UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR

Grado en Ingeniería de Tecnologías y Servicios de la Telecomunicación

TRABAJO FIN DE GRADO

Estudio del estado químico y fisicoquímico actual y futuro del Mar Menor usando algoritmos de Machine Learning

Autor: Ángel Allepuz Conesa
Tutor: Gonzalo Martínez Muñoz

enero 2025

Ángel Allepuz Conesa
Estudio del estado químico y fisicoquímico actual y futuro del Mar Menor usando algoritmos de Machine Learning

Ángel Allepuz Conesa
C\ Francisco Tomás y Valiente N.º 11

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

A mi familia, amigos y a la Región de Murcia

Men lie, women lie, numbers don't

Lil B, the Based God

Agradecimientos

A mi tutor, Gonzalo Martínez, por inspirar esta idea en mí gracias a su trabajo en el río Tajo y por aceptar mi idea con confianza.

A mis amigos, los de toda la vida y los de la nueva tierra, por ser la fuente de alegría e inspiración que me levantaba cada mañana azul.

A mis hermanos, primos y abuelos, por ser los mejores posibles ejemplos de amor y apoyo incondicional que podría haber soñado.

Y por encima de todo, a mis padres, por su paciencia conmigo, por su tiempo y atención, y por darme la oportunidad de convertirme en alguien.

Resumen

El Mar Menor, un ecosistema acuático de gran valor, ha enfrentado desafíos significativos debido a la actividad humana y los cambios ambientales, que amenazan su biodiversidad y estabilidad ecológica.

Este Trabajo de Fin de Grado con título 'Estudio del estado químico y fisicoquímico actual y futuro del Mar Menor usando algoritmos de Machine Learning**',** presenta el desarrollo de un modelo predictivo basado en técnicas de Machine Learning, diseñado para estimar y predecir variables críticas que influyen en el estado del Mar Menor. Utilizando un conjunto de datos compuesto por parámetros físico-químicos, biológicos y ambientales recogidos desde 2019 hasta la actualidad.

El objetivo principal del proyecto es desarrollar un algoritmo capaz de predecir con alta precisión variables importantes que, según el Real Decreto 817/2015 [2], indican la calidad de un cuerpo de agua, como son la clorofila α y las concentraciones de nitratos y fosfatos.

Los resultados obtenidos intentan demostrar la capacidad de este tipo de modelo para ofrecer predicciones precisas que pueden facilitar la toma de decisiones proactivas en la gestión y conservación del Mar Menor.

Este estudio no solo pretende subrayar la importancia de las tecnologías avanzadas en la gestión ambiental, sino que también trata de proponer direcciones futuras para la investigación continua que podrían incluir la integración de modelos más complejos y el análisis de nuevas variables emergentes.

Palabras clave

Mar Menor, Machine Learning, Modelo Predictivo, Conservación Ambiental, Gestión de Ecosistemas.

Abstract

The Mar Menor, an aquatic ecosystem of great value, has faced significant challenges due to human activity and environmental changes, threatening its biodiversity and ecological stability.

This final degree project, titled ‘Study of the current and future chemical and physicochemical state of the Mar Menor lagoon using Machine Learning algorithms, presents the development of a predictive model based on Machine Learning techniques, designed to estimate and predict critical variables that influence the state of the Mar Menor. It utilizes a dataset comprised of physicochemical, biological, and environmental parameters collected from 2019 to the present.

The main objective of the project is to develop an algorithm capable of predicting with high accuracy important variables that, according to Real Decreto 817/2015 [2], indicate the quality of a body of water, such as chlorophyll α and the concentrations of nitrates and phosphates.

The results obtained attempt to demonstrate the capability of this model type to provide accurate predictions that can facilitate proactive decision-making in the management and conservation of the Mar Menor or water bodies alike.

This study not only aims to underscore the importance of advanced technologies in environmental management but also seeks to propose future directions for ongoing research that could include the integration of more complex models and the analysis of new emerging variables.

Keywords

Mar Menor, Machine Learning, Predictive Modeling, Environmental Conservation, Ecosystem Management

Índice

Índice de ecuaciones

Índice de figuras

Índice de tablas

1

INTRODUCCIÓN

En la era actual, donde el impacto ambiental de las actividades humanas se hace cada vez más palpable, la preservación de nuestros ecosistemas acuáticos emerge como una prioridad crítica. Los cuerpos de agua, especialmente aquellos de pequeña escala como el Mar Menor, son hábitats cruciales para una diversidad biológica significativa. No obstante, estos sistemas están frecuentemente expuestos a presiones antropogénicas que amenazan su equilibrio y la supervivencia de las especies que albergan.

El Mar Menor, situado en el Mediterráneo español, provincia de Murcia y el mayor lago de agua salada de Europa, ha sido testigo de dramáticos cambios ecológicos que comprometen su salud y biodiversidad. La eutrofización, el cambio en la salinidad, y la contaminación son solo algunos de los desafíos que enfrenta este valioso ecosistema. En este contexto, la necesidad de herramientas eficaces para la gestión y conservación del Mar Menor es más urgente que nunca.

Este Trabajo de Fin de Grado se centra en la aplicación de técnicas de Machine Learning para desarrollar un modelo predictivo capaz de estimar y predecir variables clave que reflejan el estado actual del Mar Menor. A través de este enfoque, buscamos ofrecer una solución innovadora que permita a los gestores ambientales tomar decisiones informadas y oportunas para la protección y restauración de este cuerpo de agua.

Adoptar un enfoque basado en datos para la gestión ambiental no solo mejora la precisión de las intervenciones, sino que también facilita un monitoreo continuo y adaptativo del estado del ecosistema. Mediante la integración de la ciencia de datos en la ecología, este proyecto aspira a contribuir a la conservación del Mar Menor, dando una visión más analítica a los problemas que atacan a su sostenibilidad y biodiversidad.

1.1. Motivación

Habiendo nacido en Cartagena, mi conexión con el Mar Menor es algo personal. Con el tiempo, experimenté de primera mano su deterioro, viendo cómo el lugar que una vez disfruté se convertía en un ambiente cada vez menos agradable para estar. Esta transformación de mi apreciación del Mar Menor ha sido un poderoso catalizador en mi decisión de enfocar mi Trabajo de Fin de Grado en este ecosistema.

Este proyecto busca explorar cómo las técnicas de Machine Learning pueden aplicarse para entender mejor el estado de este cuerpo de agua y quizás mitigar algunos de los problemas que enfrenta el Mar Menor. La elección de este tema se inspira en la idea de que, aunque la acción de un individuo no puede revertir los daños, a través de mi formación en ingeniería y ciencias de la computación, puedo contribuir a un mejor entendimiento de su estado actual y al posible estado al que podría llegar si se le continúan sometiendo a determinadas presiones.

Con este Trabajo de Final de Grado mi intención es aportar modestamente un estudio al cuerpo de conocimiento que los gestores ambientales pueden utilizar para tomar decisiones informadas. Aunque soy consciente de que los desafíos son grandes y las soluciones complejas, espero que mi estudio pueda servir como un pequeño paso hacia la comprensión y eventual mejora del estado ambiental Mar Menor.

Y se debe mencionar, que este trabajo no se hubiese ni siquiera comenzado, si no fuera gracias a la inspiración que supuso el trabajo previo sobre el rio Tajo de Gonzalo Martínez-Muñoz, tutor de este trabajo [1] .

1.2. Objetivos

El objetivo principal de este proyecto es conseguir un modelo basado en algoritmos de Machine Learning, capaz de asimilar datos pasados y actuales, para obtener un modelo certero de valores futuros de las concentraciones de elementos químicos y fisicoquímicos que componen las aguas del Mar Menor. Como objetivos derivados encontramos:

O-1.- Plasmar la capacidad de este tipo de modelo de medir correctamente valores actuales de características importantes en la evaluación de cuerpos de agua.

O-2.- Evaluar el estado actual de Mar Menor, en correspondencia con los umbrales de calidad de los cuerpos de agua, asignados en el BOE [2].

O-3.- Entender la optimización de parámetros de algoritmos de regresión como el Random Forest.

O-4.-Obtener un conocimiento cimentado de los procesos bioquímicos importantes encontrados en el Mar Menor, además de sus consecuencias.

1.3. Organización de la memoria

En este apartado se dará una breve explicación de cada sección de la memoria:

INTRODUCCIÓN. En este apartado se presenta brevemente el proyecto, junto con la motivación del mismo. También se incorpora una lista de los objetivos planteados que se intentan alcanzar durante su realización.

ESTADO DEL ARTE. En esta sección se plantean los conceptos clave para entender el proyecto en su totalidad. Se explican tanto los procesos bioquímicos, la naturaleza del cuerpo de agua a estudiar como los conceptos técnicos relacionados con la utilización de un modelo de Machine Learning.

DISEÑO. Aquí, se desarrollan los pasos realizados para la instalación y uso del modelo utilizado, como son: la creación del entorno, la forma de la base de datos utilizada, Además, se da la descripción del entrenamiento del mismo y una evaluación final.

EXPERIMENTOS Y RESULTADOS. En este apartado se presentan y visualizan los resultados obtenidos a lo largo de la instalación del modelo. Se presentan en forma de datos, figuras y gráficas, con el fin de hacer más fácilmente entendibles los resultados.

CONCLUSIONES Y TRABAJO FUTURO. En esta última sección se intenta resumir y dar las conclusiones finales del estudio, ofreciendo, al mismo tiempo, posibles mejoras futuras al modelo empleado.

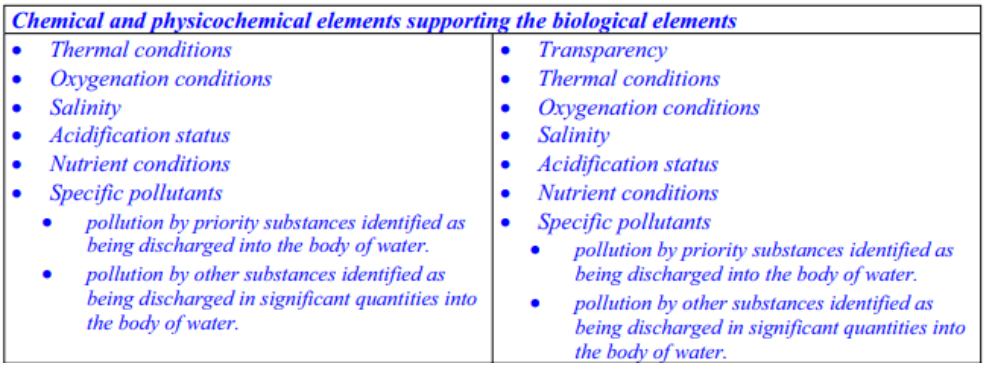
2

ESTADO DEL ARTE

En este capítulo, primero se definirán los conceptos biológicos y bioquímicos importantes para evaluar el estado del ecosistema del Mar Menor. A continuación, se desarrollaron los términos, conceptos y técnicas de Machine Learning empleadas en este proyecto.

2.1. Estudio químico y fisicoquímico

Este trabajo se basa en la modelización y predicción de la presencia de los elementos químicos y fisicoquímicos presentes en las aguas del Mar Menor. Esto lo distingue de un estudio de los elementos biológicos o hidro morfológicos que apoyarían a los elementos biológicos (Figura 2.1).



Chemical and physicochemical elements supporting the biological elements	

• *Thermal conditions*
• *Oxygenation conditions*
• *Salinity*
• *Acidification status*
• *Nutrient conditions*
• *Specific pollutants*
  ◦ *pollution by priority substances identified as being discharged into the body of water.*
  ◦ *pollution by other substances identified as being discharged in significant quantities into the body of water.*
	• *Transparency*
	• *Thermal conditions*
	• *Oxygenation conditions*
	• *Salinity*
	• *Acidification status*
	• *Nutrient conditions*
	• *Specific pollutants*
	  ◦ *pollution by priority substances identified as being discharged into the body of water.*
	  ◦ *pollution by other substances identified as being discharged in significant quantities into the body of water.*

Figura 2.1: Elementos que componen el tipo de estudio realizado en este proyecto (estudio químico y fisicoquímico). Extraído de: [3].

