



Pontificia Universidad
JAVERIANA
Cali

**PREDICCIÓN A CORTO PLAZO DE LA DEMANDA ELÉCTRICA EN UNA SUBESTACIÓN USANDO
MODELOS DE SERIES DE TIEMPO**

Fredy Alejandro Ordoñez Márquez

Código 8982484

Oscar Fernando Velásquez Chala

Código 9026734

José Luis Sánchez Ceballos

Código 9026303

*Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos*

Director
David Arango Londoño

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JUNIO 16 DE 2026

TABLA DE CONTENIDO

	Pág.
INTRODUCCIÓN.....	7
1. CONTEXTUALIZACIÓN DEL PROYECTO	10
1.1. DEFINICIÓN DEL PROBLEMA.....	10
1.1.1 PLANTEAMIENTO DEL PROBLEMA	11
1.1.2 FORMULACIÓN DEL PROBLEMA.....	13
1.1.3 SISTEMATIZACIÓN	13
1.2. OBJETIVOS.....	13
1.2.1. OBJETIVO GENERAL	13
1.2.2. OBJETIVOS ESPECÍFICOS	13
1.3. MARCO DE REFERENCIA.....	14
1.3.1. MARCO TEÓRICO.....	14
1.3.2. MARCO REGULATORIO	21
1.3.3. MARCO OPERATIVO	23
1.3.4. ANTECEDENTES	25
2. CARACTERIZACIÓN ESTADÍSTICA DE LA SERIE DE DEMANDA ELÉCTRICA	27
2.1. FUENTE DE DATOS Y ESTRUCTURA DEL DATASET.....	27
2.2. PREPROCESAMIENTO DE LOS DATOS	30
2.2.1. IDENTIFICACIÓN Y CARACTERIZACIÓN DE DATOS FALTANTES.....	30
2.2.2. EVALUACIÓN DE MÉTODOS DE IMPUTACIÓN.....	31
2.2.3. DETECCIÓN Y TRATAMIENTO DE OUTLIERS	33
2.3. ANÁLISIS EXPLORATORIO DE DATOS	35
2.3.1. ESTADÍSTICAS DESCRIPTIVAS Y EVOLUCIÓN TEMPORAL	35
2.3.2. DESCOMPOSICIÓN DE LA SERIE, STL VS. DESCOMPOSICIÓN CLÁSICA.....	36
2.3.3. ANÁLISIS DE ESTACIONALIDADES MÚLTIPLES.....	37
2.3.4. FUNCIÓN DE AUTOCORRELACIÓN Y ANÁLISIS DE MEMORIA TEMPORAL	40
2.3.5. PRUEBAS DE ESTACIONARIEDAD ADF-KPSS.....	42
2.3.6. ANÁLISIS DE CORRELACIÓN CON VARIABLES CLIMÁTICAS	43
3. AJUSTE Y VALIDACIÓN DE MODELOS.....	47

3.1. FAMILIA 1: MODELO ESTADÍSTICO: SARIMAX CON AUTOARIMA	48
3.2. FAMILIA 2: MACHINE LEARNING: XGBOOST Y PROPHET	50
3.2.1. XGBOOST CON INGENIERÍA DE CARACTERÍSTICAS TEMPORAL.....	50
3.2.2. PROPHET CON CONFIGURACIÓN EXTENDIDA PARA EL CONTEXTO COLOMBIANO	52
3.3. FAMILIA 3, DEEP LEARNING ESPECIALIZADO.....	54
3.3.1. N-HITS: INTERPOLACIÓN JERÁRQUICA NEURONAL.....	54
3.3.2. N-BEATS EN TRES VARIANTES	56
3.3.3. TEMPORAL FUSION TRANSFORMER (TFT), VERSIONES CPU Y GPU.....	56
3.4. FAMILIA 4, MODELO FUNDACIONAL PRE-ENTRENADO: CHRONOS.....	58
3.5. FAMILIA 5, MODELOS HÍBRIDOS DE RESIDUAL BOOSTING	59
3.6. ENSEMBLE PONDERADO 1/MAE	60
4. EVALUACIÓN DEL DESEMPEÑO Y COMPARACIÓN DE MODELOS	62
4.1. RANKING GLOBAL DE DESEMPEÑO (PERÍODO DE TEST 2025).....	62
4.2. CLASIFICACIÓN POR CUADRANTES: EFICIENCIA COMPUTACIONAL VS. PRECISIÓN	66
4.3. ANÁLISIS DE SESGO SISTEMÁTICO POST-CAMBIO ESTRUCTURAL 2023.....	68
4.4. ANÁLISIS TEMPORAL DEL ERROR: POR HORA DEL DÍA Y POR MES	69
4.5. ANÁLISIS DE CONSISTENCIA TSCV VS. TEST (DETECCIÓN DE SOBREAJUSTE).....	71
5. CONCLUSIONES Y TRABAJOS FUTUROS	74
5.1. CONCLUSIONES.....	74
5.2. LIMITACIONES DEL ESTUDIO	77
5.3. TRABAJOS FUTUROS	78
6. REFERENCIAS BIBLIOGRÁFICAS.....	79

