tercer conjunto mostrando los distintos horizontes.

También se quiere destacar que, durante el proceso de entrenamiento, ya se habrá observado al menos un *shock* en los datos.

Los porcentajes para entrenamiento y validación quedan como se ve en la Tabla 5.3.

Tabla 5.3: Porcentajes de entrenamiento y validación para los distintos horizontes del tercer conjunto de datos.

Horizonte	Entrenamiento	Validación
5 días	94'32	5'68
16 días	81'07	18'93
17 días	80'12	19'88

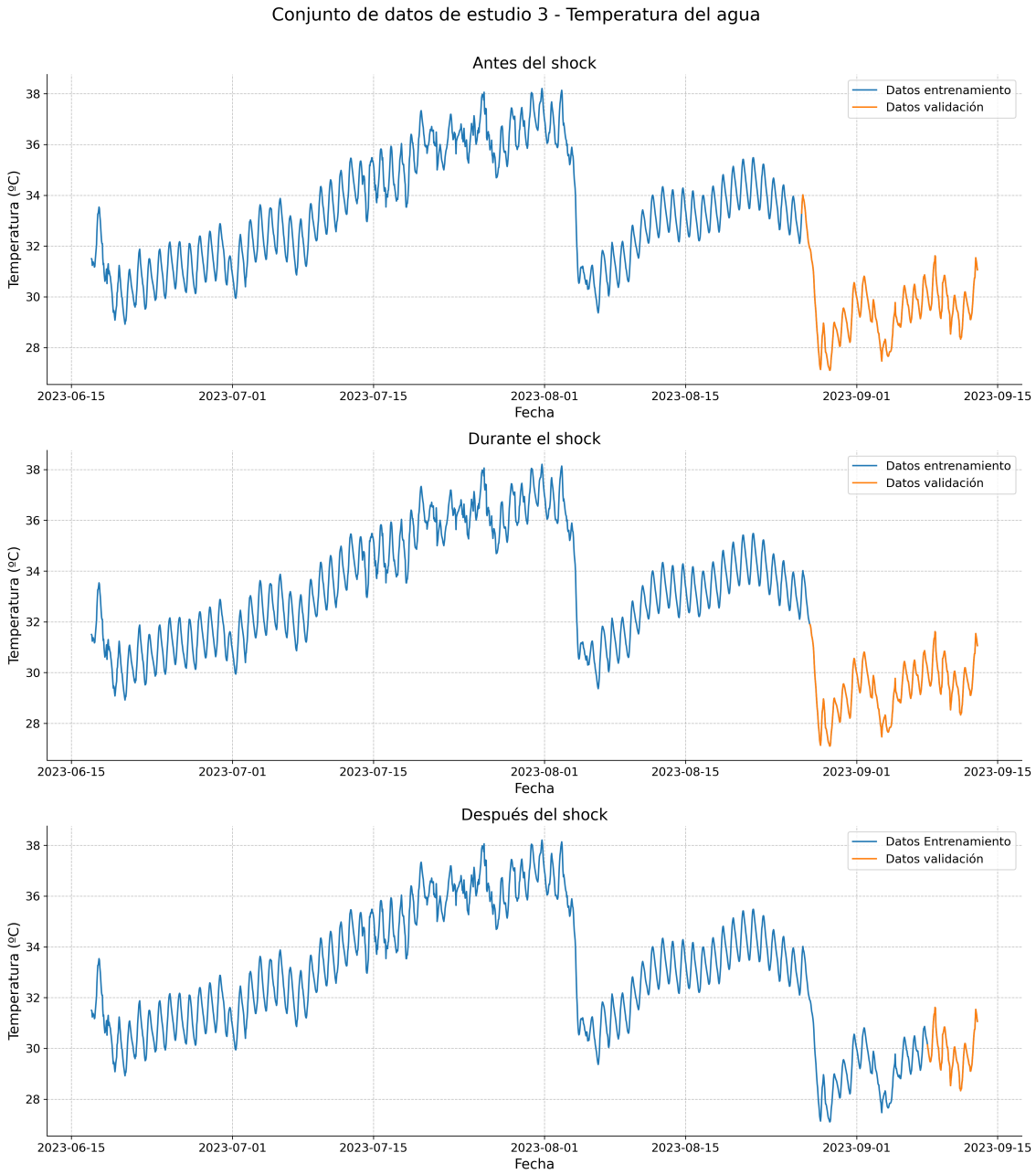

Figura 5.3: Tercer conjunto de datos de entrenamiento (solo la temperatura del agua) separando por los distintos horizontes.

En la Tabla 5.4 se puede observar un resumen de los conjuntos de datos con los que se va a trabajar.

Tabla 5.4: Resumen general de los conjuntos de datos de entrenamiento.

Conjunto	Fecha		Total	Horizontes
	Inicio	Fin		
Primero		02-08-2023	1.110	1, 6, 12 horas
Segundo	17-06-2023	18-08-2023	1.500	1, 2, 5 y 10 días
Tercero		14-09-2023	2.113	5, 16 y 17 días

5.2 Experimentos

Se han desarrollado un total de cinco experimentos enfocados en la predicción de la temperatura del agua, cada uno adoptando enfoques metodológicos distintos que varían desde modelos univariantes hasta enfoques multivariantes avanzados. Los experimentos abarcan desde el uso exclusivo de la temperatura del agua como variable de interés, implementando modelos autoregresivos en el primer experimento, hasta incorporar variables exógenas como la temperatura ambiente y el nivel del agua en modelos más complejos en los experimentos subsiguientes. Las metodologías empleadas incluyen técnicas estadísticas clásicas y redes neuronales convolucionales.

5.2.1. Primer experimento

El primer experimento se centra en el uso de AR para predecir la temperatura del agua, utilizando únicamente un enfoque univariante. Es decir, se considerará solo el valor de la temperatura del agua, lo cual representa el caso más simple.

En este experimento se han utilizado los métodos AR, ARIMA y SARIMA. La implementación de estos se ha llevado a cabo mediante la librería *statsmodels*, utilizando específicamente los métodos *AutoReg* [74], *ARIMA* [73] y *SARIMAX* [75].

Para determinar el ajuste de los modelos ARIMA y SARIMA, se ha empleado la función *auto_arima* de la librería *pmdarima* [59]. Esta función automatiza el proceso de ajuste de modelos ARIMA, incluyendo la identificación de los mejores parámetros del modelo ARIMA *(p, d, q)* y, si es necesario, los parámetros estacionales *(P, D, Q, m)*, con el fin de ajustar mejor los datos y proporcionar predicciones precisas. De esta manera se determina que el mejor ajuste es para la ARIMA(6,1,1), y para la SARIMA(3,1,0)(1,0,1).

Se van a emplear todos los conjuntos de datos definidos en la sección anterior.

5.2.2. Segundo experimento

En este segundo enfoque, se adopta un modelo multivariante que incorpora la temperatura ambiente y el nivel del agua como variables exógenas, mientras que la temperatura del agua se establece como la variable endógena, y es el parámetro que se pretende predecir.

Los modelos analizados incluyen: LR [50], RF [52], KNN [48], MLP [51], Lasso [49], DT [47], SVR [53], CNN [13] y SARIMAX.

La estructura de esta CNN inicia con una capa convolucional unidimensional (*Conv1D*) con 64 filtros y una función de activación *ReLU*, configurada para recibir entradas de forma (3, 1). Luego, la salida de esta capa se aplana mediante una capa *Flatten*, que la prepara para ser procesada por capas densas subsecuentes. A continuación, hay una capa densa con 32 unidades y activación *ReLU*, seguida de una capa de salida con una unidad, adecuada para tareas de regresión. El modelo se compila con el optimizador *Adam* y la función de pérdida *MSE*.

Para finalizar, en este análisis se utilizará exclusivamente el tercer conjunto de datos, abarcando sus diversos horizontes de predicción.

5.2.3. Tercer experimento

En el tercer experimento, mantenemos el enfoque multivariante, y se introduce la temperatura del agua de la hora anterior como variable exógena.

Hasta la fecha, se ha aplicado un modelo de predicción de horizonte fijo, también conocido como predicción directa de múltiples pasos. Este modelo está diseñado para mapear directamente las entradas a las salidas en el horizonte de predicción deseado. Sin embargo, en este nuevo experimento, se va a sustituir este enfoque por uno de predicción iterativa o recursiva de paso fijo. En lugar de realizar una única predicción para el horizonte establecido, este método implica realizar tantas predicciones como horas comprenda dicho horizonte. Por ejemplo, para un horizonte de diez horas, en lugar de emitir una sola predicción, se generan diez predicciones sucesivas.

La predicción iterativa de paso fijo consiste en que el modelo realiza predicciones secuenciales de un paso adelante sin recurrir a las salidas anteriores como parte de las entradas para futuras predicciones. Cada predicción se basa únicamente en el conocimiento adquirido durante el entrenamiento inicial del modelo. Esto implica que, en predicciones diarias, cada una se realiza de manera independiente, sin ajustar el modelo a nuevas tendencias o patrones que puedan surgir de las predicciones recientes.

Por otro lado, la predicción recursiva de paso fijo permite una retroalimentación en la que se utilizan las predicciones anteriores como parte de la entrada para futuras predicciones. Después de cada predicción, el modelo incorpora esta nueva información a sus datos, y se ajusta o reentrena antes de proceder con la siguiente predicción. Este proceso puede mejorar la capacidad del modelo para adaptarse a cambios o tendencias emergentes, aunque también incrementa el riesgo de propagar errores si las predicciones iniciales son inexactas.

La principal distinción entre las técnicas iterativa y recursiva radica en que, en la recursiva, el modelo se actualiza con cada nueva predicción, es decir, se reentrena.

En relación con este experimento, se enfrenta el desafío de predecir la temperatura para la próxima hora sin conocer la temperatura ambiental ni el nivel del agua en ese momento. Por ello, se proponen dos variantes:

- Utilizar los valores de la hora anterior como base para las predicciones futuras; en otras palabras, se hereda el último valor. Este es el enfoque más simple, en el que se toma la temperatura ambiente y el nivel del agua de la hora anterior, y se proyecta la temperatura del agua para la semana siguiente sin alterar la temperatura ambiental.

- Estimar los valores futuros mediante un modelo autorregresivo. Tanto la temperatura ambiente como el nivel del agua muestran una periodicidad que permite modelarlos efectivamente con técnicas AR, como los modelos AR o SARIMA.

Ambas estrategias buscan abordar la falta de datos futuros, ya sea mediante la herencia de datos anteriores o la estimación mediante modelos avanzados de series temporales.

