Implementación de modelos de aprendizaje |automático para predicción de abastecimiento de agua: un enfoque de pensamiento basado en el riesgo en la agroindustria

Diego Fernando Osorio Chávez
Ingeniero en Recursos Hídricos y Gestión Ambiental. Especialización en Big Data e Inteligencia de Negocios. Especialización en Inteligencia Artificial, Especialización en Métodos y Técnicas de Investigación Social, Magister en Administración de Riesgos. Servicio Nacional de Aprendizaje SENA, dfosorio@sena.edu.co, Medellín.

RESUMEN

La gestión adecuada del recurso hídrico en las organizaciones es crucial para evitar consecuencias negativas, como la interrupción de operaciones por escasez de agua. Un enfoque sistemático y sostenible es imperativo para asegurar la continuidad y eficiencia empresarial, así como para proteger este recurso para las futuras generaciones. La falta de gestión eficiente del recurso hídrico podría desencadenar crisis catastróficas, lo que subraya la urgencia de actuar. El pronóstico del consumo de agua, especialmente mediante el análisis de series temporales, es fundamental para tomar decisiones informadas y gestionar eficazmente este recurso. Las empresas enfrentan el desafío de optimizar operaciones frente a la incertidumbre del suministro de agua, lo que hace vital la implementación de modelos de aprendizaje automático para la predicción precisa del consumo. Este artículo propone un enfoque basado en el riesgo y el uso de modelos de aprendizaje automático para desarrollar un sistema de alerta temprana, con el fin de anticipar y mitigar interrupciones en el suministro de agua, garantizando así la continuidad operativa y la gestión efectiva de este recurso vital.

Palabras clave: Gestión del agua, recurso hídrico, interrupción de operaciones, aprendizaje automático, alerta temprana, riesgo.

ABSTRACT

The proper management of water resources within organizations is critical to avoid negative consequences, such as operational interruptions due to water shortages. A systematic and sustainable approach is imperative to ensure business continuity and efficiency, while also safeguarding this resource for future generations. The absence of efficient water resource management could lead to catastrophic crises, underscoring the urgency of taking action. Forecasting water consumption, particularly through time series analysis, is crucial for making informed decisions and effectively managing this resource. Companies face the challenge of optimizing operations amid water supply uncertainty, making the implementation of machine learning models for accurate consumption prediction vital. This article proposes a risk-based approach and the use of machine learning models to develop an early warning system, aiming to anticipate and mitigate interruptions in water supply, thereby ensuring operational continuity and effective management of this vital resource.

Keywords: Water management, water resource, operational interruption, machine learning, early warning, risk.

1. INTRODUCCIÓN

La valoración del recurso hídrico dentro de las organizaciones es una tarea crítica que, si se maneja de manera deficiente, puede desencadenar una serie de consecuencias negativas. Entre estas, la más preocupante es el riesgo de interrupción de las operaciones debido a problemas de abastecimiento de agua, un escenario que ninguna empresa desea enfrentar. Este riesgo se magnifica al considerar la importancia del agua, un tesoro oculto que sostiene una amplia gama de actividades industriales. La materialización de estos riesgos a menudo se debe a debilidades dentro del sistema organizacional que no consideran adecuadamente la importancia de este elemento (Osorio, 2018)

Por lo tanto, es imperativo adoptar un enfoque de gestión sistemática y sostenible para el recurso hídrico, asegurando así la continuidad y eficiencia de la operación de las empresas, y a su vez, protegiendo este recurso vital para las generaciones futuras.

El estudio reciente realizado por Tehreem y Tzanakakis pone de relieve la urgencia y gravedad de la problemática relacionada con la gestión de las aguas subterráneas. Advierten que la falta de una gestión eficiente y la incapacidad para controlar la sobreexplotación de estos recursos podrían desencadenar crisis catastróficas en el futuro (Tehreem y Tzanakakis, 2024). Esta advertencia no solo destaca la importancia de abordar de manera sostenible el recurso hídrico, sino que también subraya la necesidad imperativa de actuar de inmediato. Como lo expresa Tehreem de manera elocuente, "Es hora de dejar de esperar", lo cual representa un llamado urgente a la acción para prevenir tales crisis y asegurar la sostenibilidad de nuestros recursos hídricos.

En este contexto, el agua subterránea desempeña un papel crucial en la industria, representando aproximadamente el 5% del uso mundial de agua (Secretaría de Agricultura y Desarrollo Rural, año de publicación, 2022). Su relevancia se destaca aún más al considerar que constituye el recurso principal para la producción de alimentos, siendo utilizado en un 67% del total destinado a este fin (Ferrovial, 2023).

Esta realidad resalta la importancia de contar con pronósticos precisos y confiables sobre el consumo de agua, que permitan establecer una relación clara entre la producción y el consumo de este recurso. Dichos pronósticos no solo son esenciales para evaluar la disponibilidad cuantitativa del agua, sino también para informar las decisiones y estrategias de gestión que garanticen su uso sostenible y equitativo (Wunsch et al. 2021).

Como señala Shen (2018), las ciencias relacionadas con el agua están adoptando cada vez más modelos de aprendizaje automático, lo que está generando una transformación en la forma en que se abordan los desafíos en las industrias tradicionales. Estos modelos ofrecen la capacidad de analizar las complejas relaciones entre las variables del sistema y los parámetros de los modelos de predicción, lo que se vuelve fundamental para una gestión más eficaz de los recursos hídricos en los modelos de producción.

Basándonos en lo mencionado anteriormente, el pronóstico del consumo de agua implica principalmente el análisis de series de tiempo. Según lo definido por Guizzardi et al. (2021), una serie temporal se describe como una secuencia de puntos indexados en orden cronológico, que abarca datos como producción diaria, horas de

proceso diarias, lecturas de parámetros climatológicos y consumo de agua.

Este proceso de pronóstico implica examinar datos históricos para prever valores futuros de la serie, lo que implica la utilización de patrones y tendencias identificados en los datos previos para estimar los valores futuros de la serie.

La importancia de prever los valores futuros del consumo de agua, como señala Guizzardi et al. (2021), radica en su capacidad para permitir a la organización de estudio establecer medidas de tratamiento necesarias, generar alertas tempranas, asignar recursos de manera eficiente y mejorar la eficiencia en los procesos.

Esta práctica no solo proporciona información crucial sobre tendencias y valores futuros de variables relevantes, sino que también se convierte en una característica crítica para el funcionamiento normal del sistema.

En el contexto empresarial actual, surge el desafío de cumplir con las expectativas de los clientes y proveedores, mientras que el suministro adecuado de agua se convierte en un factor limitante para la capacidad operativa. A pesar de la necesidad de un abastecimiento suficiente para alcanzar los objetivos actuales, este no garantiza una proyección precisa de la demanda a corto y mediano plazo. En estas circunstancias, la capacidad de optimizar las operaciones a mediano plazo se vuelve esencial para gestionar las reservas de agua y el uso eficiente de la capacidad instalada.