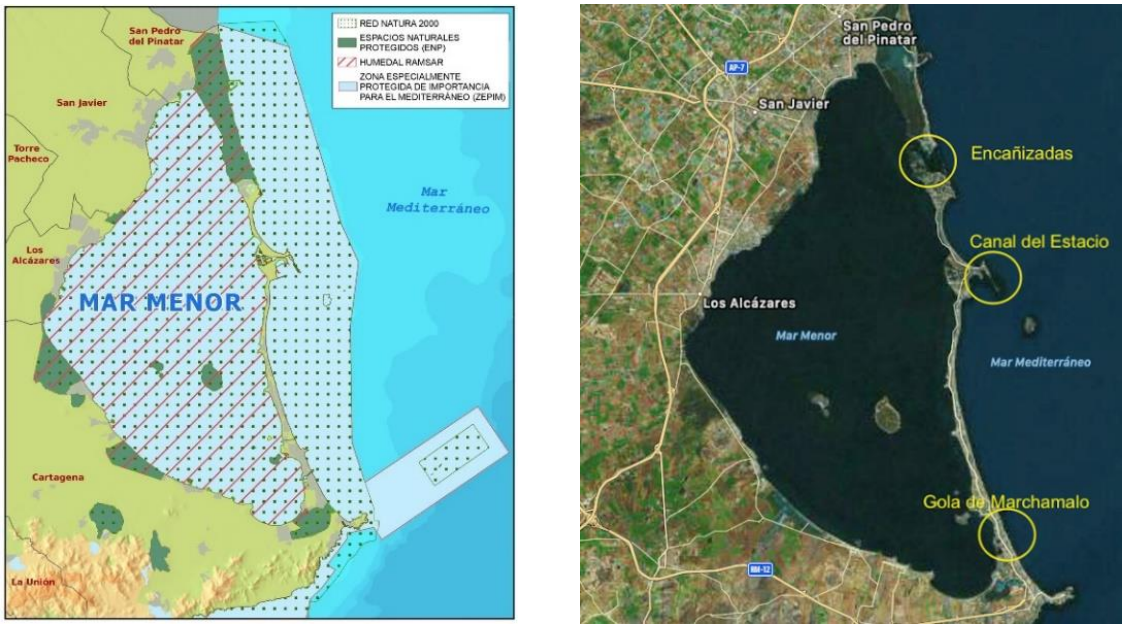
2.2. Mar Menor

El Mar Menor es la laguna salada más grande de Europa, con una superficie aproximada de 135 km^2. situada en el litoral de la Región de Murcia.

El Mar Menor es un ejemplo destacado de una laguna costera que alberga una gran diversidad de especies y hábitats. Estos ecosistemas son conocidos por su alto nivel de

productividad biológica, donde se encuentran una variedad de especies marinas adaptadas a sus características particulares, como la gran concentración salina y las aguas poco profundas.

Una parte importante de su estado actual ha dependido y sigue dependiendo de factores naturales como sus dimensiones, características como su alta salinidad y posición geológica, como la falta de lluvia y temperatura ambiente de la Región de Murcia [4].



(a) Litoral del Mar Menor y las diversas áreas protegidas. Extraído de: [4]

(b) Los tres canales de entrada y salida de agua del mediterráneo. Extraído de: [7]

Figura 2.2: En estas imágenes vemos las zonas protegidas del Mar Menor 2.2(a) y las localizaciones de los canales de salida y entrada de agua al mediterráneo 2.2(b)

Sin embargo, en las últimas décadas, el Mar Menor ha enfrentado unos crecientes factores antropogénicos que han afectado a su calidad ambiental y su estabilidad ecológica.

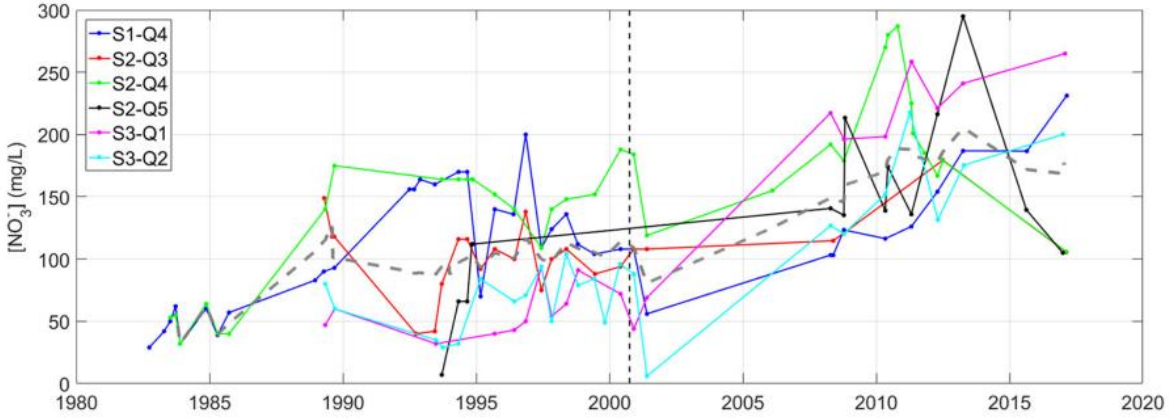


Figura 2.3: Evolución de la concentración de NO_3 (referido como nitratos en este proyecto) del Mar Menor en el periodo de 1980-2017, para 6 puntos de medida situados en boyas marinas. Extraído de: [5]

Entre otros factores como las precipitaciones, el control de los canales de entrada/salida al Mediterráneo (ver

Figura 2.2:) o la cercanía de excavaciones mineras, el factor más responsabilizado por esta rápida pérdida de calidad de las aguas es por el creciente vertido de nutrientes externos. Este vertido se suele asociar con el sector agrícola muy cercano al litoral [5] [6]. Aunque este aumento de la concentración de nutrientes externos no se puede asignar únicamente al sector agrícola, al mismo tiempo, no se puede negar este aumento, como refleja la figura 2.3.

Aunque el debate sobre la gestión de daños sigue activo [7], y se proponen entre otras medidas de descarga, es decir, de retirada de estos excesos nutrientes de las aguas y, por tanto, de la materia orgánica muerta derivada [8]. El problema de raíz, que sería el del control de vertidos y su consecuencia en la eutrofización, no parece atacarse con certeza y transparencia.

2.3. Eutrofización

Uno de los principales problemas que enfrenta el Mar Menor es la eutrofización, un proceso causado por el exceso de nutrientes, principalmente nitratos y fosfatos [9]. Estos nutrientes promueven el crecimiento excesivo de fitoplancton y algas, lo que conduce a varios efectos adversos en el ecosistema acuático.

El crecimiento extraordinario de fitoplancton y algas provoca un enturbiamiento de las aguas, impidiendo la llegada de la luz a la vegetación del fondo marino. Como consecuencia, esta vegetación muere al no poder realizar la fotosíntesis, lo que a su vez alimenta a bacterias y microorganismos con la materia orgánica muerta [10]. Este proceso incrementa la demanda biológica de oxígeno, reduciendo significativamente los niveles de oxígeno disuelto en el agua, lo que priva de oxígeno a los peces y moluscos autóctonos del Mar Menor. Se puede visualizar este proceso en la figura 2.4.

Las causas de la eutrofización pueden ser naturales; sin embargo, en el caso del Mar Menor, existe un consenso sobre la naturaleza antropogénica de los motivos. Las principales fuentes de estos nutrientes son de origen humano, provenientes tanto del sector agrícola como del entorno urbanizado.

A continuación, se enumerarán y explicaran los principales indicadores de eutrofización en las aguas, por lo que, a su vez y entre otras, serán las variables que se

intentará que el modelo entienda y prediga.

2.3.1. Clorofila-a (Clorofila α)

La clorofila α es un pigmento fotosintético fundamental que se encuentra en todo organismo capaz de hacer la fotosíntesis ya que posibilita la transformación de energía lumínica en energía química [11]. En el contexto de la eutrofización, la clorofila α es un indicador crucial del crecimiento de fitoplancton, lo cual puede desencadenar una serie de eventos negativos, como la eutrofización en los ecosistemas acuáticos [12].

La cantidad de clorofila α en un cuerpo de agua está directamente relacionada con la biomasa del fitoplancton, ya que cada célula de fitoplancton contiene clorofila α. Medir la concentración de este pigmento en el agua proporciona una estimación de la biomasa del fitoplancton presente.

Durante la eutrofización, los niveles elevados de nutrientes llevan a un crecimiento acelerado del fitoplancton, lo que resulta en un aumento de la concentración de clorofila α. En grandes cantidades, puede tener otras consecuencias ya que reduce la transparencia al incrementar la turbidez, variable que también tendremos en cuenta en la base de datos. Esta disminución de la claridad del agua afecta a las plantas de los fondos al reducirse la cantidad de luz penetrante, requerida para hacer la fotosíntesis.

2.3.2. Nitratos y fosfatos

Los nitratos, junto con los fosfatos, son responsables del aumento de nutrientes en el agua, lo que favorece el crecimiento descontrolado de algas. Este fenómeno reduce la transparencia del agua y afecta la fotosíntesis de las plantas acuáticas.

Aunque los nitratos no interfieren directamente en la adsorción de fosfatos en materiales como dolomita e hidroxiapatita, que suelen ser los adsorbentes elegidos en las mediciones de la eutrofización en cuerpos de aguas, su presencia es indicativa de una alta carga de nutrientes [13]. Esta alta concentración de nitratos facilita la creación de condiciones de hipoxia, debido al consumo de oxígeno disuelto durante la descomposición de las algas muertas, afectando negativamente a la biodiversidad acuática.

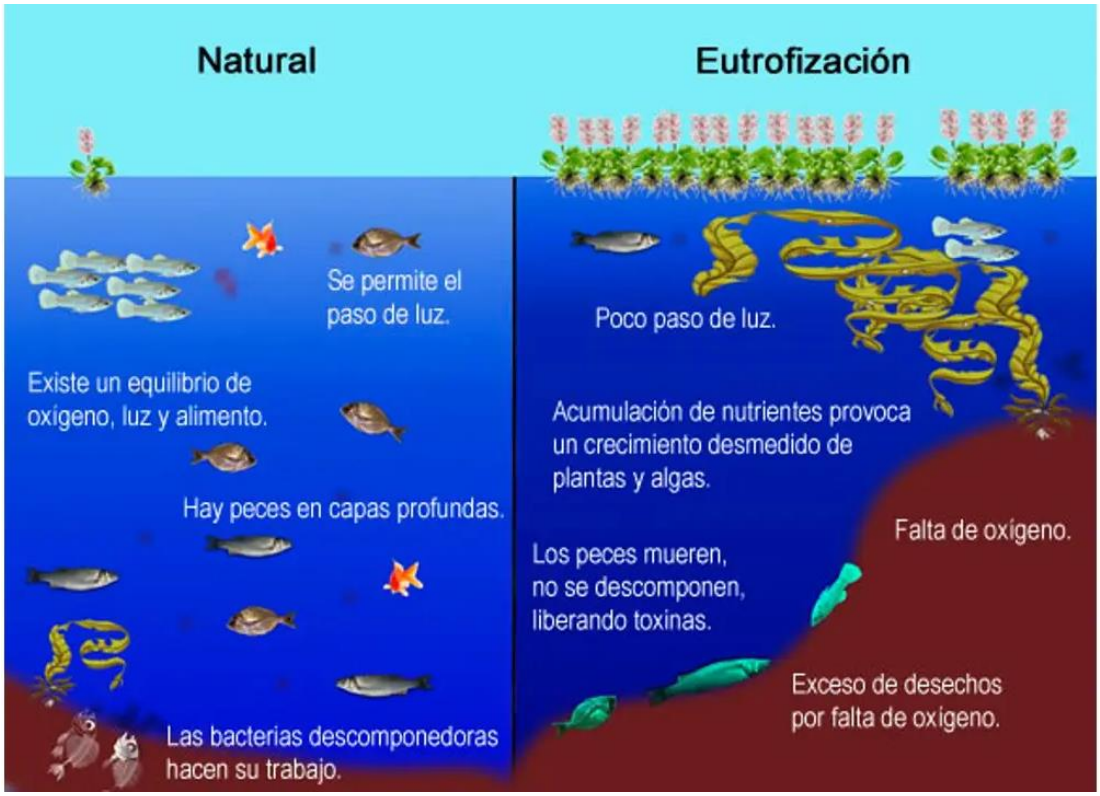


Figura 2.4: Ilustración descriptiva del proceso de eutrofización. Extraído de: [14]

2.4. Legislación sobre los cuerpos de agua

Para estimar la calidad de las aguas del Mar Menor en este trabajo se tendrá en cuenta los criterios de seguimiento y evaluación del estado de las aguas superficiales de España [2]. Este marco normativo establece valores umbral para diversas variables que indican la calidad del agua.

En el modelo desarrollado en este trabajo, intentará predecir las variables de Clorofila-a y las concentraciones de Nitratos y Fosfatos (Chl-a, Nitratos y Fosfatos en la figura 2.5 respectivamente), ya que son los datos a los que se ha tenido acceso a nivel de base de datos.