Los modelos que se analizarán en este estudio serán los mismos que en el experimento previo, con la excepción del modelo SARIMA, debido a que su tiempo de ejecución aumentaría considerablemente al realizar predicciones individuales. A pesar de que la CNN también incrementa el tiempo de ejecución, se continuará utilizando para determinar si ofrece mejores resultados que los modelos tradicionales. Además, se aplicarán los modelos AR y SARIMA para prever futuros parámetros de la temperatura ambiente y el nivel del agua.

Es importante mencionar que, en el caso de la CNN, se adoptará el enfoque de reentrenamiento con nuevos datos.

El conjunto de datos que se empleará será el tercero.

5.2.4. Cuarto experimento

En el cuarto experimento, se empleó una LSTM [38] para predecir la temperatura utilizando el tercer conjunto de datos con tres horizontes temporales definidos.

El enfoque será multivariante utilizando los mismos parámetros que en el Experimento 3. Es decir, se considerarán la temperatura ambiente, el nivel del agua, y la temperatura del agua de una hora antes para predecir la temperatura actual del agua.

Primero, los datos se normalizaron utilizando la técnica Min-Max, escalando los valores entre 0 y 1. Esta normalización es crucial para el rendimiento de las redes neuronales, ya que asegura que todos los datos estén en la misma escala. A continuación, se crearon las secuencias de entrada y salida: cada secuencia de entrada contenía una serie de días consecutivos de datos, mientras que cada secuencia de salida correspondía a la temperatura de los días siguientes.

Para el entrenamiento del modelo, se definen varias devoluciones de llamada (*callbacks*):

- **EarlyStopping**: Detiene el entrenamiento si la pérdida de validación no mejora después de 50 épocas, evitando así el sobreentrenamiento.
- **LearningRateScheduler**: Ajusta la tasa de aprendizaje de manera exponencial en cada época para mejorar la convergencia del modelo.

El modelo LSTM se configura con las siguientes características:

- Una capa LSTM con 64 unidades y una función de activación de tangente hiperbólica.
- Una capa densa con tantas salidas como días en la secuencia de entrada.
- El optimizador Adam, con una tasa de aprendizaje inicial de 0'001 y un valor de recorte de gradiente de 1'0, se utilizó para prevenir la explosión de gradientes.
- Se añade una capa Dropout para evitar el sobreajuste.

Para evaluar el modelo, se utilizó la técnica de validación cruzada con 10 divisiones mediante *TimeSeriesSplit*, permitiendo dividir los datos en diferentes conjuntos de entrenamiento y prueba.

Por último, se realizaron algunas modificaciones en los datos para añadir la temperatura del ambiente en horas anteriores. Es sabido que, por ejemplo, en una piscina a principios de verano, cuando comienzan a registrarse altas temperaturas, el agua de la piscina sigue estando fría hasta pasadas quizás un día o más. Por ello, se ha decidido añadir estos nuevos valores para evaluar si mejorarían la predicción de la temperatura del agua. Concretamente, se probarán las siguientes combinaciones:

- 6, 12, 24 y 36 horas antes.
- 12, 24, 36 y 48 horas antes.
- 24, 36 y 48 horas antes.

Estas modificaciones se probarán en el tercer conjunto de datos, específicamente para el horizonte de 420 horas, que es el más complejo.

5.2.5. Quinto experimento

En el último experimento, se utilizará AutoGluon, un marco AutoML de código abierto muy popular, conocido por entrenar modelos de ML de alta precisión en conjuntos de datos tabulares sin procesar. A diferencia de otros marcos AutoML, que se centran principalmente en la selección de modelos e hiperparámetros, AutoGluon-Tabular destaca por ensamblar y apilar múltiples modelos en varias capas [71].

AutoGluon puede predecir los valores futuros de múltiples series temporales a partir de datos históricos y otras covariables relacionadas. este entrena varios modelos para generar pronósticos probabilísticos precisos, eliminando la necesidad de manejar manualmente tareas complicadas como la selección de modelos y el ajuste de hiperparámetros.

En su núcleo, AutoGluon combina varios algoritmos de previsión de última generación. Estos incluyen métodos estadísticos establecidos como ETS y ARIMA de StatsForecast, pronosticadores eficientes basados en árboles como LightGBM de AutoGluon-Tabular, modelos de aprendizaje profundo flexibles como DeepAR y Temporal Fusion Transformer de GluonTS, y un modelo de previsión sin disparo previo, Chronos [3].

El enfoque será nuevamente multivariante, utilizando el tercer conjunto de datos, añadiendo la temperatura de la hora anterior para cada fila, como en el experimento anterior. Además, el estudio se centrará exclusivamente en un horizonte de 420 horas.

5.3 Resultados

5.3.1. Primer experimento

En la Tabla 5.5 se puede ver, para cada horizonte de cada conjunto de datos, el modelo que mejor resultado ha dado.

En general, todos los modelos tienden a empeorar en su desempeño a medida que el horizonte de predicción se alarga. Esto se evidencia en el aumento del MSE y del MAE con horizontes más largos, un resultado previsible para la mayoría de los modelos de series de tiempo.

SARIMA demuestra ser el modelo más eficaz y consistente, sobresaliendo particularmente en el conjunto de datos 2 con excelentes valores de R^2 y bajos errores en todos los horizontes. AR, por su parte, muestra buen desempeño en horizontes cortos y condiciones menos volátiles, pero su rendimiento cae drásticamente bajo condiciones más complejas y en horizontes largos. En contraste, ARIMA es el menos efectivo, especialmente en condiciones volátiles del conjunto 3, indicando su inadecuación para manejar dinámicas altamente volátiles, y destacando la necesidad de enfoques más robustos para tales escenarios.

En el análisis por conjuntos de datos, el primer conjunto muestra ausencia de anomalías, reflejada en bajos valores de MSE y MAE, y facilita la predicción. El segundo conjunto experimenta un cambio abrupto y, a pesar de ello, muestra resultados comparables e incluso superiores en algunos casos, salvo por un deterioro más marcado en las predicciones a largo plazo, probablemente debido a una mayor cantidad de datos disponibles para el entrenamiento. En contraste, el tercer conjunto, que incluye horizontes enfocados en manejar *shocks*, plantea desafíos significativos, evidenciados por altos valores de MSE y MAE. Esto sugiere que los métodos tradicionales podrían ser insuficientes para abordar dinámicas altamente volátiles, posiblemente requiriendo ajustes o la implementación de modelos más avanzados para mejorar la precisión en tales situaciones.

También es relevante señalar que, en el segundo conjunto de datos, y utilizando el modelo SARIMA, se observa una notable disminución en el rendimiento cuando se extiende el horizonte de predicción a 2 días. Curiosamente, el motivo de este deterioro es desconocido, especialmente dado que, para un horizonte de 5 días, el rendimiento del modelo mejora considerablemente y vuelve a ser bastante bueno.

Es importante añadir que el cálculo de R^2 no es factible cuando solo se dispone de una muestra, por lo que no se reportan estos valores para el horizonte de una hora.

Tabla 5.5: Resultados para el Experimento 1.

Modelo	Dataset	Horizonte	R^2	MSE	MAE
SARIMA	1	1	-	0'00	0'01
AR	1	2	-0'94	0'00	0'05
AR	1	12	0'32	0'21	0'43
SARIMA	1	24	0'75	0'08	0'22
ARIMA	1	48	0'20	0'33	0'48
SARIMA	1	120	-0'36	0'64	0'71
SARIMA	1	240	0'27	0'42	0'52
AR / SARIMA	2	1	-	0'00	0'00
AR	2	2	0'26	0'01	0'09
SARIMA	2	12	0'94	0'01	0'07
SARIMA	2	24	0'98	0'01	0'08
AR	2	48	0'05	0'29	0'50
SARIMA	2	120	0'91	0'04	0'15
AR	2	240	-0'15	0'84	0'78
SARIMA	3	120	0'57	0'25	0'37
SARIMA	3	400	-10'10	11'21	3'22
ARIMA	3	420	-8'06	14'51	3'61

Además de los resultados numéricos, el análisis visual de la Figura 5.4 muestra que, aunque la forma de la curva predicha es en general acertada, está notablemente influenciada por la tendencia de los datos previos. La predicción comenzó desde un punto con una tendencia negativa, que ha continuado manifestándose en los resultados. Esto indica que tanto las tendencias negativas como las positivas se perpetúan a lo largo de las predicciones, afectando tanto a los datos retrospectivos como al ajuste del modelo, conocido como *loopback*. Este término describe el uso de periodos de tiempo anteriores para hacer predicciones actuales o futuras. Por lo tanto, se concluye que no es factible predecir la temperatura a largo plazo con un modelo AR univariante en presencia de estos *shocks*, dado que las tendencias históricas tienden a influir significativamente en las predicciones futuras.

5.3.2. Segundo experimento

Dado que un modelo AR muestra buen desempeño en horizontes cercanos, pero enfrenta dificultades con horizontes lejanos y cambios abruptos debido al fenómeno de la tendencia, en investigaciones futuras se empleará el tercer conjunto de datos. Este conjunto es el que presentó los resultados más desfavorables en el primer experimento, lo cual lo convierte en el foco ideal para ajustes y mejoras metodológicas.

Además, se ha observado que el enfoque univariante con modelos AR no es suficiente para abordar la complejidad de los datos. Por lo tanto, se plantea adoptar un enfoque multivariante que incorpore métodos de ML y redes neuronales simples, además de AR que incorporen variables exógenas.

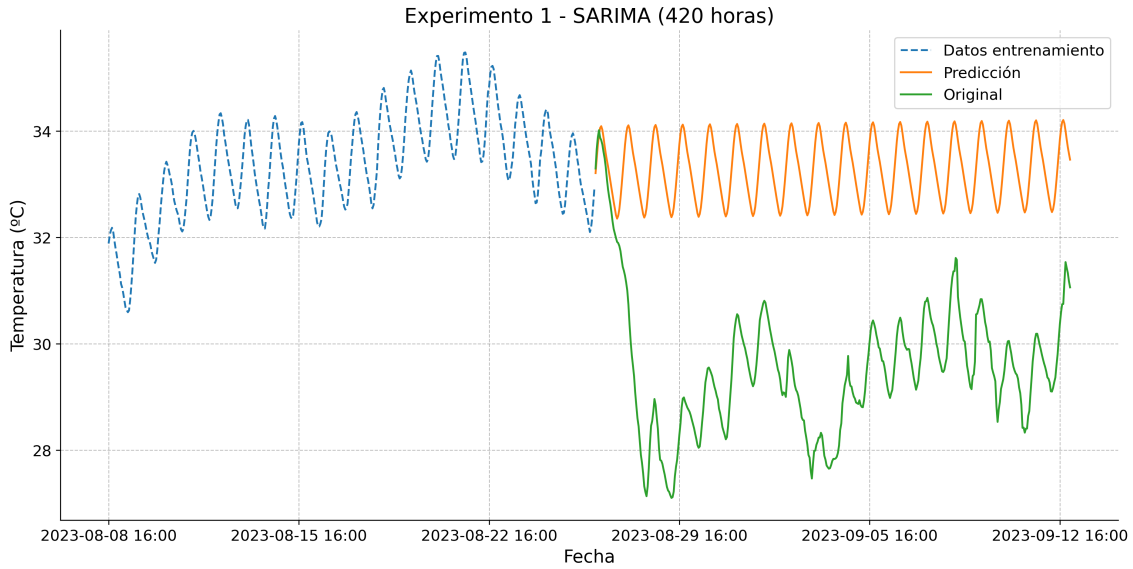

Figura 5.4: Experimento 1: SARIMA con horizonte de 420 horas.