Estas empresas tienen como objetivo satisfacer la demanda de los consumidores de manera eficiente día tras día. Sin embargo, muchas de ellas dependen fundamentalmente del suministro de agua y operan sin emplear técnicas avanzadas de predicción del consumo. Esto las deja en una posición vulnerable, cubriendo su producción en un entorno de alta incertidumbre. Frente a esta realidad, la optimización de las operaciones a corto y mediano plazo emerge como una herramienta crucial para garantizar la continuidad de la producción y establecer una gestión eficaz de las reservas de agua. En este contexto, los métodos de aprendizaje automático enfrentan el desafío de procesar y agrupar los datos de consumo de agua, permitiendo que los modelos de aprendizaje profundo realicen predicciones con la mayor precisión posible.

Por ende, el enfoque del presente artículo se orienta hacia la implementación de modelos de aprendizaje automático con el propósito de desarrollar un sistema de alerta temprana ante posibles interrupciones en las operaciones de abastecimiento de agua. Este enfoque se fundamenta en un pensamiento basado en el riesgo, que reconoce la importancia de anticipar y gestionar los riesgos asociados a la disponibilidad y suministro de este recurso vital. La aplicación de modelos de aprendizaje automático ofrece una herramienta innovadora y eficaz para identificar patrones y tendencias en los datos, lo que permite prever de manera anticipada posibles escenarios de interrupción en el abastecimiento de agua y tomar medidas preventivas de forma proactiva. Este enfoque busca no solo mitigar los riesgos potenciales, sino también optimizar la gestión del recurso hídrico y garantizar la continuidad operativa en situaciones adversas.

Para lograr este propósito, se plantean una serie de objetivos específicos que guiarán el desarrollo y la implementación de dicho

modelo. En primer lugar, se busca comprender a fondo el negocio, reconociendo la importancia estratégica del abastecimiento de agua para garantizar la continuidad operativa y la satisfacción del cliente en la empresa. A partir de esta comprensión, se realizará un análisis exploratorio de datos, que incluirá tareas como limpieza, manejo de valores perdidos y eliminación de outliers, con el fin de comprender la naturaleza y la calidad de los datos disponibles.

Una vez comprendidos los aspectos del negocio y los datos disponibles, se procederá a la preparación de los datos, transformándolos de manera adecuada para su utilización efectiva en modelos de aprendizaje automático. Esta etapa es crucial para garantizar la calidad y la relevancia de los datos utilizados en el proceso de modelado.

El siguiente paso consistirá en experimentar con diferentes tipos de modelos y configuraciones, con el objetivo de identificar el modelo que mejor se ajuste a los datos y que tenga la capacidad de predecir con precisión posibles interrupciones en el abastecimiento de agua Este proceso estará respaldado por una evaluación rigurosa, que comparará y seleccionará el modelo óptimo basado en una métrica de evaluación apropiada. Finalmente, una vez seleccionado el modelo más adecuado, se procederá a su integración en un sistema de alerta temprana, que permitirá predecir y alertar sobre posibles interrupciones en el abastecimiento de agua de manera oportuna y eficiente. De esta manera, la implementación de esta metodología completa proporcionará a la organización las herramientas necesarias para gestionar de manera efectiva este recurso vital y garantizar la continuidad de sus operaciones.

2. METODOLOGIA

Con el objetivo de ofrecer una panorámica detallada de la metodología propuesta, en esta sección se exponen todos los procedimientos que se han desarrollado.

2.1 REVISIÓN DE LITERATURA

En este capítulo, abordaremos una revisión de algunos estudios publicados sobre la predicción del consumo de agua, dichos enfoque se han divido en métodos estadísticos, de aprendizaje automático y de aprendizaje profundo.

2.1.1 Métodos estadísticos

Los métodos estadísticos son ampliamente reconocidos por su simplicidad y facilidad de interpretación, lo que los convierte en herramientas valiosas en la toma de decisiones. Su capacidad para trabajar con conjuntos de datos relativamente pequeños los hace accesibles incluso en entornos con recursos limitados. Además, su tolerancia a los valores atípicos brinda estabilidad a las predicciones, lo que es fundamental en situaciones donde la precisión y la fiabilidad son esenciales. Esta combinación de características hace que los métodos estadísticos sean una opción atractiva para una variedad de aplicaciones, desde análisis de mercado hasta pronósticos climáticos y gestión de riesgos. A pesar de las ventajas que ofrecen los métodos estadísticos, su dependencia de supuestos simplificados limita su capacidad para capturar patrones complejos, especialmente en conjuntos de datos con relaciones no lineales. Además, los modelos estadísticos pueden exhibir una capacidad predictiva baja en comparación con métodos más avanzados como el aprendizaje profundo. Esto se debe a que losresultados de los modelos estadísticos pueden ser sensibles a sesgos cuando no se cumplen los supuestos subyacentes (Gil-Gamboa et al., 2024).

En su estudio titulado "Extended linear and non-linear auto-regressive models for forecasting the urban water consumption of a fast-growing city in an arid region" (2019), Mohammad y Mousavi-Mirkalaei abordan la predicción de series de tiempo en el contexto del consumo de agua urbana en una ciudad de rápido crecimiento en una región árida. Exploraron varios métodos, tanto lineales como no lineales, para abordar esta tarea crucial. Entre los métodos lineales, se destacan la media móvil integrado autorregresiva (ARIMA) y la media móvil autorregresivo (ARMA), mientras que, en el ámbito no lineal, se centran en la red neuronal artificial (RNA). Estos enfoques, aunque diferentes en su naturaleza y aplicación, comparten el objetivo común de prever el consumo de agua, proporcionando a los investigadores y responsables de políticas herramientas valiosas para la gestión sostenible de este recurso vital.

Por su parte, en su investigación realizada en 2014, Chang, Chen, Lu, Huang y Chang se concentraron en la predicción del nivel del agua para el control de inundaciones urbanas, empleando tanto modelos estáticos como dinámicos de redes neuronales artificiales (RNA). Sus hallazgos revelaron que el modelo NARX, un tipo específico de modelo de RNA, demostró ser particularmente efectivo en la predicción del nivel del agua. Este descubrimiento resalta la utilidad y el potencial de las RNA en la gestión de desastres naturales, proporcionando una herramienta valiosa para la prevención y mitigación de inundaciones urbanas.

2.1.2 Métodos de aprendizaje automático

Estos métodos se destacan por su capacidad de adaptación a los patrones cambiantes y su complejidad, generando como resultado automatización en sus procesos de predicción, y optimización en tiempo y recursos.