TIPOS AGUAS COSTERAS	INDICADOR	UNIDADES	VALOR ABSOLUTO	Indicadores biológicos e hidromorfológicos: RCE Indicadores químicos y biológicos (ChlA): CONCENTRACIÓN			
			Condición de referencia/ Condición específica del tipo	Límite muy bueno/ bueno	Límite bueno/ moderado	Límite moderado/ deficiente	Límite deficiente/ malo
AC-T11	Chl-a	µg/L	0,9	1,1	1,8		
AC-T11	CARLIT		Valor establecido para cada una de las situaciones ambientales definidas según tipo de costa y sustrato	0,75	0,60	0,40	0,25
AC-T11	BOPA		Fauna únicamente compuesta por especies sensibles (anfípodos excepto género Jassa) y ausencia de poliquetos oportunistas. BOPA: 0	0,95	0,54		
AC-T11	Amonio	µmol NH_4/L			4,60 (CP) 2,30 (CM)		
AC-T11	Nitritos	µmol NO_2/L			0,92 (CP) 0,46 (CM)		
AC-T11	Nitratos	µmol NO_3/L			12,90 (CP) 6,45 (CM)		
AC-T11	Fosfatos	µmol PO_4/L			0,76 (CP) 0,38 (CM)		

Figura 2.5: Tabla de los valores umbral de calidad del Mar Menor (AC-T11). Extraído de: [2]

En la tabla de la figura 2.5 se muestran los umbrales de calidad de cada una de las variables a predecir. Por columnas se describe de que cuerpo de agua se trata (AC-T11 para el Mar Menor), el tipo de indicador o variable de la que se indica el umbral, las unidades del indicador y un nivel de referencia de la variable, respectivamente.

Por último, las últimas cuatro columnas describen los umbrales de calidad para cada variable, antes mencionadas, que se usaran como mediciones de calidad. Dentro de los valores, “CP” se refiere a campo próximo, esto es, mediciones de 0 a 200 metros de la costa y “CM” a campo medio, a más de 200 metros de la costa. Se tendrán en cuenta los valores “CP”, debido a la posición de las boyas de medición de la base de datos.

Se tuvieron en cuenta también los objetivos y fines de la ley asignados por la propia Dirección General del Mar Menor de la Consejería de Medio Ambiente, Universidades, Investigación y Mar Menor de la Región de Murcia. Además de otros estudios sobre diferentes criterios de calidad de las aguas en la evaluación final [15][16].

2.5. Aprendizaje automático

El Aprendizaje Automático, también conocido como Machine Learning (ML), es una subdisciplina de la Inteligencia Artificial que se enfoca en el desarrollo de algoritmos y técnicas que permiten a las computadoras aprender y mejorar automáticamente a partir de la experiencia plasmada en forma de datos pasados.

Aunque dependiendo de la aplicación, usualmente, en lugar de ser explícitamente programadas para realizar una tarea, los algoritmos de aprendizaje automático utilizan datos y ejemplos para identificar patrones en los datos. Estos patrones se utilizan para la construcción de modelos matemáticos que pueden hacer predicciones o tomar decisiones basadas en datos de entrada nuevos, no presentes en las bases de datos, durante el proceso de creación de los modelos.

En este sentido, este tipo de aprendizaje tiene una dependencia mayor de las personas expertas en el campo quienes pueden indicar qué variables puede tener más relevancia. Este tipo de modelo se conoce como ML supervisado, definido por su uso de un conjunto de datos etiquetados previamente. En cada iteración el modelo, a su vez, optimiza sus pesos hasta llegar a un punto que considera correctamente ajustado.

En la realidad, los datos tabulares, es decir, ejemplos de mediciones (filas) con el mismo compuesto de características (columnas) (ver Figura 3.3**:**) es el formato de datos más usado en aplicaciones reales, donde algoritmos como Random Forest o Gradient Boosting supervisado siguen superando en capacidad a otros tipos de modelado supervisado, como el Deep Learning [17][18]. El ML basado en datos tabulares es dependiente, en aún mayor medida, de la necesidad de crear y definir el conjunto de características que se estima más relevante para entender las segregaciones de los datos.

Durante el entrenamiento de modelos de aprendizaje automático, se deben evitar los denominados problemas de regresión, como el **sobreajuste** y el **subajuste**. El **sobreajuste** ocurre cuando un modelo de aprendizaje automático se ajusta demasiado bien a los datos de entrenamiento, capturando incluso el ruido o las peculiaridades específicas de estos datos. Como resultado, aunque el modelo puede mostrar un desempeño excepcional en el conjunto de entrenamiento, su capacidad para generalizar a nuevos datos es pobre, llevando a predicciones inexactas cuando se aplica a conjuntos de validación o prueba. Este problema se puede identificar observando las gráficas de pérdida o exactitud, donde una gran disparidad entre el rendimiento en el conjunto de entrenamiento y el conjunto de validación/test indica sobreajuste.

Por otro lado, el **subajuste** ocurre cuando el modelo es demasiado simple para capturar la complejidad subyacente de los datos. En estos casos, el modelo no logra aprender los patrones relevantes, resultando en un desempeño pobre tanto en el conjunto de entrenamiento como en el de validación/test. El subajuste puede deberse a diversas razones, como la falta de suficientes características, un modelo demasiado básico o insuficientes datos de entrenamiento.

2.5.1. Validación cruzada

El objetivo de esta técnica es evaluar la capacidad de generalización de un modelo, es decir, su rendimiento en datos no vistos durante el entrenamiento. Esta técnica es fundamental para prevenir el sobreajuste donde, en esencia, el modelo funciona bien en los datos de entrenamiento; pero falla en datos nuevos.

La validación cruzada divide el conjunto de datos en múltiples subconjuntos (*folds*) y utiliza algunos de ellos para entrenar el modelo y otros para probarlo, repitiendo este proceso varias veces para obtener una evaluación robusta del rendimiento del modelo. Esto permite una estimación más precisa y confiable del desempeño del modelo, comparado con la simple partición de datos en un conjunto de entrenamiento y otro de prueba.

Se pueden encontrar muchos métodos de aplicación de esta técnica. Entre las más utilizadas encontramos la validación cruzada “k-fold” en el cual, el conjunto de datos se divide aleatoriamente en k grupos de mismo tamaño aproximado.

El modelo se entrena “k” veces, utilizando cada vez “k-1” *folds* como conjunto de entrenamiento y el *fold* restante como conjunto de prueba. Este proceso se repite “k” veces, asegurando que cada *fold* se utilice exactamente una vez como conjunto de prueba. Los resultados de cada iteración se promedian para obtener una estimación final del rendimiento del modelo.

Sin embargo, para el caso del Mar Menor, las mediciones son dependientes del tiempo, al ser sensibles a variables meteorológicas, estacionales y eventos específicos, por lo que, el método elegido, es uno cuya separación en subconjuntos no es aleatoria, llamado validación cruzada con series temporales.

2.5.2. Validación cruzada con series temporales

La validación cruzada con series temporales, a diferencia de la validación cruzada tradicional, enfrenta retos únicos debido a la naturaleza secuencial y dependiente del tiempo de los datos. En un entorno de series temporales, es crucial preservar el orden temporal de las observaciones para evitar la introducción de sesgos y garantizar que el modelo pueda generalizar correctamente a datos futuros no vistos [19]. Este proceso implica dividir el conjunto de datos en varios subconjuntos de entrenamiento y prueba.

Una de las técnicas más comunes para realizar la validación cruzada en series temporales, y la utilizada en este proyecto, es la validación cruzada expansiva o "*expanding window cross-validation*" [20]. En este método, el tamaño del conjunto de entrenamiento crece con cada iteración mientras que el conjunto de prueba se desplaza hacia adelante. Aquí, en lugar de una ventana deslizante fija, se empieza con un pequeño conjunto de entrenamiento y se va añadiendo más datos de entrenamiento en cada iteración, manteniendo siempre el orden temporal. [21]

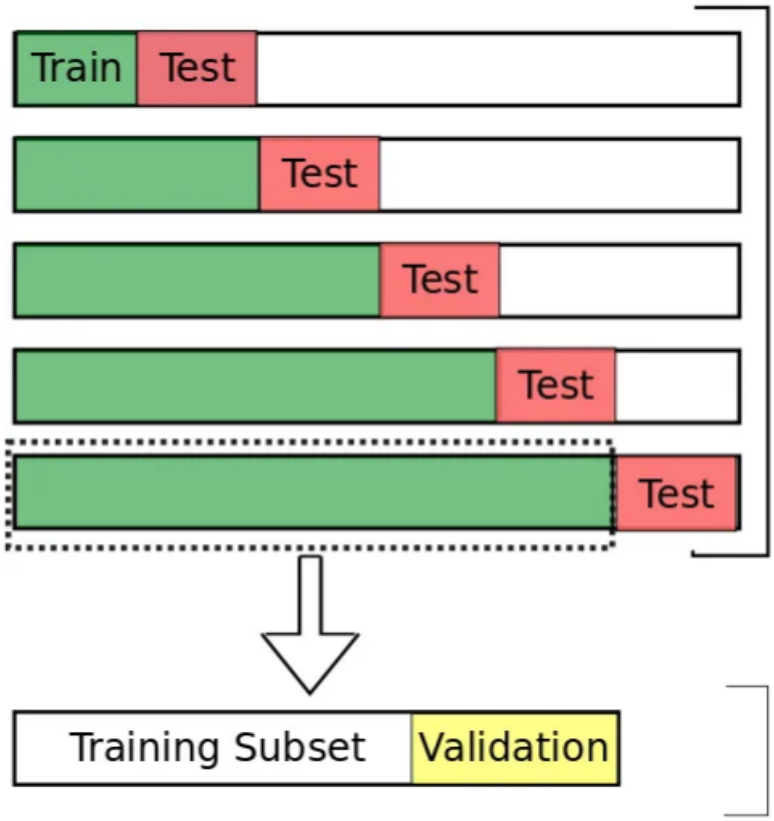


Figura 2.6: En esta imagen se visualiza el método de correlación cruzada con series temporales “expanding window cross-validation”. Extraído de: [21]

2.6. Random Forest

Al ser la referencia en capacidad de entrenamiento [18] y por su facilidad de instalación se escogió el algoritmo Random Forest para el entrenamiento de este modelo. Este algoritmo utiliza una técnica de aprendizaje automático que usa un enfoque de conjunto (*ensemble learning*) para mejorar la precisión de las predicciones. Este método implica la creación de múltiples árboles de decisión, cada uno entrenado con diferentes subconjuntos de los datos originales, generados mediante la técnica de *boostrap.*

El muestreo *bootstrap* crea subconjuntos de datos, lo que introduce variabilidad y reduce la correlación entre los árboles, mejorando así la robustez y precisión del modelo. Los nodos en un árbol de decisión representan puntos donde se evalúa una característica específica, y las ramas son las rutas que derivan de esos puntos de decisión, llevando a otros nodos o a hojas finales con predicciones.

En contraste, el *boosting* es otra técnica de *ensemble learning* que crea modelos secuenciales, cada uno corrigiendo los errores de los anteriores [22]. Mientras que el *bagging* en Random Forest entrena árboles en paralelo e independientes entre sí, el *boosting* ajusta cada modelo de manera secuencial, enfocándose en los errores cometidos por los modelos previos. Esta técnica acumulativa permite que el modelo final sea más preciso al adaptarse iterativamente y mejorar continuamente las predicciones. Así, ambos métodos utilizan conjuntos de modelos para mejorar la precisión, pero difieren en cómo manejan la interdependencia y el enfoque en los errores [23].

La naturaleza del *boostrap sampling* y su elección aleatoria de subconjuntos de los datos (con reemplazo) para la creación de cada árbol de decisión, implica una rotura parcial del orden temporal de los datos. Para solucionar este problema, cuando se le pidió al modelo final la predicción de valores futuros, se estructuraron los datos como un problema supervisado, obligando a que las características provengan únicamente de valores pasados [33].

A pesar de implementarse esta solución, se comparó la capacidad del modelo basado en Random Forest con otros modelos construidos a partir de otros algoritmos de regresión y con modelos especializados en series temporales, independientes al Machine Learning, con el objetivo de asegurarse de haber creado el modelo más capaz en la predicción de las variables.

2.7. Regresión lineal

En comparación con otros modelos de regresión, la literatura parece clara: Random Forest presenta mejores resultados en todas las métricas con las que se compara con algoritmos de regresión lineal [18]. Sin embargo, con la intención de comprobar esta conclusión y conseguir el mejor modelo posible, se construyó un modelo con un algoritmo de regresión lineal.

La regresión lineal es un algoritmo de aprendizaje supervisado utilizado para modelar la relación entre una variable dependiente y y una o más variables independientes x. La forma más básica de regresión lineal es la regresión lineal simple, que utiliza una sola variable independiente [24]. La ecuación principal de la regresión lineal simple es:

$$y = \beta_0 + \beta_1 x + \epsilon y \tag{2.1}$$

donde β_0 es la intersección con el eje y (el valor de y cuando x es cero), β_1 es la pendiente de la línea de regresión (el cambio en y por cada unidad de cambio en x), y ϵ es el término de error que captura las desviaciones no explicadas por el modelo.

