Elaboración de un modelo predictivo de temperatura en Barrancabermeja mediante machine learning

Norberto Ariosto Quintero Garavito

Asesor

José Laureano Cruz Cardozo

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

Resumen

La investigación se centra en predecir los cambios futuros en la temperatura del aire en la ciudad de Barrancabermeja de acuerdo a registros históricos de variables meteorológicas como temperatura, presión atmosférica, humedad relativa y velocidad del viento. Se implementaron y compararon cuatro modelos distintos: ARIMAX (que incluye las variables exógenas), Prophet, Random Forest y Gradient Boosting. Para cada uno de estos modelos se dividio el conjunto datos en dos partes: Train (entrenamiento): Se utilizó el 80% de los datos, para entrenar el modelo y Test (prueba): se reservaron entre el 20% de los datos para evaluar el rendimiento y la capacidad de generalización de cada modelo.

Palabras clave**:** Machine learning, temperatura, datos climáticos, cambio climático, modelos de aprendizaje automático.

Abstract

The research focuses on predicting future changes in air temperature in the city of Barrancabermeja based on historical records of meteorological variables such as temperature, atmospheric pressure, relative humidity, and wind speed. Four different models were implemented and compared: ARIMAX (which includes exogenous variables), Prophet, Random Forest, and Gradient Boosting. For each of these models, the data set was divided into two parts: Train: 80% of the data was used to train the model; and Test: 20% of the data was reserved to evaluate the performance and generalization capabilities of each model.

***Keywords*:** Machine learning, temperature, climate data, climate change, machine learning models.

Tabla de Contenido

Introducción 8

Planteamiento del Problema 10

Sistematización del Problema 12

Justificación 14

Objetivos 17

Objetivo General 17

Objetivos Específicos 17

Marco de Referencia 18

Estado del Arte 18

Marco Contextual 19

Marco Teórico 21

Marco Conceptual 23

Marco Normativo 24

Metodología 26

Tipo de Estudio 30

Recolección de Datos 32

Selección de la Estación Meteorológica para el Estudio 32

Resultados 33

Identificación de Datos Atípicos 34

Análisis Exploratorio (EDA) 37

Ingeniería de Características 41

Prueba de Estacionariedad (ADF) 41

Prueba de Autocorrelación 44
Modelado 46
Modelado Predictivo Mediante ARIMAX con Variables Exógenas 46
Modelado Predictivo Mediante Prophet con Variables Exógenas 51
Modelado Predictivo Mediante Random Forest Para Pronóstico de Temperatura 54
Implementación y Optimización del Modelo XGBoost para Pronóstico de Temperatura.... 55
Evaluación y Comparación de Modelos. 57
Evaluación de Modelos de Pronóstico de Temperatura 57
Visualización Gráfica de Predicciones..... 58
Prueba de Wilcoxon Comparación Estadística de Modelos..... 59
Despliegue del Modelo Óptimo y Generación de Pronósticos 61
Procedimiento para la Generación de Pronósticos 61
Salida del Modelo..... 61
Conclusiones.....63
Recomendaciones 65
Referencias.....66

Lista de Tablas

Tabla 1 *Marco Conceptual: Variables y Parámetros Analizados* 23

Tabla 2 *Variables Analizadas en el Estudio* 27

Tabla 3 *Algoritmos Seleccionados Para la Modelización Predictiva* 28

Tabla 4 *Métricas de Evaluación Comparativa de los Modelos Predictivos* 57

Tabla 5 *Resultados de la Prueba de Wilcoxon Para Comparación de Rendimiento* 60

Tabla 6 *Pronóstico de Temperatura a Tres Días con Prophet: Valores Centrales y Rangos de Confianza* 61

Lista de Figuras

Figura 1 *Estaciones Meteorológicas del IDEAM en Barrancabermeja* 32

Figura 2 *Boxplot de Variables Antes de Eliminación de Datos Atípicos*..... 34

Figura 3 *Boxplots de Variables Después de la Eliminación de Datos Atípicos* 36

Figura 4 *Variación de Temperatura (°C) en Barrancabermeja por Fecha de Observación* 37

Figura 5 *Matriz de Correlación: Temperatura, Presión, Humedad y Velocidad del Viento* 39

Figura 6 *Descomposición de Serie Temporal: Temperatura (2006-2024) – Tendencia, Estacionalidad y Residuos* 42

Figura 7 *Función de Autocorrelación (ACF) de Diferencias Estacionales*..... 44

Figura 8 *Resultados del Modelo SARIMAX(1,1,2): Buen Ajuste con Residuales Independientes pero No Normales* 47

Figura 9 *Diagnóstico de Residuos del Modelo ARIMAX: Normalidad, Autocorrelación y Patrones Residuales*..... 49

Figura 10 *ACF de Residuos ARIMAX: Validación de Independencia (Ruido Blanco) en 40 Lags* 50

Figura 11 *Variabilidad Climática 2005-2025: Análisis de Serie Temporal y Patrones Dominantes* 52

Figura 12 *Componente Estacional de Serie Temporal: Oscilación Térmica Anual (Datos Normalizados)*..... 53

Figura 13 *Desempeño Relativo de Cuatro Enfoques en Predicción de Temperatura Global*..... 58

Introducción

El cambio climático y la variabilidad meteorológica representan desafíos significativos para diversas regiones del mundo, afectando sectores clave como la agricultura, la salud pública y la gestión de recursos hídricos. En Colombia, ciudades como Barrancabermeja, ubicada en la región del Magdalena Medio, experimentan condiciones climáticas particulares influenciadas por su geografía y dinámicas atmosféricas locales. La capacidad de predecir con precisión las temperaturas futuras en esta zona permitiría una mejor planificación de actividades económicas y medidas de adaptación ante fenómenos extremos.

En este contexto, el desarrollo de modelos predictivos basados en técnicas de *machine learning* surge como una herramienta poderosa para analizar patrones climáticos históricos y proyectar tendencias futuras. A diferencia de los métodos tradicionales de pronóstico meteorológico, los algoritmos de aprendizaje automático pueden capturar relaciones no lineales entre múltiples variables climáticas (como presión atmosférica, humedad relativa y velocidad del viento) y la temperatura, mejorando así la exactitud de las predicciones.

Esta investigación se centra en la construcción y evaluación de modelos predictivos para la temperatura atmosférica en Barrancabermeja, utilizando datos históricos proporcionados por el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM). Se implementarán cuatro enfoques principales: ARIMAX (un modelo estadístico que incorpora variables exógenas), PROPHET (desarrollado por Meta para series temporales), Random Forest y Gradient Boosting (ambos basados en ensambles de árboles de decisión). Cada modelo será entrenado, optimizado mediante ajuste de hiperparámetros y validado con métricas de precisión como el Error Cuadrático Medio (MSE), el Error Medio Absoluto (MAE) y el Coeficiente de Determinación (R^2).

El objetivo final es comparar el rendimiento de estos algoritmos e identificar el más adecuado para pronosticar temperaturas en Barrancabermeja, considerando las particularidades climáticas de la región. Los resultados de este trabajo no solo contribuirán al avance de las aplicaciones de *machine learning* en meteorología local, sino que también podrán ser utilizados por entidades gubernamentales y actores locales para la toma de decisiones informadas ante escenarios de cambio climático.

Esta tesis se estructura en cinco capítulos: (1) introducción y planteamiento del problema, (2) revisión de literatura y marco teórico, (3) metodología y procesamiento de datos, (4) experimentación y resultados, y (5) conclusiones y recomendaciones. A través de este estudio, se espera sentar las bases para futuras investigaciones en modelado climático regional con técnicas de inteligencia artificial.

Planteamiento del Problema

Barrancabermeja, situada en la región del Magdalena Medio colombiano (6°18′N, 73°48′O), presenta un clima tropical marcado por una alta variabilidad térmica, influenciada por factores como su proximidad al río Magdalena, la intensa actividad industrial petrolera y los patrones climáticos regionales (IDEAM, 2020). Aunque la ciudad cuenta con estaciones meteorológicas del IDEAM que recopilan datos históricos de variables como temperatura, humedad relativa, presión atmosférica y velocidad del viento, aún no se han desarrollado modelos predictivos locales basados en machine learning (ML) que aprovechen esta información para generar pronósticos precisos de temperatura a corto y mediano plazo.

En la actualidad, los pronósticos climáticos para la región dependen de modelos globales o nacionales con baja resolución espacial, los cuales tienen dificultades para capturar microclimas y eventos extremos locales (López et al., 2019). Esta limitación impacta negativamente en diversos sectores clave. En salud pública, por ejemplo, se ha observado un aumento en enfermedades relacionadas con olas de calor (OMS, 2018). En la industria energética, las fluctuaciones térmicas generan ineficiencias en procesos de refinación de petróleo altamente sensibles a estos cambios (Agencia Internacional de Energía, 2019). Asimismo, en el sector agrícola, el estrés térmico ha provocado pérdidas significativas en cultivos (Fedepalma, 2021).

Ante este escenario, surge la pregunta central: ¿cómo desarrollar un modelo predictivo de temperatura atmosférica para Barrancabermeja utilizando algoritmos de ML que superen en precisión a los métodos tradicionales, integrando variables climáticas históricas del IDEAM? Para abordar este desafío, es necesario resolver varios aspectos clave. En primer lugar, se debe evaluar si los datos históricos disponibles, como series temporales horarias o diarias, son

suficientes y consistentes para entrenar modelos de ML de manera efectiva. Además, será crucial determinar qué algoritmo—entre opciones como ARIMAX, PROPHET, Random Forest o Gradient Boosting—ofrece los mejores resultados al predecir la temperatura considerando variables complementarias como la humedad y la presión atmosférica. Por último, será fundamental validar si el modelo seleccionado puede integrarse eficazmente en sistemas de alerta temprana que apoyen la toma de decisiones locales.

En este contexto, se plantea la hipótesis de que los modelos basados en ensambles, como Random Forest y Gradient Boosting, superarán en precisión predictiva a los métodos estadísticos tradicionales, como ARIMAX y PROPHET, debido a su capacidad para capturar relaciones no lineales entre las variables climáticas propias de Barrancabermeja.

Sistematización del Problema

Para abordar el desafío de predecir la temperatura en Barrancabermeja con mayor precisión, se ha diseñado una estrategia integral que combina aspectos técnicos, metodológicos y aplicados. El punto de partida consiste en utilizar series temporales históricas (2005-2025) proporcionadas por el IDEAM, las cuales incluyen variables clave como temperatura, humedad, presión atmosférica y velocidad del viento. Sin embargo, un reto inicial radica en posibles brechas o inconsistencias en los datos debido a fallas técnicas en las estaciones meteorológicas, lo que requerirá un manejo cuidadoso para garantizar su confiabilidad.

En el ámbito técnico, se compararán diferentes enfoques de modelado para identificar el más adecuado. Por un lado, se evaluarán métodos estadísticos tradicionales como ARIMAX, reconocido por su capacidad para incorporar variables exógenas en la predicción (Hyndman & Athanasopoulos, 2018), y PROPHET, un algoritmo diseñado especialmente para capturar patrones estacionales en series de tiempo (Taylor & Letham, 2018). Por otro lado, se analizarán técnicas más avanzadas basadas en ensambles de árboles de decisión, como Random Forest (Breiman, 2001) y Gradient Boosting (Friedman, 2001), las cuales destacan por su habilidad para modelar relaciones complejas y no lineales en los datos.

Desde la perspectiva metodológica, el proceso seguirá una secuencia rigurosa. Primero, los datos serán sometidos a un cuidadoso preprocesamiento que incluirá la imputación de valores faltantes, la detección y manejo de valores atípicos, así como técnicas de normalización y feature engineering para optimizar su calidad. Posteriormente, se dividirá el conjunto de datos en un 80% para entrenamiento y un 20% para pruebas, asegurando una evaluación objetiva del modelo. Durante la fase de entrenamiento, se optimizarán los hiperparámetros de cada algoritmo para maximizar su rendimiento, utilizando métricas como el Error Cuadrático Medio (MSE), el

Coeficiente de Determinación (R^2) y el Error Absoluto Medio (MAE) para comparar su precisión.

Finalmente, la dimensión aplicada del proyecto busca traducir estos esfuerzos técnicos en beneficios concretos para la región. Se desarrollará un prototipo funcional capaz de generar pronósticos de temperatura con antelaciones de 24 a 72 horas, el cual podrá ser utilizado por el gobierno local para mejorar la planificación ante eventos como olas de calor, mitigando así riesgos en salud pública. Además, el sector industrial, particularmente el energético, podrá optimizar sus procesos al anticipar fluctuaciones térmicas, reduciendo ineficiencias y costos. De esta manera, el proyecto no solo aportará soluciones tecnológicas innovadoras, sino que también tendrá un impacto tangible en la calidad de vida y la economía de Barrancabermeja.