De igual forma, son escalables a grandes volúmenes de datos, sin embargo, entre sus desventajas, esta, por un lado, la ambigüedad en sus resultados, la sensibilidad a los datos atípicos y la complejidad en los ajustes de los hiperparámetros (Gamboa et al., 2024).

Por su parte, Karamaziotis, Raptis, Nikolopoulos, Litsiou y Assimakopoulos (2020) en su estudio "An empirical investigation of water consumption forecasting methods" exploraron siete modelos de pronóstico y técnicas de combinación para prever el consumo de agua a mediano plazo. Estas técnicas, que incluyeron métodos como ARIMA y suavizado exponencial, se combinaron utilizando enfoques como promedio simple, promedio ponderado y promedio recortado.

El método de promedio recortado, junto con ARIMA, se destacó como el más preciso en las predicciones. Estos hallazgos permitieron a las empresas de agua detectar posibles aumentos repentinos en la demanda en áreas específicas y tomar medidas preventivas para reducir el estrés en el sistema de distribución. Este estudio complementa la investigación anterior al ofrecer nuevas perspectivas sobre las metodologías de pronóstico del consumo de agua y su aplicabilidad en contextos vulnerables.

2.1.3 Métodos de aprendizaje profundo

En los últimos años, ha surgido un aumento significativo en el interés y la aplicación de algoritmos de aprendizaje profundo, especialmente en el contexto de la predicción utilizando series temporales. Este avance ha sido notable, ya que los algoritmos de aprendizaje profundo, en particular las redes neuronales de memoria a largo plazo (LSTM), han demostrado

ofrecer resultados notablemente mejorados en comparación con enfoques tradicionales. La capacidad de las LSTM para capturar y modelar patrones complejos a lo largo del tiempo ha sido fundamental para su éxito en una variedad de aplicaciones, incluida la predicción de series temporales. Esta popularidad creciente de los algoritmos de aprendizaje profundo ha abierto nuevas oportunidades para explorar y desarrollar enfoques más precisos y efectivos para la predicción y el análisis de datos temporales en una amplia gama de campos, desde las finanzas hasta la meteorología.

El estudio "Developing an LSTM model to forecast the monthly water consumption according to the effects of the climatic factors in Yazd", realizado por Niknam, Zare, Hosseininasab y Mostafaeipour (2023), ofrece una valiosa perspectiva sobre el potencial de las redes neuronales de memoria a largo plazo (LSTM) en la predicción de datos temporales. En su investigación, los autores resaltan la capacidad inherente de las LSTM para capturar y retener información relevante a lo largo del tiempo, lo que las convierte en una herramienta poderosa para pronosticar series temporales, particularmente en el contexto de la influencia de factores climáticos en el consumo mensual de agua en Yazd.

Al considerar el estudio de Niknam y sus colegas, se destaca la importancia de las LSTM en la modelización de datos complejos y dinámicos como el consumo de agua, que está intrínsecamente ligado a variables climáticas. Esta investigación no solo resalta el papel crucial de las LSTM en la predicción precisa de fenómenos ambientales, sino que también subraya la necesidad de comprender y adaptarse a las complejas interacciones entre factores climáticos y comportamientos de consumo de recursos naturales.

Otro ejemplo convincente de los avances en el uso de redes neuronales de memoria a largo plazo (LSTM) proviene del estudio realizado por Kuhnert, Gonuguntla, Krieh, Nowak y Thomas (2021). Estos investigadores analizaron meticulosamente datos provenientes de dos plantas de distribución de agua con el objetivo de evaluar y comparar el rendimiento de diferentes sistemas de predicción. Sus hallazgos revelaron de manera concluyente que las redes LSTM superan significativamente al sistema actualmente utilizado en empresas de agua. Al profundizar en su investigación, Kuhnert y su equipo no solo demostraron la eficacia de las redes LSTM, sino que también proporcionaron una visión crítica sobre cómo estos modelos pueden mejorar la gestión del recurso hídrico.

2.2 Formulación del problema

En esta sección, se establece la definición del problema de pronóstico y se describe el conjunto de datos que se empleará en la investigación. En primer lugar, se presenta el problema a abordar en la Sección 3.1, centrándose en las predicciones del consumo de agua en una empresa del sector agroindustrial. En segundo lugar, se proporciona una descripción detallada del conjunto de datos utilizado, acompañado de un análisis exploratorio de los datos en la Sección 3.2.

2.2.1 Definición del problema

El problema se centra en las predicciones de consumo de agua de una empresa agroindustrial que depende de un pozo profundo para su suministro hídrico, el cual es esencial para su proceso de producción. Específicamente, se busca predecir el consumo diario de agua utilizando datos históricos que incluyen el consumo pasado de agua, la cantidad de fruta procesada, las

horas de operación del proceso y las horas trabajadas. En este contexto, el objetivo es desarrollar un modelo predictivo que pueda estimar con precisión el consumo diario de agua en función de las variables mencionadas. Este modelo permitirá a la empresa planificar y gestionar de manera más eficiente su uso del recurso hídrico, optimizando así sus operaciones de producción y garantizando un suministro adecuado de agua para sus procesos industriales.

2.2.2 Datos

En el presente estudio, se ha llevado a cabo la recolección de datos en una empresa perteneciente al sector agroindustrial, centrando nuestra atención en el área de producción. Los datos recopilados abarcan una variedad de variables pertinentes a la producción, tales como la cantidad de racimos de fruto fresco, horas de proceso y las horas trabajadas. La recolección de datos se realizó desde el año 2020 hasta el 2024, proporcionando un marco temporal extenso que permite un análisis exhaustivo y detallado de las tendencias y patrones en el consumo de agua y otras variables de producción.

Paralelamente, se ha procedido a recoger información acerca del consumo de agua, la cual fue registrada por el departamento ambiental de la empresa. Adicionalmente, con el propósito de enriquecer nuestro análisis, se han obtenido datos climatológicos provenientes de una plataforma de la NASA. Esta integración de múltiples fuentes de datos nos ha permitido abordar de manera integral y holística la problemática en cuestión. En este estudio, se ha llevado a cabo un análisis estadístico completo que abarca tanto el promedio como la correlación entre diversas variables cruciales para la producción en una empresa agroindustrial, así como con el parámetro climatológico más relevante, la humedad relativa. Los datos recopilados incluyen el consumo de agua, la cantidad de frutos procesados, las horas de proceso y las horas trabajadas, además de la humedad relativa y otros parámetros climatológicos.