2.8. Modelo ARIMA y SARIMA

Siguiendo con el objetivo de tener la seguridad de usar el mejor modelo posible para hacer predicciones, se comparó la capacidad del algoritmo Random Forest, con la de modelos que se definen como especializados en datos con dependencia temporal, es decir, cuando durante el año se midieron. Más concretamente, se denominan modelos de pronóstico de series temporales. Y más en especifico se usó el modelo SARIMA.
A continuación, se explican los conceptos que integra dicho modelo:

Los **modelos autorregresivos (AR)** asumen que el valor actual de una serie temporal, "x_t", se puede explicar mediante sus valores pasados x_{t-1}, x_{t-2}, …, x_{t-p}, acercamiento similar al de validación cruzada con series temporales. Un modelo AR de orden "p", expresado como AR(p), representa la relación entre el valor actual y "p" valores pasados, donde "p" es el número de observaciones previas consideradas en el modelo.

Los **modelos de media móvil (MA),** en cambio, representan el valor actual de una serie como una combinación lineal de los errores previos, conocidos como ruido blanco. Un modelo MA de orden "q", denotado MA(q), emplea los errores de los "q" períodos previos para estimar el valor presente.

La combinación de los modelos AR y MA da lugar a los **modelos ARMA (p, q)**. Los

modelos ARIMA son una extensión de los ARMA, y se caracterizan por integrar un parámetro adicional, “d”, que permite convertir una serie no estacionaria en estacionaria mediante la diferenciación, esto es: calcular las diferencias entre valores consecutivos en la serie temporal, eliminando tendencias lineales (d = 1) o cuadráticas (d = 2) de los datos de entrenamiento. En consecuencia, incorpora:

“p”: el número de términos autorregresivos.
“d”: el orden de diferenciación para la estacionariedad.
“q”: el número de términos de media móvil.

Para series temporales con patrones estacionales, se utilizan modelos **SARIMA (Seasonal-ARIMA)**, que son una versión expandida de ARIMA que permite manejar componentes estacionales:

“P”: el orden de autoregresión estacional.
“D”: el orden de diferenciación estacional.
“Q”: el orden de media móvil estacional.
“s”: la longitud del período estacional, por ejemplo, s = 12 para datos mensuales.

Como modelo final se eligió uno basado en SARIMA, pues ARIMA no tiene en cuenta los componentes estacionales de los datos, los cuales añaden otra capa de utilidad y complejidad al modelo final. Se debe recalcar que la explicación anterior de SARIMA no es completa. Se han subrayado los parámetros más importantes para entender dicho modelo [25].

Además, cabe explicar dónde se diferencian la validación cruzada de series temporales y el modelo SARIMA. SARIMA es un modelo completo y la validación cruzada de series temporales es una técnica de validación. Además, SARIMA, entrena con toda la base de datos para modelar patrones generales, esto es: no fragmenta los datos en bloques para evaluar la capacidad del modelo en diferentes segmentos. Al mismo tiempo que si admite la importancia de la autoregresión de los datos [26] [27].

2.9. Métricas

Para evaluar el rendimiento de los modelos predictivos en este Trabajo de Fin de Grado, se han utilizado varias métricas, cada una con sus propias características y aplicaciones [28].

Mean Squared Error (MSE). Es una métrica que cuantifica la diferencia promedio entre los valores predichos por el modelo y los valores reales observados. Se calcula como el promedio de los cuadrados de los errores, donde los errores son las diferencias entre los valores predichos ($\hat{y}$) y los valores observados (y_i), siendo "n" el número de observaciones:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.2}$$

Mean Absolute Error (MAE). Se calcula como el promedio de los valores absolutos de los errores:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.3}$$

Coeficiente de Determinación (R^2). Es una métrica que indica la proporción de la variación en la variable dependiente que es explicada por el modelo, $\bar{y}$ es el valor medio de las observaciones:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.4}$$

Cabe decir que, al iniciar el entrenamiento, sólo se tuvo en cuenta el MSE como parámetro de rendimiento. Pero con el interés de comprender la calidad del modelo, se incorporaron los otros métodos [29].

3

Diseño

3.1. Entorno de trabajo

Este Trabajo de Fin de Grado se ha desarrollado principalmente en Python, uno de los lenguajes de programación más utilizados en el mundo en proyectos de aprendizaje automático, gracias a su gran acceso a librerías especializadas.

Como entorno se ha usado un entorno virtual ('.venv') para gestionar las dependencias y asegurar que las bibliotecas especificas utilizadas no interfieran con otros proyectos. Permite evitar conflictos entre versiones y garantizar la reproducibilidad del entorno. Además, facilita la colaboración al permitir a otros desarrolladores recrear el mismo entorno utilizando el archivo de requerimientos.

Se usaron librerías de Python muy generalizadas como *sys*, para interactuar con la máquina; *pandas* para la lectura, tratamiento y división de los datos; *numpy*, para realizar cálculos numéricos como medias y *matplotlib* para la visualización de datos. Entre las librerías más especializadas de aprendizaje automático se usó *sklearn*.

Para la ejecución del modelo, se usó únicamente mi ordenador portátil personal con especificaciones de 16 GB de RAM, con un procesador Intel Core i7-1280p de 12ª generación y 2 GHz.

3.2. Base de datos

La base de datos utilizada fue resultado de la unión de varias bases de datos disponibles. Primeramente, se usó la base proporcionada por la Universidad Politécnica de Cartagena (UPCT) con el nombre de "Servidor de datos científicos del Mar Menor".

El Servidor de Datos Científicos del Mar Menor (SDC) recopila y distribuye información obtenida de campañas oceanográficas sobre parámetros ambientales del Mar Menor. Los datos, recogidos por diferentes instituciones, se presentan en gráficos georreferenciados y están disponibles para descarga pública. El sistema utiliza

protocolos como Unidata y OGC, y emplea el servidor THREDDS para compartir datos en formatos NetCDF y ASCII, facilitando su acceso y análisis a través de diversas herramientas científicas.

Los datos están organizados en cinco niveles: L0 a L4. Los datos brutos (L0). A continuación, esos datos son procesados mediante control de calidad (L1) y, esos datos son interpolados (L2). Los datos del nivel L3 son de los mapas elaborados por el servidor. El nivel L4 incluye productos derivados del nivel L2. Las variables registradas abarcan temperatura (Cº), salinidad (PSU), transparencia, clorofila (µg/L), oxígeno (mg/L), turbidez (NTU) e irradiancia (micro Einstein), cada una medida con instrumentos específicos en estaciones de muestreo distribuidas espacialmente (ver **Figura 3.1:**).

Cada variable tiene importancia específica en el análisis ambiental. Por ejemplo, la temperatura afecta a las propiedades físicas y químicas del agua, mientras que la salinidad mide la cantidad total de sales disueltas, factor crítico para la supervivencia de ciertas especies. La transparencia se mide usando el disco de Secchi, y la clorofila indica la biomasa del fitoplancton, cuyo aumento sería resultado de la eutrofización de las aguas. La turbidez refleja la cantidad de materiales suspendidos en el agua, lo cual, en altas cantidades, bloquearía la fotosíntesis de los organismos vegetales y la irradiancia se mide con radiómetros para determinar la cantidad de luz disponible para dicha fotosíntesis.

Para la base de datos final no se utilizaron las mediciones de irradiancia ya que sus mediciones comenzaron en junio de 2021.

Para este trabajo se utilizaron los datos de nivel L4, pues eran a los que se tiene acceso libre. En estos productos derivados específicos se proporcionan, para cada medición de cada variable: la fecha de la medición, la media de los valores medidos en cada punto de medición y la desviación de esa media, en el periodo de Mayo de 2017 a hasta la actualidad. En esta base de datos encontramos un total de 2203 mediciones [30].

Se puede observar que en esta base de datos no se proporciona datos etiquetados sobre concentraciones de nitratos y fosfatos. Para esto se usó una segunda base de datos libre, proporcionada por la fundación “Canal Mar Menor”, que se comienza para favorecer el cumplimiento de mediciones aprobadas por el Consejo de Gobierno de la Comunidad Autónoma de Murcia [31].

De esta base de datos se usaron sus mediciones de concentraciones de nitratos (mgNO3/L), fosfatos (mgPO4/L), mediciones del caudal (L/s) y la conductividad (µS/cm).

Para cada una de estas variables se obtuvieron mediciones desde 12 puntos diferentes del Mar Menor, que se identifican en la imagen 3.1. Se usaron las medias correspondientes de las 12 mediciones para cada variable y para cada fecha.

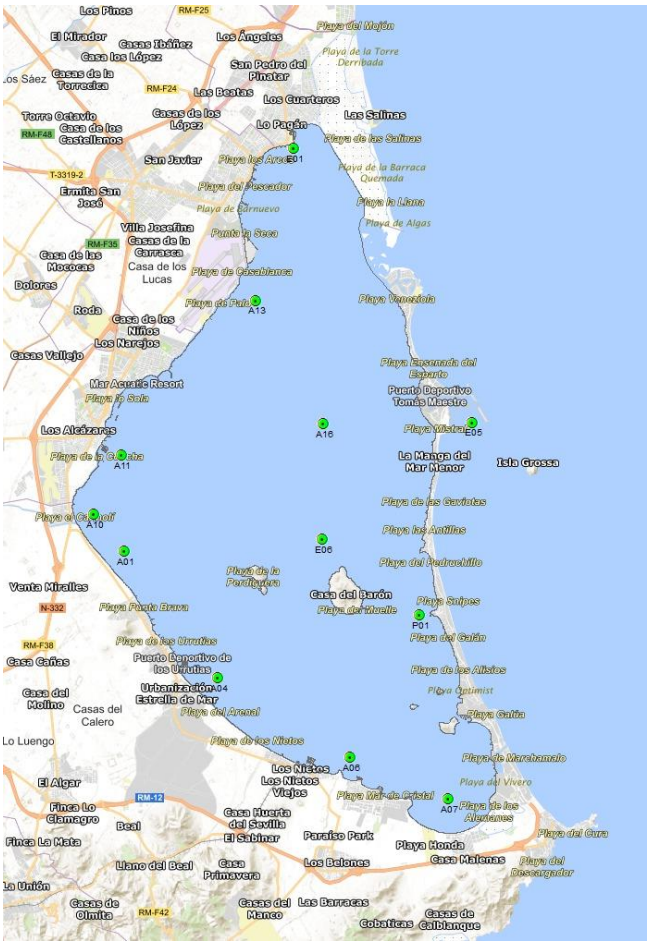


Figura 3.1: En esta imagen se ven las localizaciones de las boyas de medición de la base de datos de la fundación “Canal Mar Menor”. Extraído de: [31]

En esta base de datos, no se comenzó a medir todas las variables al mismo tiempo. Las variables más importantes que obtener de esta base eran las concentraciones de nitratos y fosfatos, por lo que serían estas quienes dictarían el comienzo de la base de datos final. Los nitratos tienen de primera fecha el “30.08.2019”, y los fosfatos “23.08.2021”.

Para solucionar este problema, se eligió la fecha de inicio de la toma de datos de los nitratos y las mediciones de fosfatos se consideraron nulas (0) hasta que llegaron a su propio inicio. Seleccionar este inicio también tuvo impacto en las mediciones de la primera base de datos pues su comienzo era del Mayo del 2017, por lo que, se acortaron esas mediciones para que comenzaran el “30.08.2019”, reduciendo el número de mediciones original de la primera base de 2203 a 1044.

Juntando las 9 variables definitivas, la base de datos final consiste de 2331 mediciones, divididas en 257 fechas.

Como se ha mencionado, para ambas bases de datos, todos los datos son etiquetados con la fecha de medición. Esto es el motivo principal de la selección de un método de validación cruzada con series temporales, y no una de separación aleatoria, esto es: que los datos son sensibles al momento en el que se tomaron, ya sea por motivos meteorológicos o estacionales.

3.3. Implementación

En este apartado se explicará en detalle la implementación del modelo que se consideró como más capaz. Antes de implementar en el modelo la capacidad de hacer predicciones futuras, se compararon los resultados con diferentes números de *splits*, se optimizaron sus parámetros y finalmente se utilizó Random Forest tras comparar su capacidad con la de otros algoritmos de regresión y con la de modelos especializados en el pronóstico con series temporales.

3.3.1. Separación de datos

Para realizar la validación cruzada en modelos de predicción de series temporales, se utiliza una técnica específica que involucra tres parámetros esenciales: el tamaño inicial de la ventana de entrenamiento “w”, el período de desplazamiento “p”, y el horizonte de predicción “h”. Este método se asegura de respetar la secuencia temporal de los datos.