En la Tabla 5.6 se puede observar los resultados obtenidos por los tres mejores modelos para cada horizonte en este experimento.

Se destaca que el modelo SARIMA exhibe un rendimiento notable en el horizonte de 120 días con un R^2 positivo de 0.29, indicando una buena capacidad para explicar la variabilidad de los datos en el corto plazo antes del *shock*. Sin embargo, en horizontes más largos de 400 y 420 días, la mayoría de los modelos muestran valores negativos en R^2, lo que sugiere dificultades para capturar y predecir las dinámicas durante y después del *shock*. En estos períodos, modelos como el CNN y el MLP presentan valores menos negativos en R^2, lo que podría indicar una mejor adaptabilidad frente a los cambios abruptos. Tras estos dos modelos, el LR se posiciona para los horizontes de 16 y 17 días, que corresponden a 400 y 420 horas, respectivamente.

El modelo CNN también se destaca por su consistencia en el rendimiento a lo largo de los diferentes horizontes, con valores relativamente estables de MSE y MAE, sugiriendo robustez frente a la variabilidad de los datos en distintas fases del *shock*. Aunque una CNN puede ofrecer mejores resultados en ciertos casos, su tiempo de ejecución es considerablemente más alto que el de un método clásico, lo que la hace menos práctica para algunas aplicaciones.

En resumen, los valores altos y en su mayoría negativos de R^2 en varios modelos reflejan la complejidad de modelar estos eventos. Los *shocks* son un reto significativo para la predicción de series temporales, como se evidencia en la incapacidad de varios modelos para explicar una gran proporción de la variabilidad durante estos periodos.

Se puede destacar un aspecto positivo al comparar los resultados entre el primer (Tabla 5.5) y el segundo experimento (Tabla 5.6) para el tercer conjunto de datos: ha habido mejoras en todas las evaluaciones de horizonte. De manera destacada, en el horizonte de 400 horas, el coeficiente R^2 mejoró significativamente, pasando de -10'10 a -3'04. Igualmente, para el horizonte de 420 horas, se observa una mejora relevante en el R^2, el cual subió de -8'06 a -2'83. Sin embargo, a pesar de estas mejoras, los resultados aún no son óptimos, y persisten altos valores en las métricas de error, lo que subraya la necesidad de continuar desarrollando y ajustando los modelos para lograr una precisión aún mayor en la predicción durante estos periodos extendidos.

Tabla 5.6: Resultados para el Experimento 2.

Modelo	Horizonte	R^2	MSE	MAE
SARIMA	120	0'29	0'41	0'53
CNN	120	-1'91	1'68	1'08
MLP	120	-2'11	1'80	1'03
MLP	400	-3'04	4'08	1'62
CNN	400	-5'37	6'44	2'08
LR	400	-8'21	9'30	2'74
CNN	420	-2'83	6'13	2'00
MLP	420	-3'23	6'77	2'15
LR	420	-4'59	8'96	2'66

A pesar de las mejoras observadas en los resultados, el fenómeno de la tendencia heredada sigue presente, como se puede apreciar en la Figura 5.5. Esta persistencia sugiere que las predicciones aún están significativamente influenciadas por los patrones anteriores.

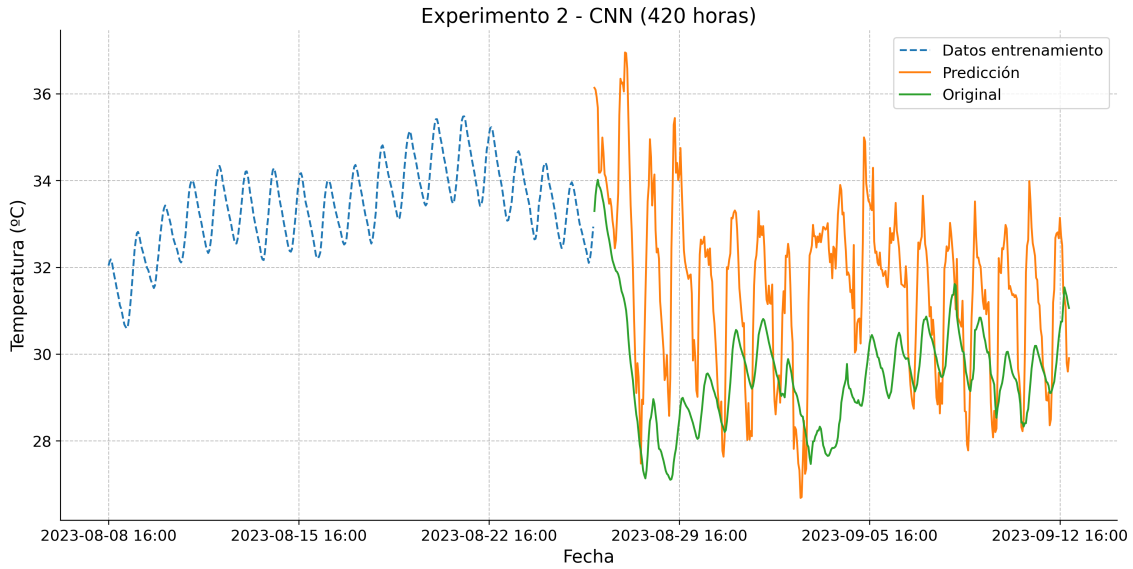

Figura 5.5: Experimento 2: CNN con horizonte de 420 horas.

5.3.3. Tercer experimento

Antes de analizar los resultados obtenidos en el Experimento 3, se revisarán los resultados generados por los modelos AR y SARIMA, así como el modelo que hereda el último valor (representado como un igual "=" en la columna "heredar" de la tabla de resultados), para los parámetros de temperatura ambiente y nivel de agua.

En los resultados para la temperatura ambiente (Tabla 5.7), el modelo AR tiene el mejor desempeño en el horizonte de 120, con un R^2 positivo de 0'16 y valores de MSE y MAE relativamente bajos. Sin embargo, su rendimiento disminuye significativamente en horizontes más largos. El modelo SARIMA muestra un desempeño deficiente en todos los horizontes evaluados, con R^2 negativos y altos valores de MSE, especialmente en el horizonte de 400 donde el MSE alcanza 17'36. El modelo que hereda el último valor tiene un rendimiento intermedio, con R^2 negativos pero valores de MSE y MAE más bajos en horizontes largos comparados con SARIMA.

Tabla 5.7: Resultados para la temperatura ambiente.

Heredar	Horizonte	R^2	MSE	MAE
AR	120	0'16	3'02	1'32
SARIMA	120	-0'21	4'33	1'75
=	120	-0'02	3'63	1'65
AR	400	-0'68	8'59	2'26
SARIMA	400	-2'40	17'36	3'65
=	400	-0'27	6'47	1'89
AR	420	-0'49	7'98	2'21
SARIMA	420	-1'79	14'95	3'31
=	400	-0'27	6'47	1'89

Para el nivel del agua (Tabla 5.8), el modelo AR muestra un rendimiento aceptable solo en el horizonte de 120, con un R^2 cercano a cero. En horizontes más largos, el rendimiento del modelo AR decae como para los valores de la temperatura ambiente. El modelo SARIMA muestra el peor rendimiento, con R^2 negativos y altos valores de MSE en todos los horizontes. El modelo que hereda el último valor, aunque también tiene R^2 negativos, presenta valores de MSE y MAE más bajos que SARIMA en horizontes largos, indicando un rendimiento ligeramente mejor en comparación.

Tabla 5.8: Resultados para el nivel del agua.

Heredar	Horizonte	R^2	MSE	MAE
AR	120	0'01	3'51	1'30
SARIMA	120	-0'03	3'65	1'31
=	120	-0'45	5'13	1'90
AR	400	-0'19	5'52	1'85
SARIMA	400	-0'38	6'43	2'09
=	400	-1'04	9'50	2'67
AR	420	-0'22	5'65	1'90
SARIMA	420	-0'52	7'03	2'23
=	420	-0'28	5'92	1'96

A continuación, se presentan los resultados para el Experimento 3.

Observando el horizonte de 120 horas, los modelos que destacan en términos del R^2 entre los diez mejores incluyen RF, LSTM, Lasso, CNN y LR en diversas configuraciones. Para un horizonte más prolongado de 400 horas, los mejores modelos continúan incluyendo a Lasso y RF, a los cuales se suman SVM y KNN. En un horizonte aún más extenso de 420 horas, Lasso sigue siendo relevante y se le une MLP como uno de los mejores modelos. Por lo tanto, Lasso es el único modelo que se sostiene en todos estos rangos de horizontes, aunque no ocupe el primer lugar en ninguno de ellos.

En cuanto a si es más efectivo heredar valores de la temperatura ambiente y el nivel del agua, o estimarlos mediante un modelo AR, y si es mejor reentrenar el modelo con los nuevos datos predichos o no, los resultados son demasiado variados para llegar a una conclusión definitiva. No obstante, es notable que, para el horizonte de 120 horas, el modelo SARIMA parece funcionar mejor, mientras que para el horizonte de 400 horas resulta más efectivo heredar los valores directamente.

En la Tabla 5.9 se pueden observar los resultados obtenidos por los tres mejores modelos para cada horizonte en este experimento.

Tabla 5.9: Resultados para el Experimento 3.