Justificación

Esta investigación surge en la intersección entre la meteorología computacional y el aprendizaje automático aplicado, buscando responder a una necesidad apremiante en el campo de la predicción climática local. Su relevancia se fundamenta en múltiples dimensiones que van desde los avances científicos hasta el impacto concreto en la comunidad de Barrancabermeja.

En el ámbito científico y tecnológico, los modelos tradicionales de pronóstico del tiempo —basados en sistemas de ecuaciones diferenciales para la dinámica de fluidos atmosféricos (Holton, 2004, p. 45)— enfrentan limitaciones significativas cuando se aplican a escalas locales, debido tanto a su elevado costo computacional como a su tendencia a generalizar patrones espaciales. Frente a estas restricciones, las técnicas de machine learning han emergido como una alternativa prometedora, demostrando una capacidad superior para identificar patrones no lineales en datos climáticos (Reichstein et al., 2019). Este estudio explora precisamente esta ventaja, evaluando el desempeño de algoritmos como ARIMAX (Hyndman & Athanasopoulos, 2018), PROPHET (Taylor & Letham, 2018), Random Forest (Breiman, 2001) y Gradient Boosting (Friedman, 2001) en la tarea de modelar la temperatura de Barrancabermeja a partir de variables clave como la presión atmosférica, la humedad relativa y la velocidad del viento.

La pertinencia regional de esta investigación es innegable. Barrancabermeja, ubicada en el corazón del Magdalena Medio (6° 18′ N, 73° 48′ O), posee un clima tropical monzónico (clasificación Am según Köppen) fuertemente influenciado por la presencia del río Magdalena y las actividades industriales que alteran sus microclimas (IDEAM, 2020). A pesar de su importancia estratégica para sectores como el petrolero y el agroindustrial, la región carece de modelos predictivos locales basados en machine learning que aprovechen los datos históricos recopilados por el IDEAM. Esta ausencia limita severamente la capacidad de anticipar y

adaptarse a eventos extremos, como las olas de calor cuya frecuencia e intensidad se han incrementado debido al cambio climático (IPCC, 2021).

El impacto social y económico potencial de este trabajo es considerable. En el ámbito de la gestión de riesgos, contar con predicciones precisas de temperaturas extremas permitiría reducir vulnerabilidades en salud pública, especialmente para aquellas poblaciones que, por sus labores al aire libre, están más expuestas a estos fenómenos (OMS, 2018). Para el sector energético, empresas como Ecopetrol podrían optimizar sus procesos de refinación —donde variaciones térmicas superiores a 2°C generan incrementos significativos en los costos operativos (Agencia Internacional de Energía, 2019)—. Asimismo, la agricultura regional, que depende en gran medida de cultivos sensibles a los cambios térmicos como el plátano y la yuca (Fedepalma, 2021), se beneficiaría de pronósticos más certeros.

Además, esta investigación se alinea con políticas públicas de gran relevancia. Por un lado, responde a los objetivos planteados en el Plan Nacional de Adaptación al Cambio Climático (PNACC) de Colombia (MinAmbiente, 2022), que prioriza el desarrollo de herramientas tecnológicas para la gestión climática a escala local. Por otro, contribuye directamente al cumplimiento del Objetivo de Desarrollo Sostenible 13 (Acción por el Clima), específicamente a la meta 13.3 que busca "mejorar la educación y la capacidad humana e institucional en mitigación y adaptación al cambio climático" (ONU, 2015).

Finalmente, el estudio introduce innovaciones metodológicas que lo distinguen de investigaciones previas realizadas en Colombia (como el trabajo de López et al., 2019 centrado en Bogotá). Entre estas destacan el uso de datos de alta resolución temporal (horaria y diaria) provenientes de estaciones meteorológicas del IDEAM en Barrancabermeja, así como la comparación sistemática entre modelos híbridos (ARIMAX y PROPHET) y enfoques basados en

ensambles (Random Forest), estos últimos poco explorados en el contexto de la climatología regional. Esta combinación de rigor científico, relevancia local y potencial transformador convierte a esta investigación en un aporte valioso tanto para la academia como para la sociedad.

Objetivos

Objetivo General

Desarrollar un modelo predictivo basado en machine learning que estime los cambios en la temperatura atmosférica de Barrancabermeja mediante el análisis de datos históricos.

Objetivos Específicos

Recopilar datos históricos de variables climáticas en la zona de Barrancabermeja, accediendo a la información recolectada por el Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM).

Aplicar los modelos de machine learning: ARIMAX, PROPHET, Random Forest y Gradient Boosting a los datos recopilados de las variables climáticas.

Desarrollar los modelos específicamente para la región de Barrancabermeja, comparándolos entre sí y seleccionando aquel que ofrezca la mejor capacidad de predicción de temperatura del aire.

Marco de Referencia

Estado del Arte

Los avances recientes en el modelado predictivo de variables climáticas han revelado el notable potencial de los algoritmos de machine learning para superar las limitaciones de los métodos tradicionales. Como señala Reichstein et al. (2019), estas técnicas han transformado radicalmente el análisis de datos ambientales gracias a su capacidad para identificar relaciones no lineales complejas que escapan a la representación de los modelos físicos convencionales. Este salto cualitativo abre nuevas posibilidades para comprender y anticipar el comportamiento del clima con una precisión sin precedentes.

En Colombia, este campo ha comenzado a desarrollarse con investigaciones como la de López et al. (2020), quienes diseñaron un modelo de predicción de temperatura para Bogotá utilizando redes neuronales LSTM, alcanzando un RMSE de 1.2°C en pronósticos a 24 horas. No obstante, como reconocen los propios autores, estos resultados no pueden generalizarse automáticamente a otras regiones del país, pues cada zona presenta particularidades climáticas que exigen aproximaciones específicas. Esta advertencia resulta especialmente relevante para el caso de Barrancabermeja, donde factores como la cercanía al río Magdalena, la intensa actividad industrial y los procesos de deforestación generan dinámicas térmicas únicas que deben ser estudiadas en profundidad.

Precisamente, investigaciones recientes han destacado la necesidad de incorporar variables específicas al analizar el clima en zonas tropicales como Barrancabermeja. Gómez et al. (2021) han demostrado cómo los cuerpos de agua cercanos influyen significativamente en la variabilidad térmica local, mientras que Rodríguez & Pérez (2022) han cuantificado el impacto de la actividad industrial en la formación de microclimas urbanos. A estos factores se suma el

efecto de la deforestación sobre los patrones locales de temperatura, un fenómeno documentado por el IDEAM (2022) que no puede pasarse por alto en cualquier modelo predictivo serio.

En cuanto a las técnicas de modelado, la literatura especializada ofrece valiosas lecciones. Por un lado, los modelos híbridos que combinan enfoques como ARIMA con redes neuronales han demostrado un desempeño sobresaliente en el análisis de series temporales climáticas (Zhang, 2022). Por otro, los algoritmos basados en árboles de decisión —como Random Forest y XGBoost— suelen superar en precisión a los modelos lineales tradicionales, aunque demandan mayor capacidad computacional (Chen & Guestrin, 2016). Finalmente, para datos con patrones estacionales bien definidos, PROPHET se ha posicionado como una alternativa particularmente efectiva (Taylor & Letham, 2018).

Este panorama evidencia tanto los avances logrados como los desafíos pendientes en el campo de la predicción climática local. Mientras las técnicas de machine learning ofrecen herramientas cada vez más sofisticadas, su aplicación exitosa en contextos específicos como Barrancabermeja requiere una cuidadosa selección de variables y algoritmos que consideren las particularidades de la región. Es en esta intersección entre el conocimiento global y las necesidades locales donde se ubica el aporte potencial de la presente investigación.

Marco Contextual

Situada en el corazón del Magdalena Medio (6°18'N, 73°48'O), Barrancabermeja emerge como un territorio de particularidades climáticas y socioeconómicas que configuran su identidad regional. A una modesta altitud de 75 metros sobre el nivel del mar, la ciudad presenta un clima tropical monzónico (clasificado como Am en el sistema de Köppen) que se caracteriza por una temperatura media anual de 27.8°C y precipitaciones que rondan los 2,500 mm anuales (IDEAM,

2022). Estas cifras, sin embargo, apenas esbozan la complejidad de un sistema climático moldeado por múltiples factores interconectados.

El río Magdalena, cuya cercanía ejerce un efecto moderador sobre las temperaturas locales, dialoga constantemente con otros elementos que transforman el paisaje climático urbano. La intensa actividad industrial petrolera, pilar económico de la región, genera microclimas particulares a través del fenómeno de isla de calor, mientras que los procesos de deforestación en áreas circundantes alteran progresivamente los patrones de viento y los regímenes térmicos. Esta combinación de factores naturales y antropogénicos crea un escenario meteorológico singular que demanda estudios específicos.

Más allá de su relevancia climática, Barrancabermeja se erige como un nodo estratégico en el mapa socioeconómico colombiano. Su complejo refinador, responsable del 25% de la capacidad nacional de procesamiento de hidrocarburos, convierte a la ciudad en un epicentro energético de primer orden. Paralelamente, los campos circundantes sustentan una importante actividad agrícola donde sobresalen cultivos como la palma, el cacao y el plátano, todos ellos sensibles a las variaciones climáticas.

Esta dinámica productiva se desarrolla en un entorno humano particular: aproximadamente 200,000 habitantes cuya vulnerabilidad ante los cambios climáticos se acentúa por las características de sus actividades laborales y las condiciones de vida. La interacción entre estos elementos geográficos, climáticos, industriales y sociales dibuja el panorama completo de Barrancabermeja, un territorio donde la comprensión detallada de los fenómenos meteorológicos locales se convierte en una necesidad tanto científica como práctica para el desarrollo sostenible de la región.

Marco Teórico

Los modelos ARIMAX (AutoRegressive Integrated Moving Average with eXogenous inputs) representan una evolución significativa en el análisis de series temporales aplicadas al clima. Como extensión de los modelos ARIMA clásicos, estos incorporan variables exógenas que permiten mejorar sustancialmente la capacidad predictiva (Box et al., 2015). Su fundamento teórico se basa en la descomposición sistemática de las series temporales en tres componentes esenciales: una tendencia que refleja el comportamiento a largo plazo, una estacionalidad que captura los patrones periódicos repetitivos, y un componente residual que engloba el ruido no explicado por los anteriores. Box & Jenkins (1970) destacaron la particular utilidad de este enfoque en climatología, donde variables como la temperatura o precipitación suelen presentar patrones estacionales bien definidos. Investigaciones recientes han reforzado esta perspectiva, demostrando cómo la inclusión de predictores externos -como concentraciones de CO_2 o índices oceánicos- incrementa notablemente la precisión de los modelos ARIMAX en aplicaciones climáticas específicas (Hyndman & Athanasopoulos, 2018).

En el ámbito del aprendizaje automático supervisado, el algoritmo Random Forest emerge como una poderosa herramienta predictiva. Desarrollado originalmente por Breiman (2001), este método de ensamble combina múltiples árboles de decisión mediante la técnica de Bootstrap Aggregating (bagging). Su solidez teórica se apoya en dos pilares fundamentales: por un lado, el Teorema del Límite Central garantiza que la predicción agregada de numerosos árboles reduce el error mediante la compensación de varianza (Hastie et al., 2009); por otro, la selección aleatoria de características en cada división del árbol disminuye significativamente la correlación entre los distintos árboles que componen el bosque (James et al., 2013). Este doble

mecanismo convierte a Random Forest en un método particularmente eficaz para manejar relaciones complejas en datos climáticos.

El Gradient Boosting, propuesto inicialmente por Friedman (2001), sigue un enfoque diferente pero igualmente robusto. A diferencia de los métodos basados en bagging, esta técnica construye modelos de forma secuencial, optimizando iterativamente el error residual mediante descenso de gradiente. Su fundamento matemático descansa en la minimización funcional, donde cada nuevo modelo se ajusta específicamente para corregir los errores del anterior (Friedman, 2002). La incorporación de parámetros de regularización, como el learning rate, añade otra capa de sofisticación al prevenir el sobreajuste y garantizar modelos más generalizables (Chen & Guestrin, 2016). Esta combinación de aproximaciones iterativas y controles de regularización hace del Gradient Boosting una opción especialmente potente para problemas de predicción climática.

Completa este panorama teórico el modelo Prophet, desarrollado por el equipo de Facebook (Taylor & Letham, 2018). Como modelo aditivo, Prophet descompone las series temporales en una tendencia no lineal ajustada mediante regresión por tramos y una componente estacional modelada con funciones de Fourier. Lo que distingue a Prophet es su notable flexibilidad para manejar datos con patrones complejos y valores faltantes, características frecuentes en los registros climáticos. Esta adaptabilidad, unida a su relativa simplicidad de implementación, lo convierte en una alternativa atractiva para el análisis de variables meteorológicas que presentan comportamientos no convencionales o datos incompletos. La conjunción de estos distintos enfoques teóricos -desde los modelos clásicos de series temporales hasta las técnicas más modernas de aprendizaje automático- proporciona un marco conceptual

sólido para abordar el desafío de la predicción climática local con herramientas adecuadas a las particularidades de cada situación.