Al analizar los promedios y los valores máximos, se observa que el consumo promedio de agua es de aproximadamente 368.23 unidades, con un máximo de 1050 unidades. La cantidad promedio de frutos procesados es de alrededor de 276,359.83 unidades, alcanzando un máximo de 888,180 unidades. Las horas de proceso y las horas trabajadas promedian alrededor de 12.05 y 15.22 horas respectivamente, con valores máximos de 31 y 34 horas respectivamente. La humedad relativa promedio se sitúa en torno al 87.64%, con un máximo de 98.25%.

Además, al evaluar la correlación entre estas variables, se han identificado relaciones significativas. Por ejemplo, se observa una correlación positiva fuerte entre las horas de proceso y la cantidad de frutos procesados (0.968566), así como entre las horas trabajadas y las horas de proceso (0.910583). También se han encontrado correlaciones moderadas entre las horas de proceso y el consumo de agua (0.503318), y entre las horas trabajadas y el consumo de agua (0.467398). Finalmente, se ha identificado una correlación positiva moderada entre la humedad relativa y el consumo de agua (0.681422), sugiriendo una posible influencia del clima en el consumo de agua en la producción agroindustrial. Estos análisis proporcionan una visión completa y detallada de las relaciones entre las variables estudiadas, lo que puede ser fundamental para comprender y mejorar la eficiencia de las operaciones en la empresa agroindustrial. Basándonos en lo mencionado anteriormente, procedemos a

presentar de forma gráfica un análisis exploratorio de los datos reales utilizados en este estudio.

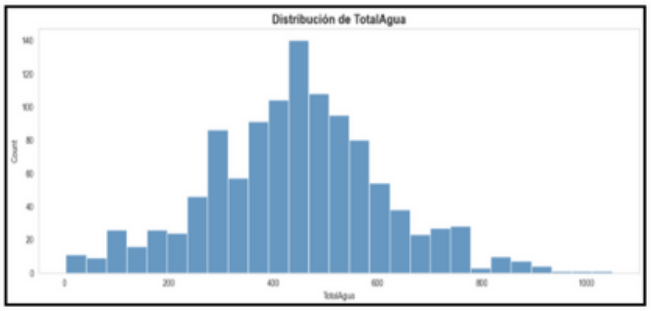


Figura 1. Distribución de Consumo Total Agua

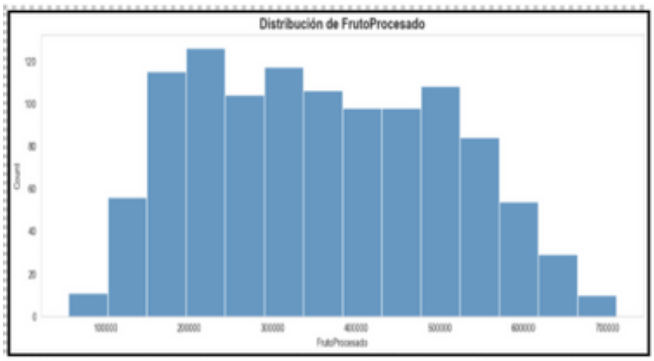


Figura 2. Cantidad de Fruto Procesado.

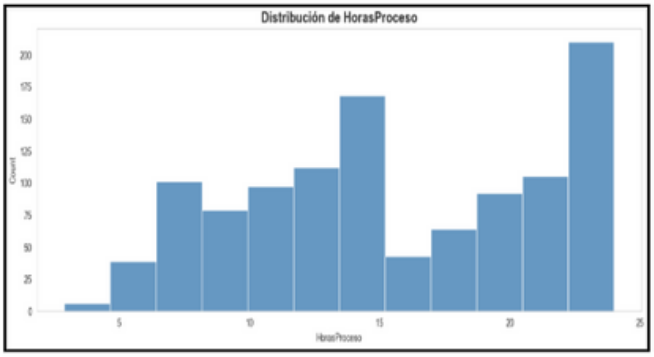


Figura 3. Distribución de Horas Proceso.

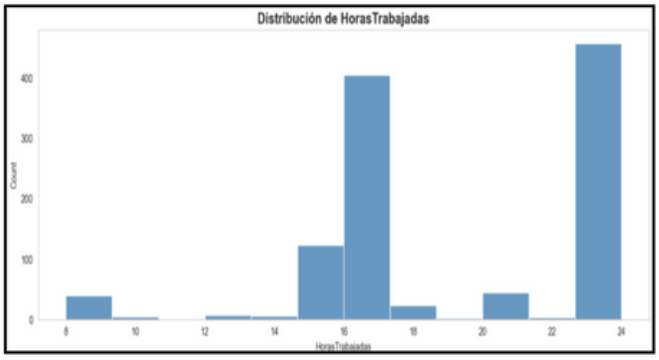


Figura 4. Distribución de Horas Trabajadas.

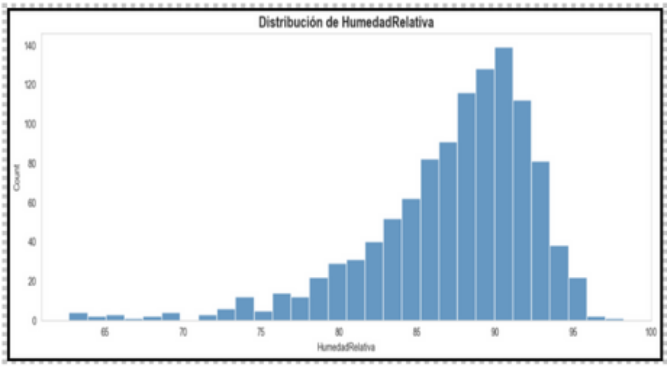


Figura 5. Distribución de Humedad Relativa

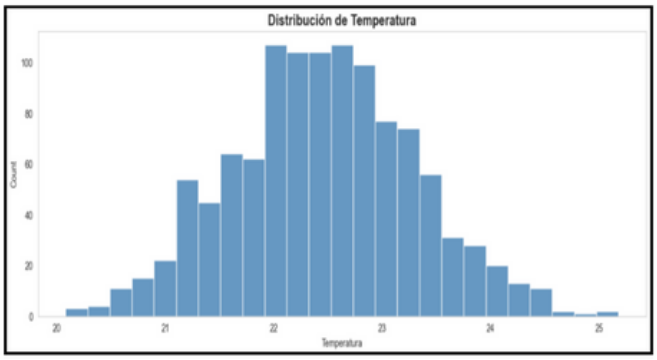


Figura 6. Distribución de Temperatura.

