PREVISIÓN DE GENERACIÓN HIDRAÚLICA EN EL CONTEXTO DE TRANSICIÓN ENERGÉTICA Y VOLATILIDAD DEL MERCADO

TRABAJO FIN DE MÁSTER
CURSO ACADÉMICO 2023-2024

MÁSTER UNIVERSITARIO EN BIG DATA SCIENCE

UNIVERSIDAD DE NAVARRA
INSTITUTO DE CIENCIA DE LOS DATOS E INTELIGENCIA ARTIFICIAL

Héctor Daniel González Vargas

Pablo Caldevilla Sánchez

Jonathan Iván Ponce Ramírez

Tutor académico: Stella Maris Salvatierra Galiano

Tutor de empresa: Marta Enesco Garrido

Cotutor de empresa: Roberto Flores Herrera

Madrid 04 de Julio de 2024

Resumen

RESUMEN: El presente proyecto evalúa diferentes técnicas de *Machine Learning* (XGBoost, LightGBM), modelado tradicional (SARIMAX) y enfoques alternativos (Prophet) en el pronóstico diario del caudal de ríos y afluentes, usando la información de variables meteorológicas, para un horizonte de predicción semanal (7 días). El punto de referencia para el análisis corresponde a la ubicación de la central hidroeléctrica de Pereruela y San Román en la provincia de Zamora en España. La extracción de los datos del caudal se realizó a través de la estación de medición más cercana, correspondiente a la estación del río Duero en Zamora. A lo largo del proyecto se establece un flujo de extracción de datos fundamentado en técnicas de *web scraping* y uso de API's *open-source* (en el caso de las variables meteorológicas). Luego de implementar técnicas avanzadas de optimización, incluyendo búsqueda Bayesiana de hiperparámetros y validación cruzada, los modelos de ensamble, en particular LightGBM, exhibieron un desempeño superior en la predicción de picos o periodos anómalos tras pruebas en entornos simulados. El modelo SARIMAX mostró una estabilidad consistente en todas las métricas de error evaluadas, logrando resultados óptimos, aunque no excepcionales, en la predicción de periodos de alto caudal. En contraste, Prophet no logró superar la estimación base de repetir el último dato observado para el periodo de estimación, obteniendo el error más elevado y resultados no satisfactorios. Además, se analizó el impacto de variables meteorológicas como la lluvia, humedad del suelo, ráfagas de viento y temperatura, resultando ser los atributos más influyentes en el nivel del caudal.

Palabras clave: Río Duero en Zamora, variables meteorológicas, *web scaping*, XGBoost, LightGBM, SARIMAX, Prophet.

ABSTRACT: This project evaluates different machine learning techniques (XGBoost, LightGBM), traditional modeling (SARIMAX) and alternative approaches (Prophet) in daily forecasting the flow of rivers and tributaries, using information from meteorological variables, for a weekly prediction horizon. The reference point for the analysis corresponds to the location of the Pereruela and San Román hydroelectric power plant in the province of Zamora in Spain. The extraction of flow data is carried out through the closest measurement station, corresponding to the Duero River station in Zamora. Throughout the project, a data extraction flow is established based on *web scraping* techniques and the use of *open-source* APIs (in the case of meteorological variables). After implementing advanced optimization techniques, including Bayesian hyperparameter search and cross-validation, ensemble models, particularly LightGBM, exhibited superior performance in predicting anomalous peaks or periods after testing in simulated environments. The SARIMAX model showed consistent stability in all the error metrics evaluated, achieving optimal, although not exceptional, results in predicting high flow periods. In contrast, Prophet failed to surpass the base estimate of repeating the last observed data for the estimation period, obtaining the highest error and unsatisfactory results. In addition, the impact of meteorological variables such as rain, soil humidity, wind gusts and temperature was analyzed, proving to be the most influential attributes on the flow level.

Keywords: Duero River in Zamora, meteorological variables, web scaping, XGBoost, LightGBM, SARIMAX, Prophet.

Índice general

Capítulo 1

Introducción

1.1. Descripción del problema

En el contexto actual, el sector energético está ganando importancia a nivel nacional e internacional debido a las recientes dinámicas geopolíticas y los efectos de la pandemia, que han afectado los precios y el funcionamiento del mercado. Además, las directrices europeas promueven una transición hacia energías renovables para mitigar el cambio climático, cuyos impactos ya se ven en fenómenos extremos como sequías y precipitaciones intensas.

La cuenca hidrográfica del Duero es una de las principales cuencas hidrográficas de España. Se extiende por gran parte del norte y noroeste de la península ibérica. Esta cuenca abarca una extensa área que incluye provincias como Soria, Burgos, Valladolid, Zamora, Salamanca, Ávila, Segovia, y parte de León, Palencia, y Cáceres. El río Duero es el principal curso de agua de esta cuenca, siendo uno de los ríos más largos de la península ibérica y uno de los más importantes de la vertiente atlántica.

Existen diversas modalidades de centrales hidroeléctricas[2], como lo son: centrales de embalse, bombeo y agua fluyente. Las centrales de embalse, que almacenan agua, tienen la flexibilidad de ajustar la generación eléctrica según factores como las condiciones del mercado energético y las precipitaciones. Las de bombeo utilizan dos embalses a diferentes alturas para regular la producción eléctrica según la demanda, bombeando agua en momentos de baja demanda y permitiendo su flujo en momentos de alta demanda. Por otro lado, las centrales fluyentes aprovechan el flujo natural de los ríos para generar electricidad mediante turbinas hidroeléctricas, sin necesidad de un embalse de almacenamiento.

Por ejemplo, La central hidroeléctrica de Pereruela y San Román[3] corresponde a una infraestructura de alrededor de 4 metros de altura situada en el paso del río Duero por la localidad zamorana de San Román de los Infantes, cuenta con dos minicentrales hidroeléctricas exteriores: La central de San Román, que comenzó a operar en 1969, tiene un salto de 14 metros y una potencia instalada de 5,6 MW. La central de Pereruela, en funcionamiento desde 1993, tiene una potencia instalada de 3,35 MW.

El paso del río Duero en Zamora representa la variable de mayor impacto en la generación de energía por parte de la central, el comportamiento estacional del caudal, y su enorme variabilidad representa un reto para la gestión y el manejo adecuado de la planta.

Por lo tanto, para la gestión eficaz de las centrales hidroeléctricas, es fundamental conocer el nivel del caudal futuro del río específico por varias razones:

- Generación de energía: La cantidad de agua disponible afecta directamente la capacidad de generación de energía de la central hidroeléctrica, en mayor medida para las centrales de agua fluyente. Cuanta más agua esté disponible, mayor será la capacidad de producir electricidad.
- Planificación operativa: Conocer el nivel de caudal futuro permite a los operadores de la central hidroeléctrica planificar la operación de las turbinas y otros equipos para optimizar la producción de energía. Esto incluye programar el flujo de agua a través de las turbinas de manera eficiente para maximizar la producción de energía eléctrica.
- Dinámica de precios en el mercado: El mercado de generación eléctrica suministra electricidad a distribuidores y comercializadores mediante contratos bilaterales físicos y a través del mercado organizado o *pool*, donde se realizan todas las transacciones. Este mercado funciona mediante la casación de oferta y demanda, siendo el mercado diario el más relevante, ya que en él se negocia la energía para cada hora del día siguiente. Por lo tanto, una estimación precisa del nivel de generación energía atado a la cantidad de agua procedente del caudal, puede optimizar los procesos de compra y venta, logrando mayor estabilidad financiera y un ajuste económico exacto en las casaciones.

En la actualidad, la disponibilidad de sistemas dedicados al pronóstico de caudales depende, en su mayoría, de las confederaciones hidrográficas, en este caso, del Duero, que, aunque pueden ser útiles, no ofrecen la versatilidad ni la robustez necesarias para hacer frente a las demandas cambiantes del entorno fluvial. Por tanto, se hace evidente la necesidad de contar con pronósticos más precisos, que sean generados a partir de diferentes arquitecturas de modelos capaces de adaptarse a diversas condiciones hidrológicas y climáticas, brindando a las empresas, autoridades y expertos en recursos hídricos la capacidad de tomar decisiones informadas y estratégicas.

Los modelos tradicionales, como ARIMA (*Autoregressive Integrated Moving Average*) y SARIMAX (*Seasonal Autoregressive Integrated Moving Average with Exogenous regressors*), suelen ser limitados debido a sus supuestos de linealidad, lo cual dificulta la detección de relaciones complejas y los hace sensibles a anomalías. Para superar estas limitaciones, los modelos de *Machine Learning* se han popularizado, especialmente los algoritmos de ensamble como XGBoost (*Extreme Gradient Boosting*) y LightGBM (*Light Gradient Boosting Machine*). Estos combinan múltiples modelos base, mejorando la precisión y capturando relaciones no lineales, lo que resulta en predicciones más robustas y precisas. Además, son más flexibles y escalables, adaptándose mejor a grandes conjuntos de datos y a variaciones en los patrones de la serie temporal. También, modelos alternativos como Prophet, desarrollado por Facebook, son cada vez más utilizados por su facilidad de uso y su robustez frente a datos faltantes y *outliers*.

1.2. Objetivos

1.2.1. General

Desarrollar un sistema integral de pronóstico diario para el caudal del río Duero en Zamora, en un horizonte semanal (7 días), utilizando técnicas avanzadas de modelado tradicional y aprendizaje automático, generando una comparativa de modelos y enfoques estadísticos.

1.2.2. Específicos

1. Diseñar un flujo de extracción, limpieza y análisis de series temporales para datos históricos del caudal del Duero en Zamora y variables meteorológicas relevantes.
2. Implementar modelos de series de tiempo, SARIMAX, Prophet, XGBoost y LightGBM para el pronóstico del caudal del Duero en Zamora. Evaluar y comparar su desempeño en términos de error de pronóstico y captura de picos o subidas repentidas.
3. Analizar e interpretar los factores internos y externos que afectan el comportamiento del caudal del río Duero en Zamora.

1.3. Motivación

La competitividad de las energías renovables está experimentando un crecimiento significativo, con la energía hidráulica consolidándose como la tercera fuente más importante en el mercado energético español. En este contexto, la integración de modelos y análisis avanzados no solo busca mejorar la capacidad predictiva y operativa de las centrales hidroeléctricas, sino también fortalecer su resiliencia frente a desafíos emergentes. Ante la creciente imprevisibilidad de las variaciones climáticas y las demandas energéticas, resulta crucial disponer de sistemas de pronóstico precisos y adaptables para asegurar la estabilidad y eficiencia de la infraestructura energética.

Este proyecto se enfoca en superar los desafíos específicos de gestión y operación que enfrentan las centrales hidroeléctricas en la cuenca del Duero, tomando como referencia la ubicación geográfica de las instalaciones de Pereruela y San Román. Esto se aborda en respuesta a los impactos de la variabilidad climática y las fluctuantes demandas energéticas. Se propone el desarrollo de un sistema avanzado de pronóstico de caudales que permita una comparativa exhaustiva entre enfoques estadísticos tradicionales y modelos predictivos avanzados, a través del uso de algoritmos de aprendizaje automático (*Machine Learning*). Esto facilitará no solo la mejora en la exactitud de las predicciones, sino también la eficiencia operativa y la gestión de riesgos.

En términos técnicos, se implementarán modelos con componentes paramétricos, como SARIMAX, que se basan en supuestos de linealidad y estructuras predefinidas. Paralelamente, se utilizarán modelos no paramétricos, como XGBoost y LightGBM, que trascienden estos supuestos y logran capturar relaciones no lineales complejas. Adicionalmente, se implementará el modelo Prophet, que descompone series temporales en componentes aditivos de tendencia y estacionalidad, empleando un enfoque Bayesiano para manejar valores atípicos y cambios en la tendencia. Esta combinación estratégica de modelos permitirá obtener predicciones más precisas y adaptativas.

1.4. Estado del arte

1.4.1. Introducción

Entre las herramientas más destacadas para el pronóstico de caudales, se encuentran los modelos hidrológicos, diseñados para simular los complejos procesos del ciclo del agua y su interacción con el entorno. Estos modelos son fundamentales para comprender y predecir el comportamiento de los caudales en diferentes condiciones y pueden clasificarse principalmente en dos tipos: empíricos y físicos. Los modelos empíricos, basados en datos históricos, generan predicciones a partir de patrones previos y son más simples y rápidos de implementar, pero pueden fallar si los datos son limitados o las condiciones cambian drásticamente. En contraste, los modelos físicos o hidrodinámicos utilizan principios fundamentales de conservación y movimiento, incorporando variables meteorológicas para mayor precisión en entornos variables, aunque son más complejos y requieren mayor esfuerzo computacional.

Los modelos de aprendizaje automático (*Machine Learning*) han revolucionado este campo, ofreciendo nuevas posibilidades para la integración y análisis de grandes volúmenes de datos. Estas arquitecturas permiten el desarrollo de modelos que pueden aprender de los datos en tiempo real y ajustarse dinámicamente a nuevas condiciones, ofreciendo pronósticos más precisos y fiables.

1.4.2. Antecedentes

Oliveira et al. (2019), a través de su trabajo: *"An intelligent hybridization of ARIMA with machine learning models for time series forecasting"*[4], utilizan un modelo ARIMA para el modelado lineal, y emplean dos técnicas de *Machine Learning*, el Perceptrón Multicapa (MLP) y la Regresión Vectorial de Soporte (SVR), para modelar los residuos del modelo ARIMA, buscando patrones no lineales que el modelo no puede capturar.

Los modelos no lineales (MLP y SVR) también se utilizaron para la combinación de pronósticos. Se exploraron dos enfoques de combinación: una combinación lineal que suma simplemente los resultados del ARIMA y los residuos modelados, y una combinación no lineal que utiliza MLP o SVR para integrar estos pronósticos. Este último enfoque buscó una función de combinación óptima que mejorara la precisión global del sistema mediante el análisis intensivo de datos.

Los experimentos realizados sobre varias series temporales complejas y conocidas mostraron que el sistema híbrido propuesto supera consistentemente tanto a los modelos individuales (lineales y no lineales) como a otros sistemas híbridos en términos de precisión, utilizando métricas como el Error Cuadrático Medio (MSE) y el Error Absoluto Medio (MAE). Esto validó la efectividad del enfoque híbrido en el manejo de diversos patrones y complejidades en los datos de series temporales.

Phan y Nguyen (2020)[5] concluyeron algo similar al desarrollar una metodología híbrida para la predicción de niveles de agua en el río Rojo, Vietnam, combinando modelos

estadísticos ARIMA y de aprendizaje automático. Iniciaron ajustando el modelo ARIMA a series temporales para capturar dinámicas lineales como tendencias y estacionalidades, y luego utilizaron los residuos de este modelo como entrada para modelos de aprendizaje automático como *Random Forest*, *Support Vector Machines* y *K-Nearest Neighbors*. Este enfoque híbrido no solo capturó las dinámicas lineales y no lineales, mejorando así la precisión de las predicciones, sino que también proporcionó un modelo robusto para aplicaciones críticas como la predicción de inundaciones y la gestión de recursos hídricos. Los resultados experimentales mostraron una mejora significativa en la precisión comparado con los modelos usados individualmente.

Si bien, la combinación de modelos ha generado resultados óptimos, es importante remarcar que, modelos ARIMA con un enfoque estacional también han demostrado tener estimaciones precisas. El estudio *"Seasonal ARIMA Prediction of Streamflow: Sobat River Tributary of the White Nile River"*[6] aborda la predicción estacional del caudal del río Sobat, un afluente del Nilo Blanco en Sudán del Sur, utilizando el modelo ARIMA estacional (SARIMA). La investigación de Kenyi et al (2023) analiza series temporales de datos históricos mensuales de flujo recolectados entre 1912 y 1982. El análisis de los datos mostró que la serie temporal no era estacionaria. Por lo tanto, teniendo en cuenta los ordenes de diferenciación, se seleccionó y ajustó un modelo óptimo SARIMA (2,1,0)(2,0,1), y se utilizó para pronosticar los valores de flujo de caudales para un período de cuatro años, obteniendo buenos resultados de ajuste con un error cuadrático medio de 1.132 y un error porcentual absoluto medio de 0.1489.

El estudio concluyó que el modelo SARIMA es efectivo para predicciones a largo plazo del flujo de caudales en el río Sobat y sugiere que podría ser una herramienta útil para la planificación y gestión de recursos hídricos.

Retomando la evaluación de algortimos de *Machine Learning*, el estudio *"Water level prediction using various machine learning algorithms: a case study of Durian Tunggal river, Malaysia"*[7] se enfocó en la predicción de niveles de agua del río Durian Tunggal en Malasia mediante el uso de seis algoritmos de aprendizaje automático. Los autores Ahmed et al. (2022) utilizaron datos diarios de 1990 a 2019 para entrenar y probar los modelos. El objetivo fue desarrollar modelos confiables para predecir los niveles de agua y mejorar la planificación y mitigación de riesgos de inundación.

Seis modelos fueron evaluados: *Gaussian Process Regression* (GPR), *Support Vector Regression* (SVR), *Linear Regression* (LR), *Tree Regression* (TR), XGBoost y *Ensemble Regression* (BOOSTER y BAGER). El GPR con kernel exponencial resultó el más preciso, capturando extremos de nivel de agua con alta precisión. El SVR cúbico fue el más efectivo en su categoría, mientras que el modelo de interacción destacó entre los lineales. XGBoost se distinguió por su rapidez y precisión. La adecuada combinación de entradas y el análisis de sensibilidad fueron cruciales para mejorar la precisión.