LISTA DE FIGURAS

	Pág.
Figura 1. Serie temporal completa de potencia aparente (VA) y variables climáticas principales, período 2019-2025	29
Figura 2. Ubicación temporal de los datos faltantes en la serie de potencia aparente	31
Figura 3. Comparación visual de métodos de imputación seleccionados sobre el intervalo de 335 horas (septiembre 2019)	33
Figura 4. Detección de outliers con STL + Isolation Forest y resultado de la imputación sobre la serie completa	34
Figura 5. Serie de potencia con medias móviles y detección del cambio estructural de enero 2023.	36
Figura 6. Descomposición STL de la serie de potencia (period=24, seasonal=25, robust=False), cuatro componentes.....	37
Figura 7. Perfil de estacionalidad horaria del componente estacional STL promediado por mes, ciclo 2019-2025.....	38
Figura 8. Gráficos estacionales de la serie de potencia a diferentes escalas temporales.	39
Figura 9. Función de autocorrelación (ACF) de la serie de potencia, lag=1 a lag=168 (una semana).	41
Figura 10. Correlación de Spearman entre temperatura y potencia por hora del día, relación dinámica y no lineal.	45
Figura 11. Importancia de variables para la predicción de potencia_va, Mutual Information	46
Figura 12. Pronóstico SARIMAX sobre las últimas 4 semanas del período de test 2025 vs. valores reales.....	49
Figura 13. Importancia SHAP de las características del modelo XGBoost, contribución marginal al pronóstico.....	51
Figura 14. Pronóstico XGBoost sobre las últimas 2 semanas del período de test 2025 vs. valores reales.....	52
Figura 15. Descomposición de componentes del modelo Prophet Full, changepoints, tendencia, estacionalidades y efecto de temperatura.....	53
Figura 16. Pronóstico N-HITS sobre el período de test 2025, últimas 4 semanas	55
Figura 17. Pronóstico N-BEATS Interpretable / Ensemble sobre el período de test 2025, primeras 2 semanas	56
Figura 18. Comparación de pronósticos TFT-CPU vs. TFT-GPU sobre una semana del test 2025	58
Figura 19. Chronos sobre el período de test 2025 – primeras 2 semanas del test.....	59
Figura 20. Protocolo de Evaluación, Partición de Datos y Esquema TSCV.....	62
Figura 21. Gráfico de barras horizontales del ranking de MAE y MASE de todos los modelos evaluados.....	64
Figura 22. Pruebas de significancia estadística, ranking promedio de Friedman y distancia crítica post-hoc de Nemenyi.....	66
Figura 23. Diagrama de dispersión de cuadrantes eficiencia-precisión (tiempo vs. MASE)	67

Figura 24. Sesgo sistemático de los modelos en el período de test 2025, diagrama de barras con dirección	69
Figura 25. MAE promedio por hora del día (0-23h) para todos los modelos evaluados, período de test 2025	70
Figura 26. Heatmap de MAE mensual por modelo, enero a agosto 2025	70
Figura 27. Inversión de rankings entre TSCV y test 2025, gráfico de puntos conectados	73