2.2.3 Procesamiento de datos

En la mayoría de las tareas de análisis de datos, es crucial realizar un preprocesamiento inicial para garantizar la calidad de las predicciones. Tras analizar los datos, se determinó que requerían un alto

nivel de preprocesamiento. En la mayoría de las tareas de análisis de datos, es esencial realizar un preprocesamiento inicial para garantizar la calidad de las predicciones.
Tras analizar los datos, se determinó que había una cantidad significativa de valores nulos. También se detectaron valores de consumo extremadamente pequeños, posiblemente errores de medición, los cuales fueron corregidos. Se aplicó un procedimiento de imputación de valores nulos y pequeños en los datos de consumo de agua. Posteriormente, se definieron las variables predictoras y la variable objetivo, y se empleó una técnica de ventana deslizante para suavizar los datos y reducir el ruido. Luego, los datos se dividieron en conjuntos de entrenamiento y prueba. Se ajustó un modelo de Bosques Aleatorios para predecir el consumo de agua basado en las horas de proceso, horas trabajadas, frutos procesados y humedad relativa. Finalmente, se graficaron las predicciones del modelo frente a los valores reales para una comparación visual.

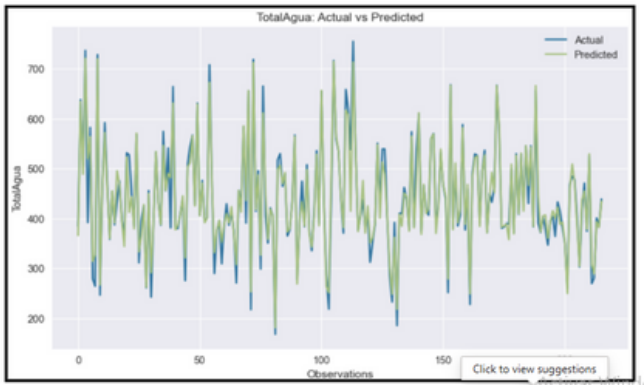


Figura 7. Modelo Random Forest Regressor

Después de ajustar el modelo de Bosques Aleatorios y calcular las métricas de evaluación, se procedió a comparar las predicciones del modelo con los datos reales de consumo de agua. La siguiente gráfica muestra esta comparación conjunta de datos de consumo de agua y las predicciones generadas por el modelo. Esto proporciona una visualización clara de la precisión del modelo al predecir el consumo de agua en función de las variables predictoras seleccionadas.

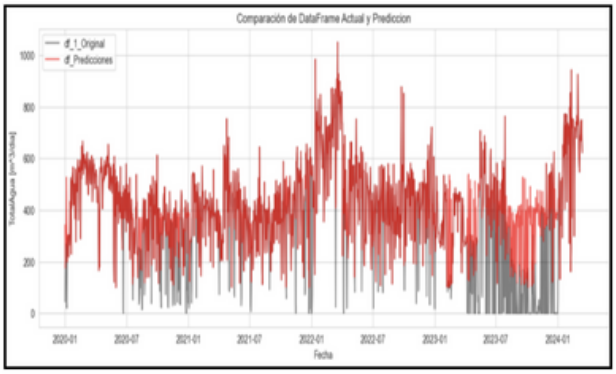


Figura 8. Grafica Predicciones del Modelo.

2.2.4 Selección y evaluación de modelos

Después del procesamiento de datos, se procedió a aplicar la metodología de Pycaret para la selección y evaluación de modelos, considerando la naturaleza de las series de tiempo en los datos. Este paso fue crucial para determinar el modelo más adecuado para predecir el consumo de agua en función de las variables disponibles. El proceso constó de las siguientes etapas: primero, se configuró el entorno de trabajo de Pycaret y se cargaron los datos procesados después del preprocesamiento; luego, se definió el experimento en Pycaret, incluyendo la variable objetivo y las variables predictoras, así como las métricas de evaluación a considerar; finalmente, se compararon varios modelos de aprendizaje automático disponibles en Pycaret utilizando un conjunto de métricas predefinidas. A continuación, se presentan los resultados obtenidos. A continuación, se relaciona los resultados obtenidos:

Tabla 1. Consolidado de Modelos

Modelo	MASE	RMSSE	MAE	RMSE	MAPE	SMAPE	TT (Sec)
Auto ARIMA	0.3296	0.2451	37.3938	37.3938	0.0518	0.0535	37.7000
dt_cds_dt	0.4500	0.3347	51.0671	51.0671	0.0766	0.0726	0.0667
exp_smooth	0.5421	0.4032	61.5158	61.5158	0.0918	0.0931	0.3233
et_cds_dt	0.5944	0.4421	67.4412	67.4412	0.0969	0.1002	0.5667

Basado en los resultados obtenidos con el modelo Pycaret, se opta por trabajar con el modelo ARIMA (AutoRegressive Integrated Moving Average) en este estudio. Esto se fundamenta en la comparación de métricas de evaluación entre los diferentes modelos considerados.

El modelo Auto ARIMA, con un MASE de 0.3296, RMSSE de 0.2451 y un MAE de 37.3938, muestra una buena capacidad para capturar los patrones y la variabilidad en los datos de series de tiempo. Aunque otros modelos como el Exponential Smoothing y Extra Trees with Conditional Deseasonalize & Detrending (et_cds_dt) ofrecen también un rendimiento aceptable, el ARIMA destaca por su equilibrio entre precisión y simplicidad, lo que lo hace adecuado para la modelización de datos temporales en este contexto específico.

Basado en los resultados obtenidos con el modelo Pycaret, se opta por trabajar con el modelo ARIMA (AutoRegressive Integrated Moving Average) en este estudio. Sin embargo, considerando la naturaleza estacional de los datos y el desempeño competitivo del ARIMA, se decide avanzar hacia un modelo SARIMA (Seasonal ARIMA) para una modelización más precisa y adecuada. Para implementar el modelo SARIMA, se utilizó un procedimiento de backtesting. Se especificaron los parámetros del modelo SARIMA, incluyendo el orden y la componente estacional, y se realizaron las predicciones sobre los datos de consumo de agua. El rendimiento del modelo se evaluó utilizando la métrica de error porcentual absoluto medio (mean_absolute_percentage_error). Los resultados del backtesting proporcionaron información valiosa sobre la capacidad predictiva del modelo SARIMA y su idoneidad para este estudio. Se obtuvo una métrica de error porcentual absoluto medio (MAPE) con un valor de 0.3638, lo que indica un buen ajuste del modelo SARIMA con respecto a los datos de consumo de agua.Principio del formulario

