VOL. 6 Nº2- 2024

IBARRA - ECUADOR

Facultad de Ingeniería en Ciencias Aplicadas
Innovation & Development in Engineering and Applied Science

IDEAS

Predicción de la precipitación pluvial para prevención aplicando modelos ARIMA, Optimización Bayesiana y Modelo XGBoost

http://doi.org/10.53358/ideas.v6i2.1004

Andrés Tapia Arévalo, Andrés Galvis Correa
Universidad Politécnica Estatal del Carchi
andres.tapia@upec.edu.ec

Fecha de envío, enero 20/2024 - Fecha de aceptación, septiembre abril 1/2024 - Fecha de publicación, julio 15/2024

Resumen: Tradicionalmente, las series temporales han sido estudiadas a través de modelos auto-regresivos debido a su naturaleza de asociación temporal, no obstante, la compleja periodicidad de las lluvias reduce la efectividad de estos métodos. Este estudio se enfoca en proporcionar un modelo estadístico para las precipitaciones en la Estación Meteorológica de la Universidad Politécnica Estatal del Carchi, cantón Tulcán, con el propósito de perfeccionar la capacidad de pronóstico y la resiliencia del sector agrícola regional. Se analiza las vulnerabilidades causadas por las variaciones meteorológicas en distintas regiones del Ecuador y resalta la importancia de la agricultura en este contexto. Para ello se propone un modelo estadístico que combina método ARIMA, optimización bayesiana y modelos XGBoost para predecir el comportamiento de la lluvia. Sobre la base de los resultados obtenidos en el estudio podemos prever el comportamiento adyacente de las precipitaciones en un margen temporal que colinda hasta el mes de diciembre del año 2024. En el modelo XGboost se optimizaron los parámetros del modelo eta, max_depth, sub_sample y colsambple_bytree por medio de optimización bayesiana minimizando la raíz del error cuadrático medio. Se concluyó que, a pesar de no cumplir con el criterio de normalidad en los residuos del modelo de precipitaciones, se pudo verificar que estos se centran en cero y su comportamiento no varía en el tiempo lo cual garantiza validez del enfoque utilizado. Esto constituye un recurso imprescindible para la elección de decisiones más favorables y así beneficiar al área agrícola en la capacidad de adaptarse a eventos climáticos desfavorables.

Palabras clave: Precipitación, ARIMA, XGBoost, Agricultura, Cambio climático.

Abstract: Traditionally, time series have been studied through autoregressive models due to their nature of temporal association; however, the complex periodicity of rainfall reduces the effectiveness of these methods. This study focuses on providing a statistical model for precipitation at the Meteorological Station of the State Polytechnic University of Carchi, Tulcán canton, with the purpose of improving the forecasting capacity and resilience of the regional agricultural sector. The vulnerabilities caused by meteorological variations in different regions of Ecuador are analyzed and the importance of agriculture in this context is highlighted. For this, a statistical model is proposed that combines the ARIMA method, Bayesian optimization and XGBoost models to predict the behavior of rain. Based on the results obtained in the study, we can predict the adjacent behavior of precipitation in a time frame that extends until December 2024. In the XGboost model, the model parameters eta, max_depth, sub_sample and colsambple_bytree were optimized. through Bayesian optimization minimizing the root mean square error. It was concluded that, despite not meeting the normality criterion in the residuals of the precipitation model, it was possible to verify that they are centered on zero and their behavior does not vary over time, which guarantees validity of the approach used. This constitutes an essential resource for choosing more favorable decisions and thus benefiting the agricultural area in the ability to adapt to unfavorable climatic events.

Keywords: Precipitation, ARIMA, XGBoost, Agriculture, Climate change.

Autor de correspondencia:
Andrés Tapia Arévalo, andres.tapia@upec.edu.ec

Introduccion

La precisión en el pronóstico de las precipitaciones es sustancial en múltiples áreas, incluyendo la agricultura, la hidrología y la gestión de desastres. El cambio climático es el punto de inflexión de la producción agrícola y la economía de un país, siendo que representa un grave problema medioambiental con importantes consecuencias para el desarrollo futuro de la humanidad. El fenómeno climático ha captado la atención de gobiernos, organizaciones no gubernamentales y académicos de todo el mundo. La capacidad de predecir con precisión y celeridad las precipitaciones no solo facilita la planificación efectiva en sectores estratégicos, sino que también desempeña un papel fundamental en la prevención y mitigación de los efectos adversos de fenómenos meteorológicos exógenos, como inundaciones y deslizamientos de tierra.

La precipitación pluvial, que consiste en la caída de agua en forma de lluvia, es un fenómeno climático fundamental cuya definición resulta esencial para comprender los patrones climáticos y los riesgos asociados, como las inundaciones urbanas y los cambios en los ecosistemas lacustres. Estudios recientes, en la península de Yucatán, resaltan su variabilidad espacial y temporal, demostrando disminuciones anuales significativas. [1]

La influencia de la precipitación pluvial en las sequías de la región, se estudió por medio de modelos de aprendizaje automático para predecir inundaciones pluviales en Shenzhen, destacando la efectividad de enfoques como las redes neuronales. Además, se exploraron soluciones basadas en la naturaleza para gestionar el riesgo de inundaciones pluviales urbanas con notables avances en la predicción en tiempo real, resaltando la precisión proporcionada por modelos combinados con datos de radar y estaciones meteorológicas, lo que representa un avance clave en la gestión y prevención de inundaciones pluviales. [2]

La humedad, es un factor climático que influye en diversos procesos ecológicos y prácticas agrícolas, es esencial para la gestión y la toma de decisiones efectivas. Diversos estudios han destacado la importancia de la humedad y su complejo comportamiento, las influencias de especies invasoras en la dinámica del agua, [3] la relación entre la humedad y el rendimiento de los cultivos, los impactos en las lesiones cutáneas, [4] las interacciones en la contaminación del suelo y las respuestas fisiológicas de las plantas a la restricción de humedad. Estos estudios revelan la variabilidad y complejidad de la humedad en diferentes contextos, subrayando su importancia en la comprensión de fenómenos climáticos, prácticas agrícolas y procesos ecológicos. [5]

La temperatura desempeña un papel crucial en la predicción del clima, ya que tiene una influencia directa en una amplia variedad de fenómenos climáticos y conlleva importantes implicaciones para los ecosistemas, la agricultura y las actividades humanas. La comprensión de los patrones y variaciones de la temperatura es fundamental para el desarrollo de modelos y pronósticos climáticos precisos. Investigaciones del estado de Goiás y el Distrito Federal en Brasil [6] en São Paulo, han contribuido a la clasificación climática de Köppen-Geiger, revelando variaciones regionales importantes en los regímenes de temperatura. Además, análisis climáticos regionales, como los realizados en Rio Grande do Sul y en Minas Gerais han destacado las variaciones de temperatura y proyectados cambios futuros, proporcionando una visión integral de los patrones climáticos en estas áreas específicas. [7]

El estudio de la presión barométrica en el contexto de la predicción climática es crucial para comprender la relación entre esta variable y los patrones climáticos. Aunque la investigación directa es limitada, hay indicios que sugieren su importancia potencial. Por ejemplo, la presión barométrica disminuye con la baja presión atmosférica, lo que podría estar relacionado con la predicción climática. [8] Similarmente, se destacan el uso de la presión barométrica para evaluar propiedades hidrogeológicas, lo que demuestra su utilidad en la comprensión y predicción de patrones climáticos. [9] Asimismo, se enfatizan la relevancia de considerar la presión barométrica en la observación del agua subterránea para mejorar los modelos climatológicos. [10]