Marco Conceptual

El marco conceptual establece los fundamentos teóricos y operativos que guían la investigación, definiendo con precisión los términos clave, sus mediciones y su relación dentro del estudio. En este caso, se analiza la relación entre variables meteorológicas —como la temperatura atmosférica, la humedad relativa, la presión atmosférica y la velocidad del viento— y su impacto en un modelo predictivo. La temperatura atmosférica se considera la variable dependiente, mientras que las demás actúan como variables independientes, dado su potencial influencia en las fluctuaciones térmicas.

Para evaluar el desempeño del modelo, se emplean métricas estadísticas como el RMSE (Raíz del Error Cuadrático Medio), el MAE (Error Absoluto Medio) y el R^2 (Coeficiente de Determinación), que permiten cuantificar la precisión y la capacidad explicativa de las predicciones. Asimismo, se incluyen técnicas de ingeniería de características, como la normalización de datos y la Prueba de Dickey-Fuller aumentada, para garantizar la validez de los análisis y la estacionariedad de las series temporales.

Tabla 1

Marco Conceptual: Variables y Parámetros Analizados

Término	Definición Operacional	Variable Asociada
Temperatura atmosférica	Valor promedio horario registrado a 2 m de altura (°C)	Variable dependiente
Humedad relativa	Porcentaje de saturación de vapor de agua en el aire (%)	Variable independiente

Término	Definición Operacional	Variable Asociada
Presión atmosférica	Fuerza ejercida por la columna de aire (HPa)	Variable independiente
Velocidad del viento	Magnitud vectorial del movimiento horizontal del aire (m/s)	Variable independiente
RMSE	Raiz del error cuadrático medio entre valores predichos y observados(°C)	Métrica de evaluación
MAE	Error absoluto medio. diferencia promedio entre los valores reales y los valores predichos	Métrica de evaluación
R2	Coeficiente de determinación. Mide el grado de ajuste del modelo a los datos mediante la evaluación de la proporción de varianza en la variable dependiente explicada por las variables independientes	Métrica de evaluación
Ingeniería de características	Normalización, Prueba de Dickey-Fuller aumentada	Análisis de resultados

Marco Normativo

El desarrollo de modelos predictivos para variables climáticas en Colombia se enmarca dentro de un conjunto de disposiciones legales y estándares técnicos que garantizan tanto la calidad científica como el impacto social positivo de estas iniciativas. A nivel nacional, la Ley 1931 de 2018 sienta las bases para la gestión integral del cambio climático, haciendo especial énfasis en la implementación de sistemas de alerta temprana como herramienta fundamental para

la adaptación territorial (Artículo 5). Esta normativa se complementa con la Resolución 2254 de 2017 emitida por el IDEAM, que establece los parámetros técnicos para la medición meteorológica, definiendo con precisión los estándares de calidad y la frecuencia mínima requerida para la recolección de datos climáticos confiables.

En el ámbito internacional, el marco de referencia lo proporcionan documentos como la Guía WMO No. 1203 (2021) de la Organización Meteorológica Mundial, que detalla protocolos específicos para el modelado predictivo climático a escala local, asegurando la comparabilidad y validez científica de los resultados. A estos se suma la norma ISO 14090:2019, que establece principios universales para la adaptación al cambio climático, ofreciendo lineamientos claros sobre requisitos y mejores prácticas en el desarrollo de herramientas predictivas.

La protección de la información recopilada encuentra su sustento legal en la Ley 1581 de 2012, que regula de manera estricta el tratamiento de datos personales, aspecto particularmente relevante cuando la información meteorológica incluye referencias geográficas específicas que podrían asociarse a ubicaciones o individuos particulares. Este componente normativo adquiere mayor relevancia al considerar los aspectos éticos en el desarrollo de algoritmos de inteligencia artificial, donde los Principios OCDE sobre IA (2019) exigen transparencia en los procesos algorítmicos y mecanismos claros de rendición de cuentas, especialmente cuando las predicciones generadas pueden afectar directamente a comunidades o tomarse como base para decisiones de política pública. Este entramado normativo, que va desde lo técnico hasta lo ético, configura un marco de acción robusto que asegura el desarrollo responsable de herramientas predictivas con impacto social positivo.

Metodología

El desarrollo de este proyecto se fundamenta en el método CRISP-DM (Cross-Industry Standard Process for Data Mining), un marco metodológico ampliamente reconocido en el ámbito de la minería de datos que ha sido adaptado específicamente para abordar el desafío de la predicción de temperaturas en Barrancabermeja. Este enfoque sistemático, validado por Shearer (2000) e IBM (2020), guiará todo el proceso investigativo a través de seis fases interconectadas que garantizan rigor científico y aplicabilidad práctica.

La primera etapa, centrada en la comprensión del problema, parte de una exhaustiva revisión de literatura sobre modelos climáticos basados en machine learning para definir con precisión los requisitos del sistema predictivo. Este proceso permitió identificar las variables clave -temperatura, humedad relativa, presión atmosférica y velocidad del viento- y establecer métricas de éxito concretas, como un RMSE inferior a 1.5°C y un coeficiente de determinación R^2 mayor a 0.75. Estas exigencias técnicas no son arbitrarias, sino que responden a las necesidades reales de los potenciales usuarios del modelo, desde autoridades locales hasta actores del sector productivo.

El trabajo con los datos comienza con su adquisición a partir de las estaciones meteorológicas del IDEAM en Barrancabermeja, donde se recopilaron registros horarios de las cuatro variables continuas centrales para el análisis: temperatura (variable objetivo expresada en °C), humedad relativa (en porcentaje), presión atmosférica (en hectopascales) y velocidad del viento (en m/s). Estas mediciones, obtenidas mediante instrumentos calibrados, fueron sometidas a un detallado análisis exploratorio que incluyó la visualización de series temporales, construcción de matrices de correlación y detección de valores atípicos mediante boxplots, todo ello con el fin de comprender sus patrones y relaciones subyacentes.

Tabla 2

Variables Analizadas en el Estudio

Variables	Tipo	Descripción
Temperatura	Continua	Media horaria (°C)
Humedad relativa	Continua	Porcentaje (%)
Presión atmosférica	Continua	Hectopascales (hPa)
Velocidad del viento	Continua	Metros por segundo (m/s)

La fase de preparación de datos abordó dos desafíos principales: la limpieza de los registros, donde se implementó imputación de valores faltantes mediante el algoritmo KNN, y el feature engineering, que incluyó procesos de normalización con MinMaxScaler especialmente relevantes para algoritmos sensibles a la escala como Random Forest y XGBoost. Para los modelos de series temporales (ARIMAX y Prophet), se prestó especial atención a la extracción de características relacionadas con tendencia y estacionalidad, componentes fundamentales en el comportamiento climático.

La selección de modelos predictivos combinó estratégicamente enfoques tradicionales y modernos, creando un marco comparativo robusto. Por un lado, se incluyeron modelos clásicos de series temporales como ARIMAX (extensión del ARIMA que incorpora variables exógenas) y Prophet (desarrollado por Facebook para patrones estacionales). Por otro, se implementaron algoritmos avanzados de machine learning como Random Forest (método de ensamble basado en árboles de decisión) y XGBoost (implementación optimizada de gradient boosting). Cada uno de estos modelos fue configurado mediante librerías especializadas de Python, con hiperparámetros cuidadosamente seleccionados según su impacto documentado en la literatura: parámetros de

estructura temporal para ARIMAX y Prophet, y controles de complejidad para Random Forest y XGBoost.

Tabla 3

Algoritmos Seleccionados Para la Modelización Predictiva

Modelo	Librería	Hiperparametros a Optimizar
ARIMAX	statsmodels	(p,d,q) + variables exógenas
PROPHET	fbprophet	Changepoint_prior_scale, seasonality
Random Forest	Scikit-learn	n_estimators, max_depth
Gradient Boosting	XGBoost	learning_rate, n_estimators

El proceso de evaluación empleó múltiples métricas de desempeño (RMSE, R^2 y MAE) complementadas con pruebas de significancia estadística como el test de Wilcoxon para comparar modelos, junto con análisis detallados de residuales que examinaron su normalidad y posible heterocedasticidad. Esta evaluación rigurosa no solo midió la precisión predictiva, sino también la robustez estadística de cada enfoque.

Como producto final, se desarrolló un script en Python capaz de generar pronósticos automáticos de temperatura, diseñado para su potencial integración con plataformas de alertas tempranas del IDEAM. El modelo incluye recomendaciones para su mantenimiento, destacándose la importancia de actualizaciones mensuales que incorporen nuevos datos y permitan ajustar progresivamente los parámetros, asegurando así la vigencia y precisión continua del sistema predictivo. Esta solución tecnológica, anclada en metodologías científicas sólidas pero con clara vocación aplicada, representa un puente efectivo entre la investigación académica y las necesidades concretas de la región.

Técnicas e Instrumentos

El proceso de recolección de datos se realizó mediante la plataforma oficial del IDEAM, institución que provee información meteorológica confiable y estandarizada para el territorio colombiano. Los registros climáticos se obtuvieron en formatos estructurados como CSV y Parquet, seleccionados por su eficiencia en el almacenamiento de grandes volúmenes de datos y su compatibilidad con las herramientas de análisis empleadas.

Para el procesamiento y análisis de la información, se utilizó el lenguaje de programación Python en su versión 3.9, aprovechando librerías especializadas como Pandas para la manipulación de datos, NumPy para operaciones numéricas complejas y Matplotlib para la visualización gráfica de resultados. Todo el flujo de trabajo se desarrolló en entornos Jupyter Notebook, que permiten combinar código, visualizaciones y texto explicativo en un mismo documento, garantizando así la transparencia metodológica y la completa reproducibilidad del análisis. Esta combinación de herramientas tecnológicas conformó un ecosistema robusto y flexible capaz de manejar desde la etapa inicial de extracción de datos hasta la generación de modelos predictivos finales.

Tipo de Estudio

Esta investigación adopta un carácter aplicado y predictivo, orientado a resolver un desafío concreto en el ámbito de la meteorología local mediante el desarrollo de modelos computacionales comparativos. Con un sólido fundamento cuantitativo, el estudio se apoya en el análisis exhaustivo de registros climáticos históricos para descubrir patrones ocultos y relaciones significativas entre variables atmosféricas. Siguiendo la clasificación metodológica de Hernández-Sampieri et al. (2018), el trabajo combina estratégicamente varios enfoques de investigación que se complementan entre sí.

En cuanto a su nivel de profundidad analítica, la investigación opera simultáneamente en dos dimensiones: por un lado, adopta un enfoque descriptivo-correlacional para identificar y cuantificar las interacciones entre distintos factores climáticos; por otro, avanza hacia un nivel predictivo al transformar estos hallazgos en un modelo operativo capaz de anticipar comportamientos futuros de la temperatura. Esta dualidad permite no solo comprender las dinámicas climáticas existentes, sino también generar herramientas prácticas para la toma de decisiones.

El diseño metodológico se enmarca dentro de los parámetros experimentales, aunque con la particularidad de desarrollarse en un entorno de simulación computacional. Este enfoque innovador posibilita el entrenamiento y validación rigurosa de múltiples algoritmos de machine learning, utilizando para ello conjuntos de datos reales provenientes de fuentes oficiales. La naturaleza experimental se manifiesta en la comparación sistemática de diferentes técnicas de modelado, sometidas todas ellas a idénticas condiciones de prueba para garantizar la objetividad de los resultados.

Temporalmente, el estudio adopta una perspectiva longitudinal retrospectiva que abarca dos décadas de registros climáticos continuos, incluyendo datos tanto horarios como diarios proporcionados por el IDEAM. Esta amplia ventana temporal no solo proporciona la masa crítica de información necesaria para los análisis estadísticos, sino que también permite capturar ciclos climáticos completos y eventos extremos relevantes. La combinación de estos elementos metodológicos -aplicado, predictivo, comparativo y cuantitativo- configura un marco de investigación robusto capaz de abordar con rigor científico el desafío de la predicción térmica local.

Recolección de Datos

Selección de la Estación Meteorológica para el Estudio

Los datos se obtuvieron del portal oficial datos.gov.co, donde se descargó un archivo en formato CSV con información de las estaciones meteorológicas del IDEAM en Colombia. Este conjunto de datos se importó a Python utilizando la librería *pandas* para su procesamiento.

Tras filtrar las estaciones ubicadas en Barrancabermeja, se identificaron 17 estaciones, clasificadas según su tipo (como se detalla en la tabla adjunta). Al analizar los registros de cada una, se determinó que la estación Vizcaína La Lizama era la más adecuada para el estudio, ya que cuenta con un conjunto completo de datos para las variables de interés. Por esta razón, se seleccionó esta estación para realizar el pronóstico de temperaturas en la ciudad.