En la primera iteración se define una ventana de entrenamiento de tamaño “w” que contiene los primeros “w” datos, y una ventana de validación de tamaño “h” que contiene los siguientes “h” datos. El modelo se entrena con los datos de la ventana de entrenamiento y se valida con los datos de la ventana de validación. Tras cada iteración, el umbral “w” se incrementa en “p” unidades, y el proceso se repite hasta que “w+h” excede el tamaño total de la serie temporal “N”.

En este contexto, el concepto de *splits* se refiere al número de veces que se divide la serie temporal para crear diferentes conjuntos de entrenamiento y validación. El número de splits “n_splits” determina cuántas veces se repite el proceso de entrenamiento y validación a lo largo de la serie temporal. Cada *split* implica un nuevo conjunto de entrenamiento y validación, desplazado según el período “p”.

La figura 2.6 es un ejemplo de cuando “h = p” [21]. Para entender mejor este algoritmo.

3.3.2. Preprocesado

Para comparar los datos de las bases de datos con los umbrales de calidad de los cuerpos de agua (figura 2.5) se cambiaron las unidades de los umbrales de la concentración de los nitratos y fosfatos.

Nitratos. El valor del umbral es de 12.9 µmol/L y se cambió a mg/L.
Se pasa de µmol a mol:

$$12{,}9\,\frac{mol}{L} = 12.9 * 10^{-6}\,\frac{mol}{L} \tag{3.1}$$

El peso molecular del nitrato es 62 g/mol. Convertimos mol a gramos:

$$12.9 * 10^{-6}\,\frac{mol}{L} * 62\,\frac{g}{mol} = 8 * 10^{-5}\,\frac{g}{L} \tag{3.2}$$

Por lo que, el umbral para los nitratos es de **0,080 mg/L**.

Fosfatos. El valor del umbral es de 0,76 µmol/L y se cambió a mg/L.
De la misma manera con un peso molecular de 95 g/mol, el umbral es de **0,072 mg/L.**

El umbral de clorofila se mantiene en **1,8 µg/L** como indica la figura 2.5.

3.3.3. Entrenamiento

El entrenamiento del modelo predictivo de Random Forest fue realizado con el objetivo de estimar con precisión las concentraciones de clorofila α, nitratos y fosfatos en el Mar Menor. Este modelo fue seleccionado debido a su capacidad para manejar datos complejos y de gran volumen, al mismo tiempo que minimiza el riesgo de sobreajuste al generar múltiples árboles de decisión basados en diferentes subconjuntos de los datos.

Dado que los datos recogidos presentan una estructura temporal, se implementó una estrategia de validación cruzada con series temporales, utilizando la técnica de expanding window, explicada en el estado del arte, apartado 2.5.1 y 2.5.2. Esta metodología asegura que el orden cronológico de los datos se mantenga, lo que es fundamental en la predicción de variables ambientales que varían con el tiempo.

Una vez construida la base de datos final, se comenzó iterando para encontrar el valor óptimo de la variable “n_splits”, responsable de indicar el número de subconjuntos de datos en los que se dividirían la base de datos, con los que se alimentaría al modelo.

Para continuar con el entrenamiento se ajustaron varios parámetros clave del modelo

de Random Forest para mejorar su rendimiento:

Número de estimadores (n_estimators): Este parámetro determina el número de árboles de decisión generados por el modelo. Aumentar el número de árboles tiende a mejorar la precisión, ya que el modelo se beneficia de una mayor diversidad de predicciones; aunque también incrementa el tiempo de procesamiento.

Profundidad máxima de los árboles (max_depth): Este parámetro limita la profundidad de cada árbol, ayudando a controlar el riesgo de sobreajuste. Los árboles con mayor profundidad tienden a captar más detalles del conjunto de datos; pero a profundidades excesivas pueden sobrentrenarse, capturando ruido en lugar de patrones relevantes.

Muestras mínimas para dividir un nodo (min_samples_split): Establece el número mínimo de observaciones necesarias para dividir un nodo. Un valor más alto para este parámetro asegura que los nodos solo se dividan cuando hayan suficientes datos, lo que ayuda a evitar divisiones poco útiles y contribuye a que el modelo sea más robusto. Con “observaciones” nos referimos a cada registro o “fila” del conjunto de datos que contiene un valor para cada variable de la base de datos.

Fecha	Turbidez	Temperatura	Salinidad	Oxigeno	Nitratos	Fosfatos	Clorofila
04/09/2019	3,334727081	28,32828921	45,58497092	5,321382563	146,00		3,240554139

Figura 3.3: Ejemplo de observación o medición. Extraído de la base de datos usada.

Muestras mínimas por hoja (min_samples_leaf): Define el número mínimo de observaciones que deben existir en un nodo hoja. Se probaron valores de 1, 2 y 4, para evitar la creación de nodos hoja con muy pocos datos, lo que podría hacer que el modelo se sobreajustara a esos casos particulares.

Características máximas para la mejor división (max_features): Controla cuántas características se consideran para dividir cada nodo. Se experimentó con las opciones ‘auto’, ‘sqrt’ y ‘log2’. Estas opciones permiten limitar el número de variables evaluadas en cada división, lo que introduce diversidad entre los árboles y mejora la capacidad del modelo para generalizar.

Muestreo con reemplazo (bootstrap): Se probó tanto con la opción True como False. El muestreo con reemplazo (bootstrap) introduce variabilidad en los árboles al entrenar cada uno con diferentes subconjuntos de los datos. Esto favorece la independencia entre los árboles y, por tanto, mejora la robustez del modelo.

Cabe explicar que no se utilizaron todos los posibles parámetros del Random Forest, ya que no eran especialmente útiles y ralentizaban gratuitamente la experimentación con el modelo [32]. Algunos ejemplos de estos parámetros serían:

class_weight: Este parámetro es útil en problemas de clasificación desbalanceada, donde ciertas clases aparecen con mucha más frecuencia que otras. Dado que el objetivo de este trabajo no involucraba un problema de clasificación multiclase desbalanceado, no fue necesario ajustar este parámetro.

max_samples: Este parámetro controla la cantidad de muestras que se extraen del conjunto de datos para construir cada árbol. En este caso, se usó todo el conjunto de datos disponible para cada iteración del modelo, por lo que no fue necesario limitar el tamaño de la muestra utilizada.

oob_score: Este parámetro habilita el uso de muestras fuera de la bolsa (*out-of-bag*) para evaluar el rendimiento del modelo. Dado que se aplicó una estrategia de validación cruzada con series temporales, no fue necesario usar este método alternativo de validación.

Una vez encontrado el modelo optimizado en su número de subconjuntos de entrenamiento y parámetros de algoritmo, se comparó su capacidad con la de otros modelos. Comenzando con un modelo basado en el algoritmo de regresión lineal. A continuación, se comparó con un modelo especializado en el pronóstico con datos con dependencias temporales llamado SARIMA.

Cuando se corroboró que la mejor capacidad la seguía teniendo el modelo basado en Random Forest optimizado, se implementó la capacidad de realizar predicciones futuras, es decir, sin referencia de base de datos. Para adaptar este modelo Random Forest a la predicción en series temporales futuras, se transformaron los datos en una estructura supervisada [33].

Las estructuras supervisadas de datos son una forma de reorganizar los datos para convertir un problema de predicción en uno supervisado, donde las entradas (características) corresponden a valores pasados (lags) y las salidas (etiquetas) son los valores futuros que se desean predecir, permitiendo que el modelo aprenda patrones temporales a partir de las observaciones previas y a su vez asegurándose de usar únicamente valores pasados en las predicciones. Solucionando en gran medida el problema planteado por la naturaleza aleatoria de los arboles de Random Forest [33].

En la predicción a futuro, el modelo genera un valor que se agrega a la ventana de entrada, desplazándola y permitiendo así predicciones sucesivas sin depender de

valores futuros reales. Esta técnica iterativa asegura que el modelo RF mantenga la secuencia temporal, facilitando predicciones basadas en patrones históricos [34] [35].

3.3.4. Evaluación

Para la evaluación de la precisión del modelo se usaron las métricas mencionadas en el estado de arte, apartado 2.8. Estas métricas permitieron evaluar el ajuste del modelo en cada iteración, y se aplicaron mecanismos de parada temprana para evitar que el modelo continuara entrenando si las mejoras en las predicciones se estabilizaban.

4

EXPERIMENTOS Y RESULTADOS

4.1. Experimentos

En este capítulo, primero vamos a describir los experimentos de las 8 iteraciones que hemos realizado y luego pasaremos a comentar los resultados

Primera iteración. Detección de la clorofila-α con un modelo inicial sencillo.

Como primer modelo se construyó una base de datos menos densa, careciente de variables como los nitratos, fosfatos, el caudal y la conductividad. El objetivo principal era testear la capacidad inicial de predicción de la **clorofila-α**, pues era de la única de la cual se tenía de datos en este momento.

Como parámetros del Random Forest, respecto a **n_estimators**, comienzo con 100 o 200 árboles para encontrar un equilibrio óptimo entre rendimiento y captura de variabilidad. Para **max_depth**, optaría por un valor entre 10 y 20 debido a las posibles complejidades temporales en mis datos. Establecería **min_samples_split** en un rango de 2 a 10 para capturar detalles sin correr el riesgo de sobreajuste.

Para **min_samples_leaf**, elegiría un valor entre 1 y 4 para evitar que el modelo sea demasiado específico. Respecto a **max_features**, consideraría sqrt o un tercio del número total de características para aumentar la diversidad entre los árboles.

Segunda iteración. Primera detección de las tres variables de interés.

El segundo experimento, fue el primer intento “completo” de predicción de las tres variables de interés, pues es el primero tras conseguir datos para todas. Se mantuvieron los parámetros del RF de la primera iteración. Sin embargo, en la predicción de cada una de las tres variables no se tuvieron en cuenta las otras dos variables a predecir, es decir, para la predicción de la clorofila-α no se utilizaron los datos de nitratos y fosfatos y viceversa. Además, se añaden barras horizontales que marcan los umbrales de calidad

establecidos en el Real decreto [2] para cada variable, mencionados en la figura 2.5.

Tercera iteración. Primera medición con la base de datos completa.

En la tercera iteración, se incorporaron nuevas variables a la base de datos, incluyendo el caudal y la conductividad, completando así la base de datos final. Esto permitió una mejora en la precisión del modelo y a la vez se pudo visualizar el impacto de la introducción de nuevas variables en la clorofila, nitratos y fosfatos. En este experimento, al igual que en la segunda iteración, se mantuvieron independientes las variables a predecir. Se mantuvieron los parámetros del RF de la primera iteración.

Cuarta iteración. Variables a predecir dependientes.

Como mejora en esta iteración se introducen como variables predictoras los valores de las otras variables a predecir. Esto implica implementar 3 modelos simultáneos. También se grafica la importancia de cada variable de la base de datos al predecir la clorofila, la concentración de nitratos y de fosfatos, para mejorar la interpretabilidad del modelo. Se mantuvieron los parámetros de RF de la primera iteración.

Quinta iteración. Optimización del número de splits.

Esta quinta iteración, al conseguir una iteración con la base de datos completa en la predicción de variables de interés, se enfocó en la búsqueda del número óptimo de splits de la base de datos, desde los 10 probados en la anterior iteración al límite de puntos temporales de la base de datos de 256. Se mantuvieron los parámetros del RF.

Sexta iteración. Optimización de parámetros del algoritmo Random Forest.

Habiendo encontrado el número de splits para cada característica, en esta iteración se estudia el impacto de los parámetros del Random Forest. Se utilizó la función GridSearchCV [37] para encontrar la mejor combinación de parámetros; la cual, en esencia, comprueba el MSE para cada una de las combinaciones de parámetros. Además, se utilizó el día, mes y año como variables de entrenamiento para mejorar con la estacionalidad de los datos.

Con la sospecha de un posible sobreajuste, se obtuvieron las gráficas de pérdidas de entrenamiento y validación. La idea era observar la posibilidad del sobreajuste, ya que continuar y construir un modelo predictivo a valores futuros acarreando sobreajuste no sería conveniente.

Séptima iteración. Comparativa de los modelos implementados.

En esta séptima iteración, una vez ajustados los hiperparámetros del modelo predictivo con RF y con intención de contrastar el modelo e intentar llegar a un mejor resultado, se comparó la capacidad predictiva del Random Forest con la de otros modelos.

Primero, se compara con un modelo basado en un algoritmo de regresión lineal. Al tratarse de un algoritmo más simple, la implementación fue de este modelo fue sencilla.

Segundo, se compara con un modelo especializado en pronóstico con series de datos temporales llamado SARIMA. Se establecieron las variables de entrenamiento. A continuación, se extrajeron de la base de datos, los datos o patrones estacionales que también se usaran para entrenar el modelo, como se ve en la figura 4.1.