Modelo	Horizonte	Heredar	Reentrenar	R^2	MSE	MAE
RF	120	SARIMA	Sí	0'39	0'35	0'42
Lasso	120	SARIMA	No	0'17	0'48	0'52
CNN	120	SARIMA	Sí	0'14	0'49	0'60
MLP	400	=	Sí	-3'29	4'4	1'86
CNN	400	=	No	-3'34	4'45	1'87
LR	400	=	Sí	-4'75	5'88	2'21
MLP	420	AR	No	-6'03	11'41	3'15
MLP	420	SARIMA	No	-8'01	14'62	3'66
Lasso	420	AR	No	-8'16	14'86	3'59

Comparando los resultados del experimento anterior (Tabla 5.6) y del experimento actual (Tabla 5.9), se observa una mejora en todos los horizontes temporales con este nuevo enfoque, con la excepción del horizonte de 400 horas. En este caso específico, la CNN del Experimento 2, que empleaba una predicción de horizonte fijo, superaba en desempeño a la MLP del experimento actual, aunque la diferencia en los resultados es mínima.

Aunque se siguen obteniendo mejores resultados, la tendencia heredada continúa manifestándose, como se puede observar en la Figura 5.6. Esta persistencia indica que las predicciones todavía están considerablemente influenciadas por los patrones anteriores.

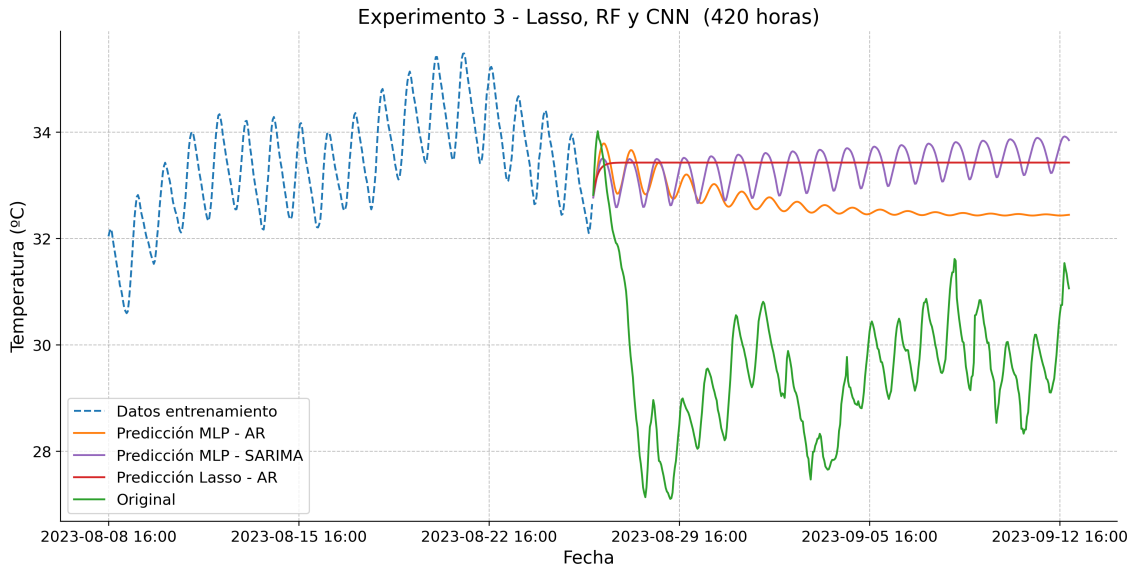

Figura 5.6: Experimento 3: RF, Lasso y CNN con horizonte de 420 horas.

5.3.4. Cuarto experimento

El uso de una LSTM en este experimento, con una estructura relativamente sencilla, ha demostrado ser altamente efectivo para la predicción de la temperatura del agua. El modelo no solo logró superar problemas de tendencia, sino que también obtuvo MSEs muy bajos, logrando los mejores resultados hasta la fecha. Sin embargo, en el horizonte de 120 horas, el tercer experimento con un RF alcanzó un MSE de 0'35, superando ligeramente al modelo LSTM, que obtuvo un MSE de 0'38. A pesar de esta pequeña diferencia, la LSTM ha demostrado ser el mejor modelo para predicciones a largo plazo y para ma-

nejar cambios abruptos en los datos, convirtiéndose en la opción más efectiva encontrada en este trabajo.

Tabla 5.10: Resultados para el Experimento 4.

Horizonte	R^2	MSE	MAE
120	0'28	0'38	0'49
400	-0'16	1'26	0'85
420	-0'25	1'33	0'92

En la Figura 5.7 se presenta el desempeño de la LSTM durante un horizonte de 400 horas, ilustrando cómo el modelo capta eficazmente un cambio brusco. A diferencia de los experimentos anteriores, este resultado demuestra que la LSTM maneja el cambio sin verse significativamente afectada por el fenómeno de la tendencia.

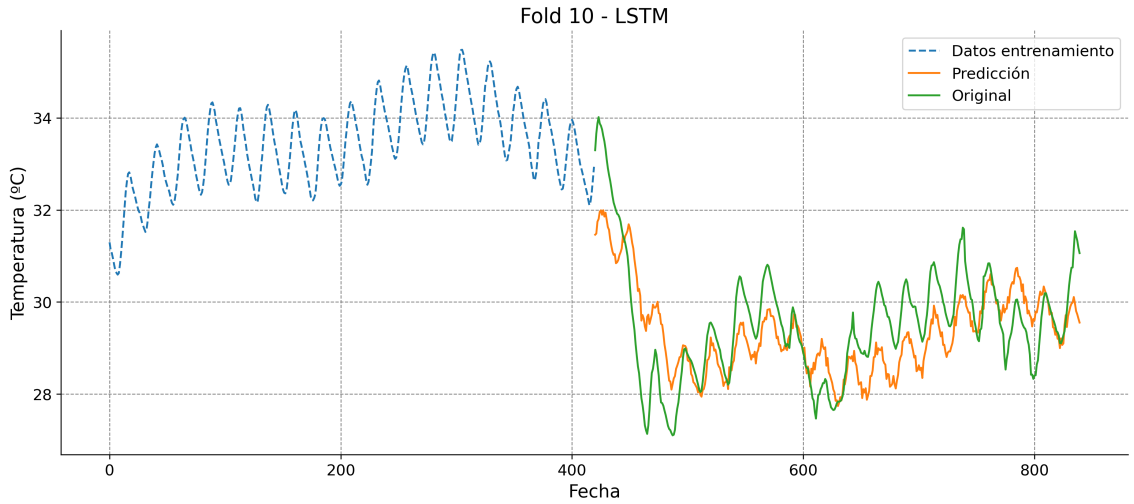

Figura 5.7: Experimento 4: LSTM con horizonte de 420 horas para el Fold 10.

Con respecto a añadir los valores de la temperatura ambiente en horas anteriores, los resultados obtenidos se pueden ver en la Tabla 5.11, pero no mejoran los anteriores.

Tabla 5.11: Resultados para el Experimento 4 añadiendo horas anteriores de Tª ambiente.

Horas Tª ambiente	R^2	MSE	MAE
6, 12, 24 y 36	-0'62	2'70	1'32
12, 24, 36 y 48	-0'21	2'04	1'09
24, 26 y 48	-0'31	2'18	1'75

5.3.5. Quinto experimento

Para evaluar los resultados obtenidos en este último experimento, se realizará una comparación con los obtenidos en el experimento anterior, que presenta los mejores resultados hasta el momento.

Los cinco modelos que probó AutoGluon y que obtuvieron los mejores resultados fueron: WeightedEnsemble, DeepAR, PatchTST y NPTS.

En la Tabla 5.12 se muestran los resultados para los modelos LSTM y AutoGluon. El modelo LSTM tiene un desempeño significativamente mejor que AutoGluon en todos

los aspectos evaluados. Mientras que el LSTM presenta un desempeño moderado, AutoGluon muestra un desempeño muy pobre en este experimento, lo que sugiere que puede no ser adecuado para este conjunto de datos o problema específico.

Tabla 5.12: Resultados para el Experimento 5.

Modelo	R^2	MSE	MAE
LSTM	-0'25	1'33	0'92
AutoGluon	-6'48	11'98	3'11

Además, viendo la Figura 5.8 se puede observar nuevamente cómo el fenómeno de la tendencia afecta las predicciones obtenidas por el modelo AutoGluon como se ha visto en todos los experimentos excepto en el cuarto.

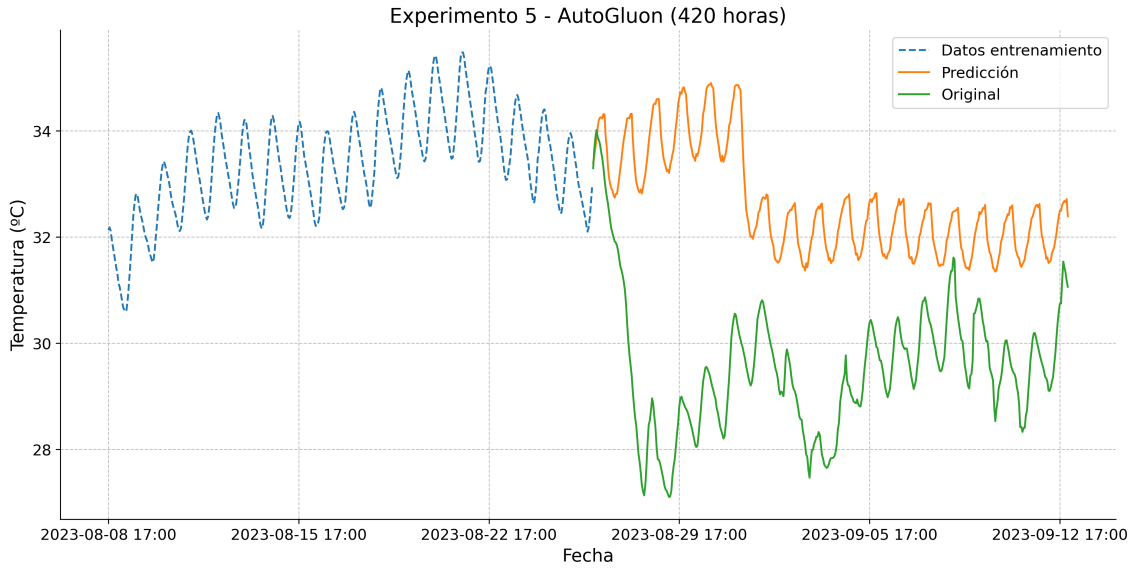

Figura 5.8: Experimento 5: AutoGluon con horizonte de 420 horas.

5.4 Conclusiones

Los modelos SARIMA y LSTM han demostrado ser eficaces para ciertos horizontes temporales en la predicción de la temperatura del agua, adaptándose cada uno a condiciones específicas. SARIMA se ha mostrado altamente eficiente en horizontes cortos, donde los datos son estables y sin cambios abruptos. En contraste, LSTM ha sobresalido en predicciones a largo plazo, adaptándose de manera notable a los cambios abruptos en los datos y mostrando una robustez significativa en estos escenarios.