Tradicionalmente, las series temporales han sido estudiadas a través de modelos auto-regresivos debido a su naturaleza de asociación temporal. Muchas aplicaciones han demostrado su eficacia en el modelado de variables relacionadas con el clima. Por ejemplo, se utilizaron estos modelos para analizar la lluvia y la temperatura mensual, [11] mientras que también se aplicaron al estudio de la lluvia mensual. Estos modelos siguen formando parte del toolkit en el contexto de interés [12]. Se llegó a la conclusión de que el modelo de pronóstico basado en ARIMA es un método altamente eficiente, interpretable y confiable para obtener pronósticos regionales de temperatura y precipitación a corto plazo (2-20 años). Estos pronósticos son de gran utilidad en una amplia gama de aplicaciones de ingeniería. [13] Un estudio desarrollado con el modelo SARIMA basado en los datos de la cantidad mensual de lluvia en West Lampung Regenc utilizó la métrica del criterio de información de Akaike (AIC) y logró obtener pronósticos confiables para los próximos 12 meses revelando que estos estarán caracterizados por precipitaciones elevadas y muy elevadas. [14] El modelo ARIMA brinda una predicción precisa de las precipitaciones mensuales para un período de 5 años en el futuro. La predicción final se basó en un modelo ARIMA con el valor AIC más bajo. [15] Se propuso una combinación del modelo ARIMA con el modelo de red neuronal de función de base radial (RBF) para la predicción de precipitaciones mensuales. El enfoque consistió en utilizar el modelo ARIMA para predecir la precipitación mensual y calcular los residuos correspondientes. Posteriormente, se empleó el modelo de red neuronal RBF para aproximar y compensar los resultados de predicción obtenidos del modelo ARIMA, y así corregir los resultados finales de predicción. [16]

A partir de los descrito, es necesario aclaras que los modelos ARIMA especialmente propuestos para el estudio de series temporales combinan información de observaciones pasadas y autocorrelación de datos para generar predicciones en diversos campos. Este modelo se basa en tres componentes principales: autorregresión (AR), integración (I) y media móvil (MA). El componente de autorregresión (AR) se refiere a la dependencia lineal de una observación actual de sus valores pasados, donde el orden de la autorregresión especifica cuántos períodos anteriores se consideran. La integración (I) se refiere a diferenciar la serie temporal para hacerla estacionaria, es decir, eliminar tendencias y hacer que la media y la varianza sean constantes a lo largo del tiempo. La media móvil (MA) se refiere a la dependencia lineal entre una observación actual y un término de error residual de un modelo de media móvil aplicado a valores pasados de la serie temporal. [17]

Estos modelos han demostrado ser efectivos en la predicción de precipitaciones, capturando patrones temporales y tendencias, como lo demuestran estudios, [18], donde obtuvieron resultados satisfactorios con bajos errores de predicción. Asimismo, la precisión de los modelos ARIMA al predecir variaciones estacionales y tendencias a largo plazo en datos de precipitaciones. No obstante, aunque los modelos ARIMA han sido una herramienta útil para modelar y predecir series temporales, tiene limitaciones frente a los modelos de

aprendizaje automático. Su incapacidad para manejar relaciones no lineales y complejas, así como su falta de flexibilidad en la inclusión de múltiples variables predictoras, lo hace menos efectivo en comparación con modelos de aprendizaje automático como el modelo XGboost, que permite capturar patrones más complejos y no lineales en los datos.

El modelo XGBoost (Extreme Gradient Boosting) es un algoritmo de aprendizaje automático utilizado para realizar tareas de clasificación y regresión. Se basa en una técnica que utiliza árboles de decisión y ha demostrado ser altamente efectiva en diversas aplicaciones. La formulación matemática del modelo XGBoost es de naturaleza compleja. Se basa en la combinación de árboles de decisión débiles, generalmente de profundidad limitada, para construir un modelo más robusto. El algoritmo utiliza la técnica de "boosting" para mejorar de manera iterativa la precisión del modelo. En el ámbito de la predicción de precipitaciones, se ha reconocido en diversos estudios que el uso del modelo XGBoost es una herramienta adecuada y efectiva para el modelado de las precipitaciones. Mediante un modelo multivariado de predicción de lluvias utilizando la técnica de XGBoost. Para ello, utilizaron datos meteorológicos históricos de 7 años. Los resultados obtenidos mostraron que el modelo es capaz de generar predicciones precisas para estimaciones diarias de lluvia. [19] Se llevó a cabo una investigación sobre la idoneidad de los modelos de machine-learning para predecir el volumen de lluvia en un período de 8 horas. Para ello, utilizaron datos de series temporales de cinco ciudades importantes del Reino Unido. Compararon diferentes modelos de pronóstico, incluyendo redes LSTM-Networks, el algoritmo de árbol de decisión XGBoost y un algoritmo resultante del uso de AutoML. [20] El modelo XGBoost es altamente adecuado para pronosticar las precipitaciones en un período de 3 a 5 años, con una precisión del 95%. Esta conclusión se basa en el análisis de datos de los últimos 30 años (1987-2017) en la región de Vishakapattanam, India. [21] Se analizaron varios algoritmos de aprendizaje automático para la predicción de lluvias, incluyendo MLR, FR y XGBoost. Utilizaron datos recopilados de la estación meteorológica de la ciudad de Bahir Dar, Etiopía. Los resultados revelaron que XGBoost demostró ser el algoritmo de aprendizaje automático más adecuado para predecir la cantidad de lluvia diaria, utilizando características ambientales seleccionadas. [22]

El modelo XGBoost, se beneficia enormemente de la optimización bayesiana para ajustar sus hiperparámetros de manera eficiente y efectiva. La optimización bayesiana, es una técnica destacable para ajustar parámetros de modelos de aprendizaje automático que ha contribuido significativamente a mejorar la precisión en la predicción de precipitaciones, de hecho, se ha demostrado mejoras significativas al aplicar la optimización bayesiana en el ajuste de hiperparámetros de un modelo ARIMA [23]. Del mismo modo, el impacto positivo de la optimización bayesiana al encontrar los mejores parámetros de un modelo XGBoost para predecir lluvias. [24] Los principales parámetros del modelo XGBoost incluyen la cantidad de árboles (nrounds), la profundidad máxima del árbol (max_depth), la tasa de aprendizaje (eta), la proporción de características (colsample) y la proporción de individuos (subsample). Estos parámetros son importantes a la hora de controlar la complejidad del modelo, la capacidad de generalización y la velocidad de entrenamiento. El enfoque usado para encontrar los mejores parámetros fue la optimización bayesiana y los algoritmos fueron desarrollados en R con ayuda principal de los paquetes xgboost y RBayessianOptimization.

La estrategia de integrar modelos ARIMA, optimización bayesiana y el modelo XGBoost ha demostrado ser efectiva, la combinación de este enfoque superó a los modelos individuales, logrando una mayor precisión en las predicciones, esta se convierte en esencia en la principal contribución del presente trabajo pues plantea una metodología factible que combina modelos paramétricos y no paramétricos a la hora de pronosticar

precipitaciones diarias. No obstante, existen áreas de conocimiento que aún no han sido exploradas, como la necesidad de investigar la aplicación a largo plazo de estos modelos, la influencia de diferentes variables de entrada y técnicas de preprocesamiento, así como la eficiencia computacional para la predicción en tiempo real. Estos temas requieren atención en futuras investigaciones. [25]

Considerando lo expuesto anteriormente, es importante destacar que la zona de la provincia del Carchi, donde se llevó a cabo el estudio, está fuertemente afectada por los cambios climáticos, como el aumento de las temperaturas promedio, las olas de calor y la excesiva evaporación durante las temporadas de verano. Estos cambios climáticos provocan una producción excesiva de lluvias, lo que afecta la economía del sector agropecuario. Por lo tanto, es crucial conocer las previsiones de precipitaciones para poder implementar acciones y estrategias que mitiguen esta problemática en toda la zona.

Debido a los cambios climáticos actuales, la zona de producción, en particular los cantones de Tulcán, Mira y Huaca, se ve afectada por condiciones climáticas adversas debido a su ubicación geográfica a más de 3200 m.s.n.m. Esto ha resultado en rendimientos de producción cada vez más bajos, asociados con la disminución de nutrientes en el suelo y una disminución en los tiempos de rotación de los cultivos. Como resultado, la productividad de la tierra se ha visto restringida.