LISTA DE TABLAS

	Pág.
Tabla 1. Estructura del dataset integrado - variables, tipos y estadísticos básicos	27
Tabla 2. Intervalos de datos faltantes en la serie de potencia, caracterización y criticidad	30
Tabla 3. Evaluación comparativa de métodos de imputación para la serie de potencia	31
Tabla 4. Comparación de métodos de detección de outliers, número de anomalías detectadas	34
Tabla 5. Estadísticos descriptivos anuales de la serie de potencia limpia	35
Tabla 6. Comparación de métodos de descomposición de la serie de potencia	36
Tabla 7. Resumen del análisis de estacionalidades múltiples de la serie de potencia	37
Tabla 8. Autocorrelaciones de la serie de potencia para rezagos clave	41
Tabla 9. Resultados del diagnóstico de estacionariedad ADF-KPSS para la serie de potencia y sus transformaciones	42
Tabla 10. Correlación de Pearson global entre potencia_va y variables climáticas	43
Tabla 11. Protocolo de evaluación estandarizado aplicado a todos los modelos	47
Tabla 12. Configuración del modelo SARIMAX con AutoARIMA	48
Tabla 13. Ingeniería de características para XGBoost, descripción y justificación	50
Tabla 14. Comparación de configuraciones de Prophet (Base vs. Full)	52
Tabla 15. Configuración de N-HITS, variantes base y completa	54
Tabla 16. Configuración de TFT, variantes CPU (arquitectura mínima) y GPU (arquitectura completa)	57
Tabla 17. Descripción de los modelos híbridos de residual boosting evaluados	59
Tabla 18. Pesos del ensemble ponderado 1/MAE, distribución de contribución entre modelos	61
Tabla 19. Ranking completo de desempeño en el período de test 2025, ordenado por MASE ascendente (menor MASE = mejor)	63
Tabla 20. Sesgo sistemático de los modelos en el período de test 2025, capacidad de captura del régimen post-2023	68
Tabla 21. Comparación MASE TSCV expanding vs. MASE test 2025, detección de sobreajuste ...	71

INTRODUCCIÓN

En el sistema eléctrico colombiano, la gestión eficiente de la demanda constituye uno de los principales retos para los Operadores de Red (OR), especialmente en un contexto de creciente variabilidad climática, aumento sostenido del consumo energético, incorporación de recursos energéticos distribuidos (DER) y mayores exigencias regulatorias en materia de continuidad, calidad y eficiencia del servicio. Las subestaciones del sistema de distribución funcionan como nodos críticos en los que convergen cargas residenciales, comerciales e industriales, y donde los sistemas SCADA (Supervisory Control and Data Acquisition) registran en tiempo real mediciones de potencia activa, potencia reactiva, tensiones y corrientes con resolución horaria o inferior.

A pesar de que estos datos representan una fuente de información de alto valor para la toma de decisiones operativas, en la práctica la mayoría de los OR colombianos no aprovechan su potencial para generar pronósticos automáticos de corto plazo. La planificación de la operación se fundamenta habitualmente en valores históricos recientes, experiencia operativa o estimaciones manuales que no capturan adecuadamente la dinámica temporal de la demanda. Esta carencia de herramientas predictivas especializadas limita la capacidad de anticipación ante contingencias, dificulta la asignación óptima de recursos de generación y transformación, aumenta el riesgo de sobrecargas en transformadores y alimentadores, y compromete el cumplimiento de los indicadores de confiabilidad SAIDI (System Average Interruption Duration Index) y SAIFI (System Average Interruption Frequency Index) exigidos por la CREG.

En este contexto, la predicción de la demanda eléctrica de corto plazo (Short-Term Load Forecasting, STLF) a nivel de subestación emerge como una necesidad operativa crítica que no está cubierta por los modelos de escala nacional que opera XM S.A. E.S.P. ni por las proyecciones de largo plazo de la UPME. Cerrar esta brecha requiere el desarrollo de modelos especializados, entrenados sobre los datos históricos específicos de cada subestación, que sean capaces de capturar las estacionalidades múltiples (diaria, semanal y anual), los cambios estructurales de nivel de demanda y la relación no lineal con variables climáticas características de series eléctricas de alta granularidad.