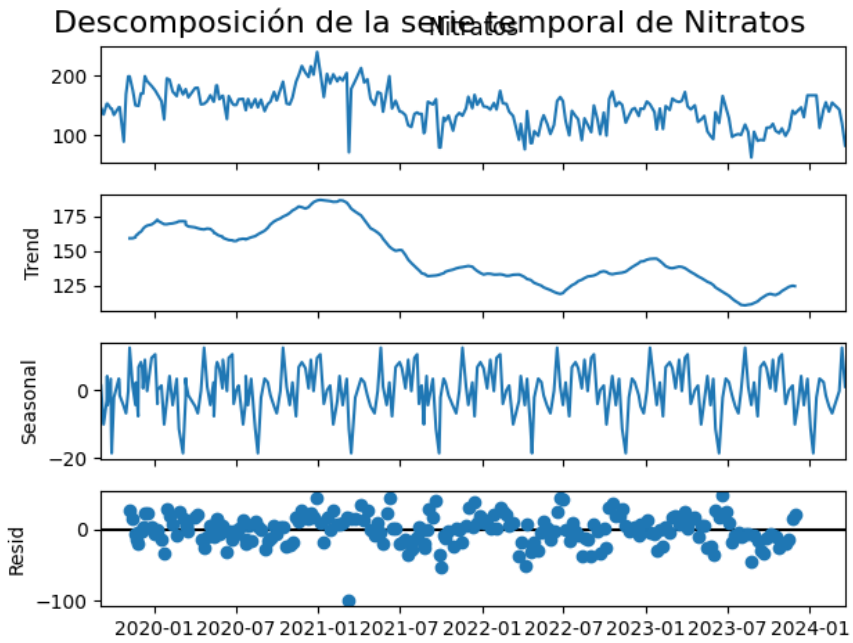


Figura 4.1: Figura que muestra los datos estacionales extraídos de la base de datos de nitratos (NO3) por el modelo SARIMA.

En la primera de cuatro gráficas de la figura 4.1.1 vemos los datos de nitratos de la base de datos. En la segunda el "trend" suavizado de los datos. La tercera es el más interesante, ya que nos hace visualizar la estacionalidad de los datos. Y la cuarta es el residuo de cada dato, respecto a la estacionalidad.

Con el objetivo de encontrar el mejor modelo SARIMA, según sus parámetros, “p”, “P”, “d”, “D”, “q”, y “Q” se ponen en un rango de 0 a 3. Aumentar este rango implica de una gran capacidad computacional. El parámetro “s” se pone en 12, señalando una estacionalidad mensual (12 meses del año).

Por último, el modelo predice sobre los valores de la base de datos y dependiendo del número de “steps” futuros asignados, predecirá ese número de datos (steps) en el futuro. Por ejemplo: para “steps” = 30, nos dará 30 valores diarios futuros. Se visualiza y devuelve el MSE, MAE y R2 de predicción del modelo.

Como apéndice del modelo central se implementó el llamado “intervalo de confianza” en los resultados finales. Este intervalo nos da un rango dentro del cual se espera que caiga el valor verdadero, es un espectro de probabilidad basado en la distribución normal [36].

Octava iteración. Predicciones futuras.

En esta última iteración, se realizarán las predicciones a futuro de las tres características estudiadas en este proyecto. A la hora de escoger el modelo a utilizar para la predicción futura de estas características se tuvo en cuenta la posibilidad de construir un modelo híbrido entre RF y SARIMA. Esto es, debido a la capacidad ya integrada de los modelos SARIMA de estimar valores futuros con su parámetro “steps” mencionado en la iteración anterior. Sin embargo, este modelo se descartó, ya que RF demostró ser más preciso en la estimación general de los datos, también “futuros” durante el entrenamiento y en este híbrido las estimaciones a futuro las realizaría el modelo SARIMA, no el modelo RF.

También se pensó dar estos parámetros estacionales extraídos por SARIMA a RF para el entrenamiento. Sin embargo, no se encontró manera de alimentarlo a la base de entrenamiento del modelo RF. Aparte que RF también parece extraerlos, quizás de forma tan exacta o visual, pero los usa para sus predicciones.

Se definió “n_lags” = 12, donde cada muestra usa los últimos 12 valores de la variable para predecir el siguiente. Este marco deslizante, o "ventana de tiempo" (last_window), se actualiza continuamente con cada predicción, de modo que siempre contiene los últimos 12 valores para anticipar el próximo. No se eligió más de 12 para evitar sobreajustes. Esta elección de 12 mediciones con lags se quitan del conjunto de entrenamiento, pasando los “splits” de 256 a 244, lo cual explica la pequeña subida en los errores de predicción.

4.2. Resultados

Resultados primera iteración

Incluso con una base datos incompleta, se puede ver la capacidad de predicción del modelo, visualizado en la figura 4.2, al obtener un valor medio de MSE de **1,276**. Se ha de tener en cuenta que este valor procede de un modelo con un número pequeño de características, lo que lo hace menos propenso al sobreajuste.

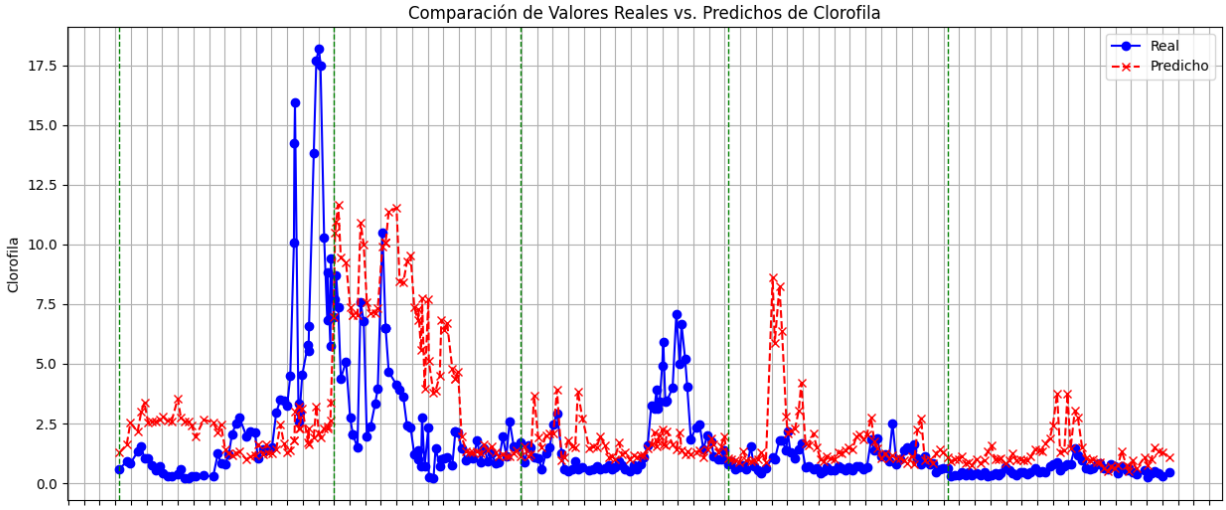


Figura 4.2: Diagrama comparativo de las muestras reales y predichas por el primer modelo. Las líneas discontinuas verticales son las muestras en las que empieza cada Split.

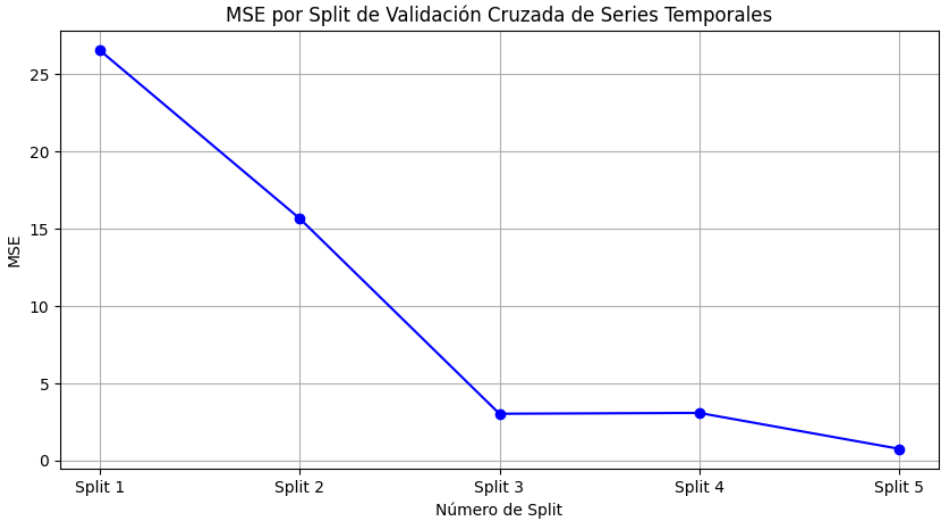


Figura 4.3: Grafica del MSE medio por Split de la primera iteración.

Los resultados de la figura 4.3 demuestran que, a pesar de los posibles desajustes, los relativamente buenos resultados del algoritmo.

Resultados comparativos de la segunda, tercera y cuarta iteración

En estas iteraciones se comprueba el aumento de rendimiento del modelo al ampliar el número de características con las que entrenar. Los resultados se muestran en la tabla 4.1(en negrita se muestra el mejor resultado para cada variable a predecir).

	Iteración 2	Iteración 3	Iteración 4
MSE Clorofila-a	9,551	9,412	**6,968**
MSE Nitratos	997,192	989,843	**762,55**
MSE Fosfatos	0,061	0,067	**0,058**

Tabla 4.1: Tabla comparativa de los MSE por iteración de cada variable.

Se puede observar una gran similitud entre los valores de la segunda y tercera iteración. Esto encajaría, pues todavía no se usaban todas las variables al predecir, por lo que al introducir las variables de nitraros y fosfatos los resultados no se vieron afectados.

Sin embargo, en la cuarta iteración, al introducir nuevas variables en el entrenamiento y al usar todas las variables para la predicción de cada una de ellas se mejoró ampliamente al modelo.

Resultados quinta iteración

En esta iteración, se comienza a modificar los parámetros de las funciones, más específicamente, la de la validación cruzada de series temporales. El propósito era conocer el impacto del parámetro de nº de “splits” (n_splits), es decir, el número de subgrupos en el que se divide el conjunto de entrenamiento.

N.º de splits	10	20	30	50	100	200	256
MSE Clorofila-a	5,622	6,586	4,739	4,686	2,482	2,134	**1,761**
MSE Nitratos	762,871	714,043	613,119	648,621	602,122	560,851	**520,882**
MSE Fosfatos	0,053	0,047	0,041	0,041	0,045	0,042	**0,0330**

Tabla 4.2: Tabla comparativa del MSE para los diferentes números de splits de cada variable predicha.

Al comienzo, se puede observar un incremento del MSE al aumentar el número de splits, que nos llevó a pensar que no sería óptimo continuar aumentando. Pero, como se podría haber predicho, al aumentar el número de splits, esto es, disminuir el tamaño de los bloques de datos de test y train en cada iteración, permites al modelo aprender más paulatina y precisamente del conjunto de datos, mejorando el MSE del modelo. Además cuantos más splits haces, las predicciones qué se hacen son más próximas a los datos de entrenamiento lo que mejora el resultado. En el límite del número de particiones igual

al número de datos entrenas un modelo para predecir el siguiente punto de la serie.

Para el MAE se obtuvieron los mismos mejores valores de error.

Resultados sexta iteración

Esta fue la iteración más costosa computacionalmente, por el uso de la función “GridSearchCV” [37]. Se usó con la intención de encontrar los mejores parámetros para cada uno de los modelos de cada una de las variables, que fueron:

- Para la **clorofila**: 'bootstrap': True, 'max_depth': 20, 'max_features': 'log2', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 100.

- Para los **nitratos**: 'bootstrap': False, 'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 100

- Para los **fosfatos**: 'bootstrap': False, 'max_depth': 20, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 200

Esta iteración se realizó con 256 splits que nos daban los mejores resultados. Los resultados se muestran en la tabla 4.3.

Variable	MSE
Clorofila-a	**1,211**
Nitratos	**252,867**
Fosfatos	**0,016**

Tabla 4.3: Tabla donde se muestra los MSE medios para cada variable para sus respectivos modelos con parámetros optimizados.

Al optimizar los parámetros se obtienen los mejores resultados hasta el momento. En este punto, se considera que hemos llegado al modelo más eficiente. Se consigue superar el MSE del modelo de la primera iteración. Para MAE se obtuvieron los mismos valores. El error R^2 no se pudo obtener, ya que no se puede definir correctamente con menos de dos datos por iteración.

Aunque en apariencia el MSE de predicción de los nitratos sigue siendo muy elevado, se puede observar en la figura 4.4, que el modelo comprende el patrón de sucesión de los datos. El elevado valor del error es debido a la alta variabilidad de las mediciones de las concentraciones de nitratos de la base de datos.