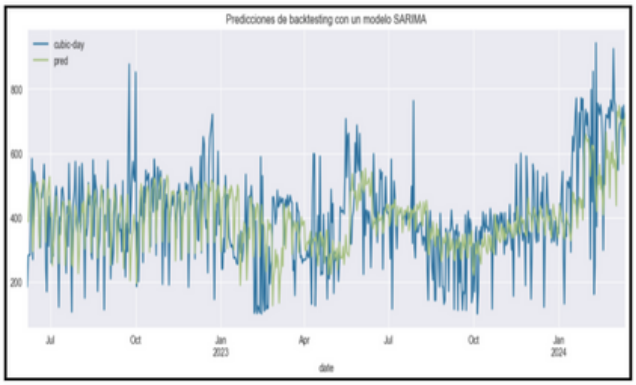
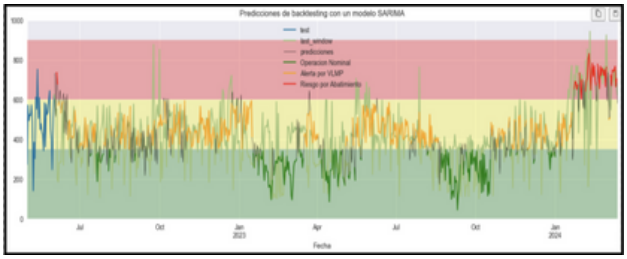


Figura 9. Resultado de Modelo SARIMA

Además, se llevó a cabo un análisis utilizando el método de backtesting para evaluar la eficacia del modelo SARIMAX en la predicción del consumo de agua, incorporando variables exógenas relevantes. Para este fin, se configuró el modelo SARIMAX con parámetros específicos, incluyendo un orden de (1, 1, 1) y una componente estacional de (1, 1, 1, 12). Utilizando datos de entrada como la cantidad de frutos procesados, las horas de proceso, la humedad relativa y la temperatura, se realizó el proceso de backtesting para evaluar el rendimiento del modelo.

Para una evaluación visual de las predicciones generadas por el modelo SARIMAX, se presenta la figura adjunta. En el gráfico, las líneas azules representan los datos reales de consumo de agua, donde la línea más clara muestra el último tramo del conjunto de datos. Las predicciones del modelo se superponen en el gráfico, mostrando diferentes colores para indicar distintos niveles de riesgo. Las áreas sombreadas en verde, amarillo y rojo representan zonas de operación nominal, alerta por volumen bajo de agua y riesgo por abatimiento respectivamente, con base en los niveles de predicción.



Final del formulario
Figura 10. Modelo SARIMA con variables endógenas.

2.2.5 Modelo Lstm

Para complementar la metodología de Pycaret, se desarrolló un modelo basado en una red neuronal de tipo LSTM (Long Short-Term Memory) con el objetivo de predecir el consumo de agua. Se seleccionaron las variables relevantes del dataset, que incluyen TotalAgua, FrutoProcesado, HorasProceso, HorasTrabajadas, HumedadRelativa y Temperatura, y se procedió a normalizar estos datos utilizando la técnica de Min-Max Scaling. Esta normalización es esencial para asegurar que los valores de las diferentes variables se encuentren en un rango comparable, facilitando el proceso de aprendizaje del modelo. Los datos se dividieron en conjuntos de entrenamiento y prueba, asignando el 80% de los datos para el entrenamiento y el 20% restante para la prueba, permitiendo evaluar el desempeño del modelo en datos no vistos durante el proceso de entrenamiento. Dado que la predicción del consumo de agua es un problema de series temporales, los datos fueron transformados en secuencias mediante la creación de ventanas deslizantes que permiten al modelo LSTM capturar las dependencias temporales de los datos. Se construyó un modelo LSTM con una capa LSTM de 54 unidades y una capa densa de salida, compilado utilizando el optimizador Adam y la función de pérdida de error cuadrático medio (MSE). Durante el entrenamiento, se empleó la técnica de Early Stopping para evitar el sobreajuste, monitorizando la pérdida de validación y deteniendo el entrenamiento si esta no mejoraba después de cinco épocas consecutivas. Tras el entrenamiento, se evaluó el desempeño del modelo utilizando los datos de prueba. Se graficaron tanto las pérdidas durante el entrenamiento como las predicciones del modelo frente a los valores reales de consumo de agua.

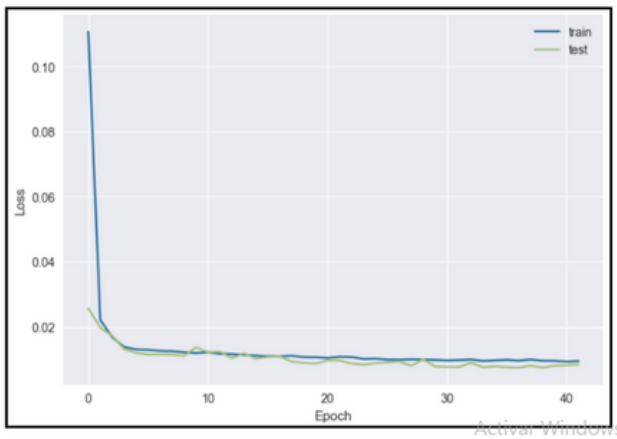


Figura 11. Modelo LSTM – Perdida y Validación

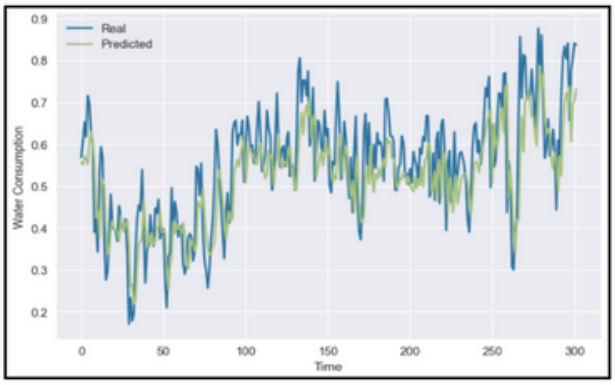


Figura 12. Modelo LSTM – Predicción

Adicionalmente, se desarrolló una función para realizar predicciones de consumo de agua utilizando nuevas entradas. Esta función normaliza los nuevos datos, crea una secuencia de entrada con la forma adecuada, realiza la predicción y luego invierte la normalización para obtener valores en la escala original. Se probó esta función con un ejemplo de nuevas entradas y se utilizaron los datos de prueba para validar la precisión de las predicciones. Esto permitió predecir el consumo de agua a partir de los datos proporcionados, demostrando la aplicabilidad práctica del modelo desarrollado y su capacidad de generalizar en diferentes contextos temporales.

3. RESULTADOS

En esta sección, se llevará a cabo un análisis exhaustivo de los resultados obtenidos por los modelos propuestos, comparando meticulosamente las métricas de error con las de otros métodos de predicción, como los bosques aleatorios, el modelo Sarimax y el modelo LSTM. Esta comparación permitirá una evaluación más profunda de la eficacia de cada modelo en la predicción del consumo de agua. Cada modelo será evaluado en función de su capacidad para capturar las complejidades de los datos de series temporales y proporcionar predicciones precisas y confiables. Además, se analizará cómo cada modelo se desempeña en términos de diversas métricas de evaluación, incluyendo el Error Porcentual Absoluto Medio (MAPE), el Error Cuadrático Medio (MSE), el Error Absoluto Medio (MAE) y el coeficiente de determinación (R^2 Score). Este análisis proporcionará una visión holística de la idoneidad de cada modelo en el contexto específico de predicción del consumo de agua, destacando sus fortalezas y debilidades relativas.