Con el propósito de responder a esta necesidad, en este proyecto se desarrolló y evaluó comparativamente un sistema de predicción de corto plazo de la demanda eléctrica en una subestación eléctrica del sistema de distribución colombiano, utilizando datos históricos horarios del período enero 2019 – agosto 2025 (58,448 observaciones). El portafolio de modelos abarcó diecisiete configuraciones de cinco familias: estadísticos clásicos (SARIMAX), machine learning (XGBoost, Prophet con configuración extendida), deep learning especializado en forecasting (N-HiTS, N-BEATS en tres variantes, TFT en versiones CPU y GPU), el modelo fundacional pre-entrenado Chronos, tres estrategias de modelos híbridos basadas en residual boosting. Todos los modelos se evaluaron bajo el mismo protocolo estandarizado: horizonte $h=24$ horas, validación cruzada temporal con esquema expanding window de 90 ventanas, y MASE como métrica

principal de ranking.

El período de prueba (2025-01-01 a 2025-08-31, 5,832 observaciones) corresponde íntegramente al régimen post-cambio estructural detectado en enero 2023 (+11.6% en la demanda media), lo que constituyó un escenario de evaluación exigente que expuso las limitaciones de los modelos estadísticos y de machine learning tabular para adaptarse a nuevos regímenes de demanda sin intervención manual. El mejor desempeño individual lo obtuvo Chronos (ctx post-2023) con $MASE=0.8645$, una mejora del 13.55% sobre el Seasonal Naïve, y solo 7 de las 17 configuraciones individuales evaluadas superaron este benchmark, evidenciando la dificultad del escenario post-cambio estructural.

Contribuciones del trabajo

Este trabajo realiza tres contribuciones específicas al campo del pronóstico de corto plazo de demanda eléctrica en subestaciones del sistema de distribución colombiano. En primer lugar, presenta una comparación sistemática y reproducible de diecisiete configuraciones de modelos pertenecientes a cinco familias. Todas las configuraciones se evalúan bajo el mismo protocolo. El conjunto de pronósticos test, configuraciones de hiperparámetros y métricas individuales y por bloque de evaluación se documenta de forma tal que permite replicar y extender el estudio.

En segundo lugar, este trabajo aporta evidencia empírica sobre el desempeño zero-shot de un modelo fundacional para series temporales (Chronos) aplicado a una subestación eléctrica. El resultado es que Chronos en modo zero-shot, con contexto restringido al período post-cambio estructural (post-enero 2023), obtiene el mejor desempeño individual del portafolio sin requerir entrenamiento específico ni ajuste de hiperparámetros sobre la subestación.

En tercer lugar, el trabajo documenta de forma transparente las limitaciones metodológicas detectadas durante el desarrollo del portafolio, sin ocultarlas tras métricas agregadas, lo cual constituye un aporte metodológico reproducible para futuros trabajos de pronóstico STLF que enfrenten el dilema entre reportar resultados optimizados y documentar el comportamiento real del portafolio bajo los protocolos efectivamente ejecutados.

El documento está organizado de la siguiente manera. El Capítulo 1 presenta la contextualización del proyecto: la definición, planteamiento y formulación del problema, los objetivos generales y específicos, y el marco de referencia que incluye el marco teórico con las bases conceptuales de cada familia de modelos y los antecedentes del campo. El Capítulo 2 desarrolla el primer objetivo específico, presentando la fuente de datos, el proceso de preprocesamiento (imputación y detección de outliers), el análisis exploratorio completo (estacionalidades, autocorrelación, pruebas de estacionariedad y análisis de correlación con variables climáticas). El Capítulo 3 presenta el segundo objetivo específico con el ajuste y configuración de cada familia de modelos. El Capítulo 4 aborda el tercer objetivo específico con la evaluación comparativa mediante el ranking de MASE, análisis de sesgo sistemático, curvas de degradación por horizonte y detección

de sobreajuste. El Capítulo 5 expone las conclusiones derivadas de los hallazgos y propone líneas de trabajo futuro.

1. CONTEXTUALIZACIÓN DEL PROYECTO

1.1. Definición del problema

El sistema eléctrico colombiano opera bajo un esquema normativo y técnico definido por tres entidades fundamentales: la Comisión de Regulación de Energía y Gas (CREG), que establece las reglas para la prestación del servicio de energía eléctrica; la Unidad de Planeación Minero-Energética (UPME), responsable de las proyecciones de demanda y la planeación de la expansión del sistema; y XM S.A. E.S.P., operador del Sistema Interconectado Nacional (SIN) y administrador del Mercado de Energía Mayorista (MEM). En este marco institucional, los Operadores de Red (OR) son los agentes responsables de garantizar la continuidad, confiabilidad y calidad del servicio en los niveles de distribución local, atendiendo directamente a los usuarios finales a través de circuitos y subestaciones distribuidas en el territorio nacional.