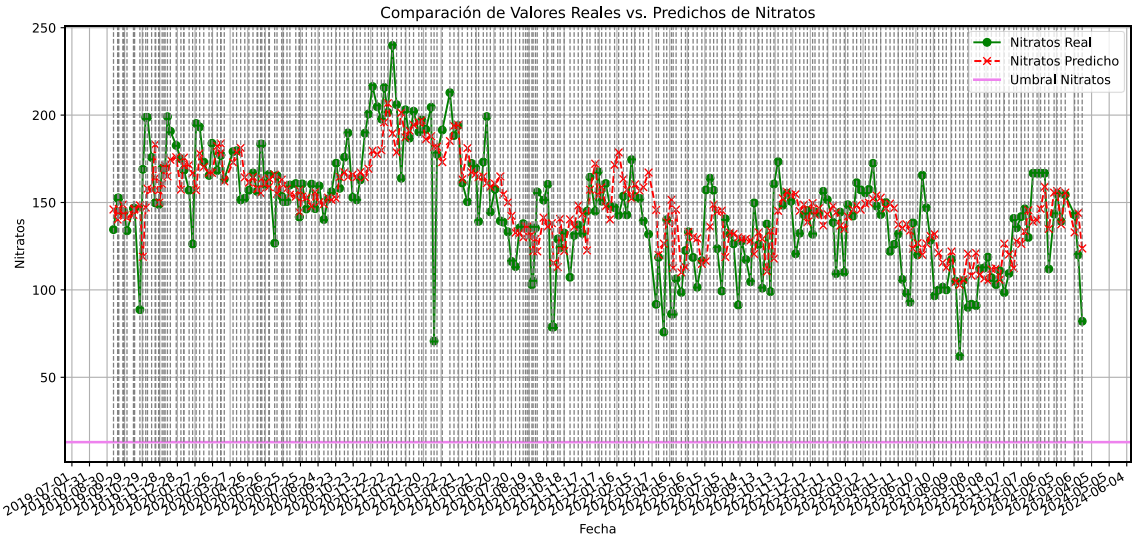


Figura 4.4: Figura comparativa entre los valores de nitratos de la base de datos y los valores predichos por el modelo. Las líneas verticales (gris) son el comienzo de cada split.

Una vez obtenido este modelo optimizado, se podría comenzar a predecir valores futuros. Con futuros nos referimos a predecir valores que no tengan medición en la misma fecha dentro de la base de datos. Sin embargo, antes de hacerlo, habría que asegurarse de no acarrear ningún tipo de sobreajuste. Para poder visualizar este posible sobreajuste se utilizaron las curvas de pérdidas para el MSE y MAE.

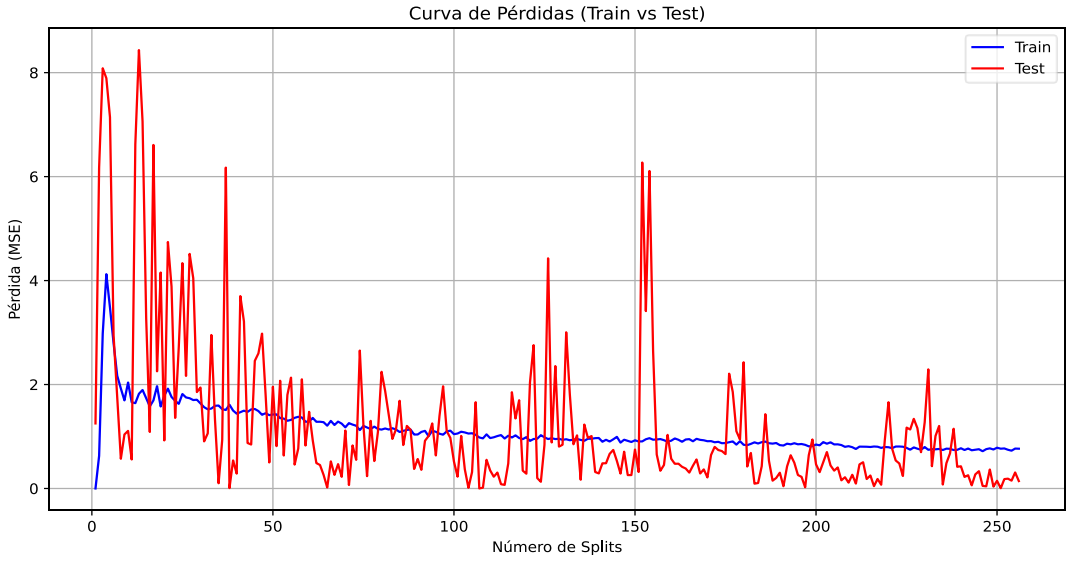


Figura 4.5: Curva de perdidas donde se representa el MSE para cada punto o valor de la base de datos a lo largo de la ejecución del modelo en la predicción de clorofila.

Según la figura 4.5, el modelo, como se podría esperar, tiene dificultades al inicio para entender y estimar los datos, pues presentan cierta variabilidad. No obstante,

conforme se le dan más datos de entrenamiento a cada “split”, parece adaptarse a los valores de clorofila sin grandes sobreajustes.

No se deduciría sobreajuste, ya que entre ambas curvas de perdida no hay una gran separación, es decir, oscilan sobre los mismos niveles.

Resultados séptima iteración

En esta iteración comprobamos la capacidad de otros modelos, con la intención de contrastar la capacidad de nuestro modelo. Primero con un modelo también de Machine Learning; pero utilizando otro algoritmo de regresión. En este caso, regresión lineal:

Variable	MSE para Regresión Lineal
Clorofila (Chl-a)	5,316
Nitratos	470,022
Fosfatos	0,194

Tabla 4.4: Tabla donde se muestran los valores de MSE para las tres variables a estimar usando el algoritmo de regresión lineal.

Como se puede observar, no se consigue ninguna mejora en ninguna de las tres variables, por lo que se corrobora lo entendido en la literatura sobre la comparación entre ambos algoritmos de regresión [18].

Como segunda comparación se utiliza un modelo SARIMA. Un modelo referencia para modelos de pronóstico con datos de series temporales. En la tabla 4.5 se muestran los mejores parámetros del modelo y parámetros estacionales derivados de los parámetros del modelo y en la tabla 4.6 los valores medios MSE de dichos modelos.

Variable	Parámetros (p, d, q)	Parámetros estacionales (P, D, Q, s)
Clorofila (Chl-a)	2, 1, 2	1, 0, 1, 12
Nitratos	1, 1, 2	0, 1, 1, 12
Fosfatos	2, 1, 2	0, 0, 0, 12

Tabla 4.5: Se muestran los parámetros del modelo y los parámetros estacionales derivados, de aquellos modelos con mejor capacidad de predicción (menor MSE).

Variable	MSE	MAE	R^2
Clorofila (Chl-a)	3,671	0,826	0,591
Nitratos	852,426	18,958	0,107
Fosfatos	0,056	0,116	-0,545

Tabla 4.6: Resultados de MSE medio, MAE medio y R^2 para el modelo de pronóstico de datos en series temporales SARIMA.

Observamos, un R^2 negativo. Esto podría ser por dos motivos principales. Podría indicar un sobre ajuste a los datos de entrenamiento, pero la figura 4.6, que enseña su predicción de los fosfatos, no parece indicarlo. Otra posibilidad es debido a una gran variabilidad de los datos, por lo que devuelve errores mayores en torno a su media. Y efectivamente, dado el grave “pico” de variación en el intervalo de confianza de las primeras predicciones, podemos confirmar esta hipótesis.

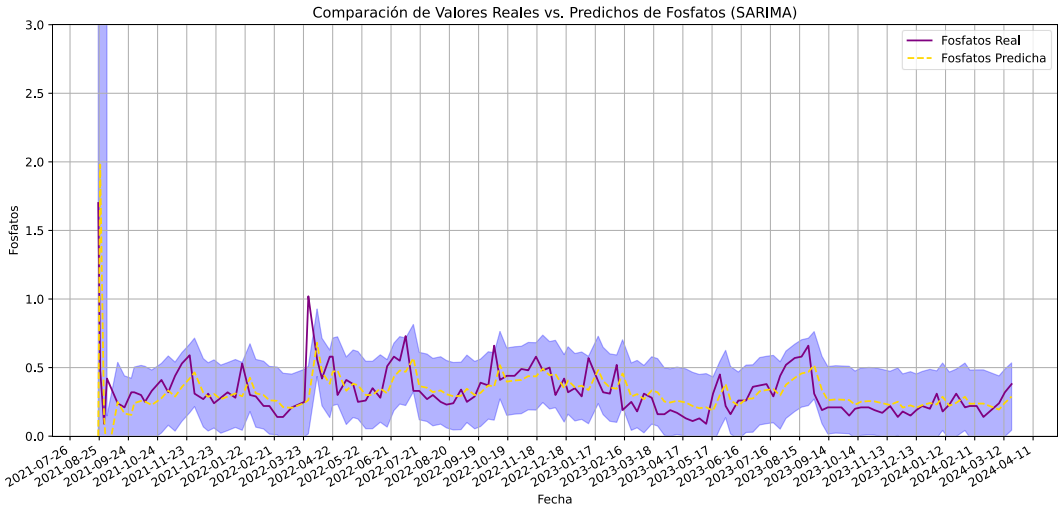


Figura 4.6: Gráfica representativa de la predicción del modelo SARIMA de los fosfatos. La sombra en azul es el intervalo de confianza para cada punto.

Con estos datos podemos concluir que, incluso con los mejores parámetros, un modelo especializado en el reconocimiento de patrones estacionales y también en datos en series temporales, no es más preciso en la predicción que un modelo Random Forest optimizado; o al menos no aplica estos patrones con la misma capacidad; aunque el modelo RF no los segregue y los pueda visualizar tan claramente.

Resultados octava iteración

Como se ha contrastado, el modelo con mayor capacidad de predicción es el Random Forest optimizado. Es este el que se utilizará para las predicciones a futuro. Estas predicciones se compararán con el umbral de calidad de las aguas del Real Decreto [2].

Para visualizar los resultados de esta predicción futura se utiliza la figura 4.7. El propósito de esta gráfica es, al mismo tiempo, visualizar la variabilidad de los datos históricos previos a la predicción (datos estacionales, a la izquierda de la línea violeta discontinua) y, a su vez, comparar los datos predichos por el modelo (datos amarillos) con las nuevas mediciones añadidas a la base de datos durante la realización de este

trabajo (datos verdes a la derecha de la línea violeta de inicio de predicciones).

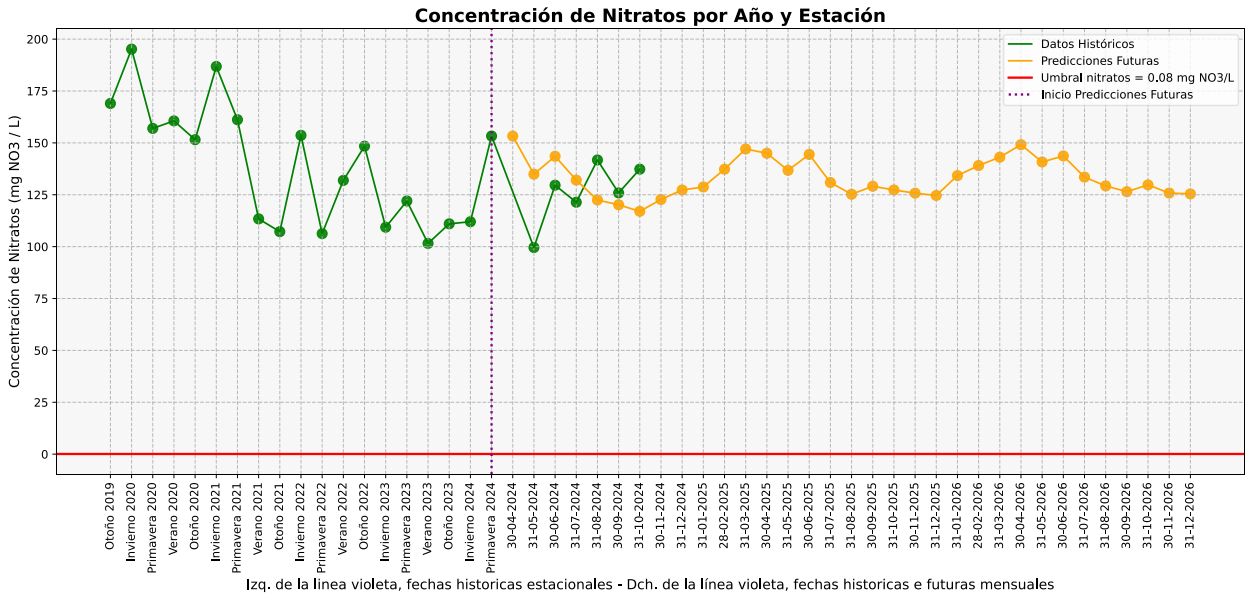


Figura 4.7: En esta gráfica se visualiza a la izquierda del umbral violeta, una representación de los datos previos a la predicción de la base de datos y a la derecha una comparación entre los datos predichos por el modelo y los datos futuros reales, todo esto de la concentración de nitratos.