3.1 Modelo Random Forest

Se realizaron cálculos exhaustivos de varias métricas de evaluación para analizar el rendimiento del modelo de Bosques Aleatorios en la predicción del consumo de agua. Entre las métricas consideradas se encuentran el Error Cuadrático Medio (MSE), el coeficiente de determinación (R2), el Error Porcentual Absoluto Medio (MAPE) y el Error Absoluto Medio (MAE).

Los resultados destacaron un MSE de 638.40, un R2 de 0.95, un MAE de 15.00 y un MAPE de 0.04. Estas cifras reflejan una predicción altamente precisa y confiable por parte del modelo.
El análisis del MAPE sugiere una precisión aceptable en las predicciones del modelo en comparación con los valores reales de consumo de agua.

El modelo Random Forest demostró ser excepcionalmente efectivo, obteniendo el menor valor de MAPE (0.0382) y el mayor valor de R2 (0.9476). Estos resultados indican que Random Forest es el más preciso en sus predicciones y el mejor en explicar la variabilidad de los datos. Su capacidad para capturar pequeñas variaciones en el consumo de agua lo convierte en una opción confiable y robusta para tareas de predicción.

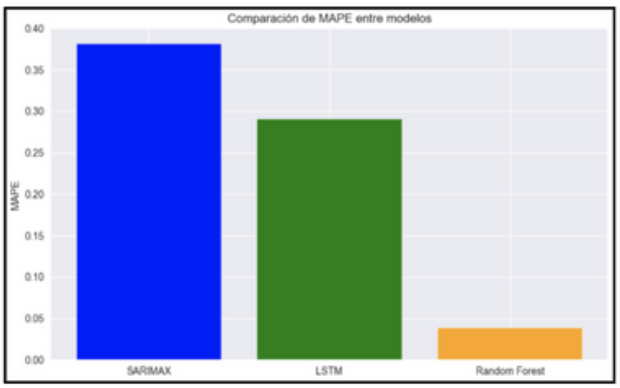


Figura 13. Métrica MAPE – Predicción

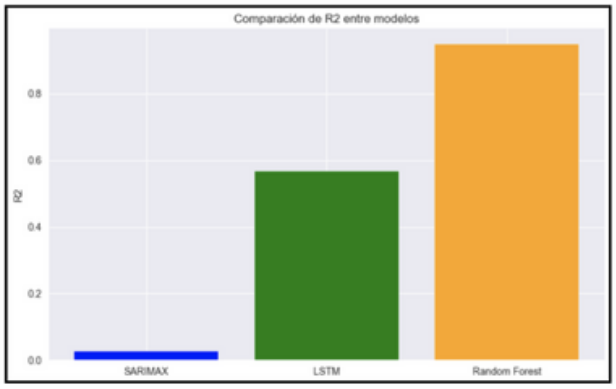


Figura 14. Métrica R2 – Predicción

Estos hallazgos consolidan la eficacia del enfoque de Bosques Aleatorios en la predicción precisa del consumo de agua.

3.2 Modelo Sarimax

Los resultados de la evaluación del modelo SARIMAX revelaron un MAPE de 0.3810, indicando una precisión razonable en las predicciones en comparación con los valores reales de consumo de agua. Esta métrica refleja el error porcentual medio absoluto entre las predicciones del modelo y los valores observados, lo que sugiere una buena capacidad del modelo para capturar las variaciones en los datos de series temporales. Durante el backtesting del modelo, se obtuvieron métricas adicionales que proporcionan una visión más completa de su rendimiento. El MSE del backtest fue de 22509.398, lo que indica la magnitud del error cuadrático medio entre las predicciones y los valores reales. Por otro lado, el MAE del backtest fue de 116.234, representando el error absoluto medio entre las predicciones y los valores reales. Además, el R2 del backtest fue de 0.0260, lo que indica que el modelo explica una pequeña proporción de la variabilidad en los datos de consumo de agua. Estas métricas complementarias ofrecen una evaluación más detallada del desempeño del modelo SARIMAX en la predicción del consumo de agua, destacando su capacidad para generar pronósticos precisos y confiables.

El modelo SARIMAX, aunque útil para capturar tendencias en series temporales, mostró un desempeño inferior en todas las métricas evaluadas. Con un MAPE de 0.3810, MSE de 22509.398, MAE de 116.234 y R2 de 0.0260, sus resultados indican una capacidad predictiva limitada en comparación con Random Forest y LSTM.

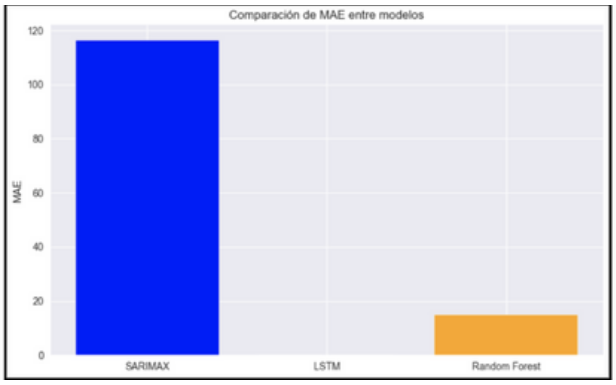


Figura 15. Métrica MAE – Predicción

Estos resultados sugieren que, si bien SARIMAX puede ser útil en ciertos contextos, no es la opción más robusta para la predicción precisa del consumo de agua.

3.3 Modelo LSTM

El análisis de los resultados del modelo LSTM revela una capacidad aceptable para predecir el consumo de agua. Utilizando

varias métricas de evaluación, se encontró que el modelo obtuvo un MAPE (Error Porcentual Absoluto Medio) del 29.04%, lo que indica un nivel de precisión razonable en sus predicciones. Además, el modelo presentó un MSE (Error Cuadrático Medio) de 0.0084 y un MAE (Error Absoluto Medio) de 0.0732. Estas métricas complementan el MAPE y respaldan la capacidad del modelo para predecir el consumo de agua con cierta precisión. El coeficiente de determinación (R2 Score) del modelo fue de 0.567, lo que indica que el modelo explica adecuadamente la variabilidad de los datos de consumo de agua. Aunque el MAPE refleja un margen de error, las gráficas de comparación entre las predicciones del modelo y los valores reales muestran una consistencia considerable, respaldando la capacidad del modelo para capturar tendencias y patrones en los datos de consumo de agua.