Los *shocks* o cambios bruscos en los datos han impactado significativamente el rendimiento de los modelos, especialmente en las predicciones a largo plazo. Esta incidencia fue particularmente evidente en el tercer conjunto de datos que incluía estos *shocks*, presentando desafíos considerables. Este fenómeno resalta la necesidad crítica de contar con modelos que puedan adaptarse o ser insensibles a tales cambios abruptos para mantener una alta precisión en las predicciones.

Adicionalmente, los enfoques multivariantes que integran variables como la temperatura ambiente y el nivel del agua han demostrado ser más efectivos que los enfoques univariantes, especialmente bajo condiciones dinámicas y volátiles. Esto resalta la im-

portancia de emplear modelos capaces de manejar múltiples influencias y variables para incrementar la precisión de las predicciones.

La retroalimentación de los modelos mediante predicciones recursivas, donde el modelo se reentrena con cada nueva predicción, ha mostrado ser ventajosa para la adaptabilidad de los modelos ante cambios en los datos, aunque esta técnica conlleva el riesgo de propagar errores si las predicciones iniciales son inexactas.

Es crucial mencionar que los modelos desarrollados, aunque efectivos en sus respectivos horizontes temporales, se basan exclusivamente en datos recogidos durante la temporada estival. Esta limitación sugiere que los resultados y la efectividad de estos modelos podrían no ser representativos ni transferibles a condiciones invernales, debido a la variabilidad climática entre las estaciones que puede influir significativamente en las condiciones del agua. Por lo tanto, la importancia de incorporar datos de múltiples temporadas para desarrollar un modelo más robusto y aplicable a lo largo del año es fundamental, asegurando que pueda manejar las variaciones estacionales que afectan las métricas y fenómenos estudiados.

Además, la tendencia inicial en los datos representó un considerable desafío para los modelos tradicionales como SARIMA, especialmente en los horizontes de predicción más largos, debido a la influencia de la tendencia de los datos previos que limitaba la capacidad de estos modelos para generar predicciones precisas y confiables. En contraste, la implementación de la red LSTM mostró una capacidad mejorada para manejar estas tendencias, ofreciendo una solución efectiva a los problemas que los modelos más tradicionales no pudieron resolver.

La inclusión de las temperaturas ambiente de horas anteriores no produjo mejores resultados. Asimismo, se probó AutoGluon, una de las herramientas más poderosas en el mercado del ML. Sin embargo, los resultados obtenidos con esta herramienta tampoco mejoraron el desempeño.

Finalmente, la capacidad de la LSTM para superar las limitaciones impuestas por las tendencias iniciales subraya la importancia de adaptar las metodologías de modelado a la naturaleza específica de los datos y a los desafíos particulares que presentan. Asimismo, los resultados enfatizan la necesidad de seguir explorando y desarrollando métodos de modelado que puedan manejar de manera más efectiva la variabilidad y los *shocks* en los datos. Esto incluye optimizar los modelos para distintos horizontes de predicción a fin de mejorar su aplicabilidad en una variedad de situaciones prácticas, subrayando la importancia de seleccionar cuidadosamente el modelo y la configuración adecuada para los tipos de datos y los contextos específicos, y de desarrollar estrategias robustas para manejar datos con variaciones abruptas en horizontes de predicción prolongados.

CAPÍTULO 6
Reconstrucción histórica

En este capítulo, se aborda el estudio para realizar una reconstrucción histórica de la temperatura del agua desde aproximadamente 1947 hasta el 2007. Esta reconstrucción permitiría acceder a datos esenciales para evaluar el impacto del cambio climático sobre las condiciones acuáticas a lo largo del tiempo, así como su efecto en la fauna y flora del parque natural de La Mata.

El principal desafío radica en que los registros disponibles de la temperatura del agua comienzan únicamente en 2022. Sin embargo, durante el estudio de los datos se ha demostrado la existencia de una correlación significativa entre la temperatura del agua y la temperatura ambiente. Basándose en esta correlación, se planea reconstruir los valores históricos de la temperatura del agua desde 1947 hasta 2007, utilizando los datos de temperatura ambiente disponibles durante ese período.

6.1 Conjunto de datos de estudio

Basándose en los datos de estudio descritos en el Capítulo 4, Sección 4.1, se definen los conjuntos de datos que se emplearán.

La Tabla 4.6 especifica dos conjuntos de datos recolectados por dispositivos IoT y AEMET, respectivamente.

Los datos obtenidos a través de dispositivos IoT se destinarán al entrenamiento de los modelos, dado que contienen la relación entre la temperatura del agua y la temperatura ambiente. Por otro lado, los datos proporcionados por AEMET se utilizarán para realizar predicciones de la temperatura del agua basadas en la temperatura ambiente.

A cada registro de estos conjuntos de datos se le añadirán las temperaturas ambiente registradas 24, 36, 48, 96 y 120 horas (1, 2, 3, 4, y 5 días) antes, con el objetivo de enriquecer la información y mejorar la precisión de las predicciones.

Por último, cabe mencionar que de los datos de IoT se eliminará el nivel del agua, ya que solo se trabajará con la temperatura del agua y la temperatura del ambiente.

6.2 Experimento

Inicialmente, se entrenará una variedad de modelos utilizando el conjunto de datos de IoT. Los modelos seleccionados incluyen: LR, Lasso, DT, KNN, RF, CNN y RNN [14]. Se experimentará con diferentes configuraciones y se aplicará validación cruzada para

identificar la configuración que ofrece el mejor rendimiento. Una vez optimizados, estos modelos se utilizarán para predecir los valores de temperatura en el conjunto histórico.

La estructura de esta CNN comienza con una capa convolucional (*Conv1D*) que tiene 128 filtros, un tamaño de kernel de 2, y una función de activación sigmoidal, tomando una entrada con forma (6, 1). A continuación, la red incluye una capa densa con 64 unidades y activación sigmoidal, seguida de otra capa densa con 32 unidades y activación *ReLU*. Posteriormente, se aplanan los datos con una capa *Flatten* y, finalmente, se incluye una capa de salida densa con una unidad. El modelo se compila utilizando la función de pérdida MSE y el optimizador Adam.

En cuanto a la estructura de la RNN, esta comienza con una capa *SimpleRNN* de 32 unidades que devuelve secuencias y toma una entrada con forma (6, 1). A esta le siguen cuatro capas *SimpleRNN* adicionales con 64, 128, 64 y 32 unidades, respectivamente, donde todas, excepto la última, también devuelven secuencias. Finalmente, la red incluye una capa densa de salida con una unidad. El modelo se compila, al igual que la CNN, utilizando la función de pérdida MSE y el optimizador Adam.

6.3 Resultados

En primer lugar, al aplicar validación cruzada a cada uno de los modelos especificados anteriormente, se obtuvieron los resultados que se muestran en la Tabla 6.1. El método KNN es el que mejor desempeño presenta, seguido de RF. Con un R^2 ligeramente menor, alrededor de 0.5, se encuentran los modelos CNN, DT y LR. Los modelos Lasso y RNN obtuvieron los peores resultados, por lo que se descartaron para el resto de las pruebas.

Tabla 6.1: Resultados para el experimento 4 añadiendo horas anteriores de T[a] ambiente.

Modelo	R^2	MSE	MAE
KNN	0’76	1’59	0’81
RF	0’64	2’33	1’14
CNN	0’49	3’27	1’38
DT	0’44	3’65	1’24
LR	0’43	3’64	1’50
Lasso	0’36	4’13	1’64
RNN	-0’01	6’41	2’13

Posteriormente se probaron distintas configuraciones para los modelos KNN, RF y DT. Para el KNN, se varió el número de vecinos; para el RF, el número de árboles y el número mínimo de muestras que debe tener un nodo para dividirse en nodos hijos; y para el DT, también se ajustaron el número mínimo de muestras y la profundidad del árbol. Los mejores resultados obtenidos fueron:

- **KNN**: El número de vecinos se establece en 3. Este parámetro define cuántos puntos cercanos se considerarán para hacer una predicción. El algoritmo KNN clasifica una muestra basándose en la mayoría de votos de sus 3 vecinos más cercanos.

- **RF**: El modelo utiliza 10 árboles de decisión. Este parámetro determina cuántos árboles compondrán el bosque. El número mínimo de muestras requeridas para dividir un nodo interno se deja en su valor predeterminado, que es 2. Esto significa que un nodo debe tener al menos 2 muestras para poder dividirse.

- **DT**: La profundidad máxima del árbol se establece en 10. Este parámetro limita el número máximo de niveles que puede tener el árbol, ayudando a prevenir el sobreajuste. El número mínimo de muestras necesarias para dividir un nodo se establece en 16, lo que significa que un nodo debe tener al menos 16 muestras para poder dividirse.

Dadas las configuraciones y los modelos indicados, se procedió a predecir los valores de la temperatura del agua en base al histórico de la temperatura ambiente. En este caso, no se disponen de datos para validar cuantitativamente la precisión de los modelos, pero, mediante la visualización de los resultados, se puede observar que los modelos CNN y LR parecen funcionar mejor (ver Figura 6.1). El resto de los métodos modelan bastante bien la época estival, pero fallan en el resto de estaciones (ver Figura 6.2). Esto es comprensible, ya que los datos de entrenamiento solo consideran el verano. Si se dispusiera de datos de invierno, es probable que los resultados fueran mejores para todos los modelos probados.