Estos datos estacionales son obtenidos, seleccionando un día representativo de cada estación, que fueron, 31/01 para invierno, 30/04 para primavera, 31/07 para verano y 31/10 para otoño. Si ya se tuvieran las mediciones específicas para estos días en la base de datos, se tomarían dichas mediciones, pero a falta del dato, igualmente se calcula una media con las 3 medidas más cercanas a la de esa fecha, siendo el resultado de esa media el dato estacional.

Como hubo un intervalo de tiempo considerable entre la instalación del primer modelo y la de este último, las bases de datos de las variables estudiadas fueron actualizándose con el tiempo, permitiendo comparar realmente los datos predichos a futuro con los datos “futuros” reales. Para hacer una justa comparación se calculó la media de las tres medidas más cercanas al último día de cada mes y se comparó con las mediciones futuras de fin de cada mes del modelo.

En el ejercicio de predicción, el modelo sólo puede devolver valores mensuales a futuro. Esto, aunque sí nos enseña a gran escala su “versión” de cómo se verá esa variable en el futuro, deja que desear, ya que se le podría pedir como a los modelos SARIMA un intervalo de confianza, donde se nos muestre la posible variabilidad de cada una de sus predicciones.

5

CONCLUSIONES Y TRABAJO FUTURO

La visión inicial de este proyecto era usar las herramientas aprendidas durante la carrera para entender un problema real y actual de un entorno que me es cercano. Construir un modelo predictor de diferentes variables críticas suficientemente capaz de mostrar el estado actual y, gracias a él, el estado futuro de la laguna de agua salada más grande de Europa: el Mar Menor.

Se implementó un modelo contrastado en la teoría [18] [26] que fue probado bajo múltiples condiciones de entrenamiento para la predicción de valores de las concentraciones de clorofila-α, nitratos (NO_3) y fosfatos (PO_4).

Se midió su capacidad (MSE) con diferentes tamaños de bases de datos y con diferentes dependencias entre las variables de dicha base de datos. Se contrastó la importancia del número de “splits” o subconjuntos de datos que alimentar al modelo en el entrenamiento. Se comprobó exhaustivamente la mejor combinación de hiperparámetros del algoritmo Random Forest. A continuación con el propósito de asegurarnos de estar usando la mejor arquitectura, se comparó el modelo RF de hiperparámetros optimizados con un modelo basado en el algoritmo de regresión lineal y con un modelo especializado en pronostico con series temporales y parámetros estacionales llamado SARIMA. Dando los mejores resultados el modelo basado en Random Forest.

Siendo estos resultados finales, unos errores MSE de **1,211**, **252,867** y **0,016** para la clorofila-α, los nitratos y fosfatos respectivamente. Estos resultados son razonablemente acertados, dándonos una visión certera del estado actual, también al ser comparados constantemente con los umbrales de calidad de las aguas [2].

Una vez configurados los modelos para reducir los posibles sobreajustes se

realizaron predicciones a futuro. Se han generado predicciones hasta 2027, donde se observa cierta correlación entre las predicciones y valores futuros obtenidos a posteriori. Teniendo como umbral de la clorofila-α **1,8 µg/L**, y de fosfatos **0,072 mg/L**, encontramos que las predicciones futuras los colocan de media sobre los **1,91 µg/L**, y **0,22 mg/L** respectivamente, que aunque no ideal, se pueden considerar en un rango aceptable, aunque lejos de los óptimo. Sin embargo, se debe señalar las mediciones de nitratos, ya que estando su umbral en **0,080 mg/L,** la media de sus predicciones es de **125,24 mg/L**. El impacto de estas mediciones queda fuera de este trabajo; pero se debe indicar que esta medición a futuro no se aleja de las ya presentes en las bases de datos disponibles actualmente.

Se han puesto de manifiesto algunas de las debilidades de este proyecto. Con interés de no alargar este proyecto aún más e inspirado por lo explicado por S. Ravid y A. Amitai, en su artículo “Tabular data: Deep learning is not all you need” [17], se decidió centrarse en el Machine Learning y no comparar el modelo final con una RNN de Deep Learning. Consideramos que este sería el siguiente paso del trabajo futuro más obvio.

También, al compararlo con otros algoritmos de regresión, se consideraron los Stochastic Gradient Boosting Decision Trees o GBDT, los cuales según el artículo “An up-to-date comparison of state-of-the-art classification algorithms” [18], eran rápidos y precisos; pero se consideraron, a pesar de sus diferencias, parecidos en forma y estructura de datos a los árboles de Random Forest.

Respecto al modelo en si mismo, se pueden encontrar posibles mejoras, como el uso de un intervalo de confianza o de variación en las mediciones futuras para observar la variabilidad de estas mediciones tan exactas y proteger las predicciones de posibles eventos extraordinarios; sobre todo, al tratarse con datos temporales dependientes de la meteorología y ser predicciones de un día en específico.

A su vez, por la naturaleza del proyecto, hay infinidad de posibles mejoras al juntar dos temas tan diferentes y en los cuales no se tiene el mismo nivel de entendimiento. Esto es: las tecnologías de comunicación y la biología y/o bioquímica.

Y respecto al formato de escritura del proyecto, se entiende que se presta mucha atención, al proceso de obtención de los resultados, enseñando paso a paso como se obtuvieron. Se considera que el valor de la investigación está en mostrar tanto los caminos exitosos como las hipótesis descartadas.

BIBLIOGRAFÍA

[1] C. Valerio, L. De Stefano, G. Martínez-Muñoz y A. Garrido, «A machine learning model to assess the ecosystem response to water policy measures in the Tagus River Basin (Spain),» *Science of the Total Enviroment,* vol. 750, nº 141252, pp. 1-14, Enero 2021.

[2] Ministerio de Agricultura, Alimentación y Medio Ambiente. (12, sept. 2015). *Real Decreto 817/2015, de 11 de septiembre, por el que se establecen los criterios de seguimiento y evaluación del estado de las aguas superficiales y las normas de calidad ambiental.* [En línea]. (Disponible)

[3] Comisión Europea, «COMMON IMPLEMENTATION STRATEGY FOR THE WATER FRAMEWORK DIRECTIVE (2000/60/EC),» [En línea]. (Disponible).

[4] Región de Murcia, «Sobre el Mar Menor,» Fondo Europeo de Desarrollo Regional. [En línea]. (Disponible). [Último acceso: 26 Marzo 2024].

[5] L. Alcolea, M. Zorrilla-Miras, J. Martínez-Paz y D. Solera, «Hydrogeological modelling for the watershed management of the Mar Menor coastal lagoon (Spain),» *Science of the Total Enviroment,* vol. 663, pp. 901-914, 2019.

[6] Oficina de Impluso Socioeconomico del Medio Ambiente, Espacios Natura Región de Murcia, «Informe integral sobre el estado ecológico del mar menor,» Comité de Asesoramiento Científico del Mar Menor, Murcia, 2017. (Disponible).

[7] Grupo de coordinación de Pacto por el Mar Menor, «pactoporelmarmenor.blogspot.com,» 2 Diciembre 2019. [En línea]. (Disponible).

[8] HydroGeoModels AG, ETHZ-EAWAG, Univ. de Murcia, «www.futurewater.es,» 2017. [En línea]. (Disponible).

[9] A. Cardoso, J. Duchemin, P. Margarou y G. Premazzi, «Criteria for the identification of freshwaters subject to eutrophication. Their use for implementation of the “Nitrates” and Urban Waste Water Directives,» 2001.

[10] J. Velasco, J. Lloret, A. Millan, A. Marin, J. Barahona, P. Abellan y D. Sanchez-Fernandez, « Nutrient and Particulate Inputs into the Mar Menor Lagoon (SE Spain) from an Intensive Agricultural Watershed,» *Springer Nature,* vol. 176, pp. 37-56, 2006.

[11] M. Acosta, «Qué es la clorofila y sus tipos,» 15 Abril 2020. [En línea]. (Disponible).

[12] P. H. Doering, R. H. Chamberlain y K. M. Haunert, «Chlorophyll a and its use as an indicator of eutrophication in the Caloosahatchee Estuary, Florida,» *Florida Scient.,* vol. 69, pp. 51-72, 2006.

[13] S. P. Boeykens, M. N. Piol, L. S. Legal, A. B. Saralegui y C. Vázquez, «Eutrophication decrease: Phosphate adsorption processes in presence of nitrates,» *Journal of Environmental Management,* vol. 203, pp. 888-895, 2017.

[14] COIAL Partners, « ¿Cuáles son las causas de la eutrofización de aguas?,» 22 Abril 2022. [En línea]. (Disponible). [Último acceso: 18 Diciembre 2023].

[15] V. R. S. Conesa, «Informe anual. PROYECTOS, PLANES, PROGRAMAS, ACTUACIONES, INICIATIVAS E INVERSIÓNES DE LA CARM,» Murcia, Septiembre 2023. (Disponible).

[16] R. Johnson, M. Lindegarth y J. Carstensen, «Establishing reference conditions and setting class boundaries,» Havsmiljöinstitutet, Sweden, 2013.

[17] S. Ravid y A. Amitai, «Tabular data: Deep learning is not all you need,» *Information*

Fusion, vol. 81, pp. 84-90, 2022.

[18] C. Zhang., C. Lui y G. A. X. Zhang, «An up-to-date comparison of state-of-the-art classification algorithms,» *Expert Systems With Applications,* vol. 82, pp. 128-150, 2017.

[19] C. Bergmeir y J. Benítez, «On the use of cross-validation for time series predictor evaluation,» *Information Sciences,* vol. 191, pp. 192-213, 2012.

[20] M. Filho, «How To Do Time Series Cross-Validation In Python,» 12 Julio 2023. [En línea]. (Disponible).

[21] H. Mirete. " Extracción y predicción de datos de series temporales de reservas de vuelo," Trabajo de Fin de Máster, ETSE., UV., Valencia, España, 2019. [En línea]. (Disponible)

[22] T. G. Dietterich, «An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization,» *Machine Learning,* vol. 40, pp. 139-157, 2000.

[23] L. Breiman, «Random Forests,» *Machine Learning,* vol. 45, pp. 5-32, 2001.

[24] D. Maulud y A. Mohsin, «A Review on Linear Regression Comprehensive in Machine Learning,» *Journal of applied science and technology trends,* vol. 1, nº 2, pp. 140-147, 2020.

[25] B. Artley, «Time Series Forecasting with ARIMA , SARIMA and SARIMAX,» 26 Abril 2022. [En línea]. (Disponible).

[26] J. KAJURU, K. ABDULKARIM y M. MUHAMMED, «Forecasting Performance of Arima and Sarima Models on Monthly Average Temperature of Zaria, Nigeria,» *ATBU J. Sci. Technol. Educ. ,* vol. 7, nº 3, pp. 205-212, 2019.

[27] F.-M. Tsenga y G.-H. T. H.-C. Yub, « Combining neural network model with seasonal time series ARIMA model,» *Technological Forecasting & Social Change,* vol. 69, pp. 71-87, 2002.

[28] DataBitAI, «Métricas de Evaluación en Machine Learning,» 17 Abril 2023. [En línea]. (Disponible).

[29] S. Varma y R. Simon, «Bias in Error Estimation When Using Cross-Validation for Model Selection,» *BMC Bioinform.,* vol. 7, nº 91, Febrero 2006.

[30] Universidad Politécnica de Cartagena, «Servidor de Datos Científicos del Mar Menor,» [En línea]. (Disponible).

[31] Canal Mar Menor, «Aforos y concentración de nutrientes,» FEDER; Región de Murcia, 7 Noviembre 2019. [En línea]. (Disponible).

[32] ScikitLearn, «RandomForestClassifier,» [En línea]. (Disponible).

[33] J. Brownlee, «Random Forest for Time Series Forecasting,» 1 Noviembre 2020. [En línea]. (Disponible)

[34] L. J. Tashman, «Out-of-sample tests of forecasting accuracy: an analysis and review,» *International Journal of Forecasting,* vol. 16, pp. 437-450, 2000.

[35] R. FIldes y S. Makridakis, «The Impact of Empirical Accuracy Studies on Time Series Analysis and Forecasting,» *Inter. Stat. Rev. ,* vol. 63, nº 3, pp. 289-305, 1995.

[36] StatsModels, «statsmodels.tsa.statespace.sarimax.SARIMAXResults.conf_int,» [En línea]. (Disponible).

[37] S. Navarro, «¿Qué es GridSearchCV?,» KeepCoding, 16 Abril 2024. [En línea]. (Disponible).