Se destaca la necesidad de continuar refinando el modelo mediante técnicas adicionales de optimización y ajuste de hiperparámetros para mejorar aún más su precisión predictiva.

	FrutoProcesado	HorasProceso	HorasTrabajadas	HumedadRelativa	Temperatura	Predicción del consumo de agua	Alerta Temprana
0	650	23	23	82	29	689.777458	Riesgo por Abatimiento
1	700	24	24	80	30	758.705902	Riesgo por Abatimiento
2	300	18	20	65	24	382.665189	Alerta por SCADA
3	720	23	23	78	31	822.699246	Riesgo por Abatimiento
4	680	18	23	75	32	881.453869	Riesgo por Abatimiento
5	630	24	24	81	25	427.997419	Alerta por SCADA
6	610	24	24	79	30	759.168977	Riesgo por Abatimiento
7	390	18	21	76	22	249.239075	Operación Nominal
8	730	24	24	77	31	824.959806	Riesgo por Abatimiento
9	670	24	24	76	32	889.133215	Riesgo por Abatimiento
10	640	18	23	83	27	554.611701	Alerta por SCADA
11	390	24	24	61	30	763.066718	Riesgo por Abatimiento
12	610	22	22	84	28	620.288602	Riesgo por Abatimiento
13	725	18	22	79	31	816.849208	Riesgo por Abatimiento
14	675	20	22	74	32	882.274866	Riesgo por Abatimiento
15	320	18	22	60	25	448.835458	Alerta por SCADA
16	710	22	22	78	30	755.949426	Riesgo por Abatimiento
17	380	18	20	67	23	320.902181	Operación Nominal
18	735	24	24	76	31	825.158784	Riesgo por Abatimiento
19	665	22	22	75	32	884.509888	Riesgo por Abatimiento

Figura 13. LSTM – Predicción

El modelo LSTM, por otro lado, mostró su fortaleza en términos de error cuadrático medio (MSE) y error absoluto medio (MAE), con los valores más bajos (0.0084 y 0.0732 respectivamente). Este rendimiento destaca la capacidad del LSTM para generar predicciones precisas y consistentes, minimizando tanto los errores de magnitud como los errores absolutos. Sin embargo, su MAPE es mayor en comparación con Random Forest, lo que sugiere que, aunque es excelente en reducir los errores absolutos, podría no ser tan preciso en términos porcentuales.

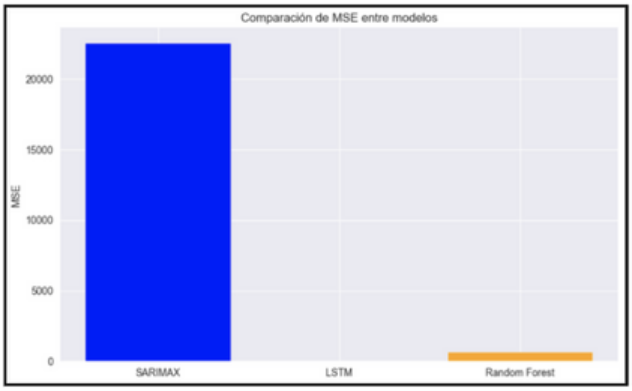


Figura 16. Métrica MSE – Predicción

4. CONCLUSIONES

El análisis comparativo de los modelos SARIMAX, LSTM y Random Forest para la estimación del consumo de agua ha revelado diferencias significativas en sus capacidades predictivas. Cada modelo fue evaluado en términos de MAPE, MSE, MAE y R2, proporcionando una visión integral de su rendimiento.

Basándonos en los resultados obtenidos, se recomienda utilizar el modelo Random Forest para la estimación del consumo de agua. Su alta precisión, reflejada en el bajo MAPE y alto R2, junto con su capacidad para explicar la mayor parte de la variabilidad en los datos, lo convierten en la opción más confiable y robusta. Sin embargo, el modelo LSTM también se presenta como una alternativa viable, especialmente si se busca minimizar los errores absolutos. El modelo SARIMAX, aunque menos preciso, puede seguir siendo útil en escenarios donde la captura de tendencias temporales sea de mayor importancia.

REFERENCIAS

Chang, F. J., Chen, P. A., Lu, Y. R., Huang, E., & Chang, K. Y. (2014). Real-time multi-step-ahead water level forecasting by recurrent neural networks for urban flood control. Journal of Hydrology, 517, 836-846.

Gil-Gamboa, A., Paneque, P., Trull, O., & Troncoso, A. (2024). Medium-term water consumption forecasting based on deep neural networks. Expert Systems with Applications, 247, 123234.

Guizzardi, A., Pons, F. M. E., Angelini, G., & Ranieri, E. (2021). Big data from dynamic pricing: A smart approach to tourism demand forecasting. International Journal of Forecasting, 37(3), 1049-1060.

Kühnert, C., Gonuguntla, N. M., Krieg, H., Nowak, D., & Thomas, J. A. (2021). Application of LSTM networks for water demand prediction in optimal pump control. Water, 13(5), 644.

Niknam, A., Zare, H. K., Hosseininasab, H., & Mostafaeipour, A. (2023). Developing an LSTM model to forecast the monthly water consumption according to the effects of the climatic factors in Yazd, Iran. Journal of Engineering Research, 11(1), 100028

Osorio Chavez, D. F. (2018). Análisis de la gestión del riesgo de desastres en cuatro instituciones educativas públicas de dos municipios del norte del Valle de Aburrá: Barbosa y Copacabana. [Master's thesis, Maestría en Administración de Riesgos].publicado en repositorio EAFIT https://hdl.handle.net/10784/13412.

Secretaría de Agricultura y Desarrollo Rural. (2022). Las aguas subterráneas tienen un papel principal en la agricultura de riego, la ganadería y el procesamiento de alimentos.

Shen, C.: A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists, Water Resour. Res., 54, 8558–8593, https://doi.org/10.1029/2018WR022643, 2018. a, b

Tehreem, H. S., Anser, M. K., Nassani, A. A., Abro, M. M. Q., & Zaman, K. (2020). Impact of average temperature, energy demand, sectoral value added, and population growth on water resource quality and mortality rate: it is time to stop waiting around. Environmental Science and Pollution Research, 27, 37626-37644.

Tzanakakis, V. A., Paranychianakis, N. V., & Angelakis, A. N. (2020). Water supply and water scarcity. Water, 12(9), 2347.

UNESCO: Los recursos mundiales de aguas subterráneas están sufriendo por una mala gobernanza, (2012).

Wunsch, A., Liesch, T., & Broda, S. (2021). Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). Hydrology and Earth System Sciences, 25(3), 1671-1687.