Las subestaciones de distribución son los nodos en los que se transforma la energía desde niveles de transmisión o subtransmisión hacia los niveles de distribución, y donde convergen cargas residenciales, comerciales e industriales con patrones de consumo heterogéneos. La demanda eléctrica en estas subestaciones presenta variaciones horarias, diarias, semanales y estacionales que determinan los requerimientos operativos del sistema. Una subestación que carece de herramientas predictivas opera de manera reactiva, respondiendo a los eventos operativos una vez que ocurren, en lugar de anticiparse a ellos con acciones preventivas.

Los sistemas SCADA instalados en las subestaciones de los OR registran con alta resolución temporal (generalmente cada hora o cada 15 minutos) variables como potencia activa, potencia reactiva, tensiones de línea, corrientes y estados de los equipos de maniobra. Esta información representa una fuente de datos de alto valor para el análisis de la dinámica de la demanda. Sin embargo, la mayoría de los OR no cuentan con modelos de predicción automatizados que procesen estos datos para generar pronósticos operativos de corto plazo. La planificación se fundamenta en valores históricos recientes o en estimaciones empíricas basadas en la experiencia operativa del personal técnico, lo cual no es suficiente para gestionar la creciente complejidad del sistema de distribución.

Esta complejidad se ha incrementado en los últimos años por cuatro factores convergentes. En primer lugar, la penetración creciente de recursos energéticos distribuidos (DER), incluyendo generación solar fotovoltaica residencial y comercial, que modifica los perfiles de flujo de potencia de manera no lineal y dependiente de las condiciones meteorológicas. En segundo lugar, la mayor electrificación residencial y comercial impulsada por el uso de equipos de climatización (aire acondicionado), cuyo consumo tiene una relación directa con la temperatura ambiental y puede elevar la demanda de manera significativa durante olas de calor. En tercer lugar, la ocurrencia de eventos climáticos extremos asociados a la variabilidad climática que aumentan la volatilidad de la demanda intradía. En cuarto lugar, la incorporación de nuevas cargas industriales y la variación

en los patrones de actividad económica local que modifican la estructura de la curva de demanda de manera no predecible con métodos empíricos.

La predicción de demanda eléctrica de corto plazo (STLF) a nivel de subestación es reconocida en la literatura internacional como una herramienta esencial para la operación eficiente de los sistemas de distribución [10]. Su implementación permite anticipar sobrecargas en transformadores y alimentadores, optimizar la programación de mantenimientos preventivos, mejorar la integración de DER sin comprometer la estabilidad de la red, y reducir las pérdidas técnicas mediante una gestión proactiva de los recursos. En el contexto colombiano, la CREG ha establecido mediante resoluciones como la CREG 015 de 2018 y la CREG 060 de 2019 que los OR deben implementar medidas que garanticen la eficiencia operativa y la reducción de pérdidas, objetivos directamente soportados por herramientas de pronóstico de demanda.

Sin embargo, los modelos de predicción que opera XM están orientados a escalas nacionales y regionales con horizontes diarios o semanales, y no están diseñados para capturar la variabilidad micro-local de subestaciones individuales. Análogamente, las proyecciones de la UPME tienen horizontes anuales o decenales y no sirven como insumo para la operación horaria. Esta brecha entre el modelamiento a escala nacional y las necesidades operativas del nivel de distribución constituye el problema central que motiva el presente proyecto.

La serie de potencia aparente analizada en este trabajo, registrada por el sistema SCADA de una subestación eléctrica entre enero 2019 y agosto 2025, exhibe todas las características que hacen complejo el problema de STLF: tendencia creciente sostenida, estacionalidades múltiples (diaria, mensual y anual), alta variabilidad residual asociada a eventos operativos, y un cambio estructural de nivel de +11.6% detectado en enero 2023 que altera fundamentalmente las condiciones del sistema. Este cambio de régimen resulta en que los datos anteriores a 2023 y los posteriores pertenecen a distribuciones estadísticas con medias diferentes, lo cual es el principal desafío metodológico del proyecto.