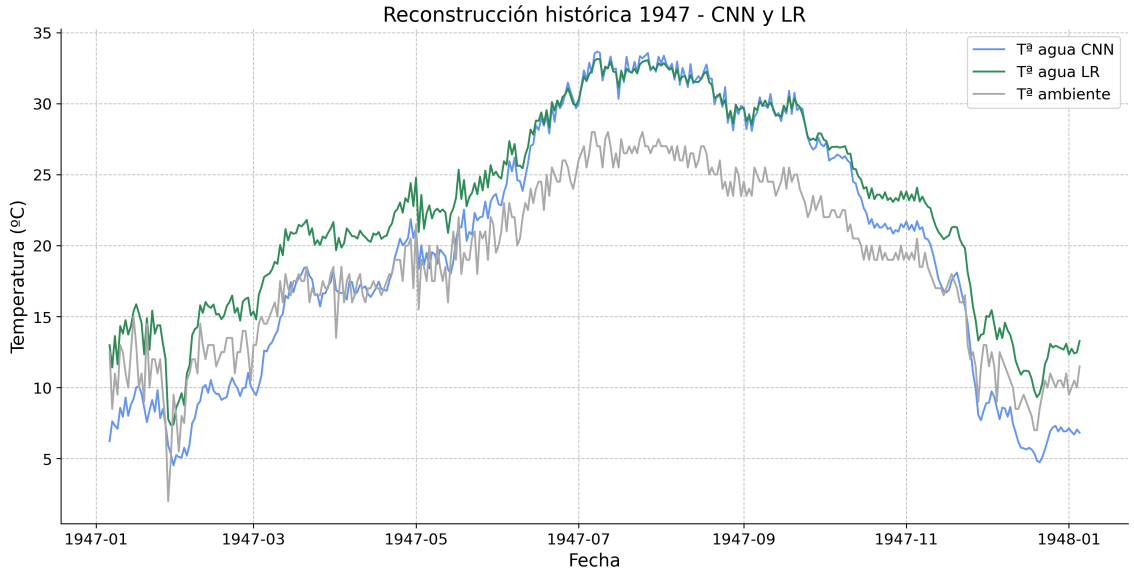

Figura 6.1: Reconstrucción histórica de la temperatura del agua con la CNN y la LR.

Viendo la Figura 6.1, ambas series de datos, CNN y LR, muestran una tendencia general similar a lo largo del año 1947. Las temperaturas aumentan desde principios de año, alcanzan un pico a mediados de año, y luego disminuyen hacia finales de año, siguiendo un patrón estacional claro. Sin embargo, las diferencias empiezan a notarse en la forma en que estas temperaturas varían. La serie de datos de CNN (en azul) tiende a tener variaciones más pronunciadas y rápidas en comparación con la serie de LR (en verde), que parece ser más suavizada. Durante algunos períodos específicos, como a mediados de año, las temperaturas de CNN pueden ser ligeramente más altas o más bajas que las de LR.

En términos de precisión y detalle, el modelo CNN parece capturar más variabilidad y detalles en los datos debido a su capacidad para aprender patrones complejos y no lineales. Esto resulta en fluctuaciones más abruptas en la serie de CNN, indicando una mayor sensibilidad a los cambios locales. Por otro lado, el modelo LR proporciona una estimación más suave y lineal de las temperaturas, con menor variabilidad. La serie de LR muestra una línea de tendencia más simplificada, ya que la regresión lineal asume una relación lineal entre las variables. Esto sugiere que mientras las reconstrucciones de CNN pueden ser más detalladas y mostrar más variabilidad, las predicciones de LR son más suaves y pueden ser más fáciles de interpretar para observar tendencias generales.

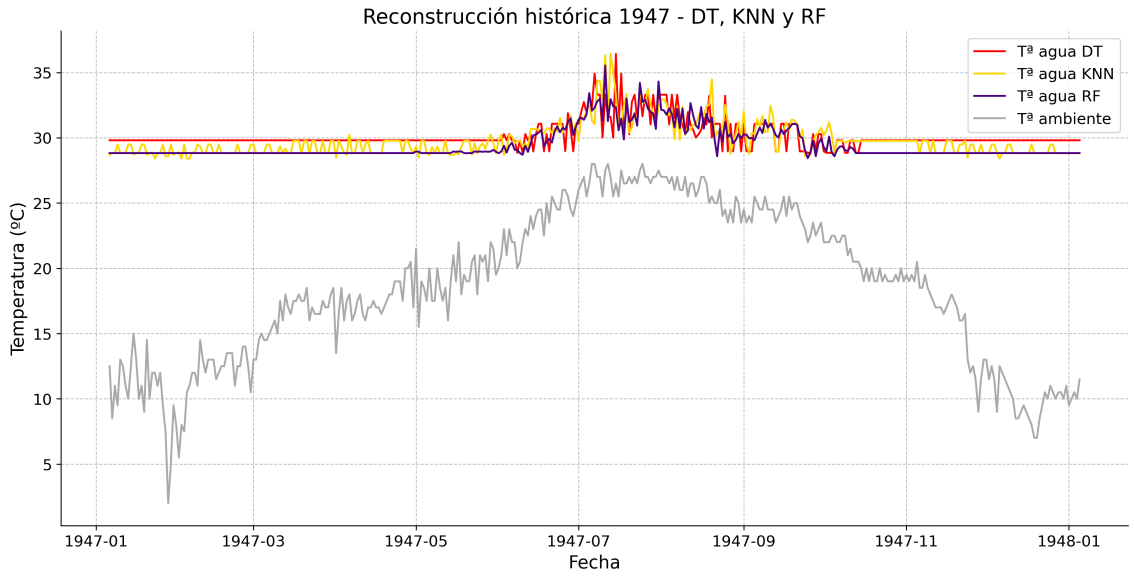

Figura 6.2: Reconstrucción histórica de la temperatura del agua con el DT, el KNN, y el RF.

Comparando con la temperatura ambiente (en gris), ambas reconstrucciones de la temperatura del agua (CNN y LR) siguen patrones similares, pero generalmente presentan valores más altos, especialmente en los picos de verano. Observando los valores reales (datos recolectados por los dispositivos IoT)

En resumen, la gráfica sugiere que las predicciones de temperatura de CNN ofrecen más detalles y varían más, mientras que las predicciones de LR son más suaves y lineales.

Se dibujó la diferencia entre 1947, el primer año con registros disponibles, y uno de los últimos, específicamente 2005, para evaluar el impacto del cambio climático (ver Figura [**?**]). Como se puede observar en la figura, los valores son bastante similares, excepto que en 2005 se alcanzan temperaturas algo más altas. Al examinar el conjunto de datos más reciente (2023), se encuentra que la temperatura ambiente más alta registrada es de 35°C, lo que indica un incremento gradual en los valores de temperatura con el tiempo.

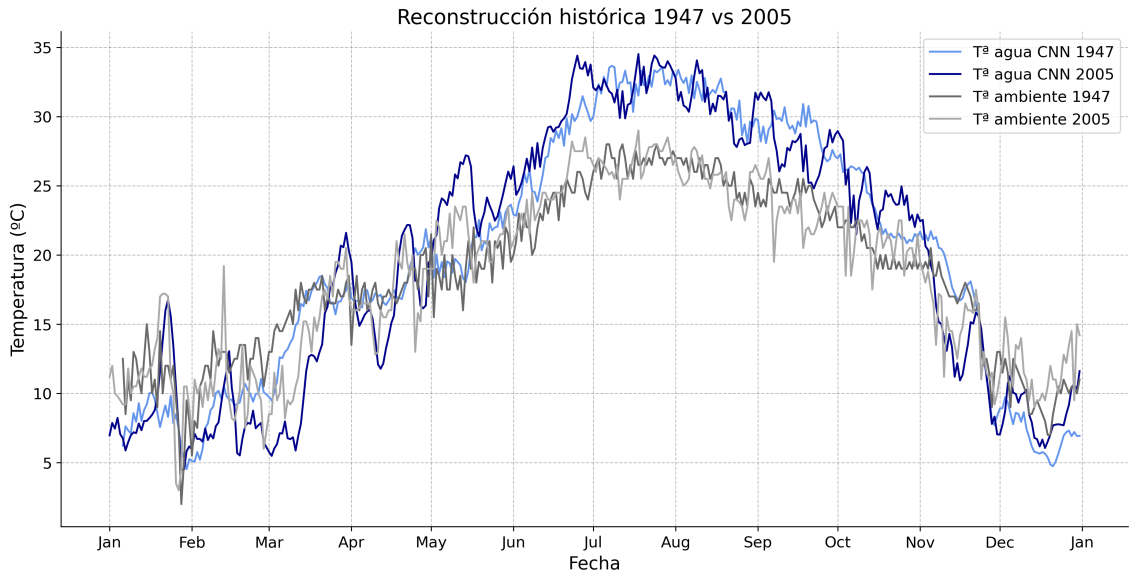

Figura 6.3: Comparativa de la temperatura ambiente y del agua en 1947 y 2005.

En los últimos años el número de aves que han visitado el parque natural ha sido mucho menor que otros. Las conclusiones que se sacaron fueron que con el aumento de las temperaturas en los últimos años, la temperatura del agua también ha aumentado, lo que ha provocado botulismo aviar y mortalidad de la artemia.

El botulismo aviar es una de las enfermedades más importantes en las aves migratorias de todo el mundo, especialmente en las aves acuáticas. Está causado por la ingestión de una neurotoxina producida por la bacteria Clostridium botulinum, que se encuentra comúnmente en la naturaleza. La epidemiología del botulismo aviar es compleja y los brotes son difíciles de predecir y prevenir, ya que intervienen múltiples factores físicos, químicos y biológicos que hacen posible que se produzcan brotes recurrentes en un humedal mientras que no aparecen en humedales vecinos de características similares. Entre estos factores, el más importante es que la temperatura del agua no sea muy elevado [16].

La preocupación es tan alta que en 2022 se declaró la emergencia para tomar medidas de control del botulismo [70]. Además de todo esto, el mismo aumento de temperatura provoca la mortalidad de las artemias salinas. Las artemias salinas son los crustáceos diminutos que viven en lagos y lagunas salobres interiores. Son importantes porque constituyen la dieta básica de la numerosa avifauna de estos humedales, que suelen vivir en aguas con alta salinidad, tan abundantes en las orillas de la laguna de La Mata [20, 17].

6.4 Conclusiones

El estudio presentado se centra en la reconstrucción histórica de la temperatura del agua, abarcando el periodo de 1947 a 2007. Aunque los registros directos de temperatura del agua solo están disponibles desde 2022, se ha logrado establecer una correlación significativa entre la temperatura del agua y la ambiente, lo cual ha permitido reconstruir los valores históricos usando datos de temperatura ambiente de ese periodo.

Para llevar a cabo este análisis, se utilizaron datos recopilados tanto por dispositivos IoT como por la Agencia Estatal de Meteorología (AEMET). Los primeros fueron clave para entrenar modelos predictivos, mientras que los segundos se usaron para realizar predicciones sobre la temperatura del agua. A estos datos se les agregaron registros de temperaturas ambiente de periodos anteriores, enriqueciendo la información y mejorando la precisión de las predicciones.

Respecto a los modelos de ML empleados, se experimentó con varios, incluidos KNN, RF, CNN, DT, LR, Lasso, y RNN, aplicando la técnica de validación cruzada para evaluar su desempeño. Aunque KNN y RF mostraron buenos resultados iniciales, y a la hora de emplear el histórico modelaban adecuadamente las temperaturas durante el verano, se verificó que su rendimiento disminuía para otras estaciones. Esto se debe a la limitación de los datos de entrenamiento, donde predominaban registros de la temporada estival. En general, los modelos que mejores resultados dieron fueron realmente la CNN y la LR.