En linea con la investigacion realizada sobre modelos de ensamble, Belyakova et al. (2022), en el artículo *"Forecasting Water Levels in Krasnodar Krai Rivers with the Use of Machine Learning"*[8], exploran el potencial de los modelos de aprendizaje automático para predecir los niveles de agua en dos ríos de montaña en el Krai de Krasnodar: el Pshish y el Mzymta. Se utilizaron tres arquitecturas de aprendizaje automático: árboles de decisión, XGBoost y una red neuronal artificial basada en perceptrón multicapa (MLP).

Los modelos fueron evaluados para tiempos de anticipación de 1 a 20 horas, destacando que el tiempo de anticipación óptimo para el río Pshish fue de 15 a 18 horas utilizando el modelo XGBoost. Sin embargo, en el río Mzymta, aunque la calidad de la simulación fue buena, no se alcanzó la eficiencia necesaria debido a la considerable afluencia lateral y la formación asincrónica del flujo en diferentes partes de la cuenca.

En el caso del río Pshish, los mejores resultados de simulación de niveles de agua se obtuvieron utilizando el modelo XGBoost, que aprovechó eficazmente la correlación dentro de la muestra. Por otro lado, para el río Mzymta, el modelo no lineal MLP fue el más efectivo, indicando su capacidad para identificar relaciones no lineales dentro del conjunto de datos. Los autores sugieren que se podrían desarrollar modelos de pronóstico más efectivos para el río Mzymta incorporando datos adicionales sobre el comportamiento de los afluentes y datos meteorológicos.

Sin lugar a dudas, representa un reto la correcta inclusión de variables explicativas en el pronóstico de este tipo de series de tiempo. Elegir el abanico adecuado de variables meteorológicas es fundamental para un resultado óptimo.

El nivel de precipataciones resultó ser una variable determinante en el estudio *"Hydrological modeling based on the KNN algorithm: an application for the forecast of daily flows of the Ramis river, Peru"*[9] realizado por Efrain Lujano et al. (2022), el cual, investiga la aplicación del algoritmo *K-Nearest Neighbor* (KNN) para pronosticar los caudales medios diarios del río Ramis en Perú. Utiliza datos de precipitación media y caudal medio diario de estaciones hidrometeorológicas, evaluando la eficacia del algoritmo mediante métricas como el error porcentual absoluto medio (MAPE) y la eficiencia de Nash-Sutcliffe (NSE). Los resultados muestran que el KNN es confiable para pronósticos con rezagos de uno y dos días en caudales y tres días en precipitación, destacando su simplicidad y robustez para pronósticos a corto plazo.

El área de estudio es la cuenca del río Ramis, en el departamento de Puno, Perú. La investigación resalta la importancia de seleccionar características relevantes para el pronóstico, utilizando el coeficiente de correlación de Pearson y el algoritmo de importancia de permutación. Los caudales rezagados y la precipitación son cruciales para mejorar la precisión del modelo KNN.

Con el fin de establecer una comparativa entre el modelo anteriormente evaluado (KNN) y algoritmos de *Machine Learning* diversos. Khan et al. (2023) en su estudio *"Monthly streamflow forecasting for the Hunza River Basin using machine learning techniques"*[10] investigan el uso de métodos de aprendizaje automático para predecir el flujo mensual del río Hunza en Pakistán, empleando datos de caudal, precipitación y temperatura del aire entre 1985 y 2013. Se evaluaron las técnicas *Adaptive Boosting (AB)*, *Gradient Boosting* (GB), *Random Forest* (RF) y *K-nearest neighbors* (KNN), y se midió su rendimiento mediante métricas como RMSE, MSE, MAE y R^2. Los resultados mostraron que AB fue la técnica más precisa, con valores de RMSE 16.8, MSE 281, MAE 6.53 y R^2 0.998, superando a las otras técnicas evaluadas.

La investigación concluye que AB es altamente eficaz para la predicción del caudal mensual del río Hunza, demostrando ser una herramienta confiable para la gestión de recursos hídricos y la mitigación de riesgos en la región.

Un desafío enorme que tienen los modelos de ensamble, tiene que ver con el nivel de interpretabilidad que pueden alcanzar. Justamente, este reto se aborda a través de los valores SHAP en el estudio *River Ecological Flow Early Warning Forecasting Using Baseflow Separation and Machine Learning in the Jiaojiang River Basin, Southeast China"*[11] , Hao Chen et al (2023) desarrollan un modelo de alerta temprana para el caudal de ríos, utilizando la separación del caudal base y técnicas de aprendizaje automático en la cuenca del río Jiaojiang, en el sureste de China. Inicialmente, emplean un método basado en la separación del caudal base y el método Tennant para calcular el caudal ecológico del río. Posteriormente, crean un modelo de alerta temprana utilizando el modelo LightGBM. Finalmente, emplean el marco SHAP (Shapley Additive Explanations) para explicar cómo diversos factores hidrometeorológicos afectan las variaciones en las condiciones del caudal ecológico.

Los resultados del estudio, centrado en las estaciones hidrológicas de Baizhiao (BZA) y Shaduan (SD), muestran que las frecuencias del caudal base mensual en la temporada seca son del 20 % (7.49 m^3/s) y 30 % (4.79 m^3/s) respectivamente. La precisión del modelo de alerta temprana del caudal ecológico alcanza cerca del 90 % en estas estaciones durante las temporadas seca y húmeda. Se observa que la evaporación y el índice de flujo base tienen el mayor impacto en el caudal ecológico del río.

Se han abordado, del mismo modo, combinaciones de modelos de ensamble con algoritmos genéticos. El estudio *Daily Scale River Flow Forecasting Using Hybrid Gradient Boosting Model with Genetic Algorithm Optimization* de Kilinc et al (2023)[12]. presenta un modelo híbrido de aprendizaje automático para la predicción del caudal diario de ríos en la cuenca del río Sakarya, en Turquía. Utilizando una combinación del modelo de CatBoost con un algoritmo genético (GA), los autores evaluaron el rendimiento de este modelo híbrido en tres estaciones de medición de caudal: Adatepe, Aktaş y Rüstümköy, durante el periodo 2002-2012. Los resultados mostraron que el modelo GA-CatBoost superó a los modelos de referencia como CatBoost, LSTM y la regresión lineal en términos de precisión de predicción, destacando especialmente en la estación Aktaş.

El artículo concluye que el modelo GA-CatBoost no solo converge más rápido que CatBoost en el proceso de aprendizaje, sino que también ofrece resultados superiores en la predicción de caudales fluviales. Se observa que, aunque el modelo CatBoost es eficaz, puede tener limitaciones cuando no se configuran adecuadamente las variables.

El algoritmo que ha hecho contrapeso a LightGBM es XGBoost, por lo tanto, es interesante evaluar su comportamiento y desempeño respecto a otro tipo de modelos. El artículo titulado *"Modeling and forecasting rainfall patterns in India: a time series analysis with XGBoost algorithm"* por Pradeep Mishra et al (2024)[13]. presenta un análisis de series temporales y técnicas de aprendizaje automático para modelar y predecir los patrones de lluvia en India. Utilizando el algoritmo XGBoost, se compararon varios modelos estadísticos y de aprendizaje automático, incluyendo ARIMA, SVM, ANN y *Random Forest*, para determinar su precisión en la predicción de la lluvia anual. Los resultados mostraron que XGBoost superó significativamente a los modelos tradicionales como ARIMA y los modelos de espacio de estado debido a su naturaleza no lineal y flexible, lo que le permitió capturar mejor los patrones complejos de los datos de lluvia.

Trasladando el enfoque hacia el *Deep Learning*, es importante resaltar que, entre

las arquitecturas mayormente usadas, se encuentran las redes LSTM, que se han popularizado en el pronóstico de series temporales debido a su capacidad única para capturar dependencias a largo plazo en los datos. A diferencia de las redes neuronales tradicionales, las LSTM están diseñadas con unidades de memoria que les permiten recordar información por períodos prolongados, lo cual es crucial al tratar con secuencias donde el contexto temporal es relevante.

Hunt et al. (2022), en su estudio *"Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States"*[14], destacan las ventajas del algoritmo LSTM para mejorar las previsiones del flujo de ríos en diez estaciones de medición en diferentes regiones climáticas del oeste de EE. UU. Los autores comparan los resultados del LSTM con sistemas tradicionales basados en física y métodos híbridos físico-estadísticos. El LSTM se entrena con variables meteorológicas e hidrológicas de los reanálisis ERA5 y GloFAS-ERA5, junto con observaciones históricas del flujo de río. El desarrollo del modelo incluye:

1. Datos de Entrada: Uso de variables meteorológicas e hidrológicas promediadas por cuenca de ERA5 y GloFAS-ERA5, y observaciones históricas del flujo de río.
2. Proceso de Entrenamiento: Adaptación del LSTM a las características de cada estación de medición, utilizando técnicas de retropropagación y ajuste de hiperparámetros clave como la tasa de aprendizaje y la longitud de la secuencia de entrada.
3. Evaluación del Modelo: Uso de métricas NSE y KGE para medir la eficiencia del modelo, comparando las predicciones con observaciones reales.

Los resultados indican que el LSTM mejora significativamente la precisión de las previsiones de flujo de ríos en comparación con modelos tradicionales y métodos híbridos, capturando variabilidades temporales y manejando complejidades hidrológicas.

Existen variaciones al enfoque LSTM, como el Bidireccional (BiLSTM), que mejora la captura de información en secuencias de datos al procesarlas en dos direcciones: hacia adelante y hacia atrás. Mientras que un LSTM estándar procesa la secuencia de manera unidireccional, un BiLSTM aprovecha tanto los datos anteriores como los futuros en cada punto de la secuencia, manejando mejor contextos complejos y dependencias.

- Granata et al. (2022)[15] compararon dos modelos de predicción de flujo diario: un modelo basado en el apilamiento (*stacking*) de *Random Forest* y *Multilayer Perceptron* con *Elastic Net* como meta-aprendiz, y un modelo basado en redes neuronales BiLSTM. Ambos modelos demostraron capacidades de pronóstico comparables, aunque el modelo de apilamiento superó al BiLSTM en la predicción de caudales máximos, siendo menos preciso en la predicción de caudales bajos.
- Ahmed et al. (2022)[16] desarrollaron un modelo híbrido avanzado llamado CVMD-CBiLSTM, que integra técnicas de descomposición de señales (CEEMDAN y VMD) con redes CNN y BiLSTM para mejorar la predicción de niveles de agua en ríos. Este modelo demostró ser superior en precisión comparado con otros métodos, esencial para la gestión de recursos hídricos en escenarios de cambio climático.

- Bak y Bae (2023)[17] introdujeron el algoritmo PNP (perceptrón positivo y negativo) para predecir el caudal de los ríos utilizando datos de precipitación afectados por el cambio climático. El PNP mostró una capacidad notable para predecir el caudal de ríos y afluentes, superando a los modelos LSTM en términos de precisión y manejo de datos atípicos, con la posibilidad de futuras mejoras mediante la integración de más variables relacionadas con el clima y operaciones de presas río arriba.

1.4.3. Conclusiones

La investigación sobre modelos híbridos de series temporales y el uso de técnicas de *Machine Learning* ha mostrado resultados prometedores. En particular, se ha destacado el uso de modelos que combinan el ARIMA para capturar patrones lineales con técnicas de aprendizaje automático como el Perceptrón Multicapa (MLP) y la Regresión Vectorial de Soporte (SVR) para modelar los residuos y capturar patrones no lineales. Este enfoque ha demostrado ser eficaz en mejorar la precisión de las predicciones en series temporales complejas.

Más recientemente, los algoritmos de ensamble como XGBoost y LightGBM han ganado importancia debido a su capacidad y velocidad para manejar datos complejos. XGBoost, conocido por su arquitectura robusta, ha demostrado ser particularmente eficaz en la predicción de caudales en temporalidad diaria. Este algoritmo se destaca por su capacidad para realizar una optimización automática y eficiente del modelo, lo que resulta en mejoras significativas en las estimaciones, superando a modelos tradicionales y otros enfoques híbridos. Por otro lado, LightGBM se ha destacado por su eficiencia y capacidad para manejar grandes volúmenes de datos, así como por su precisión en la predicción de caudales con valores anómalos. El uso del marco SHAP para la interpretación de los modelos ha proporcionado una comprensión más profunda de cómo diversos factores hidrometeorológicos afectan las predicciones, lo que es crucial para la planificación y gestión de recursos hídricos. Los estudios han demostrado que los algoritmos de ensamble como XGBoost y LightGBM no solo superan a los modelos tradicionales como ARIMA, sino que también son superiores a otros modelos de aprendizaje automático en términos de precisión y velocidad.

Es importante remarcar que las variables meteorológicas son esenciales para el pronóstico de caudales debido a su impacto directo en el ciclo hidrológico. Entre las más importantes se encuentran, la precipitación y la temperatura, ya que influyen directamente en el volumen de agua en las cuencas. La precipitación es especialmente crítica, y la incorporación de otras variables como el nivel de evaporación del agua pueden mejorar aún más la precisión de los modelos.

A pesar de los avances en modelos híbridos y algoritmos de ensamble, no se ha popularizado, para este tipo de series temporales, el modelo Prophet, este último es conocido por su facilidad de uso y capacidad para capturar tendencias y estacionalidades en series temporales, pero su rendimiento relativo frente a modelos avanzados como XGBoost y LightGBM aún no ha sido completamente evaluado.

Capítulo 2

Extracción, limpieza y análisis de datos

2.1. Definición del periodo de análisis

Se estableció un horizonte temporal de análisis y pronóstico que abarca el periodo entre octubre de 2020 y abril de 2024. Este intervalo de tiempo ha sido seleccionado con el objetivo de proporcionar un análisis exhaustivo y detallado de las variables climáticas e hidrológicas en la región, permitiendo una evaluación precisa de las tendencias y cambios en el entorno. A continuación, se detalla la justificación y las características del periodo de estudio definido:

1. Periodo de altas temperaturas y sequías: El año 2022 fue un periodo marcado por altas temperaturas y sequías, lo que afectó significativamente el comportamiento habitual del río Duero. Este fenómeno justificó la necesidad de utilizar un horizonte temporal de tres años para asegurar una evaluación precisa y detallada de las condiciones hidrológicas, considerando tanto eventos anómalos como patrones estacionales regulares.

2. Cobertura completa de ciclos estacionales: Al incluir varias temporadas completas, es posible identificar diferentes condiciones meteorológicas e hidrológicas, desde períodos de lluvias intensas hasta sequías, proporcionando una visión integral del comportamiento climático en la cuenca del Duero.

2.2. Obtención de los datos del caudal (Río Duero en Zamora)

2.2.1. Selección de la estación de aforo más cercana

Las estaciones de aforo son instalaciones diseñadas para medir y registrar el caudal de agua que fluye a través de un punto específico del río. Estas estaciones son cruciales para la gestión y el monitoreo de los recursos hídricos, proporcionando datos vitales sobre la cantidad de agua en los ríos en diferentes momentos del tiempo.

Para determinar la estación de aforo más cercana a la central hidroeléctrica objetivo (Pereruela y San Román), se utilizaron los recursos disponibles en el portal de la Confederación Hidrográfica del Duero (CHD)[18]. Las Confederaciones Hidrográficas

de los grandes ríos de España son organizaciones bajo el Ministerio para la Transición Ecológica y Reto Demográfico, las cuales tienen como objetivo cumplir con la demanda de la sociedad española de ofrecer información de las aguas de sus ríos. Para ello, se pone a disposición una variedad de datos tales como calidad del agua y la localización de las diferentes estaciones de medición.

En el caso de la Confederación Hidrográfica del Duero, se tiene a disposición pública:

- Datos en tiempo real e histórico de las estaciones de aforo, embalses y estaciones pluviométricas, así como registros de la calidad del agua en los afluentes de la cuenca.
- Situación en tiempo real de los embalses que competen a la zona del Duero.
- Petición "a la carta" de datos a través del contacto con la CHD.

Se utilizaron las librerías de extracción automatizada de datos, `beautifulsoup` y `requests`, para obtener la lista de 178 estaciones de aforo de la CHD. La información recopilada de estas estaciones incluye datos geográficos como las coordenadas UTM, datos en tiempo real referentes al caudal y datos históricos contenidos en ficheros que abarcan los últimos tres años.

Una vez localizadas las estaciones de aforo, se calcularon las distancias entre la central hidroeléctrica y cada estación de medición del caudal. Para ello, las coordenadas UTM extraídas se reformatearon a unidades de grados, minutos y fracción de segundos. Con las coordenadas estandarizadas, se efectuó el cálculo de las distancias entre ambos puntos; los factores esenciales para la elección correspondieron a cercanía y ubicación respecto al afluente, es decir, que la estación de aforo estuviese situada aguas arriba y que la distancia entre esta y la central hidroeléctrica fuese la mínima posible.

Teniendo en cuenta la elección de la central hidroeléctrica de referencia (Pereruela y San Román), la estación de aforo que cumplió los requisitos correspondió a la del "Duero en Zamora" con el código 2121.

2.2.2. Extracción de la información

En el portal de la Confederación Hidrográfica del Duero (CHD) se disponen de dos tipos de fuentes de datos para la obtención del caudal proveniente de la estación de aforo:

- Estáticas: Corresponden a ficheros `.csv`, disponibles para descargar en la página web, que contienen la información histórica del caudal en metros cúbicos por segundo desde octubre de 2020 hasta septiembre de 2023 (último año hidrológico registrado). Los datos están disponibles en formato horario.
- Dinámicas: Son datos en tiempo real, en formato horario, que proporcionan información actualizada sobre el nivel del caudal (metros cúbicos por segundo), cubriendo los últimos cuatro meses.