Las estaciones meteorológicas de la Universidad Politécnica Estatal del Carchi cuentan con datos de más de 4 años antigüedad, los cuales recopilan más de 9 variables climáticas necesarias para realizar la predicción de precipitaciones. La universidad tiene interés en llevar a cabo predicciones climatológicas particularmente sobre las precipitaciones en los cantones de Mira, Tulcán y Huaca. Sin embargo, debido a la gran cantidad de datos existentes, la universidad no cuenta con la capacidad para liderar un proyecto que pueda predecir las precipitaciones en los 3 cantones, lo que ha ocasionado los problemas mencionados anteriormente.

Esto deja en claro que el principal objetivo de esta investigación es proponer un modelo estadístico para predecir las precipitaciones en una estación meteorológica de la UPEC. Para lograrlo, es necesario cumplir con los siguientes objetivos específicos: Diagnosticar el modelo inicial de los datos meteorológicos a utilizar; Aplicar el modelo estadístico a los datos meteorológicos; y Analizar los resultados obtenidos del modelo estadístico de predicción.

Las variables inicialmente disponibles para este propósito son: WS(ave) con un rango de 0 a 32, WD(ave) con un rango de 0 a 360, WS(max) con un rango de 0 a 5, WS(inst_m) con un rango de 0 a 66.1, WD(inst_m) con un rango de 0 a 354, Max_time con un rando de 0 a 1, Radiacion_solar con un rando de 0 a 4.98, Temperatura con un rango de -1.7 °C a 21.4 °C, Humedad de 21.1 a 99.53, Rainfall de 0 a 17.5 y Bar_press de 713.7 a 722.6. Estas variables fueron delimitadas en un período de tiempo de 01-01-2019 y 28-03-2023 y fue necesario realizar algunas operaciones de preprocesamiento como: 1) agrupar los datos por fecha y suma los valores de cada columna para cada fecha; 2) completar el conjunto de datos con fechas faltantes, asegurando una observación para cada día; 3) interpolar linealmente los valores faltantes en las columnas de datos y 4) actualiza la columna de fecha para que tenga una secuencia de fechas diarias desde la fecha mínima hasta la fecha máxima del conjunto de datos original. Esto con la finalidad de garantizar una secuencia de fechas continua y completa.

Materiales y Métodos

La naturaleza de los datos y los objetivos planteados orientan la investigación hacia un enfoque cuantitativo correlacional. Se emplea la correlación entre las covariables y la variable objetivo para desarrollar un modelo que contribuya a predecir el patrón temporal de las precipitaciones pluviales

Datos, Área de Estudio y Variables

El área de estudio se ubica en Ecuador, en la ciudad de Tulcán, provincia del Carchi. La diversidad climática de Ecuador se manifiesta a través de una vasta variedad de climas, influenciados por diversos factores ambientales que le confieren características peculiares a la climatología del país. Se cuenta con datos diarios de precipitaciones, temperatura, humedad y presión desde enero de 2019 hasta febrero de 2023, lo cual se expone en la Fig 1.

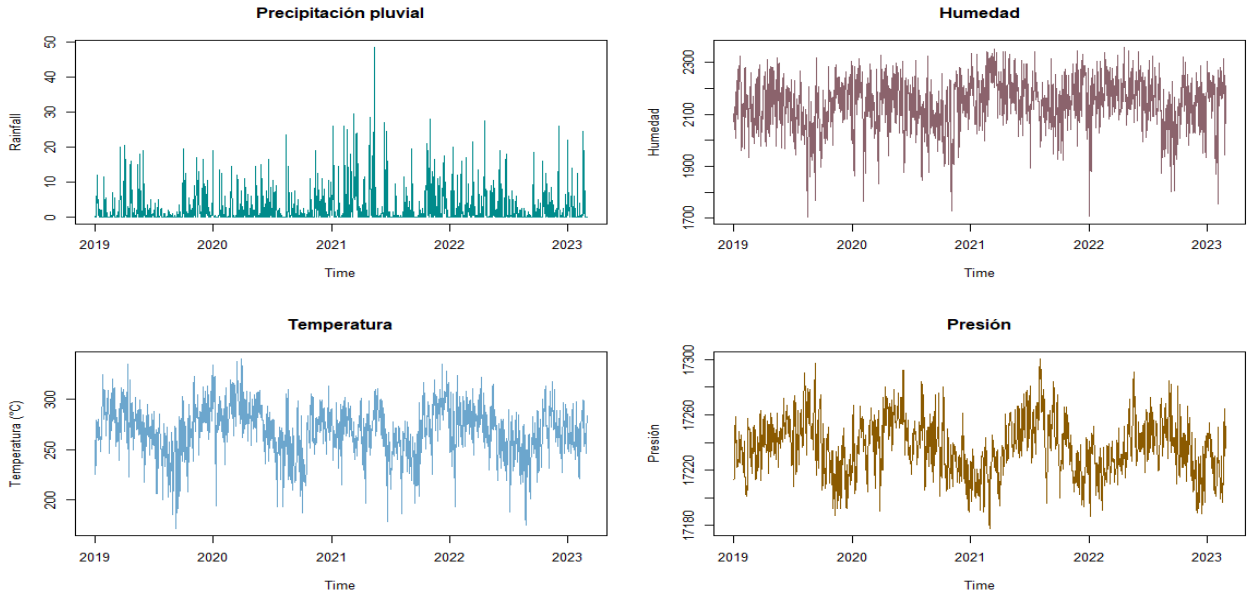

Fig. 1. Disposición temporal de las variables de estudio.

Modelos estadísticos

Modelos ARIMA

El modelo ARIMA, conocido como método de Box-Jenkins y desarrollado por Box y Jenkins (1970), es ampliamente utilizado para el ajuste y pronóstico de series temporales que presentan correlación temporal. [26] Se trata de una clase de modelos en el dominio del tiempo que ha demostrado ser efectiva en este tipo de análisis. Este modelo constituye una técnica estadística ampliamente utilizada para predecir variables aleatorias en series de tiempo. [13]

Se fundamenta en la premisa de que los valores futuros de la serie temporal son influenciados tanto por su propio historial como por las observaciones pasadas Su formulación básica se representa de la siguiente manera:

$$ARIMA(p, d, q) \tag{1}$$

Donde:

p : representa el orden del componente autorregresivo (AR).

d: representa el número de diferenciaciones necesarias para hacer que la serie sea estacionaria.

q: representa el orden del componente de media móvil (MA).

La representación matemática de un modelo ARIMA general es:

$$Y_t = c + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t - \beta_1 \varepsilon_{t-1} - \beta_2 \varepsilon_{t-2} - \cdots - \beta_q \varepsilon_{t-q}, \quad (2)$$

Donde:

Y_t : es el valor de la variable en el tiempo t.

c : es una constante.

$\alpha_1, \alpha_2, \ldots, \alpha_p$: son los coeficientes autorregresivos.

ε_t : es el término de error en el tiempo t.

$\beta_1, \beta_2, \ldots, \beta_q$: son los coeficientes de media móvil.

Modelo XGBoost

La formulación general de XGBoost se puede representar de la siguiente manera:

$$F(x) = \sum h_t(x) \quad (3)$$

Donde:

$F(x)$: es la predicción final del modelo para una instancia x.

$h_t(x)$: representa la predicción del árbol de decisión débil t para la instancia x.

El objetivo principal de XGBoost es minimizar una función de costo regularizada que combina dos componentes; *Función de costo de pérdida*: Esta función mide cuán bien el modelo se ajusta a los datos observados. Puede ser una función de pérdida como la regresión logística para problemas de clasificación o la regresión cuadrática para problemas de regresión; *Término de regularización*: Para evitar el sobreajuste del modelo, se agrega un término de regularización que penaliza la complejidad del modelo (número de nodos terminales en los árboles).