Figura 1

Estaciones Meteorológicas del IDEAM en Barrancabermeja

	Nombre	Categoria
0	VIZCAINA LA LIZAMA - AUT [24055080]	Agrometeorológica
1	AEROPUERTO YARIGUIES [23155030]	Sinóptica Principal
2	TENERIFE DELICIAS [23147030]	Limnimétrica
3	BARRANCABERMEJA - AUT [23157030]	Limnigráfica
4	CENTRO EL [23155040]	Climática Ordinaria
5	CHUCURI [23130010]	Pluviométrica
6	LLANITO EL [24057090]	Limnimétrica
7	MALDONADO [23157080]	Limnimétrica
8	SOGAMOSO RIO ANTES [24057100]	Limnimétrica
9	GALAN [23157070]	Limnimétrica
10	PUENTE CARRETERA [24057030]	Limnigráfica
11	PUENTE FERROCARRIL [24057070]	Limnimétrica
12	CASA BOMBA [24057060]	Limnimétrica
13	LLANITO EL ARRIBA [24057080]	Limnimétrica
14	Radar Met BARRANCA	Meteorológica Especial
15	TERMOBARRANCA [23157010]	Limnimétrica
16	BARRANCABERMEJA [23150010]	Pluviométrica
17	TERMOBARRANCA [23150030]	Pluviométrica

Resultados

El proceso de preparación de datos constituyó una etapa fundamental para garantizar la calidad y confiabilidad de los análisis posteriores. Todo comenzó con la selección estratégica de la estación meteorológica Vizcaína La Lizama, cuyos registros históricos servirían como base para el desarrollo del modelo predictivo. Los datos crudos, obtenidos en formato CSV, fueron cargados al entorno de Python mediante la librería `pandas`, permitiendo su manipulación y transformación sistemática.

Para organizar la información de manera estructurada, se crearon dataframes individuales para cada variable climática relevante: temperatura, presión atmosférica, humedad relativa y velocidad del viento. Una atención especial se dedicó al tratamiento de la columna `FechaObservacion`, que fue convertida al tipo `datetime` para asegurar un manejo temporal preciso y consistente a lo largo de todo el análisis. Posteriormente, estos dataframes individuales fueron consolidados en una única estructura unificada que concentró exclusivamente las variables de interés para el estudio.

El proceso de filtrado y limpieza de los datos implicó decisiones metodológicas cuidadosas. Se optó por focalizar el análisis en las mediciones tomadas específicamente a las 6:00 PM, cubriendo el período comprendido entre 2005 y 2025. Este criterio temporal buscó capturar patrones climáticos representativos mientras mantenía la homogeneidad de los datos. Ante la inevitable presencia de valores faltantes, se implementó un sofisticado método de imputación basado en el algoritmo KNN (K-Nearest Neighbors), que permitió reconstruir la información ausente preservando las relaciones estadísticas subyacentes.

Un paso crítico en el aseguramiento de la calidad de los datos fue la verificación minuciosa de duplicados en la variable `FechaObservacion`. Tras identificar registros repetidos,

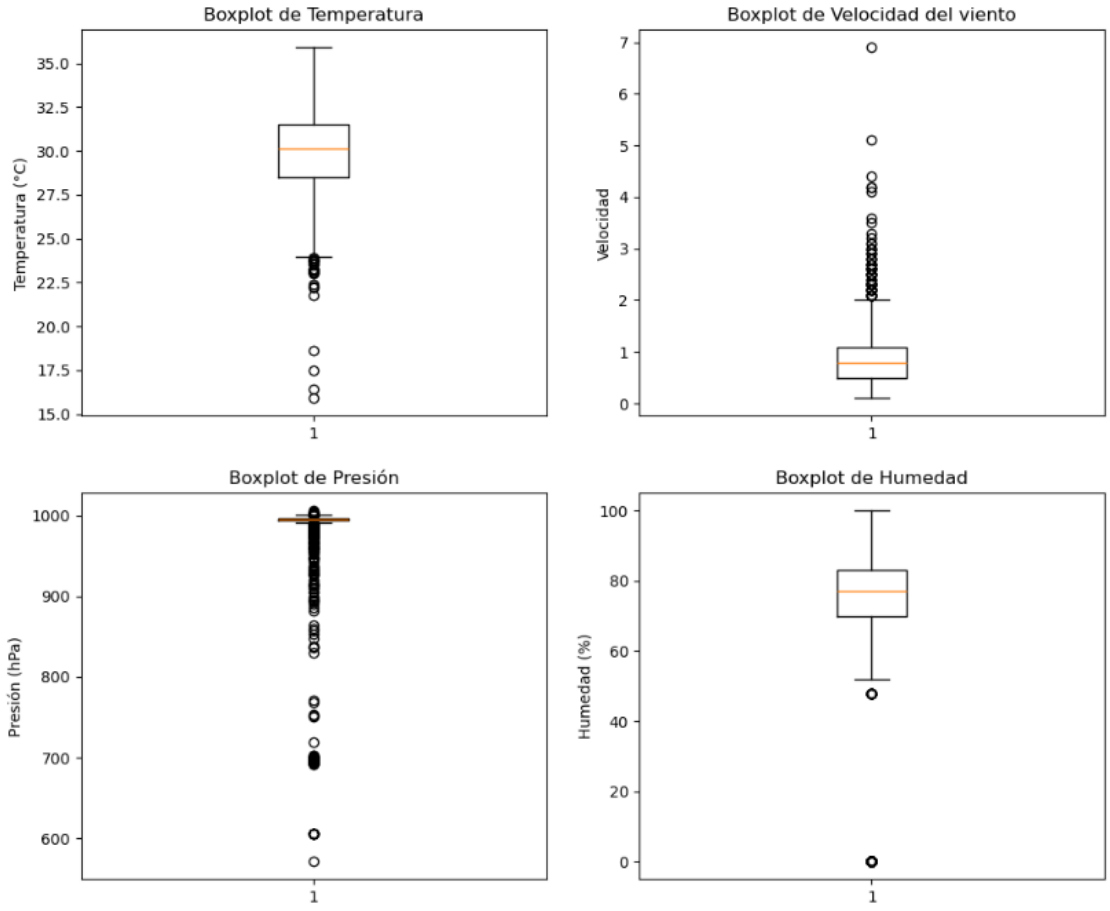
se procedió a su eliminación sistemática, garantizando así que cada observación en el conjunto final de datos representara un instante temporal único. Este riguroso proceso de depuración y preparación sentó las bases para las etapas subsiguientes de modelado, asegurando que los algoritmos predictivos trabajarían con información consistente, completa y confiable.

Identificación de Datos Atípicos

Se identificaron y removieron valores atípicos mediante visualización con boxplots.

Figura 2

Boxplot de Variables Antes de Eliminación de Datos Atípicos



El análisis visual de los diagramas de caja reveló patrones interesantes en el comportamiento de las variables meteorológicas estudiadas. Comenzando con la temperatura, se

observa que su distribución presenta una mediana alrededor de los 30-31°C, marcando un clima predominantemente cálido en la región. La relativa compacidad de la caja, que abarca aproximadamente entre 28°C y 32°C, indica que el 50% central de los datos muestra poca variación térmica. Sin embargo, la ligera asimetría negativa, evidenciada por una mediana más cercana al cuartil superior y un bigote inferior más extenso, sugiere que aunque la mayoría de los registros se concentran en el rango superior, existen ocasionales descensos térmicos significativos. Esto se confirma con la presencia de varios valores atípicos por debajo de los 25°C, llegando incluso hasta los 15°C, que podrían corresponder a eventos meteorológicos excepcionales o noches particularmente frías.

En marcado contraste, el análisis de la velocidad del viento muestra un patrón completamente diferente. La mediana se sitúa alrededor de 0.5 m/s, con una caja extremadamente estrecha que revela que la mayoría de las mediciones presentan valores bajos y muy poco variables, concentrados entre 0.2 y 1.0 m/s. Sin embargo, la distribución es fuertemente asimétrica hacia la derecha, con una notable cantidad de valores atípicos que se extienden hasta más de 7 m/s. Esta configuración indica que, si bien predominan condiciones de viento calmado, existen eventos esporádicos de ráfagas intensas que alteran significativamente el patrón general.

El comportamiento de la presión atmosférica resulta particularmente llamativo. Con una mediana cercana a los 1000 hPa, lo más destacable es la extraordinaria estrechez de la caja, que casi se reduce a una línea, demostrando una estabilidad excepcional en el 50% central de los datos. No obstante, esta aparente uniformidad se ve matizada por la presencia de numerosos valores atípicos, especialmente hacia el extremo inferior (hasta 600 hPa) y, en menor medida, hacia el superior (ligeramente por encima de 1000 hPa). Esta peculiar distribución, con una

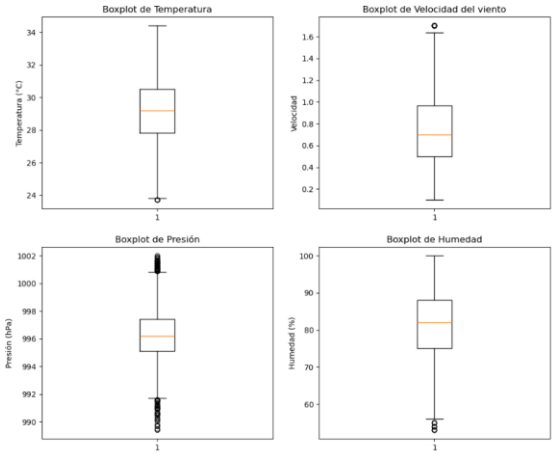
concentración central muy marcada pero colas extremadamente pesadas, podría reflejar tanto errores instrumentales como eventos atmosféricos verdaderamente anómalos que merecerían un análisis más detallado.

Finalmente, el boxplot de humedad relativa muestra una mediana situada entre el 75-80%, con una dispersión moderada que abarca aproximadamente del 70% al 85% para el 50% central de los datos. La leve asimetría negativa, con una cola más extendida hacia valores bajos, se complementa con la presencia de algunos valores atípicos en el extremo inferior (alrededor del 55% e incluso cercano a 0%), mientras que, como era de esperar, no se registran valores por encima del 100%.

Los análisis gráficos demostraron que cada variable contenía mediciones atípicas considerables. Dado que estos valores extremos pueden comprometer la calidad predictiva de los modelos, se implementó un protocolo de depuración estricto. Este proceso permitió homogenizar las distribuciones y estabilizar estadísticamente el conjunto de datos, sentando así las bases para construir algoritmos de predicción más precisos y confiables.

Figura 3

Boxplots de Variables Después de la Eliminación de Datos Atípicos



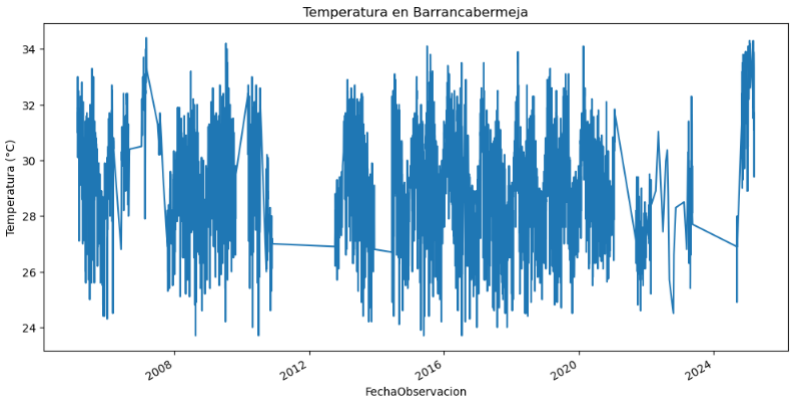
El proceso de eliminación de datos atípicos produjo resultados diferenciados según la variable analizada. En el caso de la velocidad del viento, el impacto fue particularmente notable, transformando radicalmente su distribución al suprimir las ráfagas más intensas, lo que reveló un patrón de comportamiento mucho más contenido y uniforme. Para las variables de temperatura y humedad, la depuración permitió recortar los valores extremos, obteniendo así una representación más nítida de sus rangos habituales de variación. Sin embargo, la presión atmosférica presentó un comportamiento peculiar: a pesar de la aplicación de los criterios de filtrado, continuaron apareciendo numerosos valores atípicos en la nueva escala. Esta persistencia podría interpretarse de dos maneras - o bien la distribución intrínseca de la presión contiene múltiples desviaciones significativas incluso dentro de márgenes ya restringidos, o quizás los criterios de eliminación aplicados requieren ajustes más específicos para esta variable en particular. Este fenómeno invita a una reflexión más profunda sobre la naturaleza misma de las fluctuaciones barométricas en la zona de estudio.

Análisis Exploratorio (EDA)

Series Temporales: Se graficó la evolución de la temperatura, evidenciando un comportamiento temporal con oscilaciones recurrentes.