1.1.1 Planteamiento del problema

La predicción de la demanda eléctrica de corto plazo (Short-Term Load Forecasting, STLF) constituye un componente crítico para la operación segura y eficiente del sistema eléctrico. XM, en su rol de operador del SIN, emplea modelos avanzados para la predicción nacional y regional; sin embargo, dichos modelos no están diseñados para operar a nivel de subestación, donde la volatilidad es mayor y la dinámica depende de condiciones micro-locales.

En contraste, los Operadores de Red gestionan circuitos y subestaciones cuya demanda cambia por:

- Actividad comercial e industrial local.
- Eventos climáticos de corta duración.

- Picos residenciales matutinos, vespertinos o nocturnos.
- Cambios estructurales en cargas conectadas.
- Integración de paneles solares, microgeneración y baterías.

A nivel regulatorio, la CREG exige mediante resoluciones como la CREG 015 de 2018, CREG 024 de 2005, CREG 060 de 2019 y otras, que los OR implementen medidas que garanticen la eficiencia operativa, la continuidad del servicio y la reducción de pérdidas. Para cumplir estos requisitos, la predicción de demanda es un insumo esencial.

Sin embargo, los siguientes problemas persisten:

1. Dependencia de métodos empíricos basados en valores recientes o experiencia operativa.
2. Ausencia de herramientas automatizadas capaces de procesar la información SCADA en tiempo real.
3. Dificultad para capturar múltiples estacionalidades:
 - Ciclo horario (24 h)
 - Ciclo semanal (7 días)
 - Ciclo anual (clima, temperatura, hábitos económicos)
4. Impacto del clima, reconocido por la UPME en los estudios de proyección de demanda, que no se incorpora actualmente en subestaciones a nivel operativo.
5. Valores atípicos y ruido, que complican cualquier aproximación manual.

El resultado es un déficit en la precisión del pronóstico y una menor capacidad para:

- Prevenir sobrecargas.
- Optimizar la operación de equipos (transformadores, alimentadores).
- Implementar mantenimiento predictivo.
- Integrar recursos distribuidos sin afectar la estabilidad.
- Responder a la variabilidad climática y estacional.

Por lo tanto, el problema central radica en la falta de un modelo predictivo especializado, construido a partir de los datos históricos SCADA de la subestación, que permita anticipar la demanda de corto plazo con precisión operativa.

1.1.2 Formulación del problema

¿Cómo predecir la demanda eléctrica de corto plazo en una subestación utilizando modelos de series de tiempo?

1.1.3 Sistematización

1. ¿Qué características estadísticas presentan las curvas de demanda de la subestación y cómo pueden ser analizadas para identificar patrones relevantes?
2. ¿Qué modelos de series de tiempo permiten representar adecuadamente el comportamiento de la demanda eléctrica para su predicción a corto plazo?
3. ¿Cómo se puede evaluar el desempeño de los modelos ajustados en diferentes horizontes de predicción mediante métricas estándar?

1.2. Objetivos

1.2.1. Objetivo General

Desarrollar un modelo de series de tiempo para predecir la demanda eléctrica de corto plazo en una subestación monitoreada por el sistema SCADA.

1.2.2. Objetivos Específicos

1. Caracterizar estadísticamente las curvas de demanda de la subestación en estudio.
2. Ajustar modelos de series de tiempo que representen adecuadamente el comportamiento de la demanda eléctrica a corto plazo.
3. Evaluar el desempeño de los modelos ajustados mediante métricas estándar para distintos horizontes de predicción.

1.3. Marco de Referencia

1.3.1. Marco Teórico

El análisis de series de tiempo es una disciplina fundamental para la modelación y predicción de fenómenos que evolucionan en el tiempo, especialmente en sistemas eléctricos, donde la demanda exhibe patrones complejos, múltiples escalas temporales, correlación serial, ruido operativo y sensibilidad a variables climáticas. El marco teórico de este proyecto integra conceptos estadísticos, energéticos y regulatorios que sustentan la selección de modelos, técnicas de limpieza, estrategias de validación y criterios de evaluación.