Finalmente, se constató un incremento gradual en la temperatura del agua a lo largo del tiempo, desde 1947 hasta 2023, sugiriendo un posible impacto del cambio climático sobre las condiciones acuáticas. Este fenómeno podría tener implicaciones importantes en la biodiversidad del parque natural de La Mata. Este estudio resalta no solo la utilidad de técnicas avanzadas de análisis de datos para comprender el cambio climático, sino también la importancia de disponer de registros históricos completos para realizar predicciones más precisas, y entender mejor los cambios ambientales a largo plazo.

Las conclusiones principales indican que el aumento de las temperaturas ha tenido efectos adversos significativos en el parque natural, contribuyendo a la disminución del número de aves debido a brotes de botulismo aviar y la mortalidad de artemias salinas,

esenciales para la avifauna local. Estas observaciones subrayan la importancia de monitorear y gestionar las condiciones ambientales para preservar la biodiversidad y la salud del ecosistema.

CAPÍTULO 7
Conclusiones

Este Trabajo de Fin de Máster se ha enfocado en modelizar la temperatura del agua de la Laguna de La Mata mediante técnicas avanzadas de aprendizaje automático. El objetivo es predecir cambios futuros y entender tendencias históricas, evaluando su impacto en el ecosistema. Los modelos desarrollados han demostrado ser herramientas robustas, capaces de adaptarse a la complejidad de los datos ambientales y ofrecer predicciones confiables y precisas.

7.1 Modelado predictivo

El estudio ha realizado un análisis exhaustivo de la temperatura del agua en la Laguna de La Mata, empleando diversos modelos de aprendizaje automático. Los modelos SARIMA y LSTM se destacaron por su eficacia en horizontes temporales distintos: SARIMA a corto plazo, y LSTM a largo plazo, mostrando gran robustez ante cambios abruptos.

El análisis multivariante, que incluye variables como la temperatura ambiente, resultó más efectivo que los enfoques univariantes, especialmente en condiciones dinámicas. La retroalimentación de los modelos mediante predicciones recursivas permitió ajustar los modelos a cambios en los datos, aunque presenta el riesgo de propagar errores si las predicciones iniciales son inexactas.

Los modelos desarrollados basados en datos estivales podrían no ser representativos en condiciones invernales. Es crucial incorporar datos de múltiples temporadas para crear modelos más robustos y aplicables durante todo el año.

En resumen, el uso de modelos SARIMA y LSTM ha permitido explorar con precisión las dinámicas de la temperatura del agua en la Laguna de La Mata. La incorporación de múltiples variables ambientales y el reentrenamiento continuo de los modelos mejoraron la precisión y adaptabilidad en la predicción de la temperatura del agua.

7.2 Reconstrucción histórica

Se reconstruyó la temperatura histórica del agua entre 1947 y 2007 utilizando datos de temperatura ambiente, pese a la falta de registros directos hasta 2022. Este enfoque permitió una comprensión más precisa de las tendencias históricas y su impacto en la biodiversidad del Parque Natural de La Mata.

La reconstrucción histórica reveló un incremento gradual de la temperatura del agua, probablemente influenciado por el cambio climático. Este estudio subraya la importancia de contar con registros históricos completos para realizar predicciones precisas, y para

comprender mejor los cambios ambientales a largo plazo. Aunque una CNN podría ofrecer buenos resultados a partir de datos de temperatura ambiental, no se pudo demostrar su eficacia mediante métricas estadísticas como MSE o R^2 debido a la limitación de datos en diferentes épocas del año.

En los últimos años, la subida de las temperaturas ha causado un incremento en la temperatura del agua en el parque natural, lo que ha provocado brotes de botulismo aviar y la mortalidad de artemias salinas, cruciales para la dieta de las aves acuáticas. Este fenómeno ha resultado en una notable disminución del número de aves migratorias que visitan el parque. El botulismo aviar, una enfermedad grave para las aves se ve favorecida por altas temperaturas del agua. La preocupación por este problema llevó a declarar una emergencia en 2022 para controlar la enfermedad. Además, la mortalidad de las artemias salinas debido al aumento de la temperatura ha afectado negativamente la cadena alimentaria, poniendo en riesgo la biodiversidad y la salud del ecosistema del parque.

7.3 Trabajo futuro

Este estudio sienta una base para futuras investigaciones y aplicaciones prácticas, abriendo varias vías para exploración adicional:

- **Integración de nuevos datos**: Ampliar la recolección de datos para incluir variables adicionales como precipitaciones y dirección del viento, especialmente en invierno.
- **Modelos de aprendizaje**: Experimentar con más técnicas de aprendizaje automático y profundo, como redes neuronales recurrentes complejas o transformadores, que podrían mejorar la precisión de las predicciones a largo plazo.
- **Aplicación en otros ecosistemas**: Validar y adaptar los modelos desarrollados en otros ecosistemas lagunares o entornos acuáticos diversos.
- **Desarrollo de herramientas predictivas en tiempo real**: Crear una plataforma que utilice modelos en tiempo real para proporcionar alertas tempranas sobre cambios críticos en las condiciones del agua.
- **Estudios de impacto a largo plazo**: Realizar estudios longitudinales sobre los efectos acumulativos de los cambios en la temperatura del agua sobre la biodiversidad local y la salud del ecosistema.

Estas direcciones potencian la continuidad del trabajo actual, y sugieren la posibilidad de colaboraciones interdisciplinarias para abordar las complejas interacciones entre los factores climáticos y biológicos en entornos naturales.

La continua investigación y desarrollo en este campo no solo contribuirán a la preservación de la biodiversidad en la Laguna de La Mata, sino que también proporcionarán conocimientos valiosos aplicables a otros entornos naturales, promoviendo una gestión ambiental más informada y eficaz.

Bibliografía

[1] Reza Abdi, Ashley Rust, and Terri S Hogue. Development of a multilayer deep neural network model for predicting hourly river water temperature from meteorological data. *Frontiers in Environmental Science*, 9:738322, 2021.

[2] M. Waqar Ahmed. Understanding mean absolute error (mae) in regression: A practical guide, 2023. 14-02-2024.

[3] AutoGluon. Autogluon time series forecasting tutorial. https://auto.gluon.ai/stable/tutorials/timeseries/index.html, 2023. 03-04-2024.

[4] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[5] Pritha Bose. Guide to lasso and ridge regression techniques with use cases, Aug 2023. 03-02-2024.

[6] Pritha Bose. Knn algorithm in machine learning - a complete 360-degree guide, Feb 2024. 03-02-2024.

[7] Peter J Brockwell and Richard A Davis. *Introduction to time series and forecasting*. Springer, 3 edition, 2016.

[8] Peter Bruce, Andrew Bruce, and Peter Gedeck. *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media, 2020.

[9] Daniel Caissie. The thermal regime of rivers: a review. *Freshwater Biology*, 51(8):1389–1406, 2006.

[10] Samprit Chatterjee and Ali S Hadi. *Regression analysis by example*. John Wiley & Sons, 2015.

[11] Shih-Lun Chen, He-Sheng Chou, Chun-Hsiang Huang, Chih-Yun Chen, Liang-Yu Li, Ching-Hui Huang, Yu-Yu Chen, Jyh-Haw Tang, Wen-Hui Chang, and Je-Sheng Huang. An intelligent water monitoring iot system for ecological environment and smart cities. *Sensors*, 23(20):8540, 2023.

[12] Xiaodong Chen, Frank Lupi, and Jianguo Liu. Accounting for ecosystem services in compensating for the costs of effective conservation in protected areas. *Biological conservation*, 215:233–240, 2017.

[13] François Chollet. *The Sequential model*, 2023.

[14] François Chollet. *SimpleRNN layer*, 2023.

[15] DataScientest. Sarimax model: What is it & how can it be applied to time series. https://datascientest.com/en/sarimax-model-what-is-it-how-can-it-be-applied-to-time-series, Access Year. 03-02-2024.

[16] Gobierno de España. Respuesta del gobierno a la pregunta escrita del congreso número 184/2215. Registro General, Entrada 60924, 2020. 12-05-2024.

[17] Parques Naturales de la Comunidad Valenciana. Artemia salina: un crustáceo muy salado. https://parquesnaturales.gva.es/es/web/pn-lagunas-de-la-mata-torrevieja/documents/-/asset_publisher/I56FXxph7nrG/content/artemia-salina-un-crustaceo-muy-salado, 2024. 12-05-2024.

[18] Diario Oficial de la Unión Europea. Reglamento de ejecución (ue) 2023/138 de la comisión de 21 de diciembre de 2022 por el que se establecen una lista de conjuntos de datos específicos de alto valor y modalidades de publicación y reutilización. Diario Oficial de la Unión Europea, 2023. 04/02/2024.

[19] Emmanuel. Introduction to svm - support vector machine algorithm in machine learning, Jun 2022. 03-02-2024.

[20] Food and Agriculture Organization of the United Nations (FAO). Artemia biology. https://www.fao.org/3/W3732E/w3732e0m.htm. 12-05-2024.

[21] GeeksforGeeks. Autoregressive (ar) model for time series forecasting. https://www.geeksforgeeks.org/autoregressive-ar-model-for-time-series-forecasting/?ref=header_search. 03-02-2024.

[22] GeeksforGeeks. Sarima: Seasonal autoregressive integrated moving average. https://www.geeksforgeeks.org/sarima-seasonal-autoregressive-integrated-moving-average/?ref=header_search. 03-02-2024.

[23] GeeksforGeeks. Deep learning: Introduction to long short term memory, Dec 2023. 05-02-2024.

[24] Juan Carlos Gutiérrez-Estrada, Emiliano De Pedro Sanz, Rafael López-Luque, and Inmaculada Pulido-Calvo. Estimación a corto plazo de la temperatura del agua. aplicación en sistemas de producción en medio acuático. *Ingeniería del Agua*, 12(1):77–92, 2005.

[25] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc., 2022.

[26] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[27] Tomislav Hengl, Markus G. Walsh, Jonathan Sanderman, Ichsani Wheeler, Sandy P. Harrison, and Iain C. Prentice. Global mapping of potential natural vegetation: an assessment of machine learning algorithms for estimating land potential. *PeerJ*, 6:e5457, 2018.

[28] Falk Huettmann. *Machine Learning for Ecology and Sustainable Natural Resource Management*. Springer, 2018.

[29] IBM. *Funciones de autocorrelación y autocorrelación parcial*, 2024. 10-02-2024.

[30] Carlos Iglesias. Pelando la cebolla de la gobernanza de los datos abiertos. Blog Datos.gob.es, 2022. 04/02/2024.

[31] Rana Muhammad Adnan Ikram, Reham R. Mostafa, Zhihuan Chen, Kulwinder Singh Parmar, Ozgur Kisi, and Mohammad Zounemat-Kermani. Water temperature prediction using improved deep learning methods through reptile search algorithm and weighted mean of vectors optimizer. *Journal of Marine Science and Engineering*, 11(2):259, 2023.

[32] Irene Iniesta-Arandia, Marina García-Llorente, Pedro A Aguilera, Carlos Montes, and Berta Martín-López. Socio-cultural valuation of ecosystem services: uncovering the links between values, drivers of change, and human well-being. *Ecological economics*, 108:36–48, 2014.

[33] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, et al. *An introduction to statistical learning*, volume 112. Springer, 2013.

[34] Qijun Jiang, Arnold K Bregt, and Lammert Kooistra. Formal and informal environmental sensing data and integration potential: Perceptions of citizens and experts. *Science of the total environment*, 619:1133–1142, 2018.

[35] Sara Kaffashi, Mohd Rusli Yacob, Maynard S Clark, Alias Radam, and Mohd Farid Mamat. Exploring visitors' willingness to pay to generate revenues for managing the national elephant conservation center in malaysia. *Forest Policy and Economics*, 56:9–19, 2015.

[36] Archish Rai Kapil. Advantages and disadvantages of decision tree in machine learning, Oct 2023. 03-02-2024.

[37] Vicky Katsoni and Natali Dologlou. Ict applications in smart ecotourism environments. *Smart Cities in the Mediterranean: Coping with Sustainability Objectives in Small and Medium-sized Cities and Island Communities*, pages 225–244, 2017.

[38] Keras. Keras Documentation: LSTM. https://keras.io/api/layers/recurrent_layers/lstm/, 2023. 19-03-2024.

[39] Marianne Kettunen and Patrick ten Brink. *Social and economic benefits of protected areas: an assessment guide*. Routledge, 2013.

[40] Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*, volume 26. Springer, 2013.

[41] Douglas C Montgomery, Cheryl L Jennings, and Murat Kulahci. *Introduction to time series analysis and forecasting*. John Wiley & Sons, 2015.

[42] Author's Name. Time series forecasting with arima, sarima, and sarimax. https://towardsdatascience.com/time-series-forecasting-with-arima-sarima-and-sarimax-ee61099e78f6, Access Year. 03-02-2024.

[43] Nahal Norouzi, Gerd Bruder, Brandon Belna, Stefanie Mutter, Damla Turgut, and Greg Welch. A systematic review of the convergence of augmented reality, intelligent virtual agents, and the internet of things. In Fadi Al-Turjman, editor, *Artificial Intelligence in IoT*, pages 1–24. Springer Nature Switzerland AG, 2019.

[44] Ivan Novkovic, Goran B. Markovic, Djordje Lukic, Slavoljub Dragicevic, Marko Milosevic, Snezana Djurdjic, Ivan Samardzic, Tijana Lezaic, and Marija Tadic. Gis-based forest fire susceptibility zonation with iot sensor network support, case study—nature park golija, serbia. *Sensors*, 21(19):6520, 2021.

[45] Vatsala Nundloll, Barry Porter, Gordon S. Blair, Bridget Emmett, Jack Cosby, Davey L. Jones, Dave Chadwick, Ben Winterbourn, Philip Beattie, Graham Dean, Rory Shaw, Wayne Shelley, Mike Brown, and Izhar Ullah. The design and deployment of an end-to-end iot infrastructure for the natural environment. *Future Internet*, 11(6):129, 2019.

[46] University of Pittsburgh. Lecture notes on autoregressive models. https://people.cs.pitt.edu/~milos/courses/cs3750/lectures/class16.pdf. 03-02-2024.

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. *DecisionTreeRegressor*, 2024.

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. *KNeighborsRegressor*, 2024.

[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. *Lasso*, 2024.

[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. *LinearRegression*, 2024.

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. *MLPRegressor*, 2024.

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. *RandomForestRegressor*, 2024.

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. *SVR*, 2024.

[54] Penn State. Autoregressive models - lesson 2.1. https://online.stat.psu.edu/stat501/lesson/t/t.2/t.2.1-autoregressive-models, Access Year. 03-02-2024.

[55] W. Penny and L. Harrison. Chapter 40 - multivariate autoregressive models. In KARL FRISTON, JOHN ASHBURNER, STEFAN KIEBEL, THOMAS NICHOLS, and WILLIAM PENNY, editors, *Statistical Parametric Mapping*, pages 534–540. Academic Press, London, 2007.

[56] Josef Perktold, Skipper Seabold, Jonathan Taylor, and statsmodels developers. *statsmodels.tsa.stattools.pacf*, 2023. 10-02-2024.

[57] Josef Perktold, Skipper Seabold, Jonathan Taylor, and statsmodels developers. *statsmodels.tsa.stattools.acf*, 2024. 10-02-2024.

[58] Josef Perktold, Skipper Seabold, Jonathan Taylor, and statsmodels developers. *statsmodels.tsa.stattools.adfuller*, 2024. 10-02-2024.

[59] pmdarima developers. *pmdarima.arima.auto_arima*, 2023. 11-02-2024.

[60] Sushmita Poudel. Recurrent neural network (rnn) architecture explained. *Medium*, 2023.

[61] Robin Rai. What is linear regression in machine learning?, Feb 2022. 03-02-2024.

[62] M. Rajesh and S. Rehana. Prediction of river water temperature using machine learning algorithms: a tropical river system of india. *Journal of Hydroinformatics*, 23(3):605–621, 2021.

[63] Amar Rao, Amogh Talan, Shujaat Abbas, Dhairya Dev, and Farhad Taghizadeh-Hesary. The role of natural resources in the management of environmental sustainability: Machine learning approach. *Resources Policy*, 82:103548, 2023.

[64] H. Rezapouraghdam et al. Application of machine learning to predict visitors' green behavior in marine protected areas: Evidence from cyprus. *Journal of Sustainable Tourism*, 31(5):2495–2514, 2023.

[65] Julian Rode, Heidi Wittmer, Lucy Emerton, and Christoph Schröter-Schlaack. 'ecosystem service opportunities': A practice-oriented framework for identifying economic instruments to enhance biodiversity and human livelihoods. *Journal for nature conservation*, 33:35–47, 2016.

[66] Nidhi Sahai. Mastering random forest regression: A comprehensive guide, Sep 2023. 03-02-2024.

[67] Nidhi Sahai. Convolutional neural network: Layers, types, & more, Jan 2024. 08-02-2024.

[68] Ritu Santra. Tests for stationarity in time series - dickey-fuller test & augmented dickey-fuller (adf) test, 2020. 10-02-2024.

[69] Wesley M Sarmento and Joel Berger. Human visitation limits the utility of protected areas as ecological baselines. *Biological Conservation*, 212:316–326, 2017.

[70] Cadena SER. El consell aprueba la declaración de emergencia de las acciones para controlar el botulismo en el sur de alicante. https://cadenaser.com/comunitat-valenciana/2022/07/29/el-consell-aprueba-la-declaracion-de-emergencia-de-las-acciones-para-controlar-el-botu?outputType=amp, 2022. 12-05-2024.

[71] Amazon Web Services. Amazon sagemaker developer guide: Autogluon-tabular. https://docs.aws.amazon.com/es_es/sagemaker/latest/dg/autogluon-tabular.html, 2023. 03-04-2024.

[72] Statsmodels Developers. seasonal_decompose. https://www.statsmodels.org/dev/generated/statsmodels.tsa.seasonal.seasonal_decompose.html, 2023. 10-02-2024.

[73] statsmodels developers. *statsmodels.tsa.arima.model.ARIMA*, 2023. 11-02-2024.

[74] statsmodels developers. *statsmodels.tsa.ar_model.AutoReg*, 2023. 11-02-2024.

[75] statsmodels developers. *statsmodels.tsa.statespace.sarimax.SARIMAX*, 2023. 11-02-2024.

[76] Teik Toe Teoh and Yu Jin Goh. Time series. In *Artificial Intelligence in Business Management*, Machine Learning: Foundations, Methodologies, and Applications, pages 65–85. Springer, Singapore, 2023.

[77] Akancha Tripathi. What is perceptron? introduction, definition & more, Jul 2022. 03-02-2024.

[78] Pradeep Vankayala and Manjunath Vankayala. Use of modern communication technologies iot for natural resource conservation -water. In *2019 IEEE International Conference on Communication Technologies (ICCT)*, pages 1–6. IEEE, 2019.

[79] Li Wang, Kai Wang, Jianhui Xu, and Jingjie Liu. Ecological restoration process of watershed land space with intelligent iot technology. *Water Supply*, 23(9):3881–3898, 2023.

[80] Department of Statistics Yale University and Data Science. 03-02-2024.

[81] Xinzhe Yin, Jinghua Li, Seifedine Nimer Kadry, and Ivan Sanz-Prieto. Artificial intelligence assisted intelligent planning framework for environmental restoration of terrestrial ecosystems. *Environmental Impact Assessment Review*, 86:106493, 2021.

APÉNDICE A
Entorno de ejecución

Los experimentos presentados se han llevado a cabo en un ordenador con las siguientes especificaciones de hardware:

- **CPU**: Intel(R) Core(TM) i7-8750H, con 6 núcleos a 2,20 GHz, 8 GB de RAM y 9 MB de memoria SmartCache.
- **GPU**: NVIDIA GeForce GTX 1050, con 2 GB GDDR5, 640 núcleos CUDA, PCI Express x16 3.0.
- Unidad de estado sólido de 256 GB.

En cuanto al software, el ordenador opera con el sistema operativo Windows 10 y tiene instalada la versión 3.12 de Python. Todo el código fue desarrollado y ejecutado utilizando los entornos de desarrollo Jupyter y Visual Studio Code, específicamente mediante Jupyter Notebooks.

La Tabla A.1 muestra las diferentes librerías empleadas, así como sus versiones.

Tabla A.1: Librerías junto a sus versiones empleadas en el trabajo.

Librería	Versión	Descripción
Pandas	2.2.1	Manejo de datos
Numpy	1.26.4	Manejo de datos
Statsmodels	0.14.1	Análisis y modelado de series temporales
Sklearn	1.4.1.post1	Métricas
Matplotlib	3.8.3	Gráficas
Seaborn	0.13.2	Gráficas
Keras	3.0.5	Modelado de RNAs
TensorFlow-intel	2.16.1	Modelado de RNAs