El objetivo es encontrar los árboles de decisión débiles $h_t(x)$ que minimizan esta función de costo. XGBoost también utiliza técnicas como el "Gradient Boosting" y el "Regularization" para mejorar la calidad del modelo y evitar el sobreajuste. En resumen, XGBoost es un poderoso algoritmo de aprendizaje automático que combina múltiples árboles de decisión para realizar predicciones precisas en una variedad de problemas de aprendizaje supervisado.

Optimización Bayesiana

La optimización desempeña un papel crucial en varios campos, incluida la ciencia de datos, donde estamos familiarizados con el ajuste de modelos para mejorar su rendimiento. Los algoritmos de optimización nos permiten obtener una mejor eficiencia de nuestros modelos. Una ventaja particular de la optimización bayesiana radica en su capacidad para lidiar con evaluaciones costosas de la función de aptitud, como ocurre al entrenar algoritmos de aprendizaje automático. En estos casos, resulta justificable invertir recursos adicionales en cálculos más precisos para tomar decisiones más acertadas.

El procedimiento general cuando se trabaja con Optimización Bayesiana se muestra en la Tabla 1.

Tabla 1. Procedimiento general de optimización bayesiana

Algoritmo 1 Pseudocódigo básico para Optimización Bayesiana

Coloque un proceso gaussiano antes en f

Observe f en n_0 puntos de acuerdo con un diseño experimental inicial de relleno de espacio. Establecer $n = n_0$.

Mientras $n \leq N$ **hacer**

Actualice la distribución de probabilidad posterior en f utilizando los datos disponibles

Sea x_n un maximizador de la función de adquisición sobre x. donde la función de adquisición se calcula utilizando la distribución posterior actual

Observar $y_n = f(x_n)$

Incrementar n

Fin mientras

Devuelve una solución: ya sea el punto evaluado con la f(x) más grande, o el punto con la media posterior más grande

La optimización bayesiana se compone de dos elementos principales: un modelo estadístico bayesiano utilizado para modelar la función objetivo y una función de adquisición encargada de determinar dónde se debe tomar la siguiente muestra. El proceso gaussiano es frecuentemente empleado como modelo estadístico debido a su flexibilidad y facilidad de manejo.

Proceso gaussiano

El proceso gaussiano es un modelo utilizado para aproximar la función objetivo, también denominado modelo sustituto. En la estadística bayesiana, cuando nos encontramos con un valor desconocido, suponemos que fue extraído al azar de una distribución de probabilidad previa. En el caso del proceso gaussiano, esta distribución previa se asume como una distribución normal multivariada con un vector medio determinado y una matriz de covarianza.

La distribución priori en $[f(x_1), f(x_2), \dots, f(x_k)]$ es

$$f(x_{1:k}) \sim \mathcal{N}\left(\mu_0(x_{1:k}), \Sigma_0(x_{1:k}, x_{1:k})\right) \quad (4)$$

Donde:

$\mathcal{N}(x, y)$: es la distribución normal

$\mu_0(x_{1:k})$: es la función media de cada x_i. Es común usar $m(x) = 0$ ya que el proceso gaussiano es lo suficientemente flexible como para modelar la media arbitrariamente bien.

$\Sigma_0(x_{1:k}, x_{1:k})$: es la función Kernel o función de covarianza en cada par de x_i .

El proceso gaussiano también proporciona una distribución de probabilidad posterior bayesiana que describe valores potenciales para $f(x)$ en el punto candidato x. Cada vez que observamos f en un nuevo punto, esta distribución posterior se actualiza.

La función de adquisición se emplea para elegir en qué punto de x tomaremos la muestra a continuación. El punto elegido es aquel con el valor óptimo de la función de adquisición. La función de adquisición calcula el valor que se generaría mediante la evaluación de la función de aptitud en un nuevo punto x, en función de la distribución posterior actual sobre f .

Existen diversas opciones para la función de adquisición, como la mejora esperada, el límite de confianza superior del proceso gaussiano, la búsqueda de entropía, entre otros. En este caso, nos centraremos en ilustrar la función de mejora esperada.

$$EI(x) = \begin{cases} (\mu(x) - f(x^+) - \xi)\Phi(z) + \sigma(x)\phi(z), & si\ \sigma(x) > 0 \\ 0, & si\ \sigma(x) = 0, \end{cases} \quad (5)$$

Donde:

$f(x^+)$: Mejor valor de $f(x)$ de la muestra.

$\mu(x)$: Media del predictivo posterior del proceso gaussiano en x.

$\sigma(x)$: Desviación estándar del predictivo posterior del proceso gaussiano en x.

ξ: determina la cantidad de exploración durante la optimización y los valores de ξ más altos conducen a una mayor exploración.

Φ: Función de densidad acumulada (CDF) de la distribución normal estándar.

ϕ: Función de densidad de probabilidad (PDF) de la distribución normal estándar.

Resultados y Discusión

Modelos ARIMA

Modelo ARIMA para la Humedad

El modelo ARIMA (2,1,1) que se muestra en la Figura 2, ofrece varias ventajas para la predicción diaria de la humedad. En primer lugar, al incorporar la información de las dos observaciones anteriores y aplicar una diferencia de orden 1, el modelo puede capturar efectivamente la tendencia y la estacionalidad presentes en los datos, lo que mejora su capacidad para adaptarse a los patrones subyacentes en la serie temporal de humedad. Además, al incluir un componente autorregresivo (AR) de orden 2 y un componente de media móvil (MA) de orden 1, el modelo considera tanto las relaciones lineales entre las observaciones pasadas como la influencia de los errores previos en la predicción actual, lo que proporciona una representación más completa del comportamiento de la humedad. Además, lo que sugiere que el modelo logra capturar la variabilidad no sistemática de los datos. Esta característica es esencial para validar el supuesto de normalidad, ya que los residuos normalmente distribuidos indican que el modelo no deja patrones significativos sin explicar.

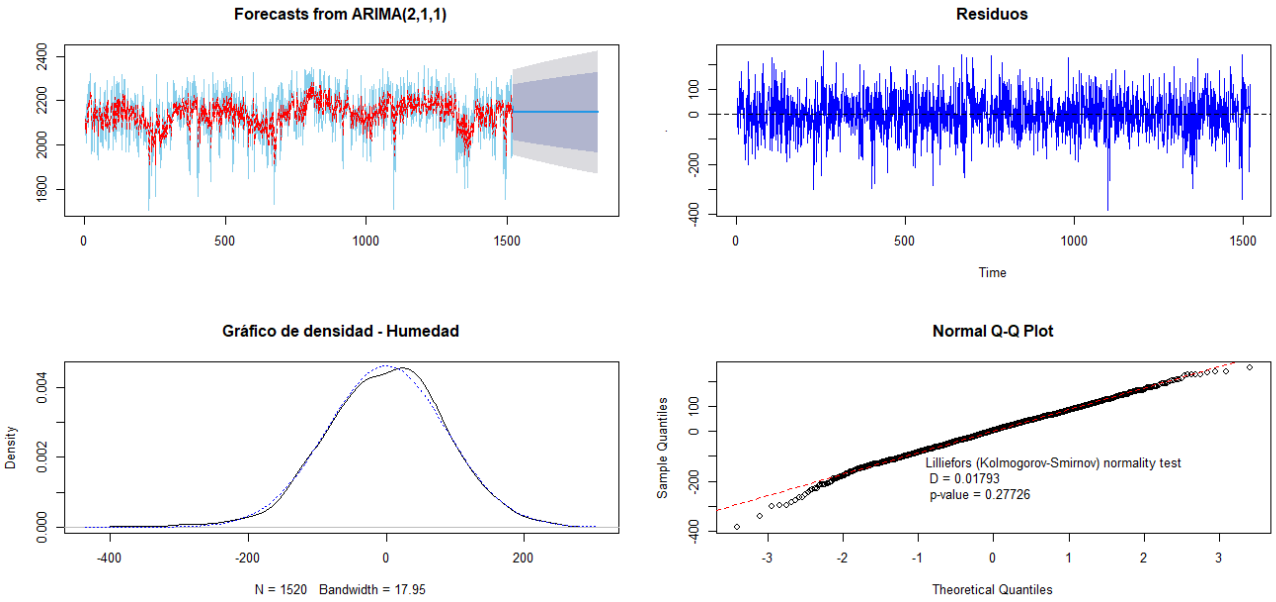

Fig. 2. Modelo ARIMA para la Humedad.

Modelo ARIMA para la Temperatura

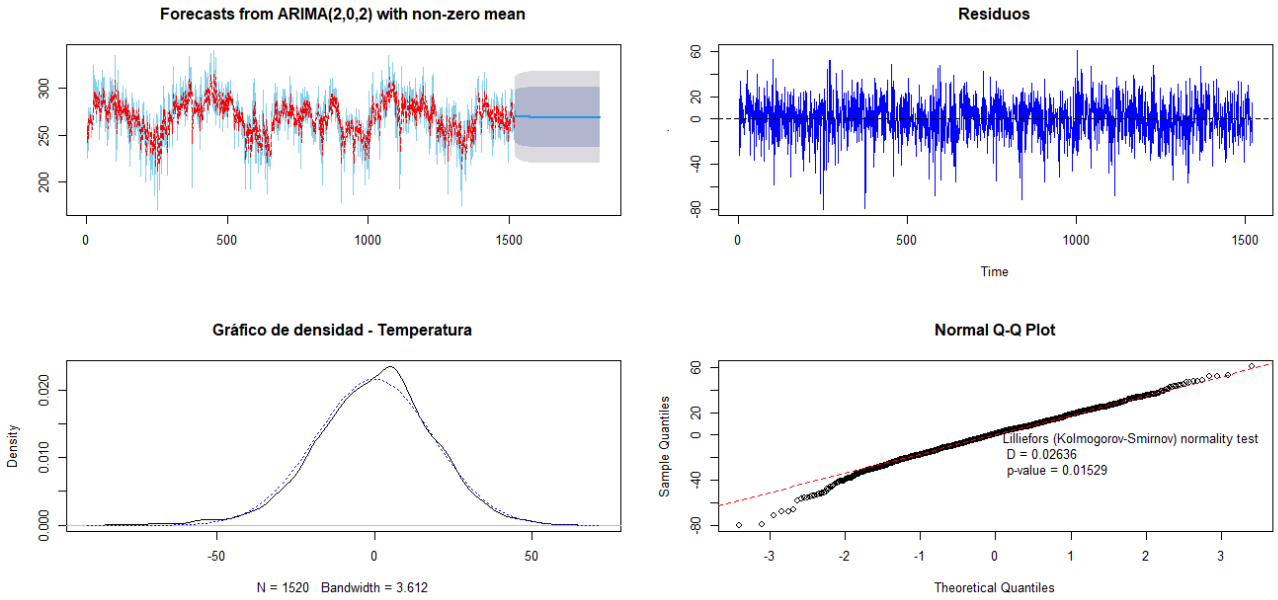

Fig.3. Modelo ARIMA para la Temperatura.

La Figura 3 muestra el comportamiento temporal de la temperatura el cual es modelado cumpliendo con los supuestos de normalidad de los residuos, con un nivel de significancia del 0,01. En este caso, el modelo ARIMA (2,0,2) sirve como una herramienta sólida para la predicción de la temperatura. La capacidad del modelo para generar predicciones coherentes se evidencia en el adecuado comportamiento de los residuos, indicando que logra capturar de manera efectiva la variabilidad no sistemática de los datos. Esta propiedad resulta crucial para validar el supuesto de normalidad, ya que los residuos normalmente distribuidos sugieren que el modelo no deja patrones significativos sin explicar.

Modelo ARIMA para la Presión Barométrica

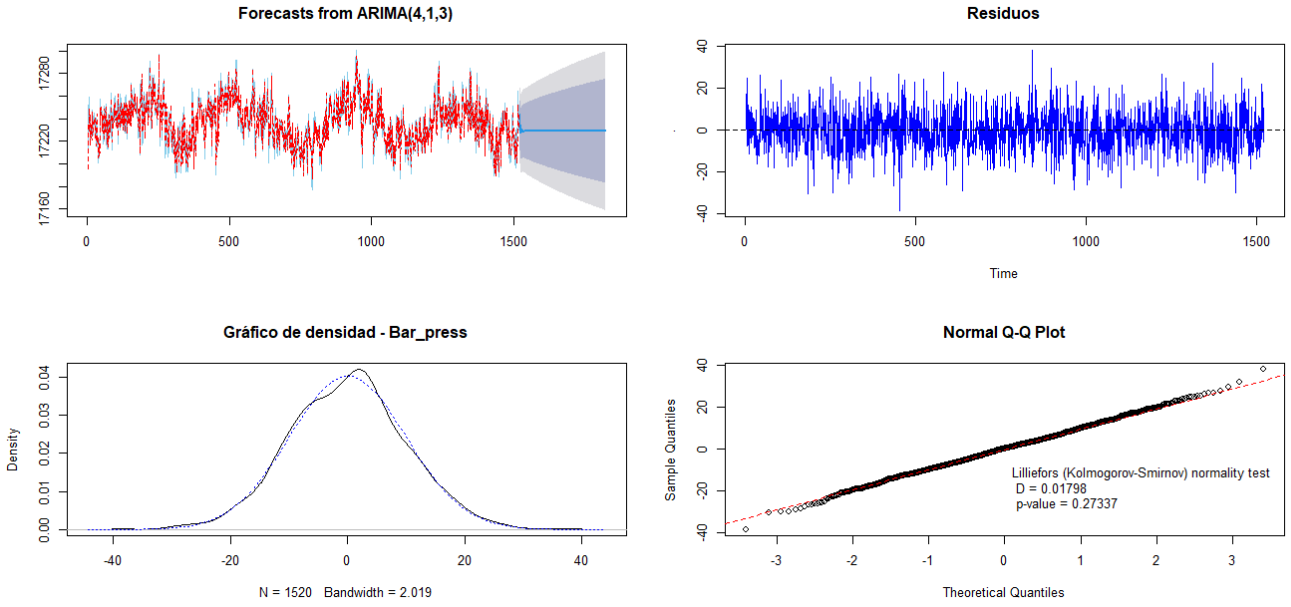

Fig.4. Modelo ARIMA para la Presión Barométrica.

El comportamiento de la presión barométrica permite obtener un modelo con un comportamiento adecuado de los residuos, ya que estos cumplen con el supuesto de normalidad (ver Figura 4). El modelo ARIMA (4,1,3) para la predicción de la presión barométrica exhibe una eficacia notable al capturar patrones complejos en la serie temporal. La inclusión de componentes autorregresivos y de media móvil con órdenes específicos (AR=4, I=1, MA=3) permite al modelo adaptarse de manera dinámica a las fluctuaciones observadas en la presión atmosférica. El análisis de los residuos permite validar el modelo mediante un comportamiento adecuado lo cual sugiere que se logra una representación precisa de la variabilidad no sistemática en los datos.