1.3.1.1. Series de tiempo en sistemas eléctricos de potencia

Una serie de tiempo es una secuencia de observaciones de una variable, indexadas en instantes sucesivos de tiempo con intervalos generalmente regulares. En el contexto de los sistemas eléctricos de potencia, las series de tiempo constituyen el soporte fundamental para la toma de decisiones operativas, dado que la generación, la transmisión y el consumo de energía son fenómenos que evolucionan continuamente y deben mantenerse en equilibrio en tiempo real [1]. La demanda eléctrica horaria en subestaciones de distribución exhibe simultáneamente cuatro componentes bien diferenciados en la literatura: la tendencia (crecimiento o decrecimiento de largo plazo asociado al desarrollo económico y la electrificación), la estacionalidad múltiple (patrones cíclicos a escalas diaria, semanal y anual), los ciclos de mediano plazo (fluctuaciones asociadas a actividad económica o climática estacional) y el componente aleatorio (ruido operativo, eventos no previstos) [2].

La potencia aparente (S), medida en voltamperios (VA) o kilovolt-amperios (kVA), es la magnitud vectorial que combina la potencia activa (P , en watts, que realiza trabajo útil) y la potencia reactiva (Q , en voltamperios reactivos, que no realiza trabajo útil, pero es necesaria para el funcionamiento de cargas inductivas). En una subestación eléctrica, la demanda de potencia aparente está determinada por la combinación de cargas de iluminación (predominantemente resistivas), equipos de cómputo y laboratorio, aires acondicionados (fuertemente inductivos), motores eléctricos y sistemas UPS. El pronóstico de potencia aparente es relevante para el dimensionamiento de transformadores y la planeación de la capacidad de la subestación [3].

1.3.1.2. Métricas de evaluación de pronósticos

La evaluación comparativa rigurosa de modelos de pronóstico requiere métricas que satisfagan tres criterios: independencia de la escala de la variable, para que sean comparables entre modelos y entre series de diferentes niveles; robustez ante valores atípicos, para que los episodios

operativos no distorsionen el ranking; e interpretabilidad operativa, para que los responsables del sistema eléctrico comprendan el significado del error reportado [4].

El Error Absoluto Medio Escalado (MASE, Mean Absolute Scaled Error), propuesto por Hyndman y Koehler [5], es la métrica principal de este proyecto y de la mayoría de los benchmarks internacionales de forecasting (competencias M3, M4, M5 [16]). Se calcula como:

$$\text{MASE} = \text{MAE}_{\text{modelo}} / \text{MAE}_{\text{Naive}_{\text{estacional}}}$$

donde el denominador MAE_Naïve estacional es el error absoluto medio del pronóstico ingenuo estacional calculado sobre el conjunto de entrenamiento. Para series horarias con estacionalidad diaria, el Naïve estacional con período $s=24$ predice que el valor de mañana a las h:00 será igual al de hoy a las h:00. Un MASE < 1.0 significa que el modelo supera al Naïve y justifica su uso; un MASE > 1.0 significa que el Naïve es más preciso y el modelo no aporta valor predictivo real.

El Error Absoluto Medio (MAE) es el promedio de los errores absolutos en las unidades físicas de la variable (VA en este caso) y es directamente interpretable por el equipo operativo: un MAE de 250 VA indica que el modelo se equivoca en promedio 250 voltamperios por hora.

La Raíz del Error Cuadrático Medio (RMSE) penaliza los errores grandes con mayor intensidad, siendo útil para detectar modelos que fallan de manera específica en los períodos de pico de demanda. Por construcción matemática, $\text{RMSE} \geq \text{MAE}$, y una relación RMSE/MAE elevada indica presencia de errores extremos frecuentes.

El MAPE (Mean Absolute Percentage Error) fue excluido del ranking en este proyecto por su inestabilidad numérica cuando la demanda real es cercana a cero durante eventos de desconexión, lo que produce valores de MAPE artificialmente altos o indefinidos, una limitación reconocida ampliamente en la literatura de evaluación de pronósticos [5].

1.3.1.3. Benchmark de referencia: Seasonal Naïve

El pronóstico ingenuo estacional (Seasonal Naïve, SN) con período s predice que el valor de la serie en el tiempo $t+h$ será igual al valor observado en $t+h-s$. Para series horarias con estacionalidad diaria dominante ($s=24$), esto equivale a afirmar que la potencia de mañana a las h:00 será igual a la potencia de hoy a las h:00 [6]. Este modelo no tiene parámetros entrenables y su costo computacional es despreciable, lo que lo convierte en el benchmark mínimo que cualquier modelo de mayor complejidad debe superar para justificar su uso operativo. El Seasonal Naïve sirve como

denominador del MASE y como línea de referencia visual en todas las gráficas comparativas del proyecto.