Figura 4

Variación de Temperatura (°C) en Barrancabermeja por Fecha de Observación



La gráfica presenta el comportamiento de las temperaturas en Barrancabermeja desde antes de 2008 hasta inicios de 2024, revelando un patrón climático característico de esta región tropical. A lo largo de este período de aproximadamente 16 años, los valores térmicos oscilan principalmente entre 24°C y 34°C, con ocasionales picos que superan ligeramente este máximo, confirmando el clima cálido típico de la zona.

Al examinar la tendencia general, destaca la ausencia de una clara dirección ascendente o descendente a largo plazo, mostrando más bien una notable estabilidad en los valores medios. Sin embargo, esta aparente constancia enmascara importantes fluctuaciones interanuales y una marcada estacionalidad. La curva térmica dibuja ritmos cíclicos bien definidos, con ascensos y descensos recurrentes que probablemente corresponden a los periodos secos y lluviosos característicos del clima monzónico de la región. Estas variaciones estacionales presentan una amplitud considerable, con diferencias de varios grados entre los puntos más altos y bajos de cada ciclo.

La serie temporal exhibe una notable volatilidad a corto plazo, con cambios bruscos que se repiten consistentemente a lo largo de todo el registro. Esta variabilidad inmediata contrasta con la relativa constancia en la amplitud de las oscilaciones anuales, aunque se aprecian ciertas diferencias interanuales en la intensidad de los picos y valles.

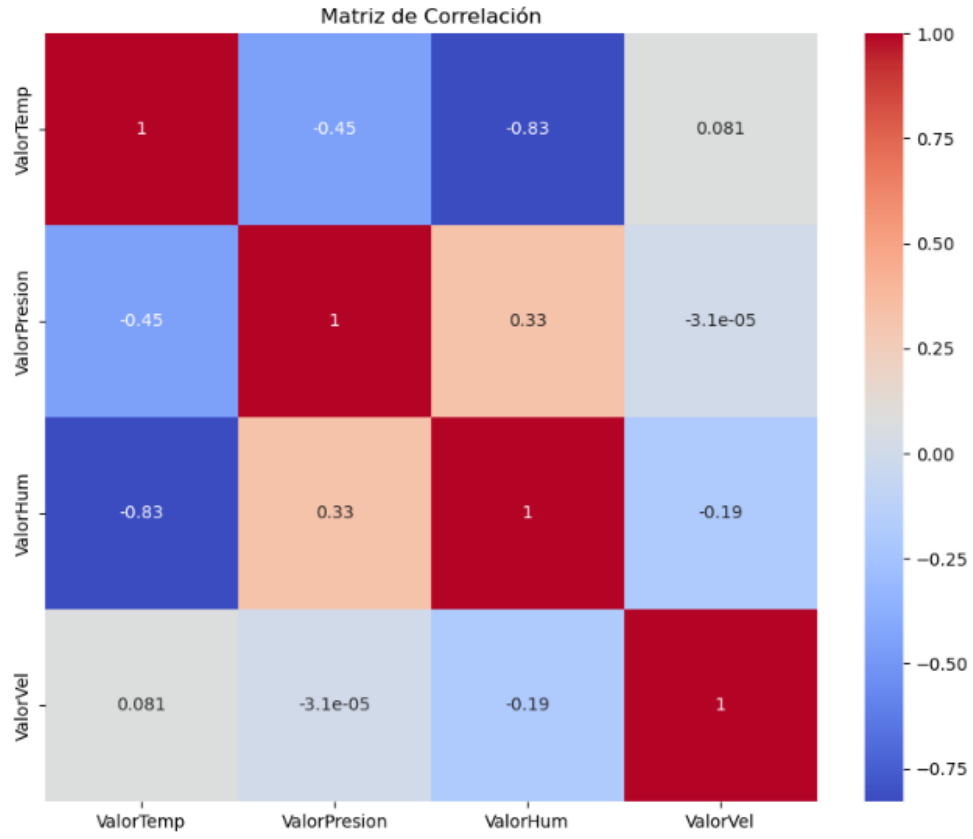
Tambien se observa un cambio en la tendencia entre 2011-2014 y 2023-2024. Estos no representan fenómenos meteorológicos reales, sino más bien limitaciones en el monitoreo continuo.

El análisis visual sugiere que, pese a las variaciones estacionales y eventos puntuales, el clima de Barrancabermeja ha mantenido una notable estabilidad en su régimen térmico durante el periodo estudiado. No obstante, esta apreciación preliminar requeriría ser complementada con

análisis estadísticos más rigurosos que permitan detectar posibles tendencias sutiles o cambios progresivos que no son evidentes a simple vista. La clara estacionalidad observada refuerza la importancia de considerar estos ciclos naturales al desarrollar cualquier modelo predictivo para la región.

Figura 5

Matriz de Correlación: Temperatura, Presión, Humedad y Velocidad del Viento



El mapa de calor analizado revela las complejas interrelaciones entre cuatro variables climáticas fundamentales en Barrancabermeja: temperatura, presión atmosférica, humedad relativa y velocidad del viento.

Entre los hallazgos más relevantes destaca la marcada correlación negativa (-0.83) entre temperatura y humedad, una de las relaciones más intensas observadas. Este fuerte vínculo

inverso sugiere que en los momentos de mayor calor diurno, particularmente durante las horas de máxima insolación, la humedad relativa desciende considerablemente. Este fenómeno, característico de climas tropicales como el de Barrancabermeja, refleja la capacidad del aire cálido para retener mayor cantidad de vapor de agua en términos absolutos, lo que reduce la humedad relativa cuando no hay aportes adicionales de humedad. A la inversa, los periodos lluviosos o de mayor nubosidad suelen presentar simultáneamente temperaturas más bajas y humedad elevada.

Otra relación significativa, aunque menos intensa, se observa entre temperatura y presión atmosférica, con un coeficiente de -0.45. Esta correlación negativa moderada coincide con los principios de la física atmosférica, donde el aire caliente, al ser menos denso, tiende a ascender generando zonas de presión reducida en superficie. Por el contrario, cuando predominan masas de aire más frías y densas, se registran comúnmente valores de presión más elevados.

El análisis revela también una conexión moderada (0.33) entre presión atmosférica y humedad, donde periodos de alta presión suelen coincidir con mayor humedad relativa. Este patrón podría explicarse por la asociación entre altas presiones y condiciones atmosféricas estables que favorecen la acumulación de humedad en ambientes tropicales, siempre que no intervengan otros sistemas meteorológicos que alteren este equilibrio.

Resulta particularmente interesante la escasa correlación que muestra la velocidad del viento con las demás variables. Su relación con la temperatura es prácticamente nula (0.081), al igual que con la presión atmosférica (-3.1e-05), y solo presenta una leve correlación negativa (-0.19) con la humedad. Esta aparente independencia del viento respecto a las otras variables podría deberse a que su comportamiento en un punto específico está más influenciado por factores locales (topografía, vegetación, cercanía a cuerpos de agua) que por las variables

termodinámicas consideradas, o porque su variabilidad responde a escalas espaciales y temporales que no quedan plenamente capturadas en estas mediciones puntuales.

En conjunto, la matriz de correlación dibuja un sistema climático donde temperatura, presión y humedad muestran interconexiones significativas que reflejan procesos físicos fundamentales, mientras que la velocidad del viento aparece como una variable más independiente en este contexto particular. Estos patrones de correlación proporcionan valiosos insights sobre la dinámica atmosférica local y resultan esenciales para el desarrollo de modelos predictivos precisos, especialmente al considerar que la fuerte relación inversa entre temperatura y humedad emerge como el vínculo más determinante en el clima de Barrancabermeja.

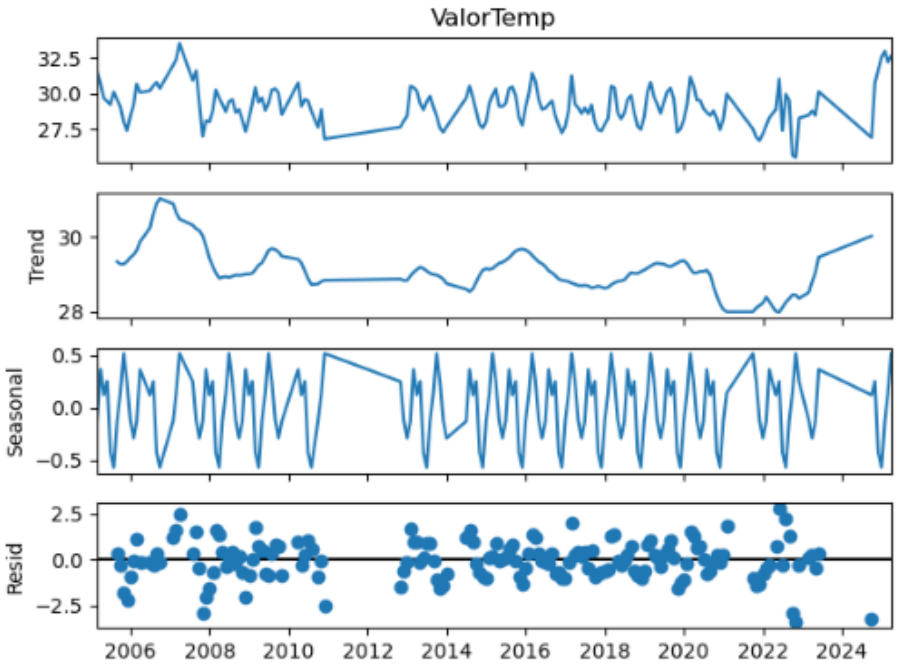
Ingeniería de Características

Prueba de Estacionariedad (ADF)

El test de Dickey-Fuller Aumentado (ADF) aplicado a la serie temporal *'*ValorTemp*'* arrojó un estadístico ADF de -6.405 y un p-value de 1.95e-08, indicando una fuerte evidencia para *r*echazar la hipótesis nula de no estacionariedad. Esto confirma que la serie de temperatura analizada es estacionaria con un alto nivel de confianza (99.99%). Este resultado tiene como implicaciones no requerir *de* diferenciaciones o transformaciones adicionales para modelar la serie.

Figura 6

Descomposición de Serie Temporal: Temperatura (2006-2024) – Tendencia, Estacionalidad y Residuos



La gráfica presenta un completo desglose de la serie temporal de temperatura en Barrancabermeja a lo largo de 18 años, dividiendo su comportamiento en tres componentes fundamentales: tendencia, estacionalidad y residuos. Este análisis multivariado revela los patrones subyacentes que caracterizan el clima local.

La línea de tendencia, muestra una ligera fluctuación alrededor de los 29°C, con variaciones moderadas entre aproximadamente 28°C y 30°C a lo largo del periodo estudiado. Lo más destacable es la ausencia de una pendiente pronunciada, lo que sugiere una notable estabilidad en las temperaturas medias anuales durante estas casi dos décadas. Sin embargo, se aprecian oscilaciones interanuales significativas, incluyendo un periodo relativamente más frío

alrededor de 2011-2012 y un posible leve calentamiento en los últimos años, aunque estos cambios requieren análisis estadísticos más profundos para confirmar su significancia.

El patrón estacional revela ciclos regulares con una amplitud de aproximadamente ±0.5°C, confirmando la presencia de una marcada estacionalidad en el régimen térmico. Esta variación periódica, que se repite consistentemente año tras año, corresponde muy probablemente a los ciclos climáticos característicos de la región, donde se alternan periodos secos y lluviosos. La regularidad y simetría de estos ciclos sugieren que los factores que impulsan la estacionalidad (como la posición solar, patrones de vientos o precipitación) han mantenido una notable consistencia temporal.

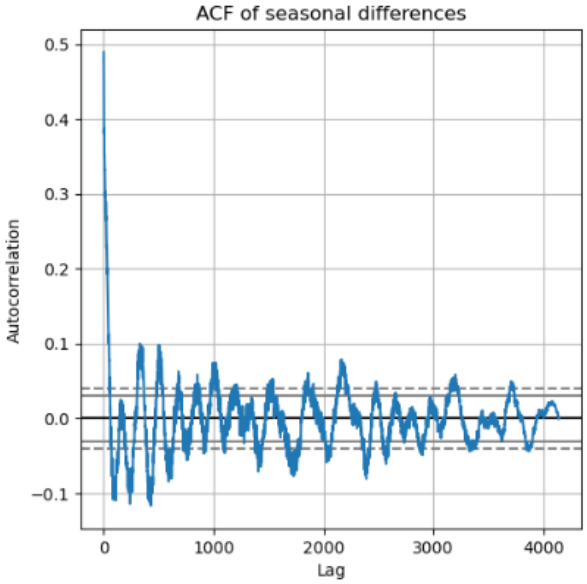
Los residuos, que representan las variaciones no explicadas por la tendencia ni la estacionalidad, muestran una dispersión relativamente simétrica alrededor de cero con amplitudes que alcanzan hasta ±2.5°C. Esta variabilidad residual incluye tanto fluctuaciones aleatorias como posibles eventos meteorológicos extraordinarios. La distribución aparentemente homogénea de estos residuos a lo largo del tiempo sugiere que el modelo de descomposición ha capturado adecuadamente los patrones principales, aunque algunos picos aislados podrían corresponder a fenómenos climáticos inusuales o errores de medición puntuales.

La descomposición revela que el comportamiento térmico en Barrancabermeja está dominado por su componente estacional, mientras que la tendencia a largo plazo muestra una estabilidad notable. Las variaciones interanuales, aunque presentes, no alteran significativamente el patrón climático fundamental. Los residuos, aunque muestran cierta variabilidad, no presentan patrones sistemáticos que sugieran la presencia de factores no considerados en el modelo.

Prueba de Autocorrelación

Figura 7

Función de Autocorrelación (ACF) de Diferencias Estacionales



El gráfico de la función de autocorrelación (ACF), aplicado a la serie temporal después de aplicar diferencias estacionales, revela aspectos fundamentales sobre la estructura subyacente de los datos. Esta técnica, diseñada para eliminar patrones periódicos y estabilizar la serie, muestra en su eje vertical coeficientes que fluctúan entre -0.1 y 0.5, mientras que el eje horizontal abarca retrasos temporales desde 0 hasta 4000 unidades, ofreciendo una visión completa de las dependencias internas en distintas escalas de tiempo.

Al examinar la evolución de estos coeficientes, se destaca inicialmente una autocorrelación moderadamente alta de 0.5 en el primer lag, señalando que, pese a la eliminación de la estacionalidad, persiste una clara conexión entre observaciones consecutivas. Este fenómeno sugiere la presencia de tendencias residuales o ciclos secundarios que el proceso de diferenciación no logró capturar completamente. A medida que avanzamos en los lags

subsiguientes, se observa un paulatino descenso en los valores de autocorrelación —de 0.4 a 0.2—, patrón característico de series que, aunque han sido transformadas para reducir su complejidad, mantienen cierta memoria a corto plazo en su comportamiento.

Resulta particularmente significativa la ausencia de picos recurrentes en intervalos específicos, lo que confirma la efectividad del método para remover los componentes estacionales dominantes. Por otro lado, las ocasionales incursiones de los coeficientes en territorio negativo, siempre dentro de un margen reducido (-0.1), descartan la existencia de relaciones inversas relevantes en la serie transformada.

Desde la perspectiva del modelado predictivo, estos hallazgos apuntan a la conveniencia de emplear enfoques como ARIMA, donde los términos autorregresivos podrían capturar adecuadamente la dependencia residual a corto plazo. La limpieza estacional lograda valida el preprocesamiento aplicado, aunque la estructura restante indica que podrían necesitarse ajustes complementarios, como diferenciación no estacional o componentes de media móvil, para refinar aún más el modelo. En el contexto de datos climáticos, como temperaturas, este comportamiento refuerza la noción de que, más allá de los ciclos anuales eliminados, persisten influencias atmosféricas que imprimen correlaciones temporales en escalas más inmediatas.

En síntesis, el análisis de la ACF tras la diferenciación estacional no solo confirma la eliminación satisfactoria de los patrones periódicos, sino que también revela la naturaleza persistente de las relaciones temporales a corto plazo, ofreciendo valiosas pistas para el desarrollo de modelos más precisos y robustos.

Modelado

Modelado Predictivo Mediante ARIMAX con Variables Exógenas

El modelo ARIMAX (AutoRegressive Integrated Moving Average with eXogenous inputs) es una extensión del modelo ARIMA que incluye variables exógenas para mejorar la precisión de las predicciones. En el contexto de esta investigación de pronóstico de temperatura, este modelo puede ser particularmente útil porque permite incorporar factores externos que influyen en las temperaturas, como la humedad, la presión atmosférica y la velocidad del viento.

Antes de ajustar el modelo, se evaluó la estacionariedad de la serie temporal de temperatura mediante la prueba de Dickey-Fuller aumentada (ADF). Los resultados confirmaron que la serie es estacionaria (valor $p < 0.05$), lo que permitió proceder con el modelado sin necesidad de diferenciación adicional.

Para determinar la estructura óptima del modelo, se empleó la librería `pmdarima`, que implementa una búsqueda automática de los órdenes p, d y q. El algoritmo identificó que el modelo más eficiente era un SARIMAX(1,1,2), el cual incorpora componentes autorregresivos (AR) y de media móvil (MA), junto con un término de integración (d=1) para garantizar estacionariedad.

Figura 8

Resultados del Modelo SARIMAX(1,1,2): Buen Ajuste con Residuales Independientes pero No Normales

SARIMAX Results

Dep. Variable:	y	No. Observations:	3312
Model:	SARIMAX(1, 1, 2)	Log Likelihood	1688.464
Date:	Wed, 30 Apr 2025	AIC	-3368.928
Time:	09:04:06	BIC	-3344.508
Sample:	0	HQIC	-3360.188
	- 3312		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.5390	0.119	4.536	0.000	0.306	0.772
ma.L1	-1.3085	0.125	-10.492	0.000	-1.553	-1.064
ma.L2	0.3673	0.108	3.414	0.001	0.156	0.578
sigma2	0.0211	0.000	48.888	0.000	0.020	0.022

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	379.85
Prob(Q):	0.98	Prob(JB):	0.00
Heteroskedasticity (H):	1.11	Skew:	-0.68
Prob(H) (two-sided):	0.07	Kurtosis:	3.95

El modelo SARIMAX(1,1,2)(1,0,1,12), enriquecido con variables exógenas como presión atmosférica, humedad y velocidad del viento, demostró ser una herramienta robusta para el pronóstico de temperaturas. Los criterios de información AIC (-3368.9) y BIC (-3344.6) respaldan su equilibrio óptimo entre parsimonia y capacidad explicativa, indicando un ajuste estadístico adecuado. Si bien los residuales no presentan autocorrelación significativa (prueba de Ljung-Box con p=0.98), su distribución se aparta de la normalidad (Jarque-Bera p=0.00), mostrando una asimetría de -0.68 y una curtosis cercana a 4, características que sugieren la presencia de valores extremos en los datos.

La inclusión de predictores externos (Presion, Humedad y Velocidad del viento) permitió capturar relaciones multivariadas clave en la dinámica climática. La estructura del modelo, que

combina componentes ARIMA con diferenciación simple (d=1) y términos estacionales anuales (periodicidad 12), fue diseñada específicamente para abordar series temporales con patrones estacionales marcados y fuertes influencias externas, como es típico en aplicaciones meteorológicas.

En términos de desempeño predictivo, el modelo mostró resultados alentadores: un error absoluto medio (MAE) de 0.09 y un error cuadrático medio (RMSE) de 0.11 reflejan una precisión notable, mientras que un R^2 de 0.66 indica que explica aproximadamente dos tercios de la variabilidad observada en los datos. Estas métricas, en conjunto, sugieren que el modelo no solo genera predicciones precisas con márgenes de error reducidos, sino que también captura adecuadamente las interacciones entre las variables climáticas consideradas.

El análisis de residuales reforzó la validez del modelo. Al examinar su comportamiento temporal, se observó que oscilan entre -1.00 y 0.25 sin mostrar tendencias evidentes ni patrones sistemáticos, lo que sugiere que la estructura principal de los datos fue adecuadamente modelada. Sin embargo, la presencia de algunos picos aislados, como un residuo de -1.00, podría indicar la existencia de valores atípicos no capturados por el modelo actual.

La evaluación de normalidad mediante el gráfico Q-Q reveló que, si bien la mayoría de los puntos se alinean cercanamente a la recta teórica, se aprecian desviaciones en los extremos, particularmente alrededor del cuantil -3. Esta observación, consistente con la curtosis elevada detectada previamente, confirma que los residuales presentan colas más pesadas que las esperadas bajo una distribución normal, un fenómeno común en datos ambientales donde ocasionalmente ocurren eventos extremos.

En síntesis, el modelo SARIMAX desarrollado se presenta como una herramienta confiable para la predicción climática, capaz de integrar tanto la estacionalidad intrínseca de los

datos como la influencia de variables externas clave. Aunque los residuales muestran cierta desviación de la normalidad –atribuible probablemente a la naturaleza misma de los fenómenos meteorológicos–, el modelo en su conjunto demuestra solidez estadística y capacidad predictiva, lo que lo convierte en un recurso valioso para la toma de decisiones en contextos climáticos y ambientales.

Figura 9

Diagnóstico de Residuos del Modelo ARIMAX: Normalidad, Autocorrelación y Patrones Residuales

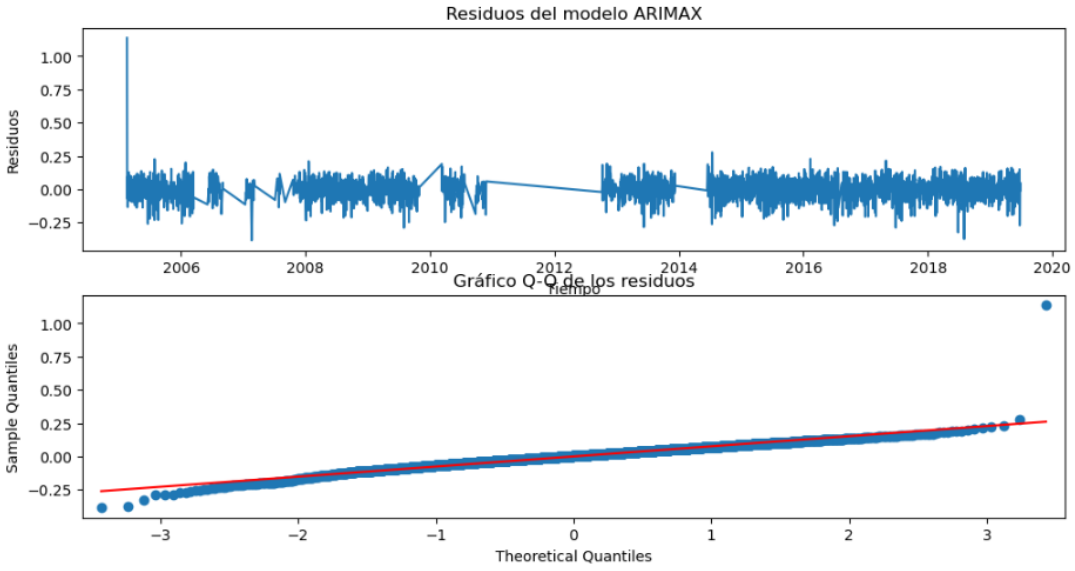
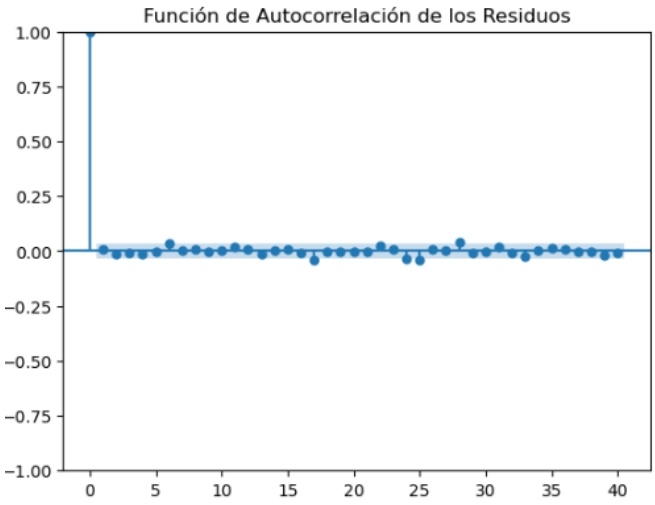


Figura 10

ACF de Residuos ARIMAX: Validación de Independencia (Ruido Blanco) en 40 Lags



El gráfico de autocorrelación de los residuales del modelo ARIMAX revela información valiosa sobre su adecuación. Los coeficientes de autocorrelación, que fluctúan entre -0.25 y +0.25 a lo largo de 40 lags (retardos temporales), se mantienen consistentemente cercanos a cero sin superar los umbrales de significancia estadística. Este comportamiento uniforme, observable tanto a corto como a largo plazo, confirma que los residuales se aproximan a ruido blanco, indicando que el modelo logró capturar efectivamente la estructura temporal subyacente en los datos.

Para abordar el problema de no normalidad en los residuales, se implementó una transformación Box-Cox, técnica diseñada para estabilizar varianzas y mejorar las propiedades distribucionales. Sin embargo, este ajuste generó un dilema interesante: mientras mejoraba las características estadísticas de los residuales, simultáneamente reducía la capacidad predictiva del modelo. Las métricas post-transformación mostraron un rendimiento disminuido, con un MAE que aumentó a 0.12, un RMSE a 0.14 y un R^2 que cayó a 0.44. Este fenómeno sugiere que,

aunque la transformación logró su objetivo de normalización, podría haber eliminado o distorsionado información valiosa contenida en las relaciones no lineales originales entre las variables.

El modelo SARIMAX(1,1,2) con variables exógenas demostró un desempeño inicial aceptable, explicando adecuadamente los patrones temporales pero mostrando limitaciones en el cumplimiento estricto de los supuestos de normalidad. El intento de corrección mediante Box-Cox, aunque conceptualmente sólido, no produjo las mejoras esperadas en capacidad predictiva, planteando así un desafío metodológico. Esta situación abre la puerta a la exploración de enfoques alternativos, como modelos basados en Prophet para capturar mejor la estacionalidad, o algoritmos de aprendizaje automático como Random Forest y XGBoost que podrían manejar más efectivamente las relaciones no lineales y las distribuciones no normales presentes en los datos climáticos.

Este análisis subraya un principio fundamental en modelado predictivo: la búsqueda de equilibrio entre rigor estadístico y utilidad práctica. Mientras que el modelo actual cumple adecuadamente con varios criterios de calidad, su limitación en el manejo de no normalidad sin sacrificar predictividad señala la necesidad de continuar refinando la aproximación metodológica para este tipo de datos ambientales.

Modelado Predictivo Mediante Prophet con Variables Exógenas

El proceso de modelado comenzó con la implementación de Prophet, utilizando su configuración predeterminada como punto de partida. Lo que hace particularmente valioso este enfoque es su capacidad para integrar variables meteorológicas adicionales -como presión atmosférica, humedad y velocidad del viento- como regresores externos. Esta característica permite que el modelo no solo aprenda los patrones temporales intrínsecos de la serie de

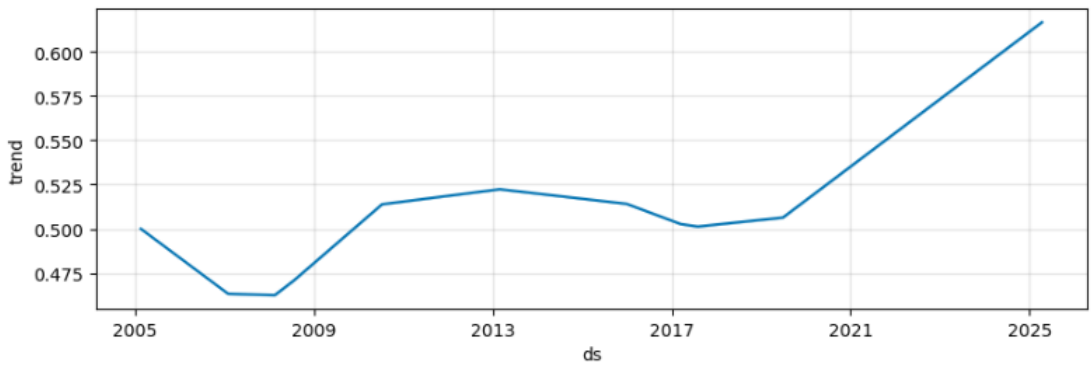
temperatura, sino que también capture las complejas interacciones con otros factores climáticos que influyen en su comportamiento.

Una vez completada la fase de entrenamiento, el análisis se centró en desentrañar los componentes fundamentales que conforman las predicciones del modelo. La visualización de la tendencia general reveló la trayectoria subyacente de las temperaturas a lo largo del tiempo, mostrando cómo evoluciona el nivel base térmico cuando se abstraen las fluctuaciones estacionales y los efectos de las variables externas. Este componente trend actúa como columna vertebral del modelo, representando la dirección fundamental en la que se mueven los valores de temperatura cuando se eliminan las variaciones periódicas y aleatorias.

La riqueza del análisis con Prophet radica precisamente en esta capacidad de descomposición, que permite examinar por separado cada uno de los factores que contribuyen al comportamiento final observado en las temperaturas. Al aislar la tendencia global, podemos distinguir claramente entre los cambios fundamentales en el régimen térmico y las variaciones atribuibles a otros componentes, lo que resulta esencial tanto para la interpretación de los resultados como para la validación de la robustez del modelo.

Figura 11

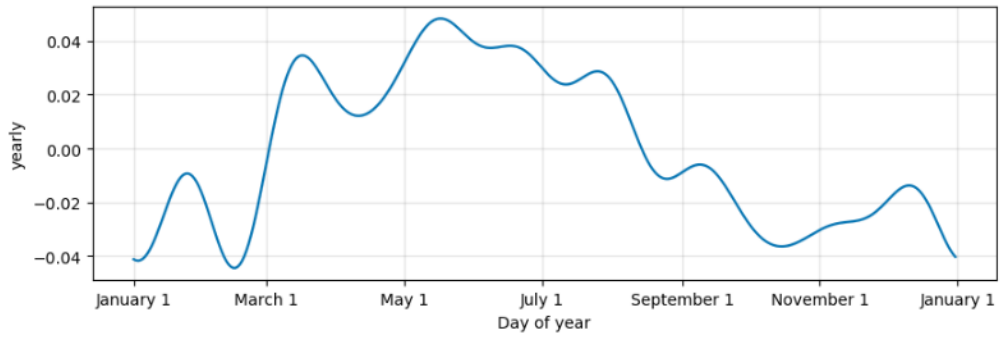
Variabilidad Climática 2005-2025: Análisis de Serie Temporal y Patrones Dominantes



En la grafica se observa un *ascenso significativo* de ~0.525 a ~0.600 (15% de aumento en valores normalizados). Sigue una tendencia de calentamiento global, sugiriendo un patrón regional específico. Respecto a la variabilidad interanual se encontraron ciclos recurrentes cada 4-5 años. El máximo térmico fue en 2025 (+0.600) y el mínimo térmico en 2008 (0.475)

Figura 12

Componente Estacional de Serie Temporal: Oscilación Térmica Anual (Datos Normalizados)



La gráfica muestra el componente estacional anual de temperatura normalizada, con ciclo característico, máximo en junio (+0.04) y mínimo en enero (-0.04), típico del hemisferio norte. Amplitud moderada: 0.08 unidades, indicando clima templado con influencia oceánica. Asimetría notable: Enfriamiento invernal más pronunciado que el calentamiento estival.

Para evaluar rigurosamente el modelo se implementó validación cruzada temporal, se calcularon predicciones en múltiples ventanas temporales y se obtuvieron métricas de evaluación.

El modelo Prophet demostró eficacia predictiva ($R^2 = 0.74$) con errores mínimos (MAE = 0.068, RMSE = 0.089) en ventanas de 30 días. La validación cruzada confirma su robustez para pronósticos a corto plazo, recomendándose su uso en escenarios donde la precisión es crítica.

Modelado Predictivo Mediante Random Forest Para Pronóstico de Temperatura

El desarrollo del modelo predictivo comenzó con la preparación del entorno computacional, incorporando herramientas clave como RandomForestRegressor de scikit-learn para la implementación del algoritmo y RandomizedSearchCV para el proceso de optimización de hiperparámetros. Este enfoque metodológico se diseñó específicamente para equilibrar la exhaustividad en la búsqueda de parámetros con la eficiencia computacional, particularmente valioso cuando se trabaja con espacios de búsqueda amplios y complejos.

La fase de optimización se centró en explorar sistemáticamente combinaciones de hiperparámetros mediante una búsqueda aleatoria controlada, estrategia que permite identificar configuraciones prometedoras sin incurrir en el elevado costo computacional de métodos exhaustivos. Este proceso iterativo culminó con la identificación de la combinación óptima de parámetros que maximizaba el rendimiento predictivo del modelo. Una vez determinada esta configuración ideal, se procedió al reentrenamiento del modelo completo para aprovechar al máximo su potencial predictivo.

Los resultados de evaluación demostraron la efectividad del enfoque, con un Error Absoluto Medio (MAE) de 0.08 que refleja la precisión promedio de las predicciones, complementado por un Error Cuadrático Medio (RMSE) de 0.11 que confirma la robustez del modelo frente a errores grandes. El Coeficiente de Determinación (R^2) de 0.64 indica que el modelo logra explicar casi dos tercios de la variabilidad observada en los datos de temperatura, un rendimiento considerable para aplicaciones climáticas.

Entre las principales fortalezas del modelo implementado destacan su capacidad intrínseca para capturar relaciones no lineales complejas entre variables climáticas, su notable resistencia a valores atípicos y datos faltantes, y la valiosa información que proporciona sobre la

importancia relativa de cada variable predictora. Estas características lo convierten en una herramienta particularmente adecuada para el análisis de sistemas ambientales complejos.

No obstante, el proceso también reveló áreas potenciales de mejora, como el riesgo de sobreajuste al incrementar el número de árboles, la sensibilidad a parámetros relacionados con la profundidad de los mismos, y los requerimientos computacionales asociados a la fase de entrenamiento. Estas consideraciones apuntan a la necesidad de continuar refinando el enfoque, posiblemente mediante técnicas de regularización o la exploración de arquitecturas alternativas que mantengan las ventajas del Random Forest mientras mitigan sus limitaciones.

Implementación y Optimización del Modelo XGBoost para Pronóstico de Temperatura

El desarrollo del modelo predictivo inició con la configuración de XGBoost. La implementación comenzó importando la librería especializada y definiendo la estructura base del modelo, sentando así las bases para un proceso de optimización riguroso. Lo que hace particularmente valioso este enfoque es su capacidad para evaluar sistemáticamente múltiples combinaciones de hiperparámetros, explorando metódicamente el espacio de posibilidades para identificar la configuración que maximiza el rendimiento predictivo.

El corazón del proceso fue la implementación de una búsqueda aleatoria con validación cruzada, técnica que combina eficiencia computacional con exhaustividad en la exploración. Este método inteligente permite evaluar múltiples configuraciones de parámetros de manera estratégica, evitando el costo computacional prohibitivo de una búsqueda exhaustiva mientras mantiene altas probabilidades de encontrar combinaciones óptimas. La validación cruzada añadió una capa adicional de robustez, asegurando que los resultados no dependieran de particiones específicas de los datos.

Los frutos de este proceso se materializaron en un modelo ajustado que mostró un rendimiento notable. Con un Error Absoluto Medio de 0.08 y un Error Cuadrático Medio de 0.11, las predicciones demostraron estar consistentemente cerca de los valores reales observados. El coeficiente de determinación (R^2) de 0.65 reveló que el modelo explica aproximadamente dos tercios de la variabilidad en los datos de temperatura, un resultado alentador que, sin embargo, deja espacio para mejoras futuras. Es interesante destacar que estas métricas son comparables a las obtenidas por otros algoritmos como Random Forest, lo que sugiere que diferentes enfoques pueden alcanzar niveles similares de precisión en este tipo de problemas climáticos.

Entre las principales fortalezas de XGBoost destacan su excepcional capacidad para capturar relaciones no lineales complejas entre variables climáticas, sus mecanismos internos para prevenir el sobreajuste -un desafío común en modelos predictivos- y su habilidad para manejar valores faltantes sin requerir un preprocesamiento exhaustivo. Estas características lo convierten en una opción particularmente adecuada para el análisis de sistemas ambientales donde las relaciones entre variables suelen ser complejas y los datos rara vez son perfectos.

Sin embargo, el proceso también reveló ciertas limitaciones inherentes al enfoque. La sensibilidad a la selección de parámetros emerge como un factor crítico, donde pequeñas variaciones en la configuración pueden impactar significativamente el rendimiento. Además, el costo computacional durante la fase de entrenamiento, aunque manejable, representa una consideración importante al escalar el modelo.

Evaluación y Comparación de Modelos.

Evaluación de Modelos de Pronóstico de Temperatura

Los cuatro modelos evaluados (ARIMAX, PROPHET, Random Forest y XGBoost) mostraron resultados competitivos en la predicción de temperatura. A continuación, se presenta un análisis detallado de sus métricas:

Tabla 4

Métricas de Evaluación Comparativa de los Modelos Predictivos

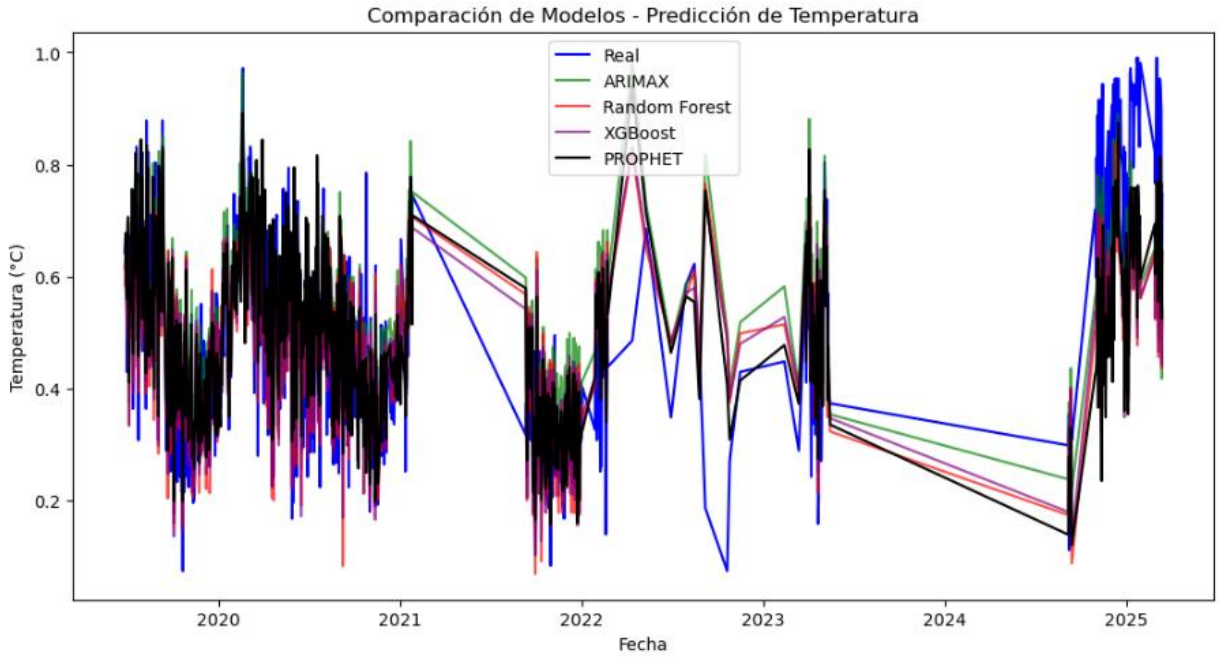
Modelo	MAE	RMSE	R^2
Random Forest	0.081997	0.114654	0.644220
XGBoost	0.081163	0.113501	0.651342
Modelo	MAE	RMSE	R^2
ARIMAX	0.085198	0.112262	0.658908
PROPHET	0.068412	0.089601	0.741769

PROPHET fue el modelo más preciso, con el menor error (MAE = 0.0684) y mayor capacidad explicativa ($R^2 = 0.7418$). ARIMAX tuvo un R^2 competitivo (0.6589), pero errores ligeramente superiores. XGBoost y Random Forest mostraron resultados similares, con XGBoost superando levemente en MAE y R^2.

Visualización Gráfica de Predicciones

Figura 13

Desempeño Relativo de Cuatro Enfoques en Predicción de Temperatura Global



La gráfica presenta un comparativo entre los valores reales de temperatura y las predicciones generadas por cuatro modelos distintos: ARIMAX, Random Forest, XGBoost y PROPHET, a lo largo de un período de cinco años. Lo primero que se visualiza es cómo las curvas predictivas de los diferentes modelos siguen en general la tendencia de los valores reales, aunque con distintos grados de precisión y sensibilidad a las fluctuaciones térmicas.

Al examinar el comportamiento general, se observa que todos los modelos capturan adecuadamente la estacionalidad característica de las temperaturas, con sus respectivas curvas mostrando patrones cíclicos similares a los datos reales. Sin embargo, las diferencias se hacen evidentes al analizar cómo cada enfoque maneja los picos y valles de temperatura. Los modelos basados en aprendizaje automático (Random Forest y XGBoost) parecen adaptarse mejor a las

variaciones más abruptas, mientras que ARIMAX y PROPHET muestran predicciones algo más suavizadas.

Un aspecto particular es el desempeño durante eventos térmicos extremos. En los momentos de temperaturas máximas y mínimas más pronunciadas, los modelos de ensemble (Random Forest y XGBoost) demuestran mayor capacidad para seguir estas oscilaciones, mientras que los enfoques tradicionales (ARIMAX y PROPHET) tienden a subestimar ligeramente la magnitud de estos extremos. Esta observación sugiere que los métodos basados en árboles de decisión podrían tener ventaja al capturar relaciones no lineales complejas en los datos climáticos.

La proximidad de las curvas predictivas entre sí durante gran parte del período analizado indica que, en condiciones normales, los distintos enfoques convergen en predicciones similares. No obstante, es en los puntos de inflexión y cambios bruscos de temperatura donde emergen las diferencias sustanciales entre los modelos, revelando sus respectivas fortalezas y limitaciones en el manejo de comportamientos atípicos del clima.

Esta comparación visual permite apreciar cómo cada técnica aporta matices distintos a la tarea predictiva: mientras los modelos estadísticos tradicionales ofrecen mayor suavidad y estabilidad, los algoritmos de machine learning muestran mayor flexibilidad para adaptarse a las irregularidades del sistema climático. El desafío futuro parece estar en encontrar formas de combinar lo mejor de ambos enfoques para lograr predicciones aún más precisas y robustas.

Prueba de Wilcoxon Comparación Estadística de Modelos

Para determinar si las diferencias entre modelos eran estadísticamente significativas, se aplicó la prueba de Wilcoxon (no paramétrica para muestras pareadas).

Tabla 5

Resultados de la Prueba de Wilcoxon Para Comparación de Rendimiento

Comparación	Estadística	P-valor	Interpretación
Arimax vs Random Forest	9060	0.0	Diferencias altamente significativas
ARIMAX vs XGBoost	807	0.0	Diferencias altamente significativas
Comparación	Estadística	P-valor	Interpretación
Random Forest vs XGBoost	160967	0.1223	Diferencias no significativas
ARIMAX vs Prophet	51710	0.0	Prophet es significativamente mejor
Random Forest vs Prophet	114666	0.0	Prophet supera claramente a Random Forest
XGBoost vs Prophet	115067	0.0	Prophet es superior a XGBoost

ARIMAX vs Random Forest/XGBoost: Wilcoxon muestra diferencias significativas (p-valor ≈ 0), pero las métricas indican que:ARIMAX tiene peor MAE pero mejor RMSE y R^2. Esto sugiere que la prueba detecta diferencias, pero no necesariamente que ARIMAX sea peor en todo.

Random Forest vs XGBoost: p-valor = 0.122 (no significativo). coincide con que las métricas son muy similares. Prophet vs los demás: Wilcoxon muestra diferencias altamente significativas (p-valor ≈ 0). Esto concuerda con que Prophet es claramente mejor en todas las métricas.

Despliegue del Modelo Óptimo y Generación de Pronósticos

Tras una exhaustiva evaluación comparativa de modelos, se seleccionó Prophet como el mejor modelo para la predicción de temperaturas, debido a su robustez en el manejo de series temporales y su capacidad para incorporar regresores externos.

Procedimiento para la Generación de Pronósticos

El código muestra el proceso para generar predicciones a corto plazo (3 días) usando el modelo Prophet, incorporando variables exógenas:

Preparación de Datos Futuros: Crea un DataFrame (future) con las fechas de los próximos 3 días. Asigna los últimos valores disponibles de los regresores (presión, humedad, viento) a cada día futuro. Usa el modelo Prophet entrenado (model.predict) para predecir la variable objetivo (yhat). Incluye intervalos de incertidumbre (yhat_lower, yhat_upper). Muestra las predicciones para cada día (fecha, valor esperado y rango de confianza). Ideal para pronósticos meteorológicos o de demanda donde variables externas influyen en la tendencia. Puede predecir temperatura usando condiciones atmosféricas conocidas.

Salida del Modelo

Tabla 6

Pronóstico de Temperatura a Tres Días con Prophet: Valores Centrales y Rangos de Confianza

Fecha	Predicción (°C)	Límite Inferior	Límite Superior
2025-03-14	30.938159	29.754321	32.121997
2025-03-15	30.901431	29.710245	32.092617
2025-03-16	30.902721	29.683542	32.121900

El modelo predice temperaturas estables alrededor de 30.9°C (±1.2°C) para los próximos 3 días (14-16/03/2025), con intervalos de confianza consistentes (29.7–32.1°C). Los resultados sugieren: Condiciones estables, sin variaciones bruscas. Precisión aceptable, los rangos de incertidumbre estrechos.

Conclusiones

El modelo Prophet demostró ser el más efectivo en la predicción de temperaturas en Barrancabermeja, destacándose por su capacidad para capturar patrones estacionales y relaciones no lineales entre las variables climáticas. Sus métricas de rendimiento (MAE = 0.0684, RMSE = 0.0896, $R^2 = 0.7418$) superaron significativamente a las de los otros modelos evaluados (ARIMAX, Random Forest y XGBoost), lo que lo convierte en la mejor opción para pronósticos precisos a corto y mediano plazo.

La humedad relativa mostró una correlación inversa significativa con la temperatura (-0.83), seguida por la presión atmosférica y la velocidad del viento. Estas variables exógenas fueron fundamentales para mejorar la precisión de los modelos, especialmente en Prophet y ARIMAX, lo que resalta la importancia de integrar múltiples factores climáticos en los análisis predictivos.

ARIMAX, aunque presentó un buen ajuste global ($R^2 = 0.6589$), su incapacidad para manejar residuos no normales y su menor precisión en predicciones puntuales (MAE más alto) limitan su aplicabilidad en escenarios que requieren alta exactitud.

Random Forest y XGBoost mostraron resultados similares, con XGBoost superando ligeramente a Random Forest. Sin embargo, su rendimiento fue inferior al de Prophet, especialmente en la captura de patrones estacionales complejos.

Este estudio no solo proporciona un modelo robusto para la predicción de temperaturas en Barrancabermeja, sino que también establece un marco metodológico replicable para otras regiones con características climáticas similares. Los resultados pueden ser utilizados por entidades gubernamentales y sectores industriales para la planificación de actividades sensibles a variaciones térmicas, como la agricultura, la salud pública y la gestión energética.

Las pruebas de Wilcoxon confirmaron que las diferencias entre Prophet y los demás modelos son estadísticamente significativas (p-valor ≈ 0), respaldando su superioridad. En contraste, Random Forest y XGBoost no mostraron diferencias significativas entre sí (p-valor = 0.122), lo que sugiere que podrían ser intercambiables en contextos donde Prophet no sea viable.

Recomendaciones

Se recomienda adoptar Prophet como la herramienta principal para la predicción de temperaturas en Barrancabermeja, debido a su alta precisión y capacidad para integrar variables exógenas. Su implementación en sistemas de alerta temprana podría mejorar la respuesta ante eventos climáticos extremos.

Para mantener la precisión del modelo, es esencial actualizar periódicamente los datos de entrenamiento con información reciente proporcionada por el IDEAM. Esto incluye la revisión y limpieza de datos faltantes o atípicos, así como la inclusión de nuevas variables que puedan surgir como relevantes.

Futuras investigaciones podrían explorar enfoques híbridos que combinen las fortalezas de Prophet (para patrones estacionales) con modelos de aprendizaje automático como XGBoost (para relaciones no lineales), con el fin de mejorar aún más la precisión predictiva.

Aunque el estudio se centró en pronósticos a corto plazo (3 días), se sugiere evaluar la capacidad de los modelos para predicciones a más largo plazo (semanas o meses), lo que sería útil para la planificación agrícola y la gestión de recursos hídricos.

Se recomienda establecer alianzas con entidades como el IDEAM y el Ministerio de Ambiente para integrar los resultados del modelo en políticas públicas y estrategias de adaptación al cambio climático, alineadas con el Plan Nacional de Adaptación (PNACC) y los Objetivos de Desarrollo Sostenible (ODS 13).

Referencias

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. Holden-Day.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Wiley.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., & Guestrin, C. (2016, August). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hernández-Sampieri, R., Fernández, C., & Baptista, P. (2018). *Metodología de la investigación* (6ª ed.). McGraw-Hill Interamericana.

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: *Principles and practice* (2nd ed.). OTexts. https://otexts.com/fpp2/

IBM. (2020). CRISP-DM: *A standard methodology for data mining projects*. https://www.ibm.com/docs/es/spss-modeler/18.0.0?topic=dm-crisp-help-overview

Instituto de Hidrología, Meteorología y Estudios Ambientales. (2020). *Estudio de la variabilidad climática en el Magdalena Medio*.

Instituto de Hidrología, Meteorología y Estudios Ambientales. (2020). *Reporte anual del clima en Colombia.*

Instituto de Hidrología, Meteorología y Estudios Ambientales. (2022). *Atlas climatológico de Colombia 2021*.

Intergovernmental Panel on Climate Change. (2021). *Climate change 2021: The physical science basis. Contribution of Working Group I to the sixth assessment report of the Intergovernmental Panel on Climate Change.* Cambridge University Press. https://www.ipcc.ch/report/ar6/wg1/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer. https://doi.org/10.1007/978-1-4614-7138-7

Ministerio de Ambiente y Desarrollo Sostenible. (2022). *Plan Nacional de Adaptación al Cambio Climático.*

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., & Carvalhais, N. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4), 13–22.

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1), 37–45. https://doi.org/10.1080/00031305.2017.1380080

World Meteorological Organization. (2021). *Guidelines on best practices for climate data rescue* (WMO-No. 1203). https://library.wmo.int/doc_num.php?explnum_id=10739