Optimización Bayesiana – Modelo XGBoost

Predicciones precipitación pluvial

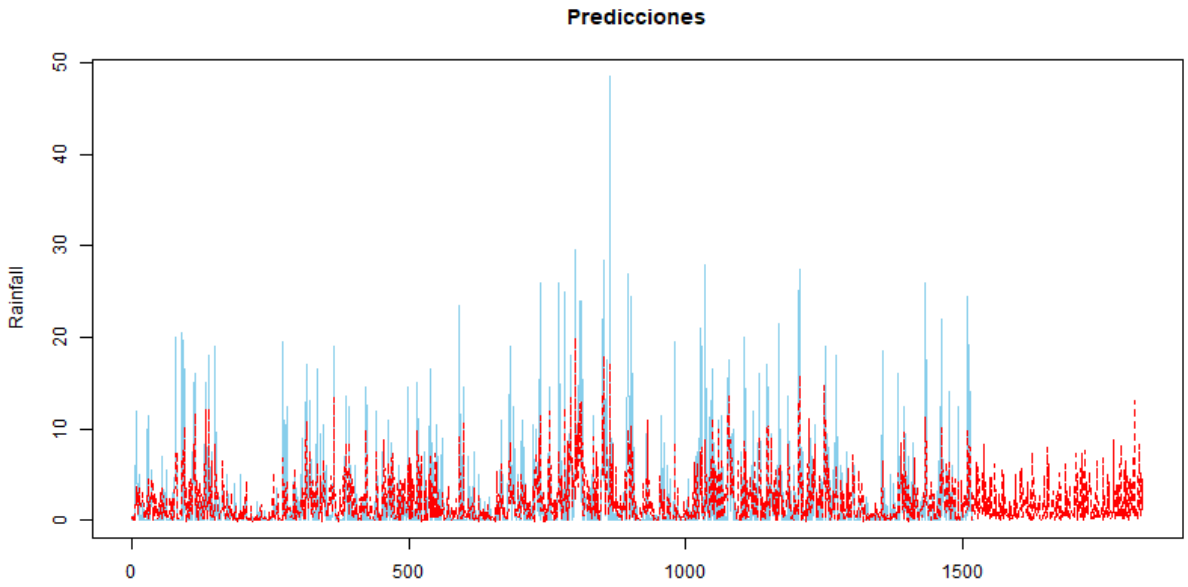

Fig. 5. Predicciones precipitación pluvial

De acuerdo con la Figura 5, el comportamiento temporal de las precipitaciones disponible no admitió un modelo paramétrico para predecir un escenario prospectivo, es por ello que se recurre a modelos no paramétricos de aprendizaje automático que no requieren supuestos de normalidad sobre los residuos. No obstante, es necesario comprobar que estos residuos se encuentren centrados en cero y sea estacionarios, lo cual se confirma con la prueba de Dickey-Fuller el cual arroja p-valor = 0,01.

Dada la naturaleza del comportamiento temporal de las precipitaciones caracterizado por su complejidad y falta de ajuste a modelos paramétricos, el enfoque no paramétrico del modelo XGBoost se presenta como una elección apropiada para la predicción de este fenómeno climático. A diferencia de los modelos paramétricos, XGBoost no depende de supuestos de normalidad sobre los residuos, lo que es particularmente relevante dada la variabilidad inherente y no linealidad de las precipitaciones. Este enfoque no solo permite flexibilidad en la modelización, sino que también proporciona resultados más precisos y adaptativos a la complejidad consustancial de las series temporales de precipitaciones. En el presente estudio, se optimizaron los parámetros del modelo XGBoost por medio de optimización bayesiana y se obtuvo que la combinación óptima de parámetros fue: eta =

0.05, max_depth = 6, sub_sample = 0.43 y colsambple_bytree = 0.3, combinación que permitió minimizar la raíz del error cuadrático medio.

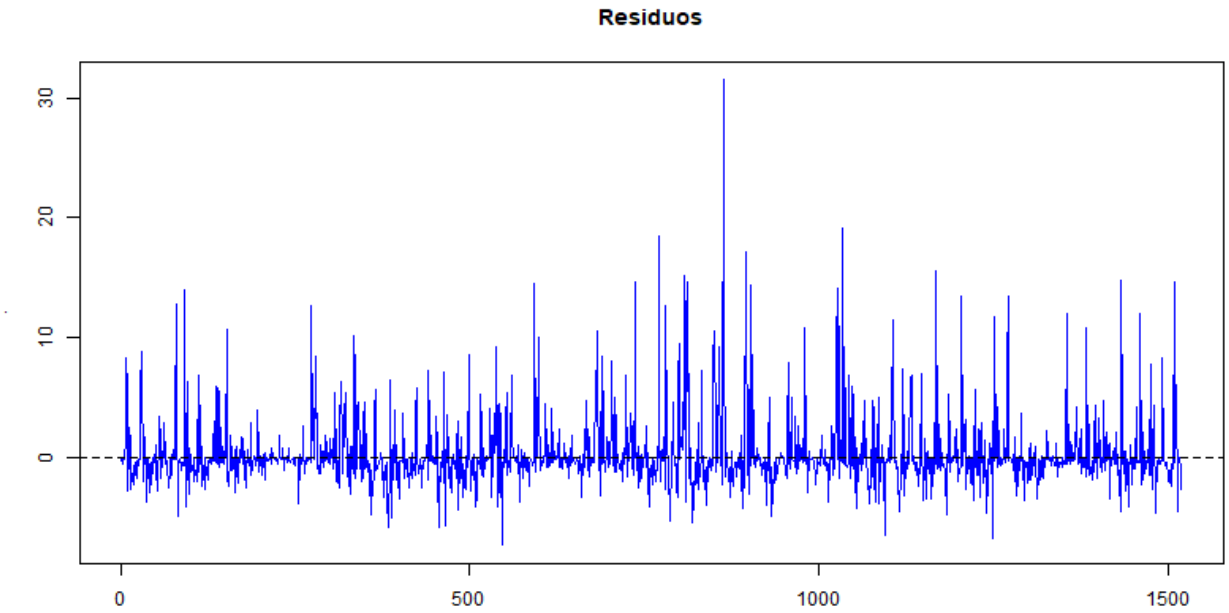

Fig. 6. Predicciones precipitación pluvial.

La Figura 6, muestra el comportamiento de los residuos del modelo XGBoost lo cual es crucial para validar su calidad y precisión. Los residuos son las diferencias entre los valores observados y los valores predichos por el modelo, y un buen comportamiento de los residuos implica que no muestran patrones sistemáticos y se distribuyen aleatoriamente alrededor de cero. Cuando los residuos exhiben este comportamiento, indican que el modelo ha capturado adecuadamente la estructura subyacente de los datos y ha modelado eficazmente las relaciones entre las variables predictoras y la variable objetivo. No obstante, dada los patrones particulares de las precipitaciones se tiene una mayor concentración de los errores por encima del cero. Sin embargo, este comportamiento sugiere que el modelo no está cometiendo errores sistemáticos y no está dejando información relevante sin explicar, lo que aumenta la confianza en las predicciones del modelo. Por lo tanto, el buen comportamiento de los residuos en un modelo XGBoost es un indicador importante de su fiabilidad y validez.

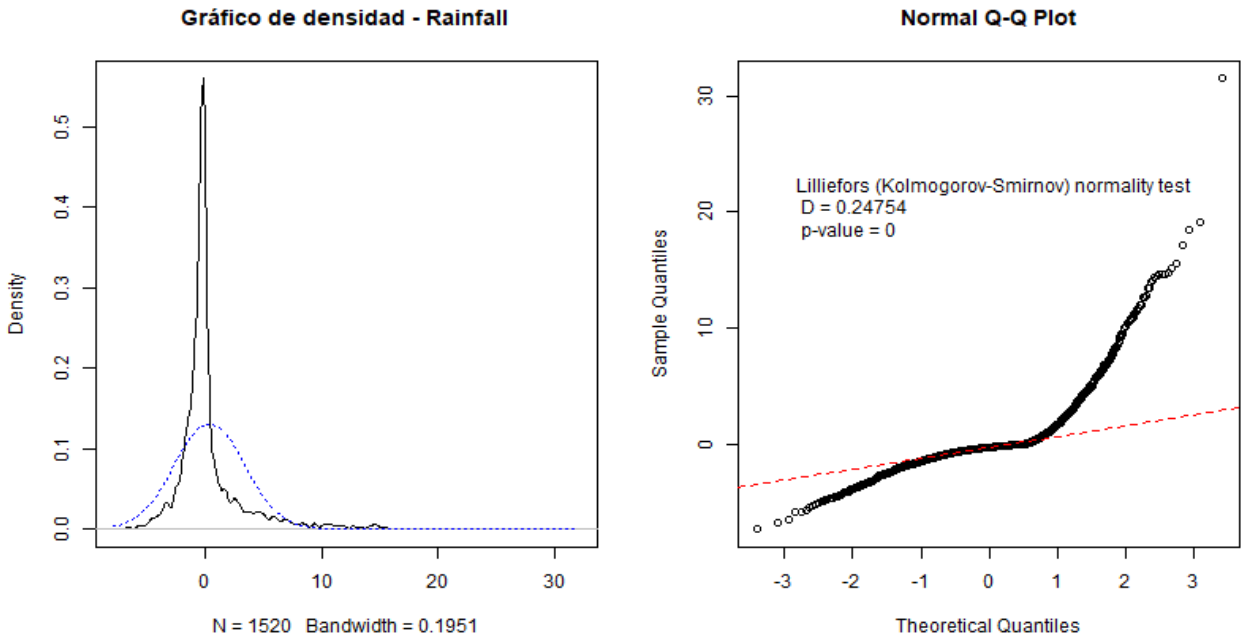

Fig. 7. Predicciones precipitación pluvial.

De acuerdo a la Figura 7, el buen comportamiento de los errores en un modelo XGBoost, incluso si no siguen una distribución normal, radica en su tendencia a acumularse alrededor de cero; es decir, aunque los errores pueden no tener una distribución simétrica típica, si mantienen esta propiedad de centrarse alrededor de cero, indican que el modelo está capturando adecuadamente la tendencia central de los datos y está realizando predicciones consistentes en promedio. Este comportamiento es fundamental para validar la calidad del modelo, ya que sugiere que, a pesar de las desviaciones individuales, el modelo generaliza bien y no comete errores sistemáticos. Por lo tanto, la acumulación de errores en torno al cero en un modelo XGBoost es un indicador crucial de su capacidad para hacer predicciones precisas y confiables.

Discusión de resultados

Las precipitaciones son un fenómeno estacional que ocurre a lo largo de un período de doce meses que muy probablemente dependen de las corrientes de aire. [26] Tradicionalmente, los investigadores han utilizado métodos ARIMA para modelar las precipitaciones en diferentes periodicidades. Sin embargo, este estudio concluye que el modelado y pronóstico de lluvias utilizando algoritmos tradicionales como ARMA, ARIMA, SARIMA, entre otros, no logra ajustarse adecuadamente, especialmente en el comportamiento de los residuos, los cuales no cumplen con los supuestos de normalidad requeridos. Estos hallazgos coinciden con lo mencionado quienes, tras una exhaustiva investigación, optaron por técnicas de machine learning, como el gradiente extremo, para el pronóstico, tal como se ha presentado en el desarrollo del presente trabajo. [21]

Los modelos autorregresivos ARIMA presentan ciertas desviaciones en la precisión de la predicción de las precipitaciones. Sugieren que es necesario combinar un modelo ARIMA con un modelo de red neuronal RBF para mejorar la precisión de las predicciones. En este sentido, proponen la combinación de modelos ARIMA para predecir las variables que influyen en el comportamiento de las precipitaciones, seguido por el uso del modelo XGBoost para generar un escenario predictivo más preciso. [16]

Es importante señalar que, en los distintos estudios citados, se hace referencia al uso de modelos ARIMA para realizar predicciones en una temporalidad de carácter mensual, sin embargo, este modelo no cumple con los requisitos necesarios para una temporalidad diaria. En este sentido, aplicaron varias técnicas de aprendizaje automático y llegaron a la conclusión de que el modelo XGBoost es capaz de predecir de manera consistente las precipitaciones diarias, lo cual se alinea con el objetivo del estudio actual. [22]

Adicionalmente, el modelo GARCH se utiliza comúnmente en el análisis de series de tiempo financieras con la finalidad de captar la volatilidad de los rendimientos, a pesar de que su aplicación se extiende a otras áreas no resultó ser útil para predecir precipitaciones debido a que estas se encuentran influenciadas por una variedad de factores meteorológicos complejos. Además, la relación entre estas variables no siempre es lineal, lo que hace que el enfoque de modelado GARCH no sea adecuado para capturar la complejidad de los procesos meteorológicos lo cual sucedió en el presente estudio al evidenciar propiedades no plausibles para el uso de este modelo en los datos disponibles.

El modelo XGBoost presenta ventajas sustanciales sobre los modelos ARIMA, como se ha demostrado en diversos estudios. Por ejemplo, el análisis de [19] resalta la capacidad del XGBoost para generar predicciones precisas de las lluvias diarias, superando las limitaciones de ARIMA en la captura de patrones no lineales y complejos. Además, [20] encontró que XGBoost fue altamente adecuado para predecir el volumen de lluvia, lo que sugiere una mayor flexibilidad y adaptabilidad del modelo XGBoost en comparación con ARIMA. Este hallazgo es respaldado por [21], donde se concluyó que XGBoost es altamente adecuado para pronosticar precipitaciones en un período de 3 a 5 años con una precisión del 95%. Del mismo modo, el estudio [22] reveló que XGBoost demostró ser el algoritmo de aprendizaje automático más adecuado para predecir la cantidad de lluvia diaria, utilizando características ambientales seleccionadas. Estas investigaciones sugieren que XGBoost, con su capacidad para capturar relaciones no lineales de manera efectiva, es una opción superior a ARIMA para la modelización y predicción de fenómenos climáticos como las precipitaciones.

En resumen, la aplicación de modelos ARIMA en estudios previos se ha centrado principalmente en temporalidades mensuales, lo cual puede limitar su capacidad para capturar patrones más finos en datos diarios de precipitación. En contraste, el modelo XGBoost, al ser especialmente robusto para manejar series temporales con frecuencias diarias, destaca por su capacidad de adaptación a la variabilidad inherente en datos climáticos diarios. Este enfoque más granular no solo se alinea mejor con la naturaleza dinámica de las precipitaciones, sino que también subraya la importancia de seleccionar modelos que se ajusten a la escala temporal específica del fenómeno estudiado. La capacidad del XGBoost para predecir consistentemente precipitaciones diarias, como respaldan estudios previos, resalta su relevancia y ventaja sobre enfoques tradicionales en escenarios donde se requiere una mayor resolución temporal para comprender y prever fenómenos meteorológicos.

Finalmente, resulta importante señalar que una de las limitaciones del enfoque utilizado es que, aunque el modelo XGBoost se utiliza para predecir las precipitaciones diarias utilizando datos de temperatura, presión y humedad obtenidos a través de modelos ARIMA, puede haber una complejidad subestimada en la relación entre estas variables y las precipitaciones. Los modelos ARIMA pueden capturar algunas relaciones lineales entre las variables meteorológicas, pero pueden no ser suficientes para reflejar las interacciones

no lineales y complejas que existen en el clima. Esto podría resultar en un sub-ajuste del modelo XGBoost en ciertos períodos de tiempo y, por lo tanto, en predicciones menos precisas de las precipitaciones.

Conclusiones

El comportamiento caótico de las precipitaciones en el tiempo de recolección de datos desde enero del 2019 hasta marzo del 2023 no permite el uso de modelos paramétricos como el ARIMA o GARCH, es por esto que se recurre a técnicas de aprendizaje automático como el XGBoost para encontrar una predicción más cercana al comportamiento de la serie temporal y logrando un rmse de 3,48; cuando en el caso de los modelos ARIMA el error siempre se mantuvo por encima de 4,00. A pesar de no cumplir con el criterio de normalidad en los residuos del modelo de precipitaciones, supuesto que no es necesario en modelos de aprendizaje automático, se pudo verificar que estos se centran en cero y su comportamiento no varía en el tiempo.

El uso del modelo XGBoost, especialmente cuando se ha ajustado mediante optimización bayesiana, demuestra ser una estrategia altamente efectiva para la predicción de variables climáticas, como las precipitaciones. La optimización bayesiana permite encontrar de manera eficiente los hiperparámetros óptimos del modelo, maximizando su rendimiento predictivo. La capacidad constitutiva de XGBoost para manejar la complejidad no lineal de las series temporales se ve potenciada por este enfoque, ya que se ajusta de manera precisa a las características específicas de los datos meteorológicos. La combinación de XGBoost y optimización bayesiana se presenta como una estrategia avanzada y significativa para abordar la complejidad de las series temporales climáticas. Esta combinación ofrece resultados robustos y mejora significativamente la capacidad predictiva del modelo, lo que la convierte en una opción muy efectiva para este tipo de análisis.

Agradecimientos:

Me gustaría expresar mi más sincero agradecimiento a todos los que contribuyeron al desarrollo y éxito de este artículo científico. Agradezco profundamente a mis colegas y colaboradores por su dedicación y valiosos aportes. También agradezco a las instituciones y recursos que hicieron posible este trabajo. Este logro no hubiera sido posible sin el apoyo constante de mis mentores y la inspiración de la comunidad científica. Cada aporte fue esencial y agradezco sinceramente el apoyo que enriqueció este trabajo. Este trabajo demuestra nuestro compromiso compartido con la investigación y el avance del conocimiento.

Referencias:

1. Bulti, E.; Abebe, T.: A review of flood modeling methods for urban pluvial flood application. Modeling Earth Systems and Environment, vol. 14, No. 6, pp. 1293-1302 (2020)

2. Huang, Y.: Nature□based solutions for urban pluvial flood risk management. Wiley Interdisciplinary Reviews Water, vol. 6, No. 4, pp. 121-146 (2020)

3. Poveda, G.: La hidroclimatología de colombia una síntesis desde la escala inter-decadal hasta la escala diurna. Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales, vol. 28, No. 107, pp. 255-272 (2023)

4. Nasca, L.: Efectos de la invasión del ligustrum, Ligustrum lucidum, sobre la dinámica hidráulica de los Yungas del noroeste argentino. Silvicultura, vol. 35, No 5, pp. 195-215 (2014)

5. Zahermand, S.; Vafaeian, M.; Bazyar, M.: Análisis de las propiedades físicas y químicas de suelos contaminados con hidrocarburos oleosos, Revista de investigación de ciencias de la tierra, vol. 24, No 5, pp. 163-168 (2020)

6. Cardoso, M.; Marcuzzo, M.; Barros, J.: Clasificación Climática de Köppen-Geiger para el Estado de Goiás y el Distrito Federal. Minutas geográficas, vol. 5, No. 6, pp. 122-151 (2015)

7. Cunha, A.; Martins, D.: Clasificación climática para los municipios de botucatu y são manuel, Revista Cientifica Sinopsis, vol. 14, No. 3, pp. 75-92 (2018)

8. Acworth, R.: Comprensión de los sistemas conectados de aguas superficiales y subterráneas mediante el análisis de Fourier de las fluctuaciones diarias y subdiarias de la carga, Revista de hidrogeología, vol. 23, No. 6, pp. 143-159 (2015)

9. Fileccia, A.: Utilizar las fluctuaciones del nivel del agua en respuesta a los cambios de la marea terrestre y la presión barométrica para medir las propiedades hidrogeológicas in situ de un acuífero sobrecargado en un yacimiento de carbón, Revista de Hidrogeología, vol. 26, No. 14, pp. 1465-1479 (2020)

10. Kamp, G.; Schmidt, R.:Revisión: Carga de humedad: la información oculta en los registros de pozos de observación de aguas subterráneas, Revista de Hidrogeología, vol. 32, No. 25, pp. 2225-2233 (2017)

11. Kaushik, I.; Singh, S.: Seasonal ARIMA Model for Forecasting of Monthly Rainfall and Temperature, J. Environ. Res. Dev., pp. 506-514 (2008)

12. Mashin, M.; Begum, M.: Modeling Rainfall in Dhaka Division of Bangladesh Using Time Series, ResearchGate, Vol. 1, No.5, pp. 67-73 (2012)

13. Lai, Y.; Dzombak, D.: Use of the Autoregressive Integrated Moving Average (ARIMA) Model to Forecast Near-term Regional Temperature and Precipitation, Weather and Forecasting (2020)

14. Nusyirwan, L.; Afriani, F.: Forecasting Monthly Rainfall Using Arima Seasonal Method: Case Study: Rainfall in West Lampung Regency, International Journal of Statistics and Applications, vol. 12, No. 4, pp. 83-86 (2022)

15. Bora, S.; Hazarika, A.: Rainfall Time Series Forecasting using ARIMA Model, 2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1) (2023)

16. Zhao, J.; Chen, R.; Xin, H.: Rainfall study based on ARIMA-RBF combined model, in 5th International Symposium on Big Data and Applied Statistics (ISBDAS 2022), Journal of Physics: Conference Series (2022)

17. Shumway,H.; Stoffer, S.: Time Series Analysis and Its Applications: With R Examples, 4th ed. Davis, CA: Springer Nature (2017)

18. Smith, J.: Predicción de precipitaciones mensuales mediante modelos ARIMA, Revista de Meteorología y Climatología Aplicadas, vol. 5, No. 6, pp. 55-73 (2018)

19. Anwar, M.; Winarno, E.; Hadikurniawati, W.: Rainfall prediction using Extreme Gradient Boosting," Journal of Physics: Conference Series. Annual Conference on Science and Technology (ANCOSET 2020), (2021)

20. Barrera, L.; Oyedele, M.; Akinosho, T.; Davila, L.: Rainfall prediction: A comparative analysis of modern machine learning algorithms for time-series forecasting, Machile Learning with Applications, (2022)

21. Poola, K.; Sekhar, H. Prediction of rainfall by using estreme gradient boost (XG boost) in Vishakapattanam area, Andhra Pradesh, International Journal of Statistics and Applied Mathematics, vol. 6, No. 3, pp. 83-86 (2021)

22. Liyew, M.; Melese, A.: Machine learning techniques to predict daily rainfall amount, Journal of Big Data, vol. 8, No. 153, p. 11 (2021)

23. Wang, Q.: Optimización bayesiana para el ajuste de parámetros del modelo XGBoost en la predicción de lluvias, Investigación de recursos hídricos, vol. 6, No. 2, pp. 91-112 (2017)

24. Li, C.; Shang, H.: Optimización bayesiana para el ajuste de parámetros del modelo XGBoost en la predicción de lluvias, Investigación de recursos hídricos, vol. 5, No. 6, pp. 67-82 (2020)

25. Chen, L.; Zhang, Y.: Predicción de lluvia integrada utilizando ARIMA, optimización bayesiana y modelo XGBoost, Revista de Hidrología, vol. 7, No. 4, pp. 145-176, (2019)

26. Rahman, A.; Mahmudul, H.:Modeling and Forecasting of Carbon Dioxide Emissions in Bangladesh Using Autoregressive Integrated Moving Average (ARIMA) Models, Technology Sylhet, Vol.7 No.4, pp. 12-22 (2017)

27. Paredes, F.; Guevara, E.; Barbosa, H.; Uzcátegui, C. et al.: Tendencia de la precipitación estacional e influencia de El Niño-Oscilación Austral sobre la ocurrencia de extremos pluviométricos en la cuenca del lago de Valencia, Venezuela, Tecnología y Ciencias del Agua, vol. 6, No. 6, pp. 33-48 (2015)

28. INAMHI.: Boletín Agrometereológico Mensual, Instituto Nacional de Meteorología e Hidrología, Quito, (2017)

29. Santillán, K. y Zamora, B.: Análisis Climático y de Dambio Climático en el Distrito Metropolitano de Quito, Universidad Politécnica Salesiana sede Quito, Quito, (2021)