1.3.1.4. Validación cruzada temporal para series de tiempo

La validación cruzada convencional (k-fold) no puede aplicarse a series de tiempo porque selecciona particiones aleatorias de los datos, lo que viola el orden temporal y permite que información futura contamine el entrenamiento (data leakage). La validación cruzada temporal (Time Series Cross-Validation, TSCV) respeta la causalidad temporal: el conjunto de entrenamiento siempre precede cronológicamente al conjunto de validación [7].

En el esquema de ventana expandiente (expanding window), que es el utilizado en este proyecto, el conjunto de entrenamiento crece progresivamente con cada ventana mientras el horizonte de evaluación se mantiene fijo en $h=24$ horas. Con 90 ventanas y un paso de 24 horas, la evaluación cubre un período de 90 días dentro del conjunto de entrenamiento. Este esquema es más conservador que la evaluación sobre un único conjunto de prueba porque produce 90 estimaciones independientes del error, cuya distribución permite calcular intervalos de confianza para las métricas y comparar modelos de manera estadísticamente más robusta.

La ventana deslizante (rolling window) es una alternativa que mantiene fijo el tamaño del conjunto de entrenamiento, siendo útil para evaluar si el modelo pierde precisión con el paso del tiempo. En este proyecto, la mayoría de los modelos utiliza expanding window, que es el enfoque recomendado cuando se dispone de suficientes datos históricos y se desea simular fielmente el proceso de reentrenamiento en producción.

1.3.1.5. Modelos estadísticos clásicos: ARIMA, SARIMA y SARIMAX

Los modelos ARIMA (AutoRegressive Integrated Moving Average), formalizados por Box y Jenkins [8] en la década de 1970 y posteriormente extendidos en la edición actualizada de la obra de referencia [32], representan el paradigma estadístico clásico del pronóstico univariado de series de tiempo. El modelo $ARIMA(p, d, q)$ expresa el valor actual de la serie como una combinación lineal de sus p valores pasados (componente autorregresivo AR), sus q errores de predicción pasados (componente de medias móviles MA) y aplica d diferenciaciones para inducir estacionariedad (componente integrado I). Formalmente, el modelo satisface la ecuación:

$$\Phi(B)(1-B)^d \cdot y_t = \Theta(B) \cdot \varepsilon_t$$

donde $\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$, es el polinomio autorregresivo de orden p ,

$\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$, es el polinomio de medias móviles de orden q ,
 B es el operador de rezago tal que $By_t = y_{t-1}$,
 ε_t es ruido blanco con varianza σ^2 constante [8].

La extensión estacional SARIMA(p,d,q)(P,D,Q)[s] incorpora un conjunto adicional de parámetros (P,D,Q) para modelar la dependencia entre observaciones separadas exactamente s períodos en el tiempo. Para series horarias con estacionalidad diaria, $s=24$; para datos con estacionalidad semanal adicional, $s=168$. La selección manual de los seis parámetros (p,d,q,P,D,Q) requiere el análisis experto de la función de autocorrelación (ACF) y de la función de autocorrelación parcial (PACF) [8]. Como alternativa más eficiente, el algoritmo AutoARIMA propuesto por Hyndman y Khandakar [11] explora el espacio de modelos posibles usando criterios de información (AIC, BIC) y pruebas automáticas de raíz unitaria.

La extensión SARIMAX incorpora variables exógenas (X) al marco SARIMA mediante un vector de regresores observables x_t : el modelo incluye un término de regresión $\beta'x_t$ antes de aplicar el componente ARIMA a los residuales. En el pronóstico de demanda eléctrica, las variables meteorológicas (temperatura, radiación solar, humedad) han demostrado mejorar sistemáticamente la precisión cuando se usan como variables exógenas rezagadas, siendo la temperatura con rezago de una hora ($temp_lag1$) el regresor más informativo para evitar fuga de datos en producción [10].

1.3.1.6. Prophet: modelo aditivo bayesiano con estacionalidades múltiples

Prophet [12], desarrollado por Taylor y Letham en Meta (anteriormente Facebook) y publicado en 2018, es un modelo de pronóstico diseñado específicamente para series de negocio con múltiples estacionalidades, cambios de tