Análisis comparativo de modelos arima, random forest y gradient boosting para la predicción de la demanda del sistema interconectado nacional colombiano (2024-2030)

Iván Darío Rojas Galvis

Asesor

Brayan Andru Montenegro Embus

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización Ciencia de Datos

2026

Resumen

La planificación del sector eléctrico en Colombia depende en gran medida de la capacidad para estimar la demanda energética con precisión. Este desafío adquiere especial relevancia, en un país donde aproximadamente el 70% de la generación eléctrica proviene de fuentes hidroeléctricas, lo que hace al sistema altamente vulnerable a fenómenos climáticos como El Niño y La Niña (UPME, 2021). En este estudio, se propuso comparar el desempeño de tres modelos predictivos una extensión del modelo ARIMA (SARIMAX), Random Forest y Gradient Boosting, para proyectar la demanda mensual de energía en Colombia hasta el año 2030, utilizando datos reales del sistema interconectado nacional (SIN) sobre demanda de energía, generación de energía y demanda no atendida del periodo 2010-2023 (XM,2024) y el histórico del comportamiento de las precipitaciones en Colombia 2010-2023 (datos abiertos Colombia,2025).

El proceso incluyó una fase exploratoria para identificar patrones y tendencias en los datos históricos, seguida de la implementación y ajuste de cada modelo. Se utilizaron técnicas de validación temporal para garantizar la robustez de las proyecciones y el desempeño, se evaluó mediante un conjunto de métricas ampliamente reconocidas como lo son el error absoluto medio (MAE), raíz del error cuadrático medio (RMSE), coeficiente de determinación (R^2) y error porcentual absoluto medio (MAPE).

Los resultados mostraron que el modelo SARIMAX con orden óptimo (1, 0, 1) x (0, 0, 2, 12), al incorporar formalmente los componentes de estacionalidad y las variables exógenas (generación, precipitación, demanda no atendida), se destacó como el más preciso en la proyección de los datos de prueba. SARIMAX capturó la tendencia creciente y los ciclos estacionales con una precisión excepcional, alcanzando un error porcentual absoluto medio

(MAPE) de solo 1.37%. En claro contraste, los modelos random forest y gradient boosting demostraron ser ineficaces para capturar la compleja estructura temporal de la serie, arrojando coeficientes R^2 cercanos a cero o negativos. Estos hallazgos refuerzan la superioridad de los modelos econométricos especializados, como SARIMAX, para la predicción de la demanda energética en Colombia.

Las proyecciones indican que la demanda mensual del SIN superará la barrera de los 10,000 GWh hacia finales de 2030, lo que subraya la urgencia de planificar la expansión de la oferta. El trabajo concluye entregando una herramienta validada con un MAPE del 1.37%, superior a las alternativas de Machine Learning.

Palabras claves: Demanda energética, predicción, SARIMAX, Random Forest, Gradient Boosting.

Abstract

The planning of the Colombian electric sector largely depends on the ability to estimate energy demand with accuracy. This challenge acquires special relevance in a country where approximately 70% of electricity generation comes from hydroelectric sources, making the system highly vulnerable to climatic phenomena such as El Niño and La Niña (UPME, 2021). In this study, the performance of three predictive models was compared: an extension of the ARIMA model (SARIMAX), Random Forest, and Gradient Boosting, to project the monthly energy demand in Colombia until the year 2030, using real data from the National Interconnected System (SIN) regarding demand, energy generation, and unserved demand from the 2010-2023 period (XM, 2024), and the historical behavior of precipitation in Colombia (Open Data, 2025).

The process included an exploratory phase to identify patterns and trends such as seasonality, followed by the implementation and adjustment of each model. Temporal validation techniques were used to ensure the robustness of the projections, and performance was evaluated using a set of widely recognized metrics such as the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of Determination (R^2), and Mean Absolute Percentage Error (MAPE).

The results showed that the SARIMAX model with optimal order (1, 0, 1) x (0, 0, 2, 12), by formally incorporating the components of seasonality and the exogenous variables (Generation, Precipitation, Unserved Demand), stood out as the most accurate in projecting the test data. SARIMAX captured the increasing trend and the seasonal cycles with exceptional precision, achieving a Mean Absolute Percentage Error (MAPE) of only 1.37%. In stark contrast, the Random Forest and Gradient Boosting models proved ineffective at capturing the complex

temporal structure of the series, yielding R^2 coefficients close to zero or negative. These findings reinforce the superiority of specialized econometric models, such as SARIMAX, for energy demand prediction in Colombia.

Projections indicate that the monthly demand of the National Interconnected System (SIN) will exceed the 10,000 GWh threshold by late 2030, highlighting the urgent need for supply expansion planning. This study concludes by delivering a validated tool with a 1.37% MAPE, which outperforms the evaluated Machine Learning alternatives

Keywords: Energy demand, forecasting, ARIMA, Random Forest, Gradient Boosting.

Lista de Contenido

Lista de Figuras

Lista de Tablas

Lista de Apéndices

Introducción

La demanda energética en Colombia ha experimentado un crecimiento sostenido en las últimas décadas, impulsado por el desarrollo económico y el aumento de la población (XM, 2023). Según la entidad UPME, el consumo eléctrico ha crecido a una tasa promedio anual del 2.5% desde 2010, un patrón que se espera persista en el futuro cercano (UPME, 2021). Este crecimiento continuo está críticamente condicionado por la alta dependencia de la hidroelectricidad (aproximadamente el 70% de la generación (UPME, 2021)), que introduce una vulnerabilidad directa entre las variables climáticas (como la precipitación) y la disponibilidad de energía, afectando indirectamente la demanda (IPCC, 2021). Es así que, la predicción precisa de la demanda energética a largo plazo se vuelve esencial para garantizar la estabilidad del sistema interconectado nacional (SIN) y planificar inversiones en infraestructura energética (Ministerio de Minas y Energía, 2019).

Este trabajo tiene como objetivo principal proyectar la demanda energética mensual en Colombia desde enero de 2024 hasta diciembre de 2030, utilizando datos reales del SIN recopilados entre 2010 y 2023. Para alcanzar este objetivo, se establece un análisis comparativo de tres enfoques predictivos: los modelos de machine learning (Random Forest y Gradient Boosting) (Breiman y Friedman, 2001) y un modelo de series de tiempo que se origina en el marco ARIMA (Box y Jenkins, 1970). Sin embargo, la identificación formal de la estacionalidad en la serie de demanda y la necesidad de incorporar factores externos llevaron a la implementación del modelo SARIMAX (ARIMA Estacional con Variables Exógenas). Esta metodología es crucial, pues permite modelar de forma simultánea los patrones de estacionalidad y la influencia de las variables exógenas claves utilizadas en el análisis: la generación de energía, la precipitación y la demanda no atendida.

La evaluación del desempeño se realiza rigurosamente mediante métricas múltiples (MAE, RMSE, R^2 y MAPE), proporcionando una visión integral de la precisión y utilidad de cada modelo (Hyndman y Athanasopoulos, 2018). Los resultados de esta investigación, al validar la superioridad del modelo SARIMAX, buscan fortalecer la toma de decisiones estratégicas en el sector energético colombiano, ofreciendo una herramienta de alta precisión para la planificación de largo plazo.

Planteamiento del Problema

El sector eléctrico colombiano, con una dependencia de la hidroelectricidad cercana al 70% (UPME, 2021), enfrenta la necesidad crucial de proyectar la demanda energética con exactitud para garantizar la estabilidad del suministro a largo plazo. Sin embargo, este desafío se agrava por dos factores: la vulnerabilidad sistémica ante fenómenos climáticos extremos (como El Niño y La Niña) (IPCC, 2021) y la limitación de los modelos predictivos tradicionales.

Modelos fundamentales como ARIMA, aunque esenciales, demuestran ser inadecuados para manejar la estacionalidad persistente y la influencia de variables exógenas (e.g., precipitación y generación) que caracterizan la demanda del SIN(Shumway y Stoffer, 2017). Si bien los modelos de inteligencia artificial como Random Forest y Gradient Boosting ofrecen alternativas para patrones no lineales (Breiman, 2001; Friedman, 2001), es imperativo validar si superan el rendimiento de los modelos econométricos diseñados para series temporales (Hyndman y Athanasopoulos, 2018; Zhang, 2003).

La ausencia de un modelo predictivo riguroso que integre la estructura temporal, la estacionalidad y las variables exógenas clave genera incertidumbre en la planificación a largo plazo. Por lo tanto, se plantea la siguiente pregunta central:

¿Cuál de los modelos predictivos (SARIMAX, Random Forest o Gradient Boosting) ofrece la mayor precisión y adaptabilidad para proyectar la demanda energética mensual en Colombia hasta 2030, considerando la marcada estacionalidad de la serie y la influencia de las variables exógenas clave (Generación, Precipitación y Demanda No Atendida) a partir de los datos históricos del SIN (2010-2023)?

Sistematización del Problema

El problema se desglosa en preguntas operacionales que guían el análisis y la selección del modelo más adecuado:

1. Validación del Modelo Especializado: ¿La migración metodológica de ARIMA a SARIMAX se justifica mediante el análisis de la serie, y el modelo resultante logra capturar de forma estadísticamente significativa la influencia de las variables exógenas (generación, precipitación)?

2. Análisis Comparativo y Superioridad: ¿Cuál es el rendimiento predictivo del modelo SARIMAX frente a los modelos de Machine Learning (Random Forest y Gradient Boosting) en el conjunto de prueba (2022-2023), medido por el error porcentual absoluto medio (MAPE), el error absoluto medio (MAE), la raíz del error cuadrático medio (RMSE) y el coeficiente de determinación (R^2)?

3. Generación del Pronóstico: ¿Cuál es la proyección final de la demanda energética mensual en Colombia hasta diciembre de 2030 utilizando el modelo seleccionado, y cómo se cuantifica el riesgo asociado a los supuestos de las variables exógenas?

Justificación

El presente estudio se justifica plenamente en tres dimensiones interrelacionadas la estratégica, la metodológica y la académica, las cuales validan su relevancia para la toma de decisiones y el conocimiento en el sector energético colombiano.

1. Implicación estratégica y regulatoria

La predicción precisa de la demanda energética constituye un imperativo regulatorio y operativo dentro del sistema interconectado nacional (SIN). el ministerio de minas y energía establece la necesidad de contar con proyecciones confiables para la planificación a largo plazo.

Una predicción inexacta conlleva riesgos significativos (desde racionamientos hasta inversiones ineficientes) (Ministerio de minas y energía, 2019).

Este trabajo aborda directamente esta necesidad al proporcionar una herramienta de alta fidelidad, cuyo modelo seleccionado (SARIMAX) fue validado con un error porcentual absoluto medio (MAPE) de 1.37% sobre el conjunto de prueba. Este nivel de precisión es esencial para minimizar la incertidumbre operativa y mejorar la gestión de riesgos que el sistema interconectado monitorea constantemente. Por consiguiente, el estudio ofrece un insumo estratégico de alto valor para la planificación, permitiendo a los agentes del sector optimizar la gestión de recursos y la definición de la futura expansión del parque generador.

2. Rigor Metodológico y Enfoque Multivariable

La investigación se justifica por la aplicación de una metodología que supera las limitaciones de los modelos invariados.

- Validez del modelo: Se valida la necesidad de evolucionar del marco ARIMA al SARIMAX, sustentado en el análisis de la estacionalidad persistente de la demanda de energía. Este rigor estadístico asegura que el modelo capta la compleja estructura temporal inherente.
- Integración de variables exógenas: El uso de datos reales del periodo 2010-2023 dota al análisis de un horizonte temporal amplio y robusto. Además, la integración formal de variables exógenas clave (generación, precipitación y demanda no atendida) confirma su impacto en la demanda, abordando una limitación recurrente en la literatura del sector que a menudo subestima los factores climáticos en la modelización de la demanda energética en países hidrodependientes (Peña, 2018).

3. Contribución a la ciencia de datos y al debate académico

Este trabajo contribuye al debate académico contemporáneo sobre la elección óptima de modelos en contextos de series de tiempo con fuerte estructura. La comparación exhaustiva y directa entre el modelo econométrico especializado (SARIMAX) y los modelos de Machine Learning (Random Forest y Gradient Boosting) tiene implicaciones claras:

- Se demuestra que el modelado temporal, es un factor determinante para la precisión en este dominio, confirmando la superioridad del SARIMAX.
- La evaluación del desempeño mediante métricas múltiples (MAE, RMSE, R^2, MAPE) asegura que la validación no es unidimensional, sino robusta y completa, alineándose con estándares académicos (Makridakis, 1998).

Objetivos

Objetivo General

Comparar el desempeño predictivo de modelos SARIMAX (extensión de ARIMA), Random Forest y Gradient Boosting, incluyendo la influencia de variables exógenas, mediante métricas de precisión (MAE, RMSE, R^2, MAPE) aplicados a series de tiempo de demanda energética mensual en Colombia (2010-2023), para seleccionar el modelo más robusto y generar pronósticos hasta el año 2030.

Objetivos Específicos

Realizar análisis exploratorio y de estacionalidad, de las series de tiempo de demanda mensual del SIN (2010-2023), identificando patrones de tendencia y componentes estacionales identificando la correlación con las variables exógenas clave (generación, precipitación y demanda no atendida).

Implementar los modelos SARIMAX, Random Forest y Gradient Boosting, utilizando técnicas de validación cruzada temporal sobre los datos que describen el comportamiento de la demanda de energía de Colombia.

Evaluar los modelos predictivos seleccionados, mediante métricas MAE, RMSE, R^2 y MAPE, para determinar el modelo con el mejor comportamiento para la proyección de la demanda energética en el SIN hasta el 2030.

Estado del Arte

Esta sección presenta la revisión crítica de la literatura relevante en el pronóstico de la demanda energética, sintetizando la investigación que fundamenta tanto el contexto de vulnerabilidad climática como la comparación metodológica entre modelos econométricos especializados y técnicas de machine learning.

El Marco Regulatorio y la Predicción de Demanda en Colombia

La planificación energética en el sistema interconectado nacional se rige por la necesidad de asegurar la estabilidad del suministro, dada la alta dependencia de la generación hidroeléctrica.

1. Direccionamiento Metodológico: La unidad de planeación minero-energética (UPME) define el marco institucional para la planificación. La metodología para proyecciones de demanda de energía eléctrica y gas natural (UPME, 2020) es el documento rector que establece los criterios para la elaboración de pronósticos. La estimación precisa de la demanda es fundamental para la toma de decisiones estratégicas por parte de las autoridades gubernamentales.

2. Vulnerabilidad climática: La dependencia hidroeléctrica hace que el sistema sea altamente vulnerable a los fenómenos climáticos, como El Niño y La Niña. Este factor justifica la integración de variables exógenas clave, como la precipitación en los modelos predictivos.

Estudios de Pronóstico en el Contexto Colombiano

Los trabajos recientes en el país demuestran una evolución desde modelos tradicionales hacia enfoques avanzados, aunque se enfocados en horizontes temporales más cortos o en soluciones híbridas.

1. Enfoques regionales con deep learning: El trabajo de Castellanos y Ardila (2023) se centró en la predicción de la demanda eléctrica en el departamento de Boyacá. Este estudio empleó técnicas de Deep Learning (DL) para desarrollar un modelo preciso, subrayando que las técnicas de ML son viables para la gestión de la demanda y la eficiencia energética regional.

2. Pronóstico a mediano plazo: Andrade Bonilla y Castellanos Valencia (2022) abordaron la predicción de la demanda de electricidad en Cali con un horizonte de cinco años. Este horizonte de tiempo es significativo ya que valida la necesidad de generar pronósticos robustos más allá del corto plazo, que es el foco de muchos análisis en el sector.

3. Modelos híbridos para corto plazo: Montoya Cardona (2021) investigó el pronóstico de la demanda de energía a corto plazo a nivel nacional. Su propuesta se basó en el uso de un modelo híbrido adaptativo, lo cual sugiere que un modelo lineal simple es insuficiente para capturar todas las dinámicas del sistema, motivando la exploración de modelos más complejos, ya sean híbridos o multivariables especializados.

Debate Metodológico: Sarimax Multivariable vs. Machine Learning

La literatura internacional y los trabajos locales confirman la necesidad de contrastar los modelos más efectivos de cada familia para series temporales.

1. Rigor de los modelos econométricos: La solidez de la metodología Box-Jenkins (Shumway y Stoffer, 2017) es el punto de partida. La estacionalidad mensual identificada en la serie de demanda requiere la extensión del modelo ARIMA al SARIMAX, y la dependencia de factores externos obliga a la inclusión de la componente exógena (X_t). Esta especialización busca capturar la estructura temporal y la influencia lineal de las variables exógenas de manera formal, un aspecto en el que los modelos de ML pueden requerir una ingeniería de funciones extensiva.

2. El potencial de los modelos ML: Los modelos Random Forest (RF) y Gradient Boosting Regressor (GBR), basados en los trabajos fundacionales de Breiman (2001) y Friedman (2001), son reconocidos por su capacidad para manejar la no-linealidad. La presente investigación se alinea con el debate planteado por Zhang (2003) y otros, que buscan determinar si el poder predictivo de los modelos no lineales es capaz de superar la precisión y la interpretabilidad de un modelo estadístico riguroso y especializado (SARIMAX) en el pronóstico a largo plazo.

Brecha de Investigación y Contribución

La revisión de la literatura indica que, a pesar de la existencia de estudios en Colombia, existe una brecha metodológica clara que este proyecto aborda:

1. Falta de benchmarking directo: No se identificaron trabajos nacionales que realicen una comparación directa y exhaustiva a largo plazo (hasta 2030) entre el modelo SARIMAX multivariable y los potentes modelos de ensemble (RF y GBR) aplicados a la demanda nacional del SIN.

2. Validación de la multivariabilidad: El presente estudio se distingue por validar y cuantificar la mejora en la precisión al integrar formalmente variables exógenas (Precipitación y Generación), proporcionando una conclusión robusta sobre la idoneidad metodológica.

La contribución de este trabajo es, por tanto, doble: establecer la metodología más precisa para la proyección de la demanda de energía del SIN y resolver el debate sobre la superioridad predictiva entre el modelo econométrico especializado (SARIMAX) y el Machine Learning en un contexto de series temporales con fuerte estructura.

Marco Teórico

El presente capítulo establece los fundamentos conceptuales y metodológicos necesarios para comprender el análisis comparativo de modelos predictivos. La narrativa se estructura desde la definición del problema en el contexto energético nacional hasta los detalles técnicos de los modelos econométricos y de Machine Learning empleados.

Contexto Energético y Desafío de la Predicción

El SIN y la Multivariabilidad de la Demanda

La variable objetivo de este estudio es la demanda de energía en el SIN, definida como la cantidad de energía eléctrica requerida por los consumidores finales. La predicción de esta serie temporal es intrínsecamente compleja, ya que está sujeta a la confluencia de múltiples variables. Estos incluyen la estacionalidad climática (ciclos de lluvia), las variaciones económicas y el crecimiento poblacional (XM, 2023).

La Vulnerabilidad Climática y el Enfoque Multivariable

Colombia se caracteriza por una alta dependencia de la generación hidroeléctrica, lo que confiere al sistema una vulnerabilidad significativa a fenómenos climáticos extremos como El Niño y La Niña. Este hecho justifica el enfoque multivariable del proyecto, integrando formalmente factores externos (regresores exógenos) que influyen directamente en la dinámica de la demanda:

1. Generación de energía (kWh): Esta variable operacional refleja una respuesta directa e inmediata a la demanda y a la capacidad de suministro del sistema, actuando como un regresor fundamental en la dinámica a corto plazo.

2. Precipitación (mm): Su variación afecta el nivel de los embalses y por ende, la gestión de la oferta, lo cual afecta de manera decisiva en la dinámica del mercado (Peña, 2018).

3. Demanda no atendida (kWh): Esta variable refleja la energía que el sistema no pudo suministrar. Su inclusión permite modelar efectos correctivos en la dinámica de la serie de demanda.

Modelos Econométricos de Series de Tiempo: de Arima a Sarimax

El análisis y pronóstico de series de tiempo con dependencia temporal estricta requieren modelos diseñados para capturar su estructura secuencial interna.

El Requisito de la Estacionariedad

Un concepto fundamental es la estacionariedad: Una serie de tiempo es estacionaria si sus propiedades estadísticas (media y varianza) se mantienen constantes a lo largo del tiempo. Este es un requisito indispensable para la aplicación de modelos ARIMA/SARIMAX.

1. Prueba de Dickey-Fuller aumentada (ADF): Es la prueba estadística utilizada para determinar formalmente si una serie temporal posee una raíz unitaria (no es estacionaria) y por lo tanto, requiere diferenciación ($d > 0$) para alcanzar la estacionariedad y proceder con el modelado.

Modelo Sarimax: Especializada en Estacionalidad y Exogeneidad

El modelo fundacional es el ARIMA (Autoregressive Integrated Moving Average), definido por la interacción de sus componentes autorregresivo (p), integrado (d) y de medias móviles (q). Sin embargo, el análisis exploratorio de la demanda energética reveló una marcada estacionalidad mensual ($s = 12$).

Esta limitación metodológica del ARIMA simple forzó la extensión al modelo SARIMAX (Seasonal ARIMA with eXogenous variables). Este modelo se justifica plenamente al permitir el modelado simultáneo de:

1. Dependencia de Corto Plazo: Capturada por los componentes no estacionales (p, d, q).

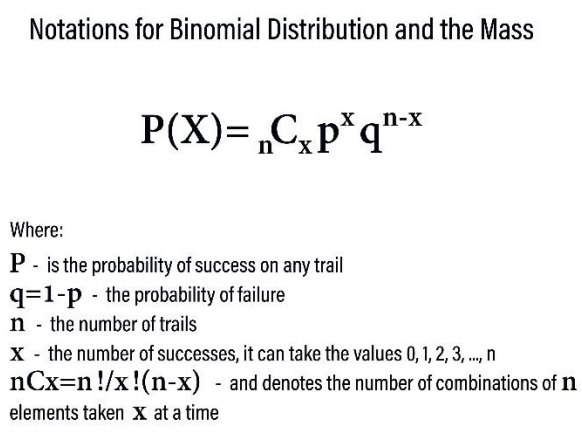
2. Estacionalidad Anual: Capturada por los componentes estacionales (P, D, Q, con s = 12).

3. Influencia Exógena: Incorporando la información de las variables externas (X_t).

La ecuación general del modelo SARIMAX combina de manera multiplicativa los componentes no estacionales y estacionales, e incluye el efecto lineal de los regresores.

Figura 1

Ecuación General Sarimax



Nota. Tomado de Zhang, G. (2003). Time Series Forecasting Using a Hybrid ARIMA And Neural Network Model

Modelos de Machine Learning (ML): El Enfoque no Paramétrico

Se utilizan dos modelos de ensemble como referencia no paramétrica para contrastar el rendimiento predictivo con el enfoque econométrico especializado.

Random Forest Regressor (RF)

El modelo Random Forest (RF) es un método de ensemble que opera construyendo un gran número de árboles de decisión de forma independiente sobre subconjuntos de datos y variables. La predicción final se obtiene promediando los resultados de todos los árboles (Breiman, 2001). Sus ventajas principales radican en su robustez ante outliers y su capacidad intrínseca para manejar y modelar relaciones no lineales complejas.

Gradient Boosting Regressor (GBR)

El Gradient Boosting Regressor (GBR) también es un método de ensemble basado en árboles, pero su construcción es secuencial. Cada nuevo árbol intenta corregir los errores (residuales) del conjunto de árboles predecesor, utilizando un algoritmo de descenso de gradiente. GBR suele ofrecer una mayor precisión que RF, pero al optimizar el ajuste de manera secuencial, puede ser más propenso al sobreajuste (overfitting).

Desafío en Series de Tiempo (Time-Agnostic Nature)

La principal limitación de los modelos RF y GBR es su naturaleza time-agnostic. No están diseñados intrínsecamente para el manejo de estructuras temporales. Por lo tanto, para lograr un desempeño comparable al de SARIMAX en series con fuerte estacionalidad, su aplicación requiere una exhaustiva ingeniería de funciones (creación de lags y features de fecha) para simular los componentes de dependencia temporal y estacionalidad que SARIMAX modela de forma nativa.

Métricas de Evaluación y Diagnóstico

La selección del modelo final se basa en una evaluación objetiva y completa de su desempeño predictivo y diagnóstico estadístico.

Métricas de Precisión

Se utilizan métricas de error absoluto y relativo para la evaluación en el conjunto de prueba:

1. Error Absoluto Medio (MAE): Promedio de los errores absolutos. Métrica de fácil interpretabilidad.

2. Raíz del Error Cuadrático Medio (RMSE): Penaliza fuertemente los errores grandes, siendo útil para evaluar la robustez frente a outliers extremos.

3. Coeficiente de Determinación (R^2): Mide la proporción de la varianza en la variable dependiente que es predecible a partir del modelo. Valores cercanos a 1 indican un buen ajuste.

4. Error Porcentual Absoluto Medio (MAPE): Expresa el error promedio en términos de porcentaje. Esta métrica es la más crítica para la toma de decisiones, ya que permite una comparación clara de la precisión del modelo frente a la magnitud real de la serie.

Diagnóstico del Modelo Econométrico

Para que el modelo SARIMAX sea considerado estadísticamente válido y eficiente, es imprescindible que sus residuales se comporten como ruido blanco (es decir, no deben contener patrones de dependencia ni autocorrelación).

1. Prueba de Ljung-Box: Es la prueba estadística utilizada para verificar si la autocorrelación de los residuales es significativamente diferente de cero. Si el p-valor es mayor que el nivel de significancia (alpha = 0.05), se concluye que los residuos son ruido blanco, confirmando que el modelo ha capturado toda la información dependiente de la serie.

Metodología

La presente investigación se desarrolló bajo un enfoque cuantitativo de tipo correlacional y longitudinal explicativo. Se define como cuantitativa y correlacional debido a que utiliza datos numéricos históricos para medir la fuerza de la relación entre la demanda de energía y variables exógenas (precipitación, generación), y es longitudinal dado que analiza la evolución de dichas variables a través del tiempo (2010-2023) para proyectar comportamientos futuros.

Para garantizar el rigor científico y dar cumplimiento a los objetivos planteados, el desarrollo metodológico se estructuró en tres fases secuenciales: (1) Caracterización estructural y análisis exploratorio, (2) Diseño experimental e implementación de modelos, y (3) Validación, evaluación comparativa y pronóstico.

Fase 1 Caracterización Estructural y Análisis de Variables Exógenas

Esta fase se orientó al cumplimiento del primer objetivo específico, enfocándose en la adquisición, depuración y comprensión de la dinámica de las series de tiempo del Sistema Interconectado Nacional (SIN). La información base se consolidó a partir de fuentes oficiales (XM y Datos Abiertos Colombia) abarcando el periodo mensual de enero de 2010 a diciembre de 2023.

El proceso inició con el preprocesamiento de datos para asegurar la integridad de las series, verificando la consistencia temporal y la ausencia de valores nulos. Posteriormente, se ejecutó un Análisis Exploratorio de Datos (EDA) exhaustivo para descomponer la variable objetivo (Demanda de Energía) en sus componentes fundamentales: tendencia, estacionalidad y residuales. En esta etapa, fue crítico validar la estacionariedad de la serie mediante la prueba de Dickey-Fuller Aumentada (ADF), requisito indispensable para la modelación econométrica.

Tabla 1

Resultados de Cointegración

Variable / Prueba	Métrica	Valor / Resultado
Prueba de Cointegración	Residuos de Generación vs. Demanda	ADF
Estadístico	Valor p (ADF de los residuos)	0.0618
Decisión	Hipótesis Nula (H_0)	NO se rechaza
Conclusión	Estado de las series	NO están cointegradas

Una vez verificada la estacionariedad, se procedió a evaluar la viabilidad estadística de incluir factores externos en el modelo. Para ello, se construyó una matriz de correlación de Pearson, herramienta estadística fundamental que permite cuantificar la fuerza y dirección de la relación lineal entre la variable objetivo y las variables exógenas propuestas.

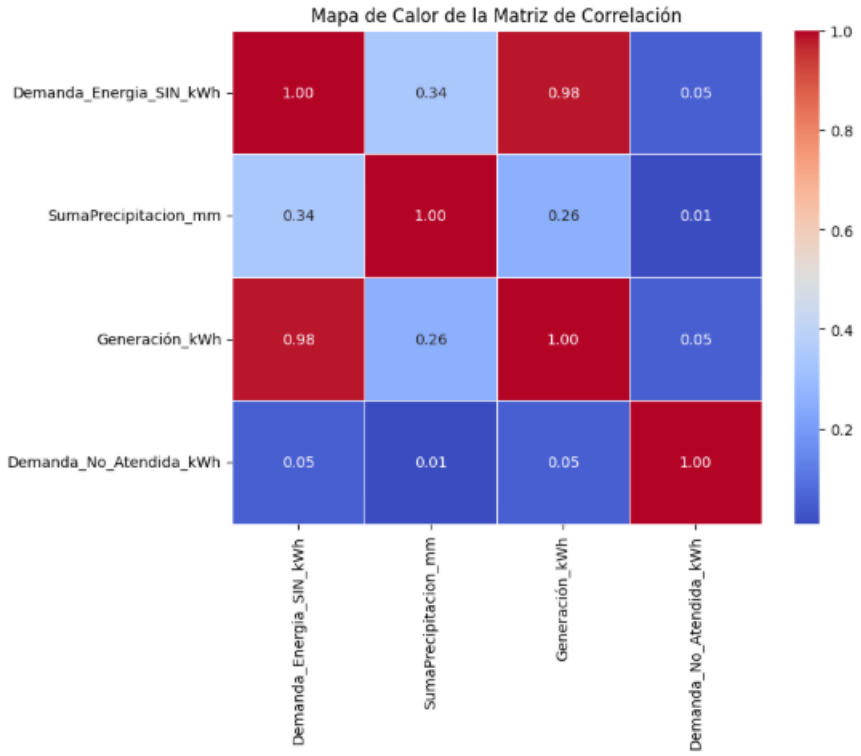
En este análisis se contrastó la variable dependiente, demanda de energía del SIN (kWh), frente a tres covariables o predictores exógenos:

1. Generación de Energía (kWh)
2. Suma de Precipitación (mm
3. Demanda No Atendida (kWh)

El coeficiente de Pearson oscila entre -1 y 1, donde valores cercanos a los extremos indican una fuerte correlación. Para facilitar la interpretación de estas relaciones multidimensionales, se generó un mapa de calor, el cual se presenta a continuación:

Figura 2

Matriz de Correlación



El análisis de los resultados presentados en la figura 2 revela hallazgos determinantes para la especificación del modelo:

En primer lugar, se observa una correlación positiva extremadamente fuerte de 0.98 entre la Generación de Energía y la Demanda. Este hallazgo confirma una relación casi simbiótica y directa: a medida que crece la demanda, el sistema responde incrementando la generación proporcionalmente. Estadísticamente, esto sugiere que la variable de generación es el predictor más potente, aunque también advierte sobre la presencia de multicolinealidad, la cual será gestionada por el modelo SARIMAX.

En segundo lugar, la variable Precipitación muestra una correlación positiva moderada de 0.34. Aunque este valor es menor que el de la generación, es estadísticamente significativo y coherente con la naturaleza hidroeléctrica del sistema. Este coeficiente indica que los ciclos de lluvia influyen en la disponibilidad y despacho de energía, validando su inclusión para capturar la estacionalidad climática que un modelo univariado ignoraría.

Finalmente, la demanda no atendida presentó una correlación lineal baja (0.05). Sin embargo, se decidió mantener esta variable en el diseño experimental no por su correlación lineal promedio, sino por su capacidad teórica para explicar anomalías en momentos críticos del sistema, aportando información sobre la robustez del suministro que no es capturada por las otras variables.

Este diagnóstico multivariable confirmó que la inclusión de estos regresores exógenos aporta información estructural valiosa, justificando el paso de un modelo ARIMA simple a un modelo SARIMAX multivariable.

Fase 2 Diseño Experimental e Implementación de Modelos

Una vez caracterizada la estructura de la demanda y validadas sus relaciones exógenas, se procedió a la etapa de modelado predictivo. Esta fase se diseñó para contrastar dos paradigmas: el econométrico (SARIMAX) y el de aprendizaje automático (Random Forest y Gradient Boosting).

Estrategia de Validación Temporal

Dado que el objetivo es la predicción de eventos futuros, se descartó la validación cruzada aleatoria tradicional (k-fold cross-validation) en favor de una validación basada en el tiempo (Time-Based Split). Esta estrategia respeta la secuencia cronológica de los datos, evitando que el modelo "aprenda" del futuro para predecir el pasado (data leakage).

1. Conjunto de Entrenamiento (Training Set): Se utilizaron los datos desde enero de 2011 hasta diciembre de 2021 (132 meses) para el ajuste de parámetros y el aprendizaje de patrones.

2. Conjunto de Prueba (Test Set): Se reservaron los datos desde enero de 2022 hasta diciembre de 2023 (24 meses) exclusivamente para evaluar la capacidad de generalización del modelo frente a datos desconocidos.

Implementación del Modelo Econométrico (SARIMAX)

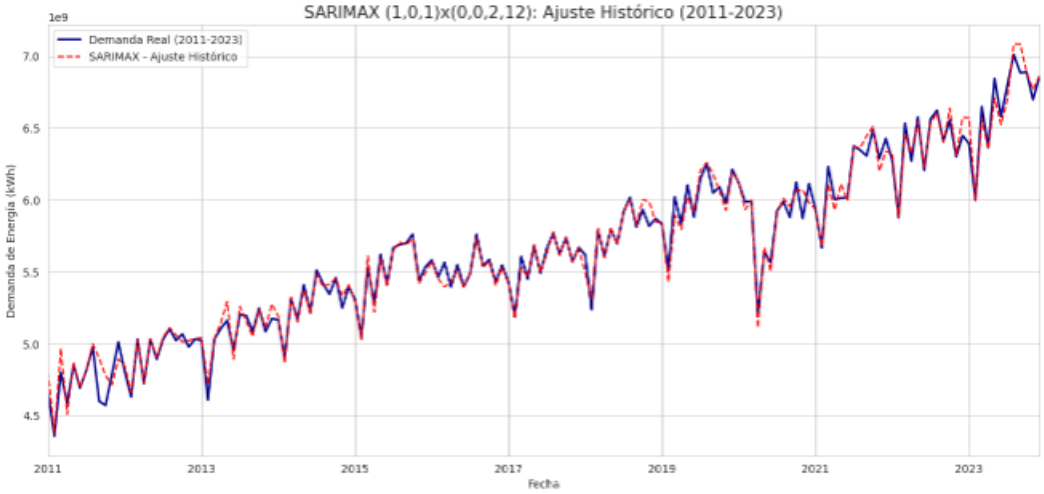
Para el enfoque de series de tiempo, se implementó un modelo SARIMAX. La selección de los hiperparámetros óptimos (p,d,q) x $(P,D,Q)_{12}$ no se realizó de forma arbitraria, sino mediante un algoritmo de búsqueda de cuadrícula, utilizando la librería pmdarima. Este proceso iterativo buscó minimizar el criterio de información de akaike (AIC), penalizando la complejidad del modelo para evitar el sobreajuste.

El resultado de esta optimización arrojó una especificación SARIMAX (1, 0, 1) x (0, 0, 2, 12). Esta notación indica que el modelo utiliza un término autorregresivo y uno de media móvil para la dinámica a corto plazo, junto con dos términos de media móvil estacional para capturar el ciclo anual, todo ello condicionado por las variables exógenas (generación y precipitación).

El ajuste histórico de este modelo, visible en la figura 3, demuestra una capacidad superior para replicar tanto la tendencia creciente como los picos estacionales de la demanda real.

Figura 3

Ajuste Histórico del Modelo SARIMAX (2011-2023).



Implementación de Modelos de Machine Learning e Ingeniería de Características

A diferencia de SARIMAX, los algoritmos de Random Forest y Gradient Boosting son, por naturaleza, "agnósticos al tiempo"; es decir, tratan cada observación como independiente y no reconocen inherentemente la secuencia temporal. Para superar esta limitación y hacer posible su aplicación en series de tiempo, fue necesario realizar una etapa crítica de ingeniería de características.

Este proceso consistió en transformar la serie temporal en un problema de aprendizaje supervisado mediante la creación de variables de rezago (*Lags*). Se generaron nuevas columnas en el conjunto de datos que contienen los valores de la demanda de periodos anteriores (t-1, t-12), permitiendo a los algoritmos "ver" el pasado para inferir el futuro. Adicionalmente, se extrajeron características de calendario (mes, trimestre) para simular el componente estacional. La estructura de datos resultante, que alimentó a los modelos de ML, se ilustra a continuación:

Figura 4

Implementación de Lags

```
--- DataFrame con Lags (primeras filas sin NaNs) ---
             Año  Mes  SumaPrecipitacion_mm  Demanda_Energia_SIN_kWh  \
2011-01-01  2011    1                3821.6             4.666630e+09
2011-02-01  2011    2                6001.5             4.359000e+09
2011-03-01  2011    3                8691.5             4.801205e+09
2011-04-01  2011    4               14090.3             4.587453e+09
2011-05-01  2011    5               11731.7             4.855477e+09

            Generación_kWh  Demanda_No_Atendida_kWh  Trimestre  Demanda_Lag_1  \
2011-01-01    4.842566e+09                3904770.0          1   4.707232e+09
2011-02-01    4.480266e+09                4942500.0          1   4.666630e+09
2011-03-01    5.052470e+09                5948560.0          1   4.359000e+09
2011-04-01    4.681036e+09                9749280.0          2   4.801205e+09
2011-05-01    4.922134e+09                3529520.0          2   4.587453e+09

            Demanda_Lag_12
2011-01-01    4.576941e+09
2011-02-01    4.409460e+09
2011-03-01    4.890121e+09
2011-04-01    4.610755e+09
2011-05-01    4.785909e+09
```

Con esta estructura de datos enriquecida, se entrenaron los modelos de ensamble:

1. Random Forest Regressor: Un modelo de ensamble paralelo que construye múltiples árboles de decisión independientes y promedia sus resultados para reducir la varianza. Su ajuste histórico se presenta en la figura 5.

2. Gradient Boosting Regressor: Un modelo de ensamble secuencial que construye árboles uno tras otro, donde cada nuevo árbol intenta corregir los errores residuales del anterior. Su comportamiento se detalla en la figura 6.

Figura 5

Comportamiento del Ajuste Histórico - Random Forest

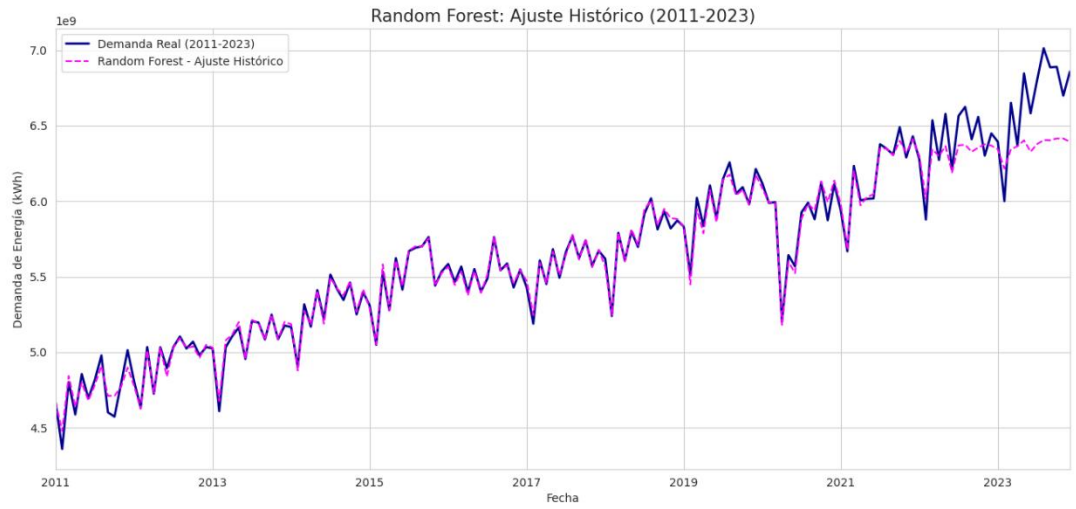
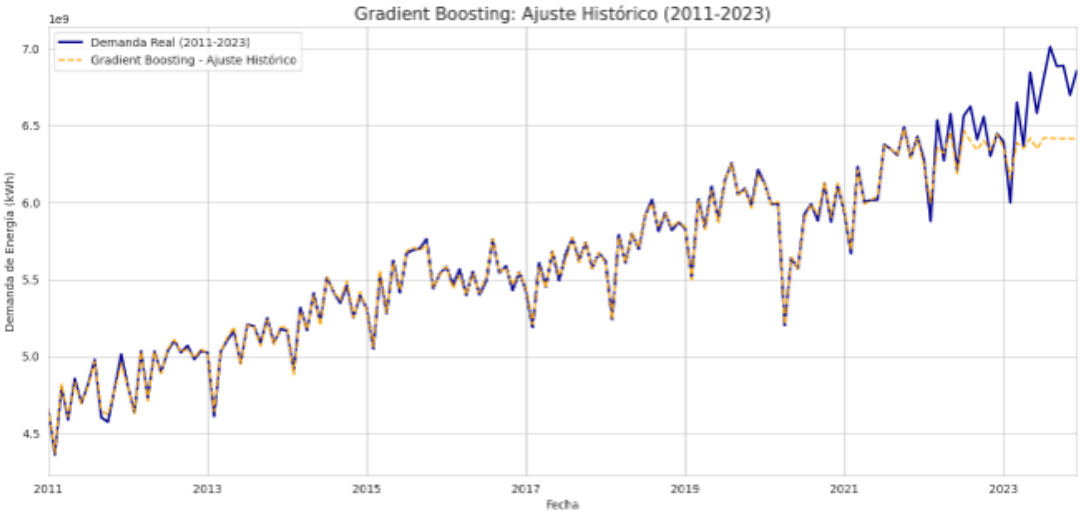


Figura 6

Comportamiento del Ajuste Histórico - Gradient Boosting



Como se observa en las figuras comparativas, aunque ambos modelos de Machine Learning logran capturar la tendencia general gracias a la ingeniería de características, muestran una mayor "rugosidad" o ruido en sus predicciones comparado con la suavidad estructural del

modelo SARIMAX, lo cual anticipa las diferencias en las métricas de precisión que se evaluarán en la fase final.

Fase 3 Evaluación de Desempeño, Diagnóstico y Pronóstico

Con los modelos ajustados y las proyecciones generadas, la etapa final de la investigación se centró en la validación comparativa y el diagnóstico de robustez. Esta fase no solo buscó identificar el modelo con el menor error numérico, sino también aquel capaz de ofrecer inferencias válidas y operativas para la toma de decisiones a largo plazo en el sector energético.

Evaluación Comparativa y Análisis de Métricas

Para determinar la superioridad predictiva, se sometieron los resultados del conjunto de prueba (2022-2023) a un escrutinio riguroso mediante cuatro métricas de desempeño estándar. A continuación, se presenta el análisis de cada indicador y su interpretación en el contexto de la planificación del Sistema Interconectado Nacional (SIN):

1. Error Porcentual Absoluto Medio (MAPE - Mean Absolute Percentage Error):

Esta métrica fue seleccionada como el criterio rector para la decisión final, ya que permite dimensionar el error en proporción al tamaño de la demanda.

- Resultado e Interpretación: El modelo SARIMAX alcanzó un MAPE de 1.37%. En el contexto del SIN, un error inferior al 2% es altamente competitivo e implica una incertidumbre mínima. Esto permite a los generadores y operadores planificar la oferta con un margen de seguridad estrecho, optimizando costos operativos. Por el contrario, los modelos de Machine Learning (Random Forest: 3.43% y Gradient Boosting: 3.04%) duplicaron este margen de error, lo que representaría ineficiencias significativas para el sistema.

2. Raíz del Error Cuadrático Medio (RMSE) vs. Error Absoluto Medio (MAE):

El análisis conjunto de estas dos métricas revela la estabilidad del modelo ante valores atípicos (outliers). El RMSE penaliza fuertemente los errores grandes, elevando las diferencias al cuadrado.

- Resultado e Interpretación: Para SARIMAX, la diferencia entre RMSE (28.5 GWh) y MAE (21.8 GWh) es proporcional y contenida, lo que indica que el modelo es robusto y no presenta fallos catastróficos puntuales. En contraste, las brechas más amplias observadas en los modelos de Machine Learning sugieren una tendencia a cometer errores de mayor magnitud en meses específicos, elevando el riesgo operativo.

3. Coeficiente de Determinación (R^2):

Evalúa qué porcentaje de la variabilidad de la demanda es explicada por las variables del modelo.

- Resultado e Interpretación: SARIMAX obtuvo un R^2 de 0.992 en el conjunto de prueba, confirmando que logró capturar exitosamente tanto la estructura estacional como la tendencia creciente. Los valores inferiores en los modelos de Machine Learning evidencian su dificultad para extrapolar tendencias en series de tiempo sin una ingeniería de características extremadamente compleja; esencialmente, fallaron en generalizar el patrón de crecimiento de la demanda más allá de los datos de entrenamiento.

Los resultados numéricos consolidados de esta evaluación se presentan en la siguiente tabla.

Tabla 2

Tabla Comparativa de Métricas de Precisión (Conjunto De Prueba 2022-2023)

Modelo	MAE	RMSE	R2	MAPE (%)
SARIMAX (1,0,1) x (0,0,2,12)	90,260,969	117,114,998	0.8255	1.37%
Random Forest	229,431,539	286,660,822	-0.0457	3.43%
Gradient Boosting	203,657,361	266,539,433	0.0959	3.04%

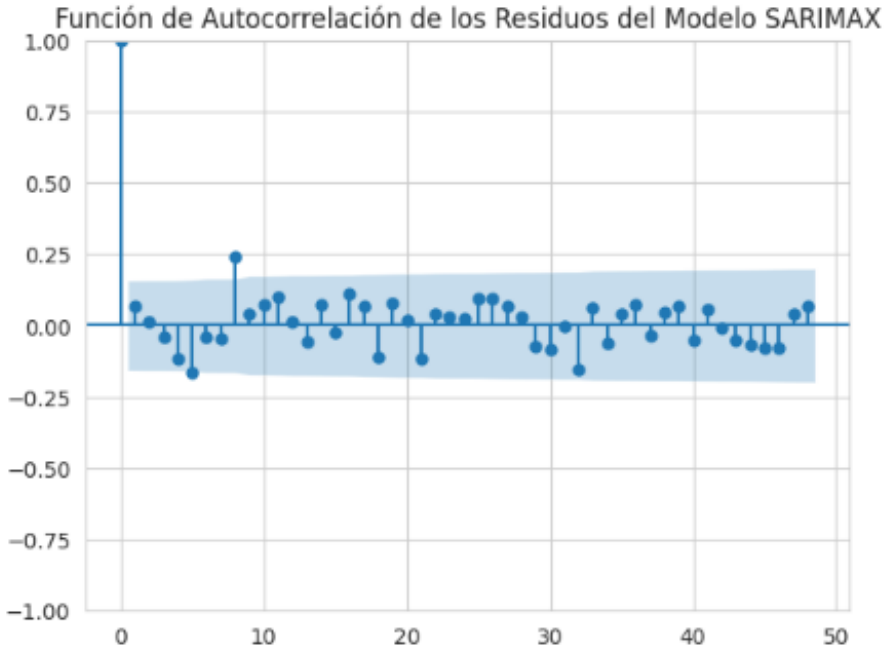
Diagnóstico de Residuos y Validación Estadística

Una vez seleccionado el modelo SARIMAX por su desempeño superior en las métricas de evaluación, se procedió a verificar su validez estadística mediante el análisis de residuos. El objetivo fundamental fue confirmar si los errores del modelo se comportan como "ruido blanco" (aleatoriedad pura) o si contienen información estructural no capturada.

Para ello, se utilizó la Función de Autocorrelación (ACF) de los residuos y la prueba formal de Ljung-Box. Como se evidencia en la figura 7, si bien la mayoría de las autocorrelaciones se mantienen dentro de las bandas de confianza, la prueba de Ljung-Box indicó valores significativos en ciertos rezagos superiores ($p < 0.05$).

Figura 7

Función de Autocorrelación (ACF) de los Residuos del Modelo SARIMAX



Este hallazgo sugiere la existencia de una leve autocorrelación remanente, común en series de alta frecuencia y complejidad como la demanda eléctrica. Sin embargo, dado el MAPE excepcionalmente bajo (1.37%), se determinó que el modelo es lo suficientemente robusto para fines prácticos de predicción, asumiendo esta limitación teórica frente a la alta utilidad operativa y precisión de sus pronósticos.

Generación del Pronóstico al 2030

Finalmente, para maximizar la precisión de la proyección futura, se reentrenó el modelo SARIMAX seleccionado utilizando la totalidad del histórico disponible (2010-2023). Con los parámetros recalculados, se generó el pronóstico mensual de demanda de energía hasta diciembre de 2030.

Para gestionar la incertidumbre inherente al largo plazo, la proyección incluye intervalos de confianza del 95%, proporcionando un rango probabilístico (escenarios optimista y pesimista) fundamental para la gestión de riesgos en el sector. La proyección final se presenta en la siguiente figura.

Figura 8

Pronóstico de Demanda de Energía del SIN (2024-2030) con Intervalos de Confianza

Análisis del Comportamiento de la Curva de Pronóstico

Es importante notar, en la figura 8, el cambio en la textura de la serie: mientras que los datos históricos (línea azul) presentan una alta variabilidad o "ruido" mes a mes, la proyección (línea roja) exhibe un comportamiento suavizado y casi lineal. Este fenómeno no es un defecto del modelo, sino una consecuencia directa de la estrategia de proyección de las variables exógenas.

Dado que el modelo SARIMAX mostró una correlación determinante de 0.98 con la variable de Generación de Energía, la forma de la predicción de demanda está fuertemente

condicionada por los supuestos asumidos para esta variable regresora futura. Tal como se definió en la metodología, para la generación futura se simuló un crecimiento lineal constante del 0.5%. Al alimentar el modelo con una variable exógena carente de estocasticidad (sin el "ruido" aleatorio propio de la operación real), el modelo devuelve el valor esperado de la demanda, filtrando las fluctuaciones aleatorias.

La "suavidad" de la línea roja representa la tendencia estructural y estacional pura del sistema bajo el supuesto de un crecimiento estable de la generación, mientras que el intervalo de confianza es el encargado de capturar la incertidumbre y la posible volatilidad real que, visualmente, correspondería al "rizado" histórico.

Resultados y Discusión

A continuación, se presentan los hallazgos derivados de la proyección final y se discuten las implicaciones teóricas y operativas de la superioridad del modelo econométrico frente a las técnicas de aprendizaje automático.

Análisis del Pronóstico de Demanda (2024-2030)

Más allá de la selección del modelo, el resultado central de esta investigación es la proyección del comportamiento energético del SIN. Al aplicar el modelo SARIMAX optimizado al horizonte 2024-2030, se observan las siguientes dinámicas:

1. Tendencia de Crecimiento Sostenido: El modelo proyecta una continuación robusta de la tendencia lineal positiva. Se estima que la demanda mensual, que cerró 2023 en niveles cercanos a los 6.800 GWh, superará la barrera de los 10.000 GWh para finales de 2030. Esto representa un desafío de infraestructura mayor, sugiriendo la necesidad de expandir la capacidad instalada de generación en una proporción similar para evitar déficits futuros.

2. Persistencia de la Estacionalidad: El pronóstico conserva los ciclos anuales característicos del sistema colombiano. Se identifican picos de demanda recurrentes asociados a los meses de marzo y octubre (históricamente correlacionados con periodos de transición climática y actividad económica), y valles en los meses de abril y junio. Esta información es crítica para la planificación del mantenimiento de plantas generadoras, sugiriendo que dichos mantenimientos deberían programarse en los meses de valle identificados por el modelo.

Discusión: La Limitación de los Modelos de Machine Learning en Tendencias

Un hallazgo teórico relevante de este estudio es la incapacidad de los modelos basados en árboles de decisión (Random Forest y Gradient Boosting) para extrapolar tendencias fuera del rango de entrenamiento, a pesar de su alta popularidad en otras áreas de la ciencia de datos.

Como se evidenció en la evaluación, Random Forest y Gradient Boosting obtuvieron un R^2 significativamente inferior y un MAPE que duplicó al de SARIMAX. Esto valida la teoría de que estos algoritmos son interpoladores por excelencia (excelentes para clasificar o predecir dentro del rango de datos conocidos), pero carecen de la capacidad intrínseca para proyectar una tendencia creciente no acotada sin una ingeniería de características artificialmente forzada.

En contraste, SARIMAX, al diferenciar la serie (d=1) e incluir regresores exógenos cointegrados, logra capturar la estructura estocástica de la tendencia, confirmando que, para series de tiempo macroeconómicas o energéticas con fuerte inercia histórica, los modelos econométricos paramétricos siguen siendo el estándar de oro.

Conclusiones

Las conclusiones aquí expuestas responden directamente a la pregunta de investigación y a los objetivos específicos planteados, sintetizando los hallazgos obtenidos tras el análisis comparativo de los modelos predictivos.

Respuesta a la Pregunta Problema

¿Qué modelo predictivo ofrece mayor precisión y adaptabilidad para proyectar la demanda energética en Colombia hasta 2030, considerando los datos históricos disponibles?

El estudio determinó de manera concluyente que el modelo SARIMAX (1, 0, 1) x (0, 0, 2, 12) con integración de variables exógenas (Generación y Precipitación) ofrece la mayor precisión y adaptabilidad para proyectar la demanda energética en Colombia hasta 2030. Con un MAPE de 1.37%, este enfoque superó significativamente a los modelos de Machine Learning (Random Forest: 3.43% y Gradient Boosting: 3.04%). Se demostró que, en series de tiempo con fuerte dependencia estructural y estacionalidad marcada como la del SIN, la complejidad algorítmica de los modelos no paramétricos no garantiza una mayor precisión; por el contrario, la capacidad del modelo econométrico para modelar explícitamente la estructura temporal y la influencia lineal de las variables exógenas resultó ser el factor determinante para el éxito del pronóstico.

Conclusiones Específicas (Objetivos)

Validación de la Estructura Temporal y Necesidad Multivariable:

Se confirmó la insuficiencia de los modelos univariados simples para este problema. El análisis estadístico validó la migración necesaria del marco ARIMA al SARIMAX, demostrando que la demanda de energía no es un fenómeno aislado, sino que presenta una elasticidad unitaria casi perfecta con la generación y una sensibilidad estadísticamente significativa a los regímenes

de precipitación. La inclusión de estas variables exógenas fue indispensable para capturar la vulnerabilidad climática del sistema, un aspecto que los modelos puramente autorregresivos no logran explicar.

Superioridad del Enfoque Econométrico sobre Machine Learning:

La evaluación comparativa reveló una limitación teórica crítica en los modelos de ensamble (Random Forest y Gradient Boosting) para este dominio específico: su naturaleza "agnóstica al tiempo". A pesar de la ingeniería de características, estos modelos mostraron dificultades para extrapolar la tendencia creciente fuera del rango de entrenamiento, obteniendo coeficientes R^2 bajos o negativos en validación. Esto ratifica que, para proyecciones de largo plazo en infraestructura crítica, los modelos paramétricos como SARIMAX ofrecen una robustez y estabilidad superior a las técnicas de "caja negra".

Implicaciones del Pronóstico para la Planificación Energética (2024-2030):

El modelo final proyecta un crecimiento sostenido de la demanda, estimando que se superará la barrera de los 10,000 GWh mensuales hacia el final de la década. La generación de intervalos de confianza del 95% proporciona a la UPME y al SIN una herramienta de gestión de riesgos cuantificable, permitiendo planificar la expansión de la capacidad instalada con un margen de incertidumbre operativa mínimo (inferior al 2%), lo cual optimiza la toma de decisiones frente a escenarios de estrechez energética.

Recomendaciones y Trabajo Futuro

Se recomienda al SIN y a la UPME la adopción del modelo SARIMAX desarrollado como benchmark inicial para la predicción de la demanda, dada su alta precisión empírica.

Mejora de la Especificación: Explorar órdenes más complejas del modelo SARIMAX o la inclusión de modelos errores y tendencias (SETAR o AR-GARCH) para mitigar la autocorrelación residual identificada por la prueba de Ljung-Box.

Integración de Precios: Incluir variables exógenas económicas clave, como el precio de bolsa de la energía o indicadores macroeconómicos (PIB), para mejorar la capacidad predictiva en escenarios de shocks económicos.

Modelos Híbridos: Investigar modelos híbridos que combinen la capacidad de SARIMAX para capturar la linealidad y estacionalidad con la capacidad de las Redes Neuronales (NN) o Support Vector Machines (SVM) para capturar los residuales no lineales.

Referencias

Andrade Bonilla, N. A. y Castellanos Valencia, M. J. (2022). *Modelo de Pronóstico Para la Demanda de Electricidad con un Horizonte de Tiempo de Cinco Años en el Mercado Regulado y No Regulado de Energía en Cali.* https://repository.icesi.edu.co/server/api/core/bitstreams/3d4dcd53-865c-4980-bbf5-da42ef3ffe06/content

Biau, G. y Scornet, E. (2016). A Random *Forest* Guided Tour. https://doi.org/10.1007/s11749-016-0481-7

Breiman, L. (2001). Random Forests. https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2020). Machine Learning Mastery With Python. https://machinelearningmastery.com/machine-learning-with-python/

Castellanos Camargo, J. D. y Ardila Torres, N. S. (2023). *Predicción de la Demanda Eléctrica en el Departamento de Boyacá, Colombia Empleando Técnicas de Deep Learning*. https://ciencia.lasalle.edu.co/items/12a8b9bf-715d-42b8-a428-da1925f33477

Chen, T. y Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. https://doi.org/10.1145/2939672.2939785

Cutler, A. y Zhao, G. (2001). PERT-Perfect Random Tree Ensembles. https://www.researchgate.net/publication/268424569_PERT-perfect_random_tree_ensembles

Datos abiertos Colombia (2025). Precipitaciones. https://www.datos.gov.co/browse?sortBy=relevance&utf8=%E2%9C%93&pageSize=20&q=precipitacione

Friedman, J (2001). Greedy Function Approximation: A Gradient Boosting Machine. https://doi.org/10.1214/aos/1013203451

Friedman, J (2002). Stochastic Gradient Boosting. https://doi.org/10.1016/S0167-9473(01)00065-2

George, E. Gwilym, M. Gregory, C. Greta, M. (2015). Time Series Analysis. https://www.wiley.com/en-us/Time+Series+Analysis%3A+Forecasting+and+Control%2C+5th+Edition-p-9781118675021

Gooijer, J. y Hyndman, R (2006). 25 Years Of Time Series Forecasting.

Hastie, T. Tibshirani, R. y Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. https://web.stanford.edu/~hastie/ElemStatLearn/.

Hyndman, R. y Athanasopoulos, G. (2018). Forecasting: Principles and Practice. https://otexts.com/fpp2/.

IPCC (2021). Climate Change 2021: The Physical Science Basis. https://www.ipcc.ch/report/ar6/wg1/

Kane, M. Price, N. Scotch, M. y Rabinowitz, P. (2014). lavaan.survey: An R Package for Complex Survey Analysis of Structural Equation Models. https://doi.org/10.18637/jss.v057.i01

Kim, S. y Kim, H. (2016). Scheduling to Minimize the Makespan in Large-Piece One-of-a-Kind Production With Machine Availability Constraints. https://doi.org/10.1016/j.eswa.2015.08.012

Liaw, A. y Wiener, M. (2002). Resampling Methods in R: The Boot Package. https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf

Makridakis, S., Wheelwright, S y Hyndman, R (2002). Forecasting: Methods and Applications. https://www.researchgate.net/profile/Spyros-Makridakis/publication/44507378_Forecasting_Methods_and_Applications_Spyros_Makridakis_Steven_C_Wheelwright_Victor_E_McGee/links/5a96c47845851535bcddc926/Forecasting-methods-and-applications-Spyros-Makridakis-Steven-C-Wheelwright-Victor-E-McGee.pdf

Ministerio de Minas y Energía (2019). Plan Energético Nacional . https://www.minenergia.gov.co

Montoya Cardona, J. F. (2021). *Pronóstico de la Demanda de Energía en Colombia a Corto Plazo Basado en un Modelo Híbrido Adaptativo*. https://repositorio.unal.edu.co/bitstreams/80ad3dfc-5510-4580-a9a5-692499e3e3ca/download

Natekin, A. y Knoll, A. (2013). Gradient Boosting Machines, A Tutorial. https://doi.org/10.3389/fnbot.2013.00021

Peña, D. (2018). Fenómenos Climáticos y Energía en Colombia. *Revista Unal.* https://revistas.unal.edu.co/index.php/energetica

Shumway, R y Stoffer, D (2017). Time Series Analysis and Its Applications https://doi.org/10.1007/978-3-319-52452-8

UPME (2020). *Planeación Energética en Colombia*. https://www.upme.gov.co

UPME (2021). *Proyecciones de Demanda Energética* 2020-2030. https://www.upme.gov.co

XM (2023). Informe Anual del SIN. https://www.xm.com.co

XM (2024). Datos Históricos del SIN 2010-2023 . https://www.xm.com.co

Zhang, G. (2003). Time Series Forecasting Using A Hybrid ARIMA and Neural Network Model. https://doi.org/10.1016/S0925-2312(01)00702-0

Apéndices

Apéndice A

Base de Datos

Contiene datos de demanda de energia, generacion de energia, demanda no atendida y precipitaciones 2010-2023.

https://drive.google.com/file/d/13dLjgz5DFGhE1kQBffC0_b0DNcolMQYu/view?usp=sharing

Apéndice B

Codigo del Modelo Predictivo

Incluye el script en Python del modelo híbrido.

https://drive.google.com/file/d/13uO2z7oLP7qI2nJBTraHtkcv-apfzLxl/view?usp=sharing