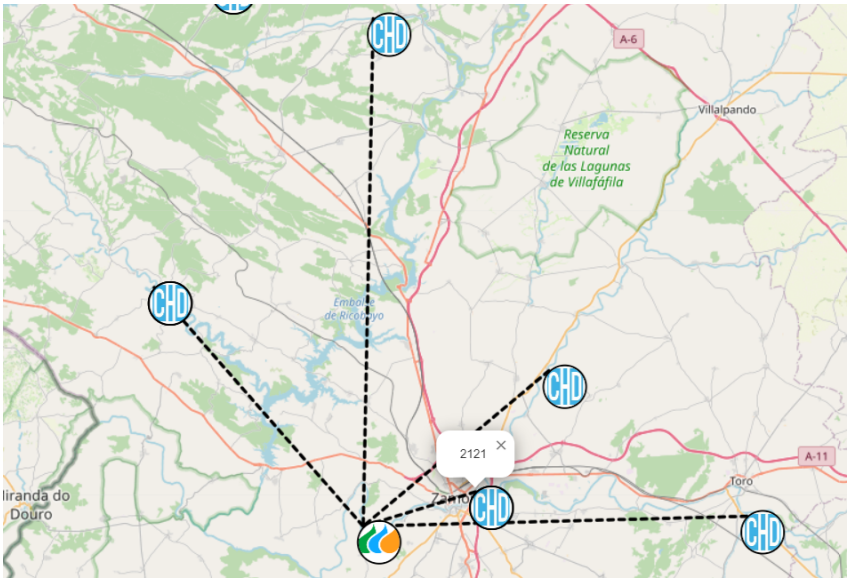

Figura 2.1: Proximidad de la estación de aforo (Zamora) a la central hidroeléctrica de Pereruela y San Román.

Existe una brecha importante entre los datos históricos y los datos en tiempo real. Esta información faltante se debe a que estos últimos no se almacenan en el portal web y, por lo tanto, no están disponibles para su consulta retrospectiva, con lo cual, se realizó la solicitud formal de la información remanente ante el servicio de atención al ciudadano de la Confederación Hidrográfica del Duero.

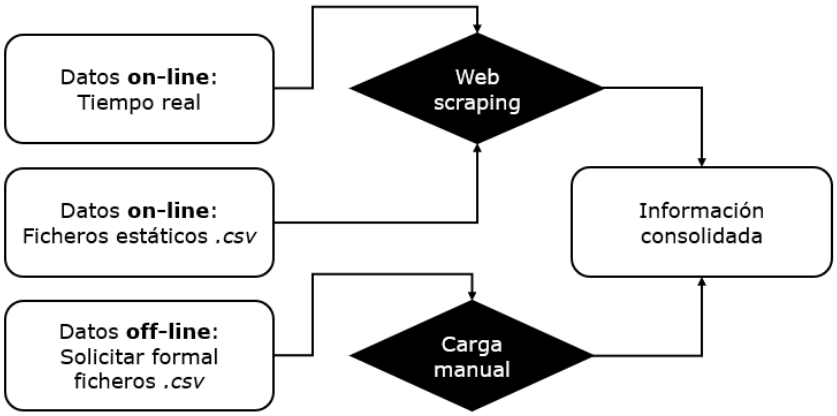

Figura 2.2: Diagrama del flujo de extracción de la información del caudal del Río Duero en Zamora.

Para la obtención directa de ficheros `.csv` (fuente estática) desde la página web, se diseñaron URLs dinámicas para variar el año de consulta (2020, 2021, 2022, 2023), posteriormente se emplearon módulos específicos de *web scraping* para su extracción:

A continuación, se detalla el proceso y las herramientas empleadas:

- Descarga de ficheros `.csv`:
  - Librería `requests`: Realiza solicitudes HTTP para descargar los archivos

`.csv` desde las URLs especificadas. Esta biblioteca se encarga de manejar la comunicación con el servidor y obtener los datos necesarios.

- Lectura y procesamiento de los datos:
  - Librería `io StringIO`: Convierte el contenido descargado en un objeto en memoria que se puede tratar como un archivo. Esto facilita la lectura del contenido directamente en un *DataFrame* de `pandas`.
  - Librería `pandas`: Lee el contenido del archivo CSV desde el objeto StringIO y lo convierte en un *DataFrame*. Posteriormente, se realiza la limpieza y manipulación de los datos para asegurar su formato y calidad.

Los datos en tiempo real están contenidos en tablas desplegadas en la interfaz de usuario del portal de la CHD. Para su extracción se utilizaron módulos de *web scraping* para interactuar directamente con la página web y extraer la información correspondiente. El proceso se describe a continuación:

1. Configuración del navegador: Utilizando *web driver* de `Google Chrome`, se configura y lanza una instancia del navegador controlada por `Selenium`.
2. Uso de la librería `Selenium`: Configura y controla el navegador para acceder a la página web de la estación, interactuar con elementos de la página y extraer los datos necesarios. Específicamente, los datos se encuentran en una tabla en la página web, y Selenium se emplea para navegar a la sección correcta de la página y seleccionar todos los registros disponibles.
3. Uso de la librería `BeautifulSoup`: Parsea el HTML extraído por `Selenium` para estructurar y extraer los datos en un formato legible y manejable. `BeautifulSoup` facilita la extracción de datos específicos desde la tabla HTML, permitiendo convertir estos datos en un *DataFrame*.

2.3. Obtención de los datos de las variables explicativas

Para el pronóstico del caudal, se emplearon variables meteorológicas debido a su impacto en el ciclo hidrológico. Las condiciones atmosféricas determinan la cantidad de agua que cae en una cuenca, influyendo en los flujos hacia ríos y arroyos. Comprender los patrones climáticos permite prever con mayor precisión las fluctuaciones en el caudal, ya que el clima afecta cuándo y cuánto del agua almacenada ingresa al sistema fluvial. Por ejemplo, en regiones con nieve, las condiciones climáticas influyen en el derretimiento de la nieve y, en regiones sin nieve, afectan la evaporación. Además, las corrientes atmosféricas pueden transportar humedad, alterando las precipitaciones y la evapotranspiración, lo que modifica los caudales. Las condiciones atmosféricas también influyen en las tormentas, que pueden causar variaciones significativas en el caudal.

Todos estos factores permiten adquirir un mayor nivel de interpretabilidad en los resultados de las estimaciones y desembocar en proyecciones más precisas.

2.3.1. Fuentes de datos evaluadas y proceso de extracción

La recolección de variables explicativas para la estimación de modelos meteorológicos se llevó a cabo a través de Open-Meteo[19], una API gratuita y de código abierto. Esta plataforma ofrece datos meteorológicos actualizados y en tiempo real para cualquier ubicación, diseñada para entregar un extenso rango de variables climáticas, facilitando así análisis detallados y mejoras en las predicciones meteorológicas.

Open-Meteo integra datos de múltiples fuentes, incluyendo satélites, estaciones terrestres, modelos numéricos de predicción, boyas oceánicas y datos de aeropuertos, lo que garantiza datos precisos y actualizados. Sus principales características incluyen previsión del tiempo detallada, acceso a datos históricos climáticos y generación de alertas meteorológicas en tiempo real. El proceso para la extracción de la información se lleva acabo en tres pasos:

- Configuración del cliente: Se configura un cliente específico que facilita el acceso y manejo de la información meteorológica. En primer lugar, se inicializa una sesión de caché para almacenar las respuestas de la API localmente, lo que reduce el número de solicitudes repetidas y mejora la eficiencia. Además, se implementa una función de reintento para manejar posibles errores de red, asegurando la robustez del proceso.
- Configuració Open-Meteo: Una vez configurada la sesión, se inicializa el cliente de Open-Meteo, que utiliza la sesión previamente creada para realizar las solicitudes a la API. Se definen los parámetros necesarios, como las coordenadas geográficas y las variables meteorológicas definidas para el modelo, y se envía la solicitud a la API. La respuesta recibida contiene los datos meteorológicos en el formato solicitado.
- Extracción de la información: Posteriormente, se extrae la información relevante de la respuesta de la API, incluyendo detalles generales como las coordenadas, la elevación y la zona horaria. Además, se procesan los datos específicos solicitados.

2.3.2. Definición del área y coordenadas para la extracción de la información

El río Duero es uno de los principales ríos de la península ibérica, con una longitud aproximada de 897 kilómetros. Su cuenca hidrográfica abarca alrededor de 97,290 km^2, de los cuales unos 78,859 km^2 están en España y el resto en Portugal. El Duero nace en los Picos de Urbión, en la Sierra de la Demanda, provincia de Soria, y desemboca en el Atlántico, cerca de Oporto, Portugal. Su caudal se incrementa con varios afluentes, entre los que destacan el Pisuerga y el Esla al norte, así como el Adaja y el Tormes al sur.

Se procedió a definir un área geográfica que abarca 145 kilómetros al este, 90 kilómetros al norte y 25 kilómetros al sur, tomando como referencia la ubicación de la central de Pereruela y San Román 15 kilómetros al oeste, con la finalidad de cubrir una extensión significativa que proporcionara datos meteorológicos relevantes sobre el río Duero en la provincia de Zamora. Esta delimitación estratégica no solo incluye una porción considerable del trayecto del río desde su origen hasta donde finaliza su recorrido, sino que también

toma en cuenta la influencia de importantes afluentes, tales como el río Pisuerga, que desciende desde el norte, y el río Eresma, que fluye desde el sur.

Figura 2.3: Mapa hidrográfico de la cuenca del Duero [20].

Con el área de estudio claramente definida, el siguiente paso consistió en la extracción de las coordenadas geográficas de las estaciones meteorológicas gestionadas por la Agencia Estatal de Meteorología (AEMET)[21] dentro de los límites establecidos. Esto con el fin de garatinzar que la elección de los puntos geográficos se basara en la disponibilidad de datos meteorológicos representativos. A continuación, se muestra el área seleccionada y sus elementos principales:

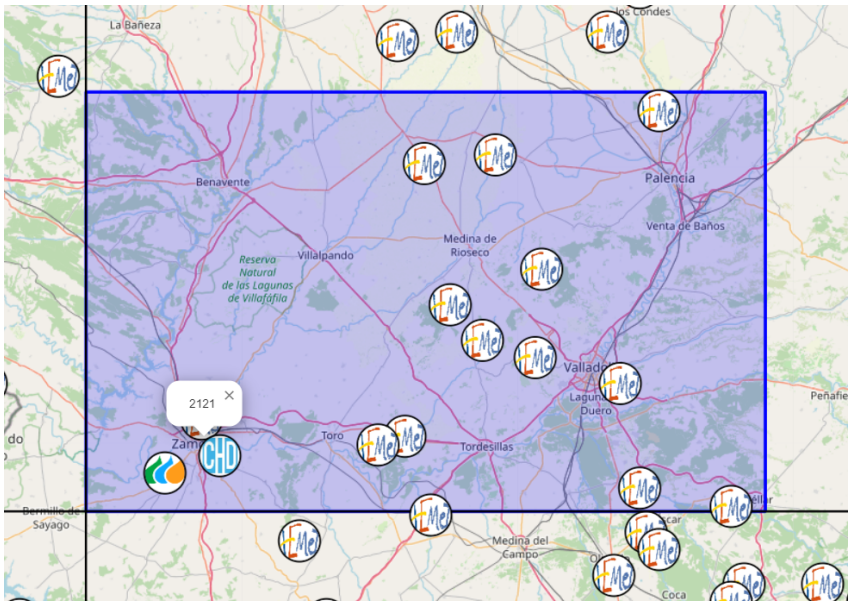

Figura 2.4: Estaciones de AEMET en el área geográfica de estudio (Río Duero).

Finalmente, se realizó la extracción de la información meteorológica a través de la API de Open-Meteo para todas las coordenadas de las estaciones definidas en el área de estudio. En total, se consideraron 14 estaciones meteorológicas, cuidadosamente seleccionadas según los criterios previamente establecidos.

2.4. Limpieza de los datos

Para garantizar la integridad de las series temporales se estableció el rango completo de fechas esperado en el conjunto de datos entre octubre del 2020 y abril de 2024. Este rango actuó como referencia para asegurar que todas las posibles fechas estuviesen representadas. La inclusión de todos los puntos temporales es crucial para evitar brechas en la serie, que podrían comprometer la calidad del análisis posterior e inconsistencias en la estimación del modelo. Luego, se efectúo la comparación de los datos obtenidos con el rango completo de fechas. Durante esta fase, se identificaron las fechas ausentes en el conjunto de datos original. Finalmente, se procedió a completar el conjunto de datos, por medio de una interpolación lineal. En ningún caso, los datos faltantes superaron el 0,1 % de la totalidad de los datos.

Para los datos meteorológicos, además, fue necesario realizar un ajuste en las fechas a la zona horaria de 'Europe/Madrid' dado que la API maneja formatos de fecha estándar, que en ocasiones no están directamente ajustados a la zona horaria de la ubicación consultada, España en este caso.

2.5. Análisis estadístico del caudal del río Duero en Zamora

2.5.1. Análisis de la serie temporal

El análisis del caudal del río Duero en Zamora es fundamental para comprender su comportamiento hidrológico y los patrones relevantes que luego se reflejan en los parámetros e inferencia del modelo.

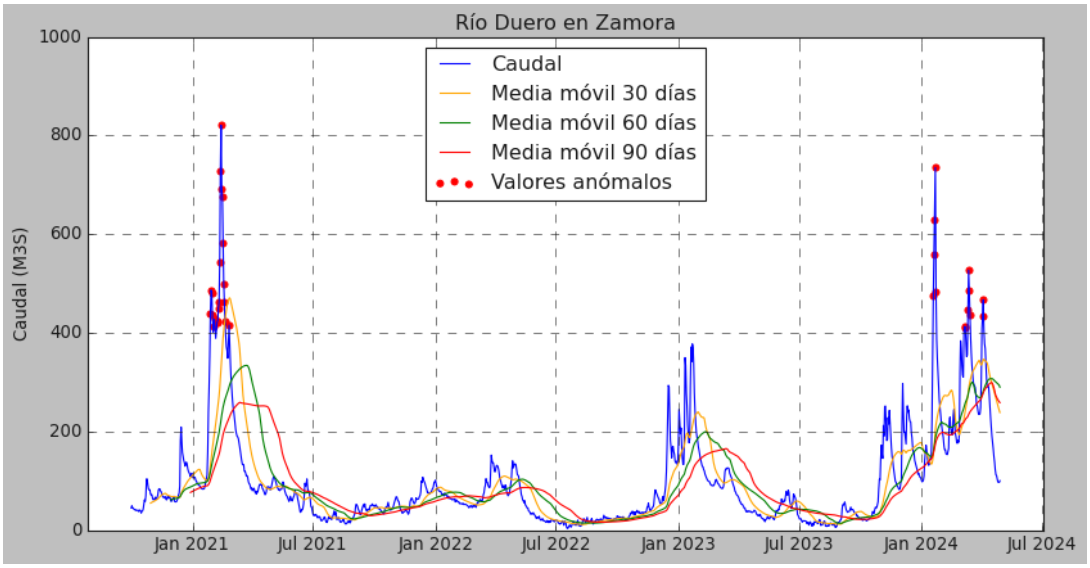

Figura 2.5: Análisis de tendencia del caudal del río Duero.

- La serie del caudal, representada en azul, exhibe varias fluctuaciones significativas. Se observan picos prominentes a finales de 2020 y principios de 2021, con un valor máximo que supera los 800 metros cúbicos por segundo, seguido de una disminución gradual. A lo largo de 2021 y 2022, el caudal permanece relativamente bajo, con algunas elevaciones moderadas.

- A finales de 2022 y principios de 2023, se observa un nuevo aumento significativo del caudal, alcanzando valores cercanos a los 400 metros cúbicos por segundo. Posteriormente, durante el resto de 2023 y principios de 2024, hay varias fluctuaciones con picos notables, pero de menor magnitud en comparación con los registrados a finales de 2020. Este comportamiento es explicado en gran medida por la estación de inverno, meses para los cuales, el nivel de lluvias es mucho más elevado.

- Las medias móviles, representadas en líneas de colores (amarillo para 30 días, verde para 60 días y rojo para 90 días), suavizan estas fluctuaciones y permiten observar tendencias más estables a lo largo del tiempo. Las medias móviles de 30 y 60 días muestran una mayor sensibilidad a los cambios bruscos en el caudal, mientras que la media móvil de 90 días proporciona una visión más suavizada y de largo plazo, resaltando las tendencias generales del flujo del río.

- Respecto a los datos anómalos, se ha decidido no eliminar o reemplazar estos puntos de la serie, dado que, representan los aumentos estacionales del caudal durante el invierno. Eliminar estos picos de la serie temporal podría conducir a una comprensión incompleta o errónea del comportamiento del caudal, especialmente durante estos períodos críticos. En cambio, al conservar y estudiar estas variaciones, el objetivo es capturar y predecir con mayor precisión los patrones estacionales y los eventos extremos que son vitales para la gestión eficaz del nivel del caudal y la planificación a largo plazo.

2.5.2. Patrones estacionales

A continuación, se expone una descripción detallada del caudal analizado a través de diferentes períodos de tiempo. Se incluyen cuatro gráficos de caja (*boxplot*) que representan la variabilidad y la distribución del caudal según el día del mes, el día de la semana, la semana del año y el mes del año.

1. Variabilidad diaria: La variabilidad del caudal es mayor en los días intermedios del mes, con valores máximos que superan los 800 m^3/s en algunas ocasiones. A medida que avanza el mes, la dispersión disminuye, indicando una estabilización del caudal. Esta variabilidad puede estar influenciada por eventos de precipitación que tienden a ser más frecuentes a mediados del mes, así como por la gestión de embalses y la liberación controlada de agua.

2. Fluctuaciones semanales: A lo largo de la semana, el caudal muestra fluctuaciones, pero la variabilidad no es tan pronunciada. Los valores medianos se mantienen relativamente constantes, con algunos picos y valores atípicos. Esto sugiere que el caudal no está significativamente influenciado por los días de la semana, pero podría

reflejar patrones de uso del agua para riego agrícola o actividades industriales que siguen un ciclo semanal.

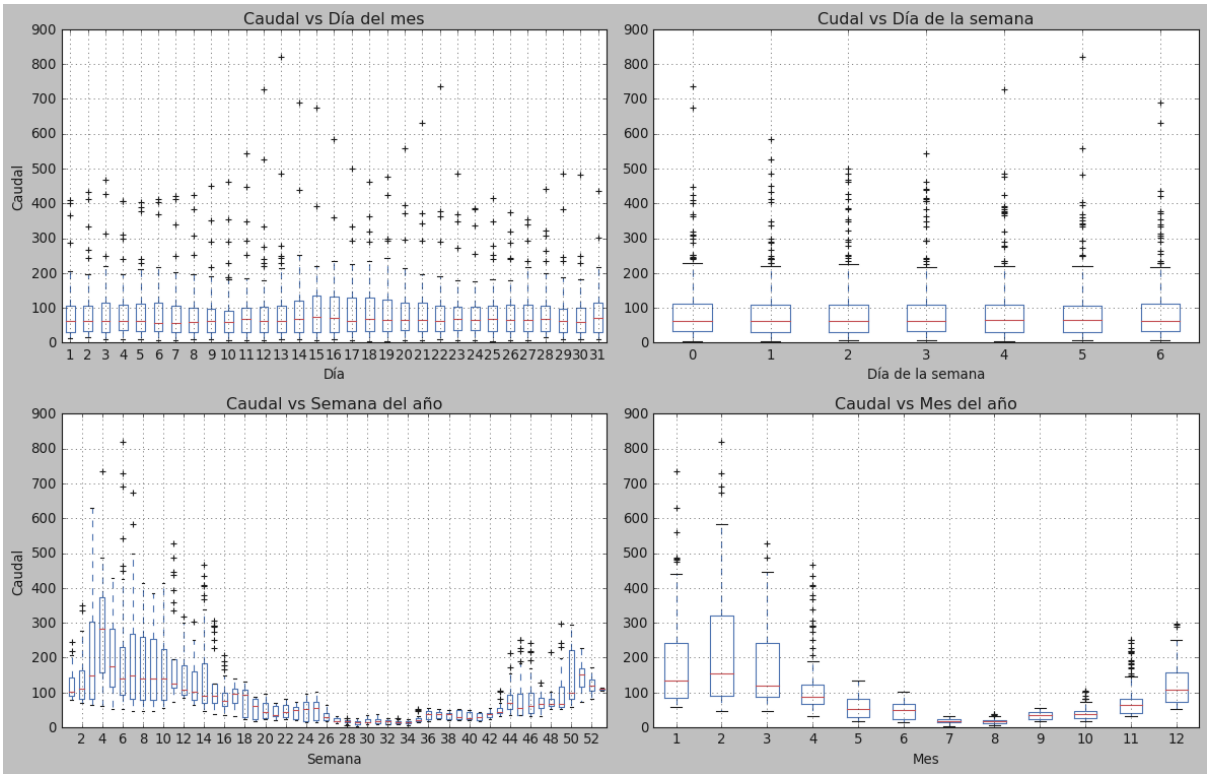

Figura 2.6: Análisis de estacionalidad del caudal del río Duero.

3. Patrón estacional: Existe un patrón estacional claro, con un aumento significativo del caudal durante las primeras semanas del año, alcanzando picos de más de 800 m^3/s. El caudal disminuye considerablemente durante los meses de verano, manteniéndose bajo hasta finales de año, con una ligera recuperación hacia el final del periodo. Este comportamiento puede explicarse por las precipitaciones invernales y el deshielo en las montañas que alimentan el río durante los primeros meses del año, mientras que los meses de verano, más secos, reducen significativamente el caudal.

2.5.3. Propiedad de estacionariedad en la serie temporal

Una serie de tiempo es estacionaria si sus propiedades estadísticas, como la media, la varianza y la autocorrelación, son constantes a lo largo del tiempo. Esto significa que, los patrones en la serie temporal no dependen del momento en el que se observan. La estacionariedad es una propiedad crucial porque muchos modelos de series temporales, como ARIMA, que requieren que los datos sean estacionarios para funcionar y obtener un resultado óptimo generalizable.

El Test de Dickey-Fuller Aumentado (ADF) se utiliza para verificar la estacionariedad de una serie temporal. Los resultados del test ADF para la serie del caudal del río Duero en Zamora, muestran que el estadístico de prueba ADF es -3.833704, con un valor p de 0.002581. Dado que el valor p es menor que los niveles de significancia comúnmente

ADF Test	
ADF test statistic	-3.8337
p-value	0.0026
# lags usados	10
# observaciones	1295
Valor crítico (1 %)	-3.4354
Valor crítico (5 %)	-2.8638
Valor crítico (10 %)	-2.5680

Cuadro 2.1: Resultados del test ADF

utilizados (0.01, 0.05, 0.10), esto indica que la hipótesis nula de que la serie no es estacionaria puede ser rechazada con alta confianza.

Además, se utilizaron 10 retardos en la prueba para capturar la autocorrelación de la serie temporal, y se analizaron 1295 observaciones. Los valores críticos para los niveles de significancia del 1 %, 5 % y 10 % son -3.43541, -2.863774 y -2.56796, respectivamente. El estadístico de prueba ADF es menor (más negativa) que todos los valores críticos, lo que proporciona una fuerte evidencia contra la hipótesis nula de no estacionariedad.

2.6. Análisis descriptivo de las variables meteorológicas

Se han recopilado diversas variables meteorológicas, a través de la API de OpenMeteo, esenciales para el análisis detallado de las condiciones atmosféricas. Estas variables permiten una comprensión profunda del comportamiento del clima y sus fluctuaciones a lo largo del tiempo. La precisión en la extracción y el seguimiento de estos atributos es crucial para la elaboración de pronósticos hidrológicos confiables y para el desarrollo de modelos predictivos precisos.

2.6.1. Integridad de la información

Para identificar anomalías en las series de tiempo, se utilizó una metodología estadística que involucró el cálculo de la media (μ) y la desviación estándar (σ) a partir de los datos disponibles. Se definieron como anomalías aquellos valores que exceden tres veces la desviación estándar por encima o por debajo de la media ($\mu \pm 3\sigma$). Del mismo modo, se evaluó el coeficiente de variación (CV), que corresponde a una medida de dispersión relativa calculada como el cociente entre la desviación estándar y la media, expresado generalmente como un porcentaje. Indica la magnitud de la variabilidad en relación con el tamaño de la media.

Variable	Media	Desv. Est.	CV	Nulos	Outliers
Temp. a 2m sobre el suelo	12,54	7,35	0,59	-	-
Humedad relativa a 2m	67,1	17,47	0,26	-	-
Punto de rocío a 2m	5,1	4,31	0,84	-	2
Temp. aparente	9,83	8,21	0,83	-	-
Cantidad de lluvia	0,04	0,11	2,95	-	31
Cantidad de nieve	0	0,01	19,67	-	3
Profundidad de la nieve	0	0	8,32	-	15
Presión a nivel del mar	1018,87	7,13	0,01	-	9
Presión en la superficie	932,38	6,27	0,01	-	15
Cobertura de nubes	38,15	29,31	0,77	-	-
Cobertura de nubes bajas	21,06	25,67	1,22	-	3
Cobertura de nubes medias	21,14	24,4	1,15	-	7
Cobertura de nubes altas	35,97	30,78	0,86	-	-
Evapotranspiración (FAO)	0,13	0,09	0,7	-	1
Déficit de presión de vapor	0,74	0,74	0,99	-	20
Vel. del viento a 10m	12,35	5,29	0,43	-	10
Dir. del viento a 10m	168,46	73,68	0,44	-	-
Ráfagas de viento a 10m	25,16	9,74	0,39	-	12
Duración del sol	1365,52	566,14	0,41	-	-
Radiación de onda corta	181,88	97,62	0,54	-	-
Radiación directa	126,22	89,53	0,71	-	-
Radiación difusa	54,93	24,76	0,45	-	8
Irradiación normal directa	226,27	128,5	0,57	-	-
Irradiación global inclinada	181,88	97,62	0,54	-	-
Radiación terrestre	310,35	119,61	0,39	-	-
Rad. de onda corta inst.	181,45	97,8	0,54	-	-
Rad. directa instantánea	126,1	89,5	0,71	-	-
Rad. difusa instantánea	54,64	24,64	0,45	-	9
Irrad. normal directa inst.	223,98	128,18	0,57	-	-
Irrad. global inclinada inst.	180,64	97,64	0,54	-	-
Radiación terrestre inst.	309,41	119,61	0,39	-	-
Temp. del suelo (0 a 7 cm)	13,91	8,34	0,6	-	-
Humedad del suelo (7 a 28 cm)	0,22	0,07	0,33	-	-

Cuadro 2.2: Descripción de las variables meteorológicas

El conjunto de variables explicativas está compuesto por 33 variables meteorológicas. No existen datos faltantes, los coeficientes de variación calculados superan el valor estándar de 0.30 y reflejan una volatilidad importante en el comportamiento de los datos. Esto último en línea con la naturaleza de la información, dado que, el comportamiento de las variables climatológicas está influenciado directamente por eventos extremos, como sequías o periodos de lluvia intensa y, a su vez, las estaciones del año repercuten en el comportamiento anómalo de las variables.

No se consideró adecuado eliminar o agregar algún tipo de tratamiento específico para

los *outliers* encontrados, ya que, se corresponden con periodos altamente volátiles de la serie temporal objetivo (caudal) y agregan valor en los tramos de subidas repentinas, para los cuales se requiere un mayor poder de predictibilidad.

2.6.2. Patrones estacionales y de tendencia

A continuación, se presentan algunas de las principales variables evaluadas, que son una muestra representativa del comportamiento general de las 33 variables meteorológicas tenidas en cuenta para el análisis. Los distintos comportamientos se podrían distinguir en tres categorías, comportamiento estacional, comportamiento esporádico y comportamiento de variabilidad continua:

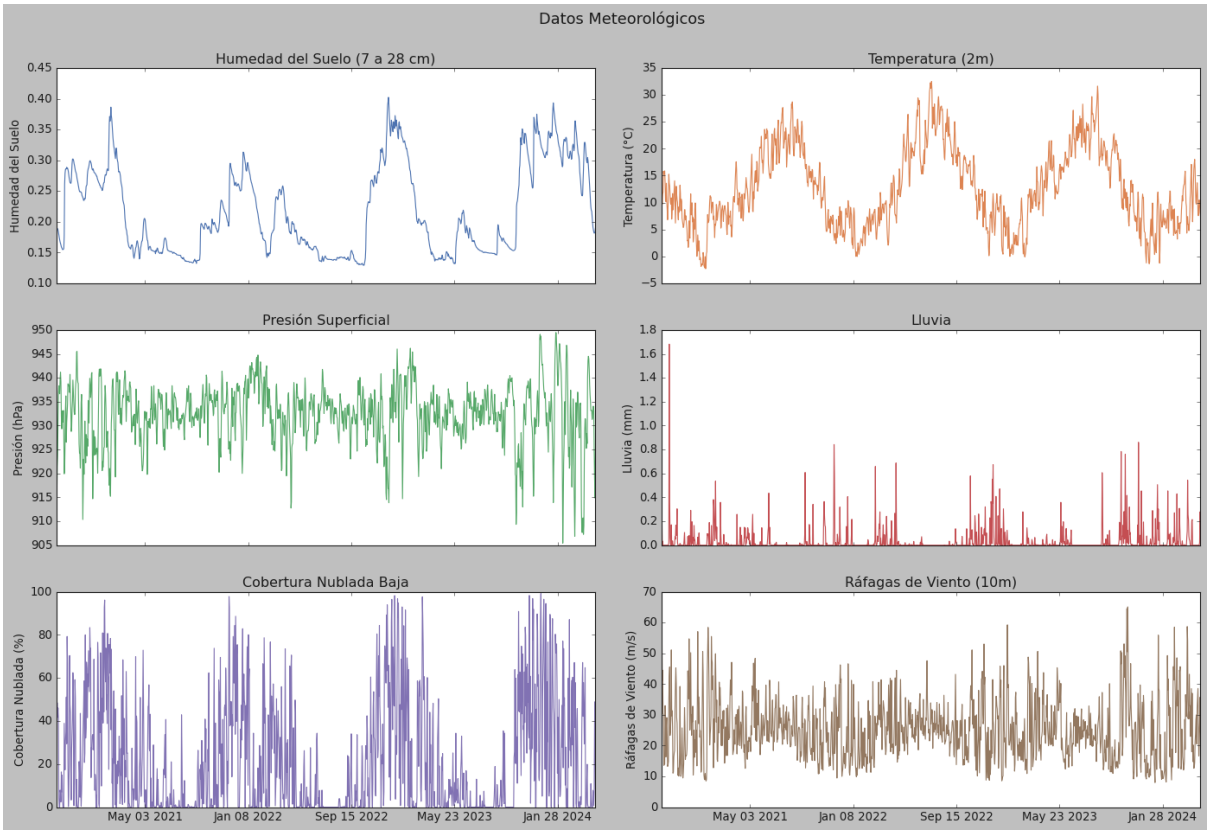

Figura 2.7: Análisis de estacionalidad para las variables meteorológicas.

1. Comportamiento estacional: Este tipo de comportamiento se caracteriza por patrones regulares y repetitivos que coinciden con los ciclos estacionales del año. Por ejemplo, las variables que encajan en esta categoría son la temperatura (2m), que muestra picos y valles regulares reflejando los cambios de estación, y la humedad del suelo (7 a 28 cm), que también presenta variaciones siguiendo un patrón estacional, aunque con mayor variabilidad.

2. Comportamiento estocástico: Las variaciones estocásticas o erráticas son fluctuaciones irregulares que no siguen un patrón predecible y parecen ser aleatorias. Algunas variables representativas de esta categoría son la presión superficial (*surface*

pressure), que tiene una banda de variación definida, pero con fluctuaciones más erráticas sin un patrón estacional claro, y las ráfagas de viento (10m), que muestran variaciones considerables y aleatorias sin un patrón estacional evidente.

3. Eventos esporádicos/picos repentinos: Los eventos esporádicos o picos repentinos se caracterizan por ocurrir de manera súbita y sin una regularidad predecible. Por ejemplo, en este tipo de comportamiento, las variables que encajan son la lluvia (*rain*), que presenta eventos de precipitación esporádicos con picos altos en ciertos puntos, y la cobertura de nubes bajas (*cloud cover low*).

2.7. Análisis de correlación

En el análisis de correlación, los umbrales comunes para interpretar la fuerza de las relaciones entre variables son: $|r| < 0{,}3$ se considera una correlación débil, $0{,}3 \leq |r| < 0{,}5$ indica una correlación moderada, $0{,}5 \leq |r| < 0{,}7$ representa una correlación fuerte, y $|r| \geq 0{,}7$ denota una correlación muy fuerte. A continuación se presenta la matriz de correlaciones para aquellas variables meteorológicas que tienen una correlación por lo menos moderada con el caudal (variable objetivo).

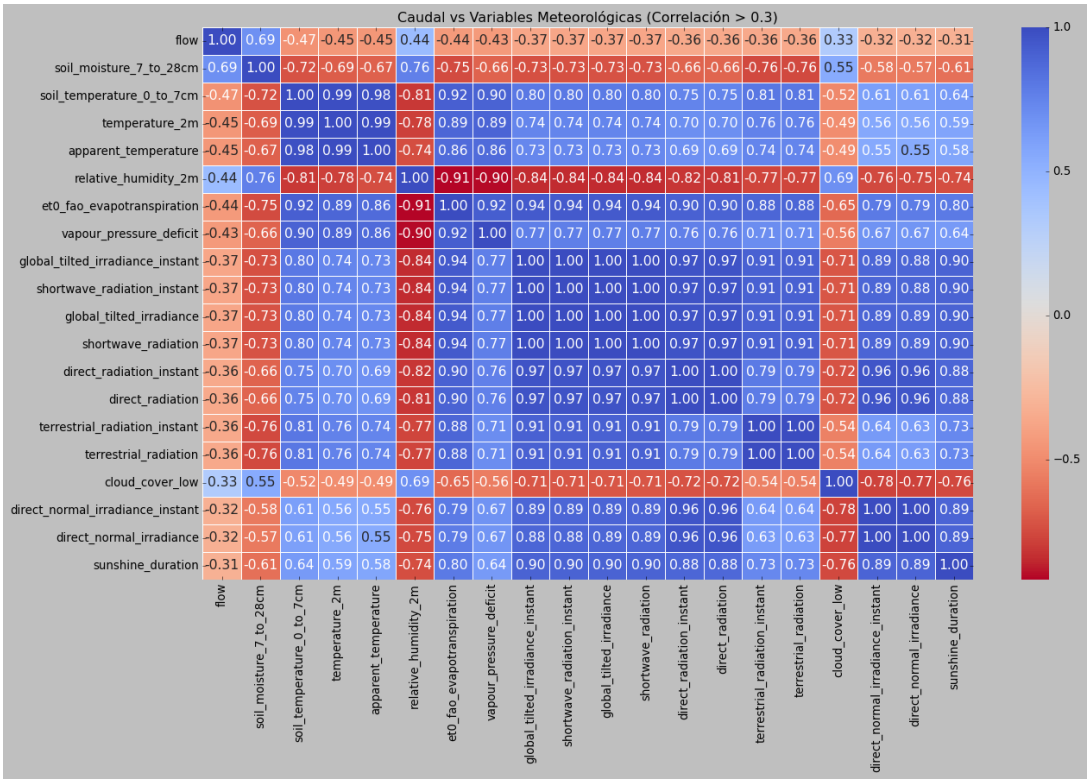

Figura 2.8: Matriz de correlaciones (Variables meteorológicas y caudal del río Duero).

La matriz muestra que *flow* (caudal) del río tiene una fuerte correlación positiva con *soil_moisture_7_to_28cm* (humedad del suelo a 7-28 cm), con un valor de 0.69, indicando que un aumento en la humedad del suelo está asociado con un incremento en el caudal. Por otro lado, *soil_temperature_0_to_7cm* (temperatura del suelo a 0-7 cm), *temperature_2m*

(temperatura a 2 metros), y *apparent_temperature* (temperatura aparente o sensación térmica) tienen correlaciones negativas con *flow* (caudal), con valores de -0.47, -0.45, y -0.45 respectivamente, lo que sugiere que un aumento en estas temperaturas está asociado con una disminución en el caudal. Finalmente, *relative_humidity_2m* (humedad relativa a 2 metros) muestra una correlación positiva moderada de 0.44, indicando que un aumento en la humedad relativa también está relacionado con un mayor caudal del río. Estos valores reflejan cómo diferentes factores climáticos y del suelo pueden influir significativamente en el comportamiento del caudal del río.

Por otro lado, se evidencia una correlación muy alta entre algunas variables del conjunto inicial. Tener variables muy correlacionadas entre sí, conocido como multicolinealidad, puede complicar la estimación de los modelos predictivos y el análisis de datos. La multicolinealidad incrementa las varianzas de los coeficientes estimados, haciendo los resultados menos confiables y difíciles de interpretar.

Capítulo 3

Metodología para la estimación de los modelos

3.1. Definición de los conjuntos de entrenamiento, validación y prueba

Para la construcción de los conjuntos de entrenamiento, validación y prueba de los modelos de pronóstico del caudal del Río Duero en Zamora, se han considerado cuidadosamente las siguientes fechas:

- Conjunto de entrenamiento: Del 31 de octubre de 2020 al 22 de octubre de 2023. Este periodo garantiza un conjunto de datos extenso, abarcando múltiples estaciones del clima y capturando variaciones estacionales y patrones relevantes en la serie temporal.
- Conjunto de validación: Del 23 de octubre de 2023 al 4 de febrero de 2024. Incluye la estación completa de invierno y parte de primavera, permitiendo evaluar y optimizar el modelo en condiciones críticas, asegurando su robustez y precisión.
- Conjunto de prueba: Del 5 de febrero de 2024 al 28 de abril de 2024. Abarca la estación de primavera mayoritariamente, sin embargo, corresponde a periodos recientes de caudal elevado, lo que permite verificar la capacidad del modelo en condiciones exigentes antes de su implementación final.

Figura 3.1: División de la serie temporal para la estimación de los modelos.

3.2. Creación de atributos (*Feature engineering*)

Con el fin de mejorar la capacidad predictiva de los modelos, se llevaron a cabo diversas transformaciones y cálculos sobre los datos originales para extraer y crear nuevas variables que capturaran mejor la información temporal y cíclica inherente a los datos meteorológicos.

1. Creación de retardos y promedios móviles: Para todas las variables explicativas, se calcularon promedios móviles con el objetivo de suavizar las series temporales y captar tendencias a corto y largo plazo. Se generaron valores a intervalos de 7 y 30 días, proporcionando una visión de las tendencias semanales y mensuales. Adicionalmente, se introdujo un retardo temporal para cada variable, permitiendo que los modelos de predicción tengan en cuenta la información del día anterior.

2. Definición de las estaciones del año: Se definieron las estaciones del año basándose en los meses correspondientes: Invierno (diciembre, enero, febrero), primavera (marzo, abril, mayo), verano (junio, julio, agosto) y otoño (septiembre, octubre, noviembre). Esta clasificación estacional permitió que los modelos capturaran las variaciones de estos periodos específicos en las variables meteorológicas.

3. Extracción de variables temporales: A partir de la columna fecha, se extrajeron variables adicionales que incluyen el mes del año, la semana del año y el día de la semana.

4. Cálculo de variables solares: Utilizando la ubicación geográfica específica de Zamora, España, se calcularon las horas de salida y puesta del sol para cada fecha en el conjunto de datos. Esto se realizó con el propósito de incorporar la variabilidad y el número de horas dirario de la luz solar en el minicipio.

5. Codificación cíclica de variables temporales: Para capturar la naturaleza cíclica de ciertas variables temporales, como el mes del año, la estación del año, la semana del año y el día de la semana, se aplicó una codificación cíclica. Esta técnica transforma estas variables en dos componentes (seno y coseno), preservando su periodicidad y facilitando la detección de patrones cíclicos por parte de los modelos de aprendizaje automático.

   Dado un conjunto de datos x y una longitud de ciclo L, la transformación se realiza de la siguiente manera:

$$\text{Componente seno:} \quad \sin\left(\frac{2\pi x}{L}\right)$$

$$\text{Componente coseno:} \quad \cos\left(\frac{2\pi x}{L}\right)$$

Donde:

- x es el valor de los datos en el conjunto (por ejemplo, una hora del día, un mes del año).
- L es la longitud del ciclo (por ejemplo, 24 para horas del día, 12 para meses del año).
- sin es la función seno.
- cos es la función coseno.

Estas transformaciones permiten que los datos con naturaleza cíclica sean representados de manera que los modelos de pronóstico puedan capturar mejor las relaciones periódicas.

Finalmente, se ajusta un método de normalización utilizando únicamente los datos de entrenamiento y aplicando las transformaciones de normalización a los conjuntos de datos de entrenamiento, validación y prueba.

3.3. Selección de atributos (*Feature selection*)

El proceso de selección de características es crucial en la construcción de modelos predictivos, ya que permite identificar y retener aquellas variables que tienen una mayor relevancia, mejorando así la precisión del modelo y reduciendo su complejidad. En este caso, se siguieron dos enfoques principales: eliminación de variables altamente correlacionadas y selección de atributos basados en la importancia calculada a través de un modelo de bosque aleatorio (*Random Forest*).

Eliminación de variables altamente correlacionadas:

- Umbral de correlación: Se definió un umbral de correlación alto, en este caso, 0.95, para identificar pares de variables que tienen una relación lineal muy fuerte.
- Cálculo de la matriz de correlación: Se calculó la matriz de correlación para las variables del conjunto de datos, excluyendo la variable objetivo (caudal) y las variables cíclicas.
- Identificación de pares altamente correlacionados: Se iteró sobre la matriz de correlación para encontrar pares de variables cuya correlación absoluta excedía el umbral definido. Luego, se calculó la correlación de cada una con la variable objetivo (caudal) y, finalmente, se seleccionó la variable del par con la mayor correlación absoluta eliminando la remanente.

Posterioremente, se utilizó un modelo de *Random Forest Regressor* con parámetros específicos (100 estimadores y profundidad máxima de 10) para determinar la importancia de las variables. El modelo se entrenó con las variables restantes tras eliminar las altamente

correlacionadas. Luego, se calculó la importancia de cada variable y se seleccionaron los 15 regresores más relevantes según los valores obtenidos por el algoritmo.

Respecto a los atributos más relevantes: Las variables relacionadas con las nubes bajas y la humedad del suelo son las que presentan una mayor importancia, lo que sugiere que estas condiciones meteorológicas juegan un papel crucial en el comportamiento del modelo. El valor atado a la semana del año también se destaca, indicando que las variaciones estacionales son un factor significativo en el análisis.

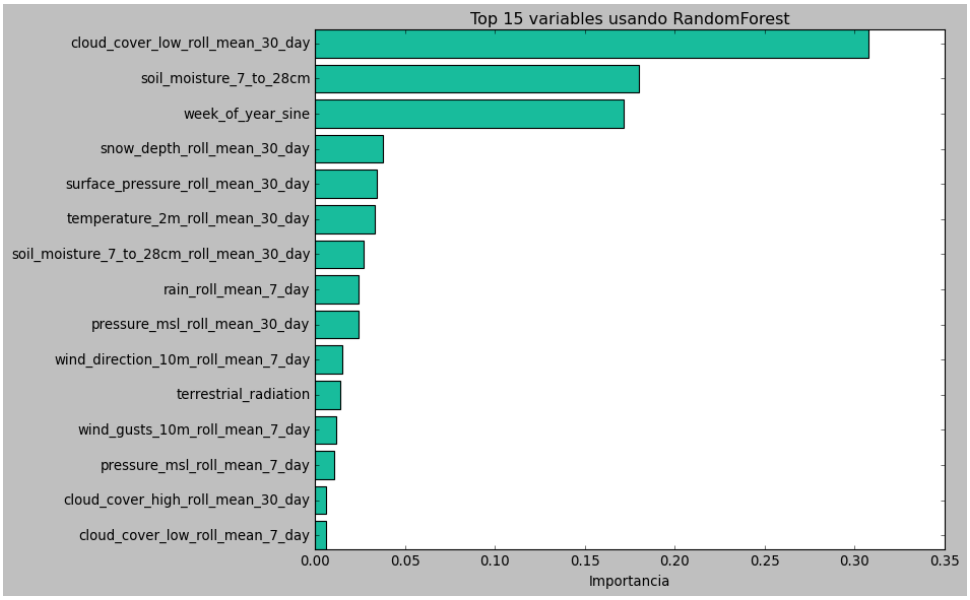

Figura 3.2: Top 15 variables más importantes según *Random Forest Regressor*.

Otras variables de importancia incluyen medidas de presión atmosférica y temperatura a lo largo de periodos de 30 días, así como la dirección del viento y la radiación terrestre. La inclusión de estas variables sugiere que una combinación de factores atmosféricos y temporales es esencial para capturar las dinámicas del caudal del río Duero.

3.4. Definición de los modelos estimados

En este análisis, se realiza una comparativa exhaustiva entre modelos de predicción con enfoque paramétrico y arquitectura tradicional, frente a modelos con enfoque no paramétrico y mayor capacidad de detección de relaciones no lineales. El objetivo de esta comparativa es evaluar y contrastar el desempeño de modelos como SARIMAX con modelos avanzados como XGBoost, LightGBM y finalmente Prophet, en la tarea de predicción del caudal del río Duero en Zamora.

a) El modelo SARIMAX (*Seasonal AutoRegressive Integrated Moving Average with Exogenous regressors*)[22] es una extensión del modelo ARIMA que incorpora componentes estacionales y variables exógenas. Es adecuado para series temporales que presentan patrones de estacionalidad y tendencias influenciadas por factores externos.

La fórmula general de un modelo SARIMAX es:

$$\begin{aligned} Y_t = {} & \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} \\ & + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \\ & + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \cdots + \beta_n X_{n,t} + \epsilon_t \end{aligned} \tag{3.1}$$

donde Y_t es la serie temporal en el tiempo t, ϕ_i son los coeficientes del término autoregresivo, θ_i son los coeficientes del término de media móvil, ϵ_t es el término de error, y $X_{i,t}$ representan las variables exógenas. Además, los componentes estacionales también se modelan mediante términos autoregresivos y de media móvil específicos para la estacionalidad.

b) XGBoost (*Extreme Gradient Boosting*)[23] es un algoritmo de boosting de árboles de decisión optimizado. El boosting es una técnica que combina múltiples modelos débiles para crear un modelo fuerte, mejorando así la precisión. XGBoost utiliza técnicas avanzadas de optimización y regularización para evitar el sobreajuste y mejorar la eficiencia en términos de velocidad y uso de memoria.

   La predicción en XGBoost se define como:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F} \tag{3.2}$$

   donde $\hat{y}_i$ es la predicción para la instancia i, K es el número total de árboles, f_k representa cada árbol de decisión en el modelo, y $\mathcal{F}$ es el espacio de funciones que contienen todos los árboles posibles. XGBoost optimiza esta función utilizando un proceso iterativo basado en el gradiente.

c) LightGBM (*Light Gradient Boosting Machine*)[24] es una implementación de *gradient boosting* que utiliza técnicas basadas en histogramas para acelerar el proceso de entrenamiento y reducir el uso de memoria. Esta técnica agrupa los valores continuos en *bins* discretos, lo que reduce el tiempo de entrenamiento y mejora la eficiencia. Se enfoca en manejar grandes volúmenes de datos.

   La función de predicción en LightGBM se expresa como:

$$\hat{y}_i = \sum_{t=1}^{T} \gamma_t h_t(x_i) \tag{3.3}$$

   donde $\hat{y}_i$ es la predicción para la instancia i, T es el número de iteraciones o árboles, γ_t es el peso asignado al árbol h_t en la iteración t, y $h_t(x_i)$ es la predicción del árbol en la iteración t. LightGBM utiliza técnicas avanzadas de discretización y selección de características para mejorar la eficiencia del modelo.

d) Prophet[25] es un modelo de series temporales desarrollado por Facebook, diseñado para manejar series temporales con fuertes efectos de tendencia, estacionalidad y datos faltantes. Es especialmente útil para datos que presentan cambios repentinos en la tendencia y múltiples ciclos estacionales. Prophet está diseñado para ser fácil de usar y permite un ajuste flexible y automático de los componentes del modelo.

La fórmula general del modelo Prophet es:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{3.4}$$

donde $y(t)$ es el valor de la serie temporal en el tiempo t, $g(t)$ es el modelo de tendencia que puede ser lineal o logístico, $s(t)$ es el componente estacional que captura las variaciones periódicas, $h(t)$ representa los efectos de días festivos o eventos especiales, y ϵ_t es el término de error. Prophet utiliza técnicas de ajuste flexible para modelar las diversas componentes de la serie temporal.

3.5. Entrenamiento de los algoritmos

3.5.1. Modelos de ensamble (XGBoost y LightGBM)

Para el entrenamiento de los modelos de ensamble se ha seguido un proceso metodológico fundamentado en el módulo `skforecast`.

`Skforecast`[26] es una biblioteca de `Python` diseñada para la predicción de series temporales utilizando modelos de aprendizaje automático. Integrándose fácilmente con `scikit-learn`[27], permite el uso de cualquier modelo de regresión para hacer pronósticos, simplificando la creación de características como reatardos (`lags`) y su uso en los modelos. Además, ofrece soporte para validación cruzada específica para series temporales y optimización de hiperparámetros a través de búsqueda de cuadrícula (*GridSearchCV*), búsqueda aleatoria (*RandomizedSearchCV*) o Bayesiana (*BayesianSearch*). También facilita la implementación de *forecasting* directo y recursivo, haciendo más accesible la modelización de series temporales con técnicas modernas de *Machine Learning*.

Los hiperparámetros definidos para los modelos de ensamble son:

- n_estimators: Número de árboles; un número mayor puede mejorar el rendimiento del modelo, pero también aumentar el riesgo de sobreajuste.
- max_depth: Profundidad máxima de los árboles; controla la complejidad del modelo y ayuda a prevenir el sobreajuste.
- learning_rate: Tasa de aprendizaje del modelo; un valor más bajo puede mejorar la precisión, pero requiere más iteraciones.
- subsample y colsample_bytree: Fracción de muestras utilizadas para entrenar cada árbol; ayudan a reducir el sobreajuste.
- gamma, reg_alpha y reg_lambda: Parámetros de regularización que ayudan a evitar el sobreajuste penalizando modelos demasiado complejos.
- max_bin (sólo para LightGBM): Número máximo de bins en los que se agrupan los datos para los histogramas.

La búsqueda de hiperparámetros se llevó a cabo mediante la optimización Bayesiana con `Optuna`, siguiendo un proceso de *backtesting* que entrena el modelo en cada iteración con diferentes combinaciones de hiperparámetros (definidos previamente) y *lags*. Luego evalúa su rendimiento con el conjunto de validación y selecciona la combinación de hiperparámetros que minimiza el error. Porteriormente, el modelo se reentrena con la mejor combinación encontrada utilizando tanto los datos de entrenamiento como los de validación.

La métrica utilizada para optimizar los modelos fue el RMSE (*Root Mean Squared Error*), que penaliza los errores grandes de forma cuadrática, está en la misma escala que la variable objetivo y es más sensible a los errores grandes en comparación con otras métricas como el MAE. Esto permite que el modelo se enfoque en minimizar los errores más significativos, mejorando tanto la interpretación como el rendimiento del modelo en contextos de alta variabilidad como el evidenciado en el caudal del río Duero.

3.5.2. Prophet

El enfoque metodológico seguido para el entrenamiento del modelo Prophet, incluye etapas desde la preparación de los datos hasta la validación cruzada y la evaluación del modelo.

Se aplicó una transformación logarítmica a la variable objetivo, el caudal, para mejorar la precisión y el rendimiento del modelo. Esta técnica contribuye a estabilizar la varianza, reduciendo la heterocedasticidad y haciendo la serie temporal más homogénea. Además, facilita la descomposición de la serie en tendencia, estacionalidad y componentes residuales al transformar la escala a una forma aditiva. También ayuda a normalizar distribuciones sesgadas y reduce el impacto de valores atípicos, lo que resulta en un modelo más robusto y efectivo.

A continuación, se detallan Los hiperparámetros definidos y el impacto asociado ante un cambio en sus valores:

- changepoint_prior_scale: Controla la flexibilidad del modelo para adaptarse a cambios repentinos en la tendencia. Valores más altos permiten que el modelo se ajuste más libremente a estos cambios, mientras que valores más bajos imponen una mayor rigidez.
- seasonality_prior_scale: Determina la fuerza de la estacionalidad en el modelo. Un valor más alto permite capturar variaciones estacionales más pronunciadas, mientras que un valor más bajo reduce la influencia de la estacionalidad.

La selección de estos hiperparámetros se realizó generando todas las combinaciones posibles dentro de un rango predefinido. Esto permite evaluar cómo diferentes configuraciones afectan la capacidad del modelo para capturar patrones en los datos.

En este proceso, se empleó un enfoque de ventana deslizante (*rolling window*) con los periodos iniciales de entrenamiento y validación definidos para evaluar las predicciones

del modelo. Para cada conjunto de hiperparámetros, se entrenó un modelo Prophet y se realizó la validación cruzada, dividiendo los datos en múltiples segmentos y evaluando el rendimiento del modelo en cada segmento. Finalmente, se eligió la configuración de hiperperametros que daba como resultado el RMSE mínimo.

3.5.3. SARIMAX

Los parámetros definidos para la optimización del modelo SARIMAX son de caracter estacional y no estacional.

- p, P: Orden de la parte autorregresiva (AR regular y AR estacional, respectivamente). Representa el número de retardos de la variable dependiente incluidos en el modelo.
- d, D: Orden de diferenciación (diferenciación regular y diferenciación estacional, respectivamente). Indica el número de veces que la serie temporal se diferencia para hacerla estacionaria.
- q, Q: Orden de la media móvil (MA regular y MA estacional, respectivamente). Representa el número de retardos del término de error incluido en el modelo.
- s: Periodicidad estacional, que define la longitud del ciclo estacional (en este caso, 30).

Estos parámetros se establecen dentro de rangos específicos, y se utiliza el producto cartesiano de estos rangos para generar todas las combinaciones posibles. Esto asegura una exploración exhaustiva del espacio de parámetros para identificar la mejor configuración posible.

Una vez definidos los parámetros, se procede a la optimización del modelo SARIMAX. En primer lugar, se itera sobre todas las combinaciones posibles de los hiperparámetros generados, y, para cada combinación, se ajusta un modelo SARIMAX utilizando los datos de entrenamiento y validación. Posteriormente, se calcula el Criterio de Información de Akaike (AIC) para cada modelo ajustado. El AIC es una medida de la calidad del modelo que penaliza tanto el error de ajuste como la complejidad del modelo. La fórmula para el AIC es:

$$\mathrm{AIC} = 2k - 2\ln(L) \tag{3.5}$$

donde k es el número de parámetros en el modelo y L es la función de verosimilitud del modelo. Los resultados, que incluyen los parámetros del modelo y el valor del AIC, se almacenan en una lista para su posterior análisis.

Después de ajustar todos los modelos posibles y calcular sus respectivos AICs, se procede a seleccionar el mejor modelo. Los resultados obtenidos se ordenan por el valor del AIC de forma ascendente, de modo que el modelo con el menor AIC, es decir, el modelo que mejor equilibra el ajuste y la complejidad se corresponda con el algoritmo empleado en la estimación.

3.6. Backtesting

Para llevar a cabo la validación de los modelos en el contexto de un ambiente productivo, se empleó el backtesting con *refit*. Este consiste en la recalibración periódica del modelo utilizando datos más recientes, garantizando que el algoritmo permanezca relevante y preciso a medida que se dispone de nueva información. El *backtesting* evalúa el desempeño de los modelos predictivos mediante la simulación de sus predicciones en datos históricos.

El proceso se realiza de la siguiente manera:

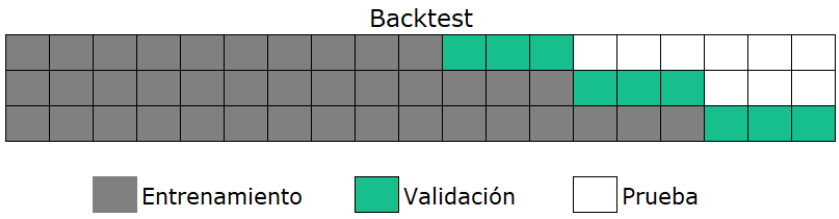

Figura 3.3: Backtesting con *refit*.

1. Definición del modelo: Se selecciona el modelo para realizar el pronóstico y se define el conjunto de datos de la serie temporal. Para este caso XGBoost, LightGBM, Prophet y SARIMAX, junto con los conjuntos de entrenamiento y validación configurados previamente.

2. Establecimiento del horizonte de pronóstico y la estrategia de refit: Se configura el horizonte de pronóstico semanal, y se determina la estrategia de refit, que puede ser en cada paso de tiempo o después de un número definido de pasos, para este caso el reentrenameinto ocurre cada 7 pasos.

3. Realización del backtesting: Se entrena el modelo inicialmente con la ventana de entrenamiento. Luego, para cada paso en la ventana de prueba, se realiza un pronóstico. Al realizar la prueba con refit, después del número predeterminado de pasos (7), se actualiza el modelo reentrenándolo con los datos más recientes, incluyendo los nuevos datos observados.

4. Evaluación del rendimiento: Se evalúa la precisión de los pronósticos utilizando la métrica RMSE. Esta evaluación ayuda a entender cómo se comporta el modelo a lo largo del tiempo y con diferentes configuraciones de *refit*.

Este enfoque garantiza que el modelo se pone a prueba con la información más reciente en cada iteración, lo cual es crucial para adaptarse a las condiciones cambiantes que podrían encontrarse en un entorno productivo. Al implementar esta metodología, se puede observar y evaluar el desempeño del modelo de manera continua y sistemática, lo que permite una comprensión más detallada de su efectividad en un entorno real.

Capítulo 4

Resultados

4.1. Entrenamiento y proceso de optimización

Durante el proceso de optimización de los modelos de pronóstico, se llevaron a cabo múltiples iteraciones con el fin de encontrar la mejor combinación en el espacio de características definido. Las iteraciones totales para cada modelo son las siguientes: 200 para LightGBM y XGBoost, 16 para Prophet, y 80 para SARIMAX. Como se mencionó en la sección anterior, los procesos de optimización empleados fueron diversos: LightGBM y XGBoost se optimizaron mediante una búsqueda Bayesiana de hiperparámetros con `Optuna`, Prophet a través de validación cruzada y SARIMAX mediante la reducción del AIC (Criterio de Información de Akaike).

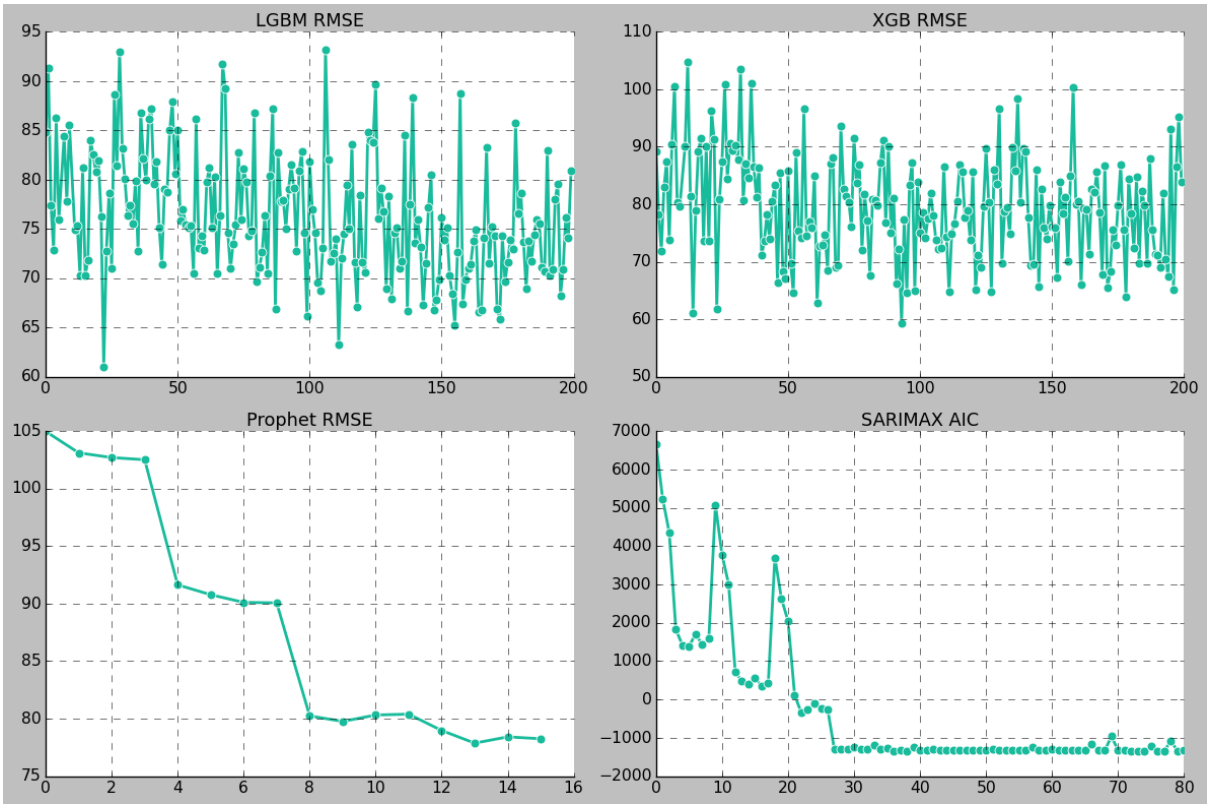

Figura 4.1: Evolución del error durante el entrenamiento.

En cuanto al comportamiento de las métricas de error y ajuste:

- LightGBM: El RMSE varía significativamente entre iteraciones, con valores que oscilan entre 65 y 95. Esta alta variabilidad se debe a la optimización Bayesiana, que explora diversas combinaciones de parámetros, algunas de las cuales pueden no ser ideales, resultando en un rendimiento fluctuante del modelo.

- XGBoost: Similar al LightGBM, el algoritmo XGBoost muestra una variación considerable en el RMSE a lo largo de las 200 iteraciones. Los valores fluctúan entre 60 y 110, indicando que la selección de parámetros tiene un impacto notable en la precisión del modelo. Esta variabilidad también puede atribuirse a la búsqueda bayesiana de parámetros, que mientras explora el espacio de búsqueda, puede seleccionar combinaciones subóptimas en ciertas iteraciones.
- Prophet: En este caso, la RMSE muestra una tendencia decreciente a medida que avanzan las iteraciones, comenzando en 105 y estabilizándose alrededor de 80. Esto sugiere que el proceso de validación cruzada está permitiendo un ajuste progresivo y eficaz del modelo.
- SARIMAX: El gráfico del AIC para el modelo SARIMAX indica una significativa reducción del valor del AIC a lo largo de las 80 iteraciones. Inicia en valores cercanos a 7000 y desciende abruptamente hasta valores alrededor de -1600. Este comportamiento refleja una mejora continua en el ajuste del modelo a medida que se optimizan sus parámetros.

4.2. Hiperparámetros óptimos

4.2.1. Modelos de ensamble (XGBoost y LightGBM)

Los modelos de ensamble, XGBoost y LightGBM, presentaron hiperparámetros que reflejan sus respectivas arquitecturas y la manera en que manejan la regularización y la selección de características.

Parámetro	**XGBoost**	**LightGBM**
n_estimators	1100	1200
max_depth	3	3
learning_rate	0.20	0.43
subsample	0.66	-
colsample_bytree	0.55	-
gamma	0.29	-
reg_alpha	0.99	0.70
reg_lambda	0.46	0.90
max_bin	-	225

Cuadro 4.1: Hiperparámetros óptimos para XGBoost y LightGBM

XGBoost, con un número de estimadores ligeramente inferior, una tasa de aprendizaje más baja y una mayor regularización L1, sugiere un enfoque más conservador y regularizado en comparación con LightGBM. Este último, utiliza una mayor tasa de aprendizaje y una regularización L2 más fuerte, lo que podría indicar una mayor capacidad para ajustar los datos más rápidamente, aunque con un mayor riesgo de sobreajuste. La ausencia de parámetros como `subsample` y `gamma` en LightGBM y la inclusión de `max_bin` resalta diferencias en cómo cada modelo maneja la selección de características y la discretización de valores continuos.

Además, se incluyeron siete retardos de la serie temporal como regresores en el modelo para incorporar el comportamiento pasado de la serie. Este enfoque permite capturar las dinámicas inherentes de la serie temporal, proporcionando una mejor comprensión y predicción de los patrones subyacentes.

4.2.2. Prophet

En el caso del modelo Prophet, los parámetros óptimos obtenidos indican un ajuste muy conservador para los puntos de cambio (`changepoint_prior.scale`) y la estacionalidad (`seasonality_prior.scale`). Esto sugiere que el modelo se beneficia de ser menos sensible a las variaciones abruptas y a los componentes estacionales.

Parámetro	Valor
changepoint_prior_scale	0.001
seasonality_prior_scale	0.01

Cuadro 4.2: Hiperparámetros óptimos para Prophet

Los hiperparámetros indican que el modelo Prophet favorece la estabilidad y la robustez ante cambios repentinos en la serie temporal y fluctuaciones estacionales. El valor bajo de `changepoint_prior.scale` (0.001) sugiere una penalización alta para la detección de puntos de cambio, resultando en un modelo que evita sobreajustarse a cambios bruscos. Por otro lado, el valor relativamente más alto de `seasonality_prior.scale` (0.01) permite una mayor flexibilidad en la captura de patrones estacionales sin comprometer la precisión.

4.2.3. SARIMAX

El modelo SARIMAX obtenido tiene parámetros que indican una estructura autoregresiva y de media móvil tanto para el componente regular como para el componente estacional. En el contexto regular, el modelo incluye dos términos autoregresivos ($AR(2)$) y dos términos de media móvil ($MA(2)$), sugiriendo que las observaciones actuales están influenciadas por los valores y errores de las dos observaciones anteriores. Para la componente estacional, que considera un periodo de 30 unidades de tiempo, también se emplean dos términos autoregresivos ($AR_s(2)$) y dos términos de media móvil ($MA_s(2)$), lo que sugiere una influencia similar pero en un contexto cíclico o repetitivo cada 30 unidades de tiempo.

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \Phi_1 y_{t-30} + \Phi_2 y_{t-60} + \Theta_1 \epsilon_{t-30} + \Theta_2 \epsilon_{t-60} + \epsilon_t \quad (4.1)$$

Es importante resaltar que los órdenes de diferenciación son cero tanto en el componente regular como en el estacional. Esto indica, como se evalúo en secciones anteriores, que la serie de tiempo utilizada (caudal del río Duero) para ajustar el modelo, es estacionaria, es decir, sus propiedades estadísticas como la media y la varianza son constantes a lo largo

del tiempo, eliminando la necesidad de realizar transformaciones adicionales para lograr el cumplimiento de esta propiedad.

4.3. Backtesting

Se ha estimado un modelo de referencia (*baseline*) que consiste en repetir el último valor observado para el horizonte de pronóstico (7 días). Este enfoque sencillo permite establecer una base comparativa para evaluar el rendimiento de los modelos más complejos y determinar, en primera instancia, si logran ser mejores que la estimación *naive*.

Para evaluar los resultados del *backtest* de los modelos, se han empleado las métricas MAPE (*Mean Absolute Percentage Error*), RMSE (*Root Mean Squared Error*) y MSE (*Mean Squared Error*). Además, se ha definido una métrica personalizada para medir qué tan bueno es el modelo pronosticando picos repentinos. Esta métrica, denominada MAE Peaks, aplica el MAE (*Mean Absolute Error*) específicamente a esos picos, proporcionando una evaluación más precisa de la capacidad del modelo para manejar variaciones abruptas en los datos.

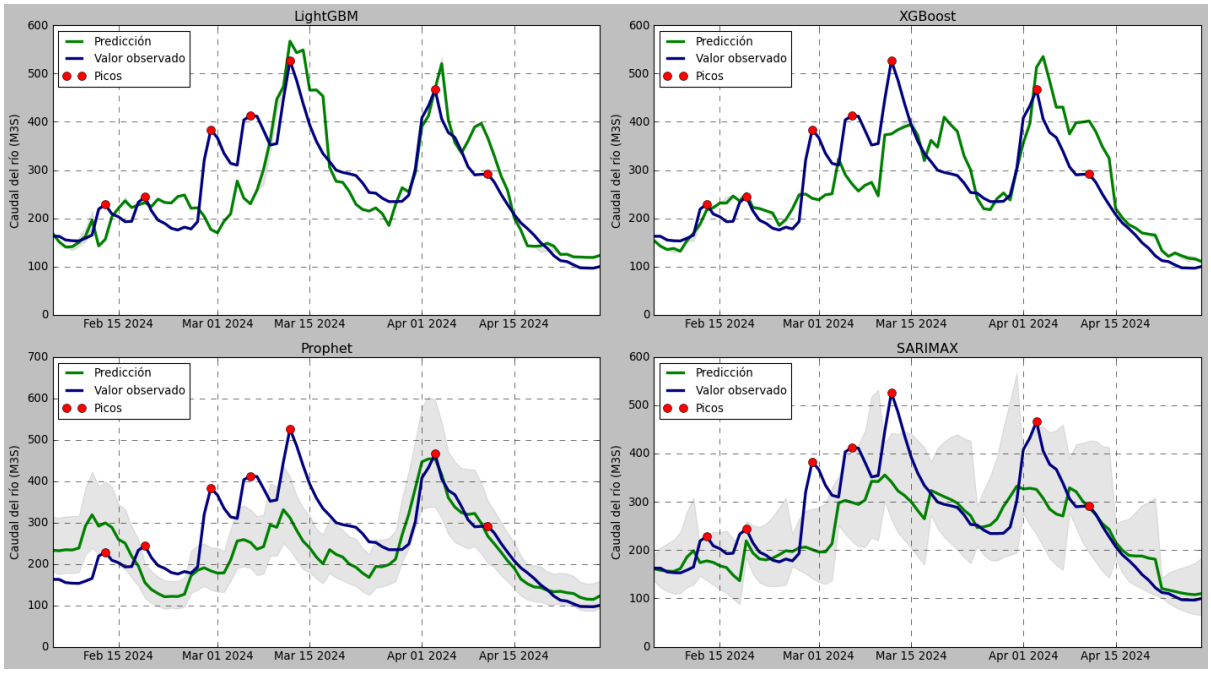

Figura 4.2: *Backtesting* para los algortimos evaluados.

En la métrica de MAE Peaks el modelo LightGBM se destaca por tener el menor error absoluto medio en los picos de la serie temporal, lo que indica que es el más efectivo para predecir valores extremos en los datos. Le sigue de cerca XGBoost, que también muestra un buen rendimiento en la predicción de picos, y SARIMAX, que tiene un rendimiento intermedio. Los modelos *baseline* y Prophet presentan los mayores errores en esta métrica, sugiriendo que son menos efectivos para capturar las variaciones extremas en los datos.

Metric	Baseline	LightGBM	XGBoost	Prophet	SARIMAX
MAE Peaks	108.64	84.48	87.16	110.40	100.15
MAPE	0.21	0.18	0.18	0.26	0.15
MSE	5155.51	4511.77	4332.46	7721.39	4350.58
RMSE	71.80	67.17	65.82	87.87	65.96

Cuadro 4.3: Comparación del desempeño de los modelos.

En cuanto a las métricas de MAPE, RMSE, y MSE, el desempeño de los modelos varía. Para MAPE, el modelo SARIMAX tiene el menor error porcentual absoluto medio, seguido por LightGBM y XGBoost, lo que indica una mayor precisión en términos relativos. El modelo Prophet, por otro lado, tiene el mayor error porcentual, sugiriendo una menor precisión.

Para RMSE y MSE, XGBoost presenta el menor error, seguido muy de cerca por SARIMAX y LightGBM, lo que indica una mayor precisión en la predicción de valores absolutos. En ambas métricas, Prophet nuevamente tiene el mayor error, destacándose como el modelo menos preciso. En resumen, SARIMAX y XGBoost son los modelos con los menores errores y las predicciones más precisas a nivel general, LightGBM es el mejor modelo a nivel de precisión en datos anómalos, mientras que Prophet consistentemente muestra resultados deficientes.

Gráficamente se expone a continuación:

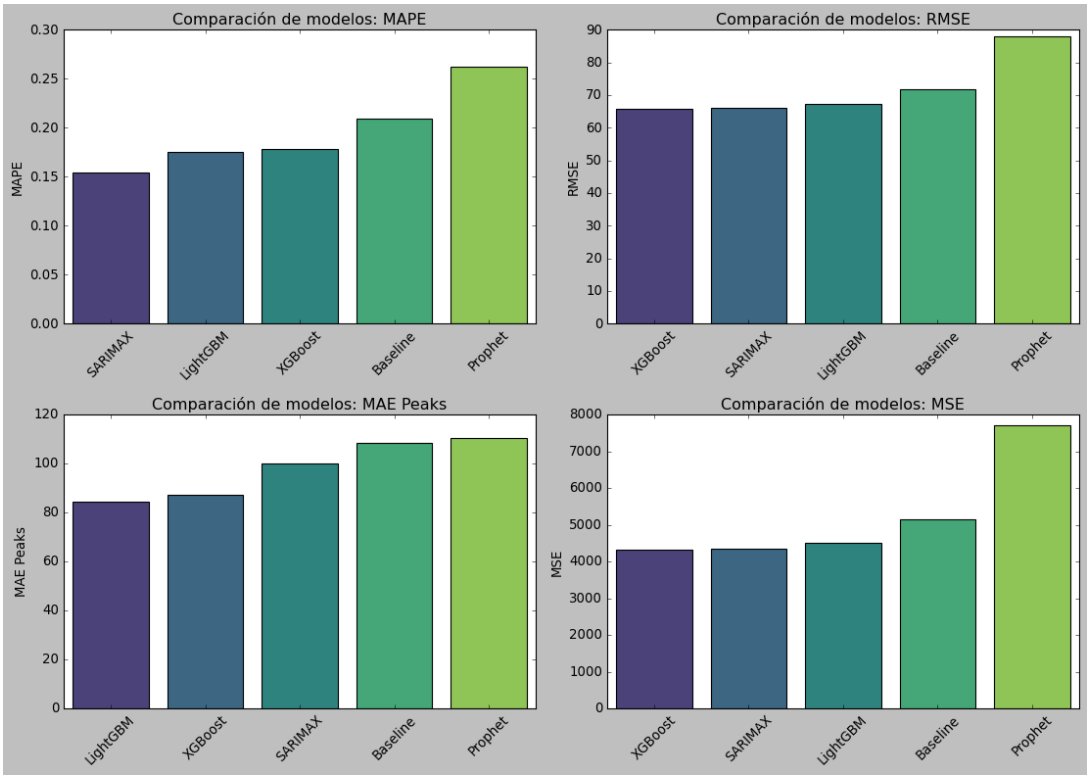

Figura 4.3: Resultados de las métricas de error evaluadas.

4.4. Sobreajuste

Es fundamental comparar la métrica de error obtenida durante el proceso de optimización con la métrica de error del *backtest* para evaluar el sobreajuste en los modelos. El sobreajuste ocurre cuando un modelo tiene un desempeño óptimo en los datos de entrenamiento o validación, pero no generaliza bien a datos no vistos, lo que se puede evidenciar si hay una gran disparidad entre las métricas de error de optimización y de *backtest*. En el contexto del proyecto, se evaluó el RMSE para la mayoría de los modelos, mientras que para SARIMAX se utilizó el AIC (*Akaike Information Criterion*), que aunque no es una métrica de error, proporciona una medida relativa de la calidad del modelo ajustado.

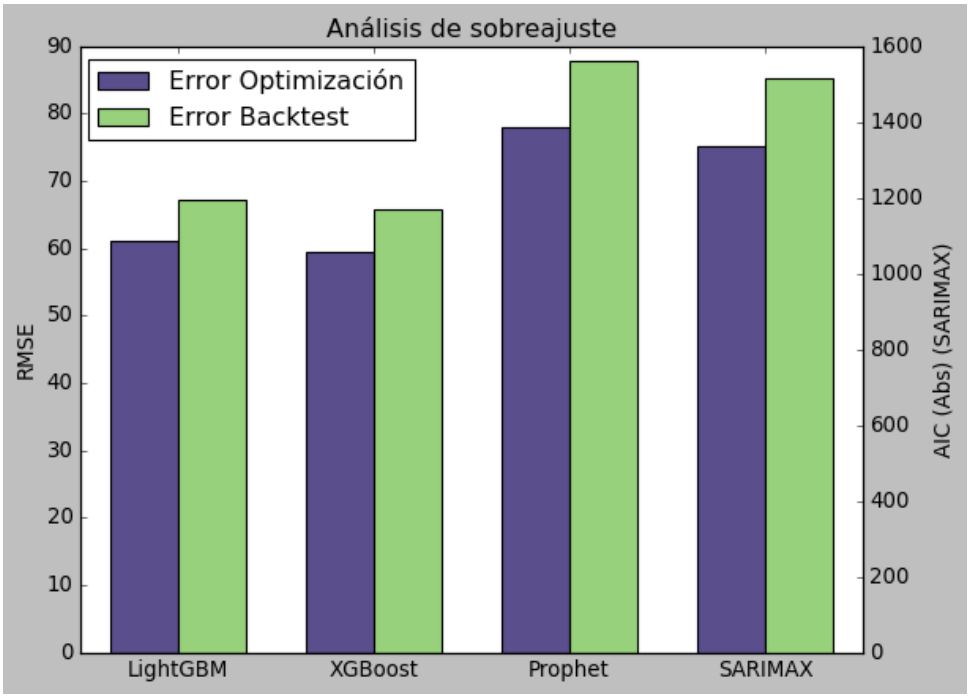

Figura 4.4: Comparación de métricas de error. hiperparametrización y *Backtesting*.

En el caso de LightGBM y XGBoost, los valores de RMSE durante la optimización y el backtest son relativamente cercanos, lo que sugiere un buen rendimiento generalizado de estos modelos. Prophet, sin embargo, muestra un incremento notable en el RMSE durante el backtest en comparación con la optimización, indicando un posible sobreajuste.

Para SARIMAX, se analizó el valor absoluto del AIC para efectos gráficos. Se observa que el valor del criterio estadístico es mayor en el backtest (posee una métrica óptima), sugiriendo que aunque no se basa en una métrica de error tradicional, el modelo final estimado es ligeramente más eficiente que el entrenado.

4.5. Interpretabilidad

Para llevar a cabo un análisis robusto del impacto de las variables explicativas en los modelos, se ha dispuesto de la información propia del algoritmo y sus parametros estimados, como en el caso de Prophet y SARIMAX. Luego, se ha hecho uso de técnicas mas avanzadas para los modelos de ensamble, esto último dada la necesidad de evaluar impactos y relaciones directas o inversas entre los regresores y la variable objetivo.

SHAP (*Shapley Additive Explanations Values*) es una técnica de explicación de modelos en *Machine Learning* que se basa en la teoría de juegos, específicamente en el valor de Shapley. Este valor, originalmente ideado para distribuir equitativamente la ganancia de una coalición entre sus jugadores, se adapta en el contexto de *Machine Learning* para distribuir la predicción de un modelo entre las distintas características de entrada. De esta manera, los valores de SHAP permiten cuantificar la contribución de cada característica a la predicción de un modelo para una instancia particular. La principal ventaja de esta técnica es su capacidad de ofrecer explicaciones consistentes y justas, ya que asegura que las características que contribuyen más a la predicción recibirán mayores valores SHAP.

Los valores de SHAP se calculan de manera que la suma de las contribuciones de todas las características más el valor promedio de la predicción del modelo es igual a la predicción del modelo para esa instancia específica. Esto facilita la interpretación y la identificación de las características que influyen en las decisiones del modelo.

4.5.1. Componentes autorregresivos

A continuación, se evalúa la relevancia de los componentes autorregresivos en los modelos y su impacto en las proyecciones.

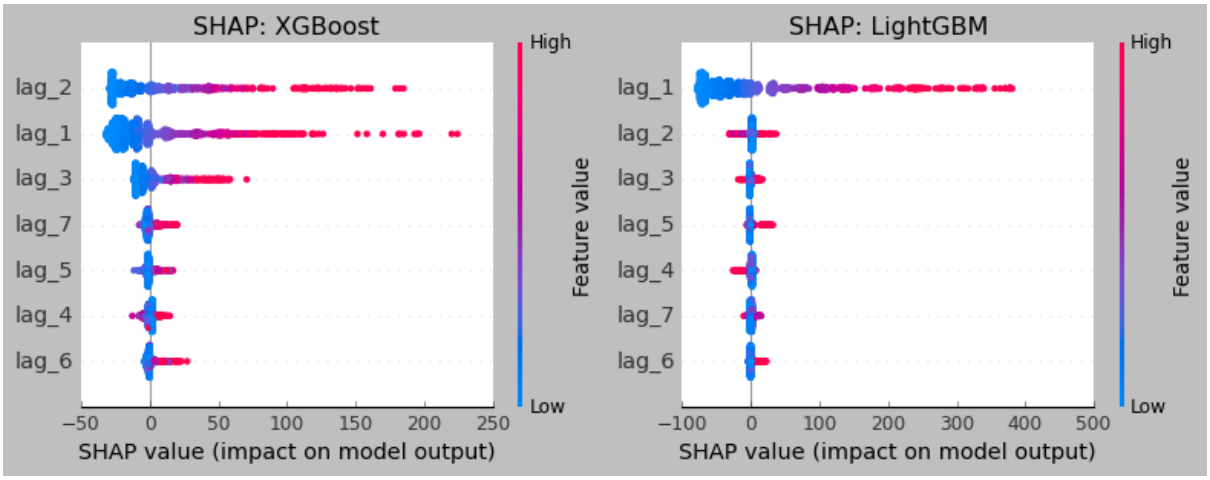

Figura 4.5: Valores SHAP para componentes autorregresivos.

- Modelos XGBoost y LightGBM: Los gráficos de SHAP muestran que los retardos más recientes (`lag_1`, `lag_2` y `lag_3`) son los más importantes para la predicción.

En XGBoost, el retardo `lag_2` tiene un impacto significativo y positivo, seguido de `lag_1` y `lag_3`. Valores altos de `lag_2` y `lag_1` tienden a aumentar el nidel del caudal, mientras que los valores bajos lo disminuyen. En LightGBM, `lag_1` es particularmente destacado, mostrando un patrón similar donde los valores altos de los primeros retardos tienen un impacto positivo. Los retardos mayores (`lag_6` y `lag_7`) tienen menor importancia, indicando que los valores más recientes son más relevantes para ambos modelos.

Variable	Coef	P-valor
ar.L1	1.3865	0.000
ar.L2	-0.3876	0.002
ma.L1	-0.2366	0.056
ma.L2	-0.1947	0.000
ar.S.L30	0.0299	0.953
ar.S.L60	0.9683	0.070
ma.S.L30	-0.0377	0.970
ma.S.L60	-0.9619	0.348
sigma2	0.0168	0.057

Cuadro 4.4: Coeficientes autorregresivos y *p-values* del modelo SARIMAX

- Modelo SARIMAX: Los resultados indican que el primer retardo (`ar.L1`) es altamente significativo (coeficiente: 1.3865, p-valor: 0.000) y positivo, sugiriendo que el valor de la serie en el primer retardo es un fuerte predictor del valor actual. El segundo retardo (`ar.L2`) también es significativo (coeficiente: -0.3876, p-valor: 0.002) pero con un efecto inverso. Otros componentes autorregresivos, como `ar.S.L30` y `ar.S.L60`, no son tan significativos, con p-valores altos, lo que sugiere que los retardos estacionales de mayor alcance no tienen un impacto fuerte en la predicción. Los componentes de promedios móviles (`ma.L1` y `ma.L2`) también influyen significativamente en la predicción, aunque los estacionales (`ma.S.L30` y `ma.S.L60`) no llegan a tener representatividad.

El modelo Prophet no se utilizó en este análisis porque maneja la información de los retardos de manera diferente, enfocándose más en componentes de tendencia y estacionalidad explícita. Esto hace que la interpretación no sea tan clara como en los modelos XGBoost, LightGBM y SARIMAX, los cuales permiten una mejor comprensión del impacto de los retardos específicos en la predicción.

4.5.2. Variables exógenas

Para evaluar los resultados atados a las variables exógenas, se ha examinado nuevamente los valores SHAP (*Shapley Additive Explanations*) para los modelos de ensamblaje. La capacidad de los valores SHAP para descomponer las predicciones en las aportaciones de cada variable nos permite entender mejor las dinámicas subyacentes en los modelos complejos.

Además, hemos evaluado los coeficientes de los modelos Prophet y SARIMAX para determinar el impacto y la importancia de las variables meteorológicas en estas metodologías específicas. Analizando estos coeficientes, es posible identificar cuáles variables tienen una influencia significativa sobre el caudal, permitiéndo inferir su relevancia en la modelización y pronóstico.

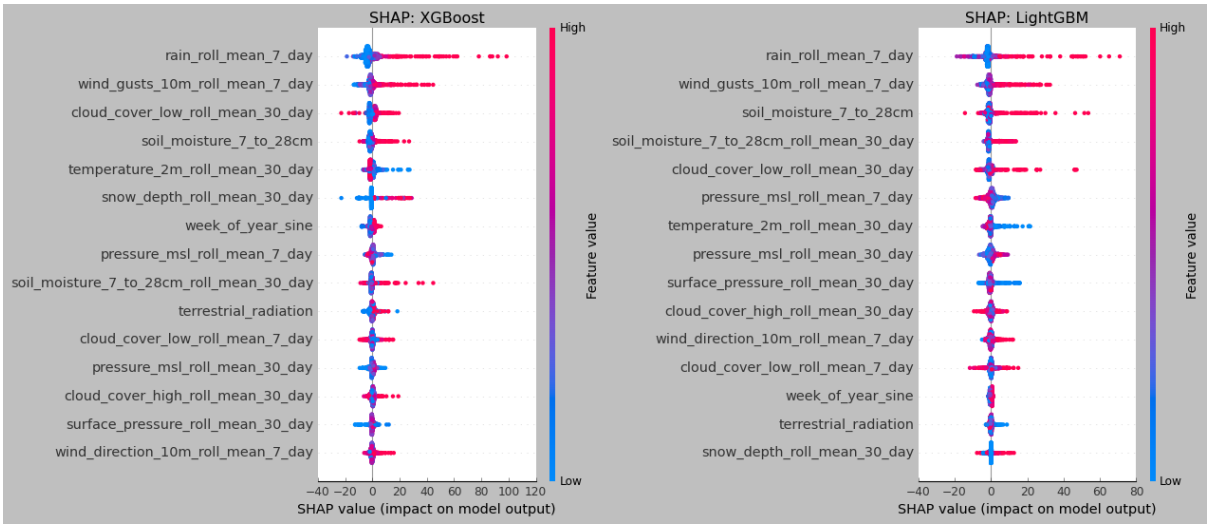

Figura 4.6: Valores SHAP para las variables meteorológicas (Modelos de ensamble).

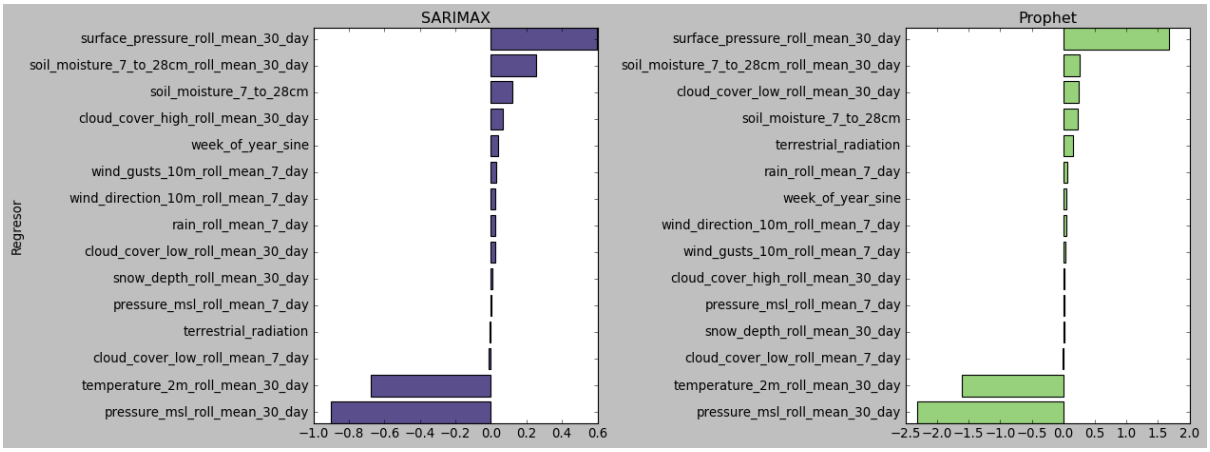

Figura 4.7: Magnitud de los coeficientes para las variables meteorológicas (SARIMAX y Prophet).

- En los modelos de ensamble las variables *rain_roll_mean_7_day*, *wind_gusts_10m_roll_mean_7_day* y *soil_moisture_7_to_28cm* tienen una alta influencia en los pronosticos estimados, como se muestra por sus altos valores SHAP. El impacto o correlación que tienen con la variable objetivo es positiva; un incremento en estos atributos generalmente se traduce en un aumento en el nivel del caudal.

- En los modelos SARIMAX y Prophet, las variable *surface_pressure_roll_mean_30_day* y *soil_moisture_7_to_28cm_roll_mean_30_day* son los atributos con los coeficiente más significativos, reflejando, además, una correlación positiva con la variable objetivo. Por otra parte, las variables *temperature_2m_roll_mean_30_day* y

pressure_msl_roll_mean_30_day destacan como los atributos que mayor influencia tienen en la reducción del nivel del caudal, dado el valor negativo de sus coeficientes.

Finalmente, se calcula la importancia global de las variables en las estimaciones multiplicando su posición de relevancia en cada algoritmo por un peso asignado en función del desempeño del modelo en la métrica personalizada (MEA Peaks). Los pesos otorgados a cada modelo son: LGBM 40 %, XGBoost 30 %, SARIMAX 20 % y Prophet 10 %. Estas asignaciones garantizan que los modelos con mejor *performance* tengan una mayor influencia en la importancia combinada de las variables. Luego, los atributos se ordenan según este factor ponderado, permitiendo identificar cuáles son las variables explicativas más relevantes al considerar las contribuciones individuales de todos los modelos involucrados.

Variable meteorológica	LightGBM	XGBoost	Prophet	SARIMAX	Factor
rain_roll_mean_7_day	1	1	8	10	3.5
soil_moisture_7_to_28cm	3	4	6	5	4.0
wind_gusts_10m_roll_mean_7_day	2	2	11	8	4.1
temperature_2m_roll_mean_30_day	7	5	3	2	5.0
soil_moisture_7_to_28cm_roll_mean_30_day	4	9	4	4	5.5
cloud_cover_low_roll_mean_30_day	5	3	5	11	5.6
pressure_msl_roll_mean_30_day	8	12	1	1	7.1
surface_pressure_roll_mean_30_day	9	14	2	3	8.6
pressure_msl_roll_mean_7_day	6	8	14	15	9.2
week_of_year_sine	13	7	9	7	9.6
cloud_cover_high_roll_mean_30_day	10	13	6	12	10.3
wind_direction_10m_roll_mean_7_day	11	15	9	10	11.7
cloud_cover_low_roll_mean_7_day	12	11	10	12	11.8
snow_depth_roll_mean_30_day	15	6	15	13	11.9
terrestrial_radiation	14	10	7	14	12.1

Cuadro 4.5: Importancia global de las variables meteorológicas

A continuación, se detallan las cinco variables con mayor impacto según el factor de ponderación calculado. Estos atributos no sólo destacan por su aporte cuantitativo sino que también por su contribución en aspectos cualitativos, estos últimos, esenciales para la comprensión de la variabilidad del caudal del río Duero.

1. *rain_roll_mean_7_day*: Este indicador calcula el promedio móvil de siete días de la cantidad de lluvia registrada. La relación entre las precipitaciones y el aumento del caudal fluvial es directa; cuando se registra un aumento en la cantidad de lluvia, se produce una acumulación correspondiente de agua en la superficie. Esta agua se canaliza hacia arroyos y ríos, incrementando su caudal.

2. *soil_moisture_7_to_28cm*: Este índice mide la cantidad de agua absorbida en una capa específica del suelo, entre 7 y 28 cm de profundidad. Cuando el suelo alcanza un alto nivel de humedad, su capacidad para absorber más agua se reduce significativamente, lo que puede resultar en un aumento del flujo superficial. Este exceso de flujo superficial es dirigido hacia los ríos y arroyos cercanos, aumentando su caudal.

3. *wind_gusts_10m_roll_mean_7_day*: Esta medida refleja el promedio móvil de siete días de las ráfagas de viento a 10 metros sobre el suelo. Los vientos fuertes aceleran la evaporación del agua en ríos y suelos, afectando los niveles de caudal. Además, pueden modificar la distribución espacial de las precipitaciones en una región, alterando así los patrones habituales de lluvia y potencialmente impactando en los recursos hídricos disponibles.

4. *temperature_2m_roll_mean_30_day*: Esta variable representa el promedio móvil de treinta días de la temperatura a dos metros sobre el suelo. La temperatura tiene un efecto directo sobre la tasa de evaporación del agua y la transpiración de las plantas. Temperaturas más altas pueden intensificar estas tasas, lo que resulta en una disminución del caudal de los ríos, ya que menos agua permanece disponible para fluir hacia ellos.

5. *cloud_cover_low_roll_mean_30_day*: Este indicador refleja el promedio móvil de treinta días de la cobertura de nubes bajas. Un valor alto en este atributo generalmente se asocia con un aumento en la precipitación debido a la mayor retención de humedad en las nubes, lo que puede llevar a un aumento en el caudal de los ríos. Además, la presencia de nubes también puede reducir la evaporación del agua superficial, contribuyendo así a mantener mayores niveles de agua en el río.

Las cinco variables descritas evidencian una influencia significativa tanto cuantitativa como cualitativa en la variabilidad del caudal del río Duero. Los promedios móviles de la cantidad de lluvia, humedad del suelo, ráfagas de viento, temperatura y cobertura de nubes bajas ofrecen una visión integral de los factores que impactan el flujo de agua en la cuenca. Cada uno de estos atributos interactúa de manera compleja con el entorno, subrayando la importancia de un enfoque multifacético para la gestión del recurso hídrico.

Capítulo 5

Conclusión

El uso de técnicas de *web scraping* en el proyecto, para la extracción de los datos del caudal del río Duero en Zamora, han resultado altamente efectivas. Estas permitieron la recolección, almacenamiento y consolidación eficiente de la información en tiempo real y de los ficheros estáticos disponibles en el portal web. Además, la API de Open-Meteo ha sido esencial para obtener una amplia variedad de datos meteorológicos. La definición de puntos geográficos específicos, basada en la información de las estaciones de AEMET en la región, ha sido crucial para garantizar que los datos meteorológicos proporcionados están alineados con los puntos relevantes para la agencia meteorológica.

El desarrollo avanzado en la ingeniería de características y la selección de regresores ha enriquecido el conjunto de datos base. La inclusión de nuevos atributos, como las medias móviles de las variables meteorológicas, ha sido fundamental para los modelos predictivos. Asimismo, la codificación cíclica ha mejorado la capacidad de los modelos para integrar y procesar la información temporal de manera efectiva. Estos atributos no solo han mejorado la precisión de los modelos, sino que también han facilitado la interpretación de los resultados. Por ejemplo, de las quince variables externas con mayor importancia global en las estimaciones, sólo dos se corresponden con los datos originales, mientras que las restantes se derivan del cálculo de promedios móviles.

En términos de evaluación de algoritmos, los modelos de ensamble han demostrado una notable capacidad para identificar picos y anomalías en la serie temporal, capturando la mayoría de los eventos atípicos en el conjunto de datos de prueba. Aunque el modelo SARIMAX exhibió métricas de error sobresalientes en general, no se destacó en la predicción de eventos atípicos. Por su parte, el modelo Prophet no alcanzó un rendimiento especialmente bueno en comparación con el modelo *baseline*, que replicaba los valores de los últimos siete días. Los modelos de ensamble, XGBoost y LightGBM, se beneficiaron de un proceso de optimización Bayesiana con `Optuna`, permitiendo una rápida y efectiva optimización de hiperparámetros. Por el contrario, los modelos con procesos de optimización menos sofisticados, basados en validación cruzada simple o en la minimización del AIC como SARIMAX, mostraron mayor susceptibilidad al sobreajuste. La capacidad de los algoritmos de ensamble para detectar relaciones no lineales y su resistencia a la multicolinealidad también influyeron positivamente en los resultados.

Es importante destacar que la interpretación de los modelos mediante valores SHAP proporcionó una visión detallada de los efectos de los componentes autorregresivos y las variables exógenas. Se encontró que los retardos más recientes tienen el mayor impacto sobre el caudal, y que las variables meteorológicas, como la lluvia, la humedad del suelo, las ráfagas de viento y la temperatura, resultaron ser los atributos más influyentes en el nivel del río a lo largo del tiempo. Esto permitió superar los desafíos de la complejidad inherente a los modelos de ensamble y facilitó una comprensión más profunda y técnica de los factores que influyen en la dinámica del caudal y las condiciones meteorológicas

asociadas.

Finalmente, contar con modelos avanzados que puedan pronosticar con precisión el caudal de ríos y afluentes, así como detectar comportamientos anómalos y subidas repentinas (modelos de ensamble) con un alto grado de fiabilidad, es de suma importancia para las compañías de energía hidroeléctrica. Estas capacidades permiten a las centrales anticipar el nivel de energía que podrán generar en días futuros, lo cual es crucial para posicionarse estratégicamente en el mercado de energías renovables. Además, al disponer de esta información anticipada, las centrales obtienen un mayor margen de maniobra en el establecimiento de precios, permitiéndoles ajustar sus estrategias comerciales para maximizar ingresos y asegurar una oferta competitiva. Mediante la implementación de estos modelos predictivos, estas compañías pueden optimizar su producción, gestionar mejor los recursos y contribuir de manera más eficiente y sostenible al suministro energético.

Bibliografía

[1] NTT Data Project UNAV(2024). Disponible en: https://github.com/hectordgv15/NTTDataProjectUNAV.git

[2] Iberdrola (2024). Energía hidroeléctrica. Disponible en: https://www.iberdrola.com/conocenos/nuestra-actividad/energia-hidroelectrica

[3] Iberdrola España (2024). Centrales hidroeléctricas en la cuenca del Duero. Disponible en: https://www.iberdrolaespana.com/conocenos/lineas-negocios/energia-hidroelectrica/cuenca-duero

[4] Santos Júnior, D. S. de O., de Oliveira, J. F. L., & de Mattos Neto, P. S. G. (2019). An intelligent hybridization of ARIMA with machine learning models for time series forecasting. *Knowledge-Based Systems*, 175, 72-86. https://doi.org/10.1016/j.knosys.2019.03.011

[5] Phan, T.-T.-H., & Nguyen, X. H. (2020). Combining statistical machine learning models with ARIMA for water level forecasting: The case of the Red river. *Advances in Water Resources*, 142, 103656. https://doi.org/10.1016/j.advwatres.2020.103656

[6] Kenyi, M. G. S., & Yamamoto, K. (2023). Seasonal ARIMA Prediction of Streamflow: Sobat River Tributary of the White Nile River. *Conference Paper*. https://www.researchgate.net/publication/377766498

[7] Ahmed, A. N., Yafouz, A., Birima, A. H., Kisi, O., Huang, Y. F., Sherif, M., Sefelnasr, A., & El-Shafie, A. (2022). Water level prediction using various machine learning algorithms: A case study of Durian Tunggal River, Malaysia. *Engineering Applications of Computational Fluid Mechanics*, 16(1), 422-440. https://doi.org/10.1080/19942060.2021.2019128

[8] Belyakova, P. A., Moreido, V. M., Tsyplenkov, A. S., Amerbaev, A. N., Grechishnikova, D. A., Kurochkina, L. S., Filippov, V. A., & Makeev, M. S. (2022). Forecasting Water Levels in Krasnodar Krai Rivers with the Use of Machine Learning. *Water Resources*, 49(1), 10–22. https://doi.org/10.1134/S0097807822010043

[9] Lujano, E., Lujano, R., Huamani, J. C., & Lujano, A. (2022). Hydrological modeling based on the KNN algorithm: an application for the forecast of daily flows of the Ramis river, Peru. *Tecnología y Ciencias del Agua*, 14(2), 5. https://doi.org/10.24850/j-tyca-14-2-5

[10] Khan, S., Khan, M., Khan, A. U., Khan, F. A., Khan, S., & Fawad, M. (2023). Monthly streamflow forecasting for the Hunza River Basin using machine learning techniques. *Water Practice & Technology*, 00(0), 1-11. https://doi.org/10.2166/wpt.2023.124

[11] Chen, H., Huang, S., Xu, Y.-P., Teegavarapu, R. S. V., Guo, Y., Nie, H., Xie, H., & Zhang, L. (2023). River ecological flow early warning forecasting using baseflow separation and machine learning in the Jiaojiang River Basin, Southeast China. *Science of the Total Environment*, 882, 163571. https://doi.org/10.1016/j.scitotenv.2023.163571

[12] Kilinc, H. C., Ahmadianfar, I., Demir, V., Heddam, S., Al-Areeq, A. M., Abba, S. I., Tan, M. L., Halder, B., Marhoon, H. A., & Yaseen, Z. M. (2023). Daily Scale River Flow Forecasting Using Hybrid Gradient Boosting Model with Genetic Algorithm Optimization. *Water Resources Management*, 37, 3699–3714. https://doi.org/10.1007/s11269-023-03522-z

[13] Mishra, P., Al Khatib, A. M. G., Yadav, S., Ray, S., Lama, A., Kumari, B., Sharma, D., & Yadav, R. (2024). Modeling and forecasting rainfall patterns in India: a time series analysis with XGBoost algorithm. *Environmental Earth Sciences*, 83, 163. https://doi.org/10.1007/s12665-024-11481-w

[14] Hunt, K. M. R., Matthews, G. R., Pappenberger, F., & Prudhomme, C. (2022). Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States. *Hydrology and Earth System Sciences (HESS)*, 26(21), 5449-5472. https://doi.org/10.5194/hess-26-5449-2022

[15] Granata, F., Di Nunno, F., & De Marinis, G. (2022). Stacked machine learning algorithms and bidirectional long short-term memory networks for multi-step ahead streamflow forecasting: A comparative study. *Journal of Hydrology*, 613(Part A), 128431. https://doi.org/10.1016/j.jhydrol.2022.128431

[16] Ahmed, A. A. M., Deo, R. C., Ghahramani, A., Feng, Q., Raj, N., Yin, Z., & Yang, L. (2022). New double decomposition deep learning methods for river water level forecasting. *Science of The Total Environment*, 831, 154722. https://doi.org/10.1016/j.scitotenv.2022.154722

[17] Bak, G., & Bae, Y. (2023). Deep learning algorithm development for river flow prediction: PNP algorithm. *Soft Computing*, 27, 13487–13515. https://doi.org/10.1007/s00500-023-08254-1

[18] Confederación Hidrográfica del Duero (2024). Disponible en: https://www.chduero.es/

[19] Open-Meteo (2024). Disponible en: https://open-meteo.com/

[20] iAgua: ¿Cuáles son los afluentes del río Duero? (2024). Disponible en: https://www.iagua.es/respuestas/cuales-son-afluentes-rio-duero

[21] Agencia Estatal de Meteorología (AEMET) España (2024). Disponible en: https://www.aemet.es/

[22] Statsmodels: Statistical modeling and econometrics in Python (2024). Disponible en: https://www.statsmodels.org/

[23] XGBoost: Scalable and Flexible Gradient Boosting (2024). Disponible en: https://xgboost.readthedocs.io/

[24] LightGBM: A Highly Efficient Gradient Boosting Decision Tree (2024). Disponible en: https://lightgbm.readthedocs.io/

[25] Prophet: Forecasting at Scale (2024). Disponible en: https://facebook.github.io/prophet/

[26] Skforecast: Forecasting with Scikit-learn and XGBoost (2024). Disponible en: https://skforecast.org/

[27] Scikit-learn: Machine Learning in Python (2024). Disponible en: https://scikit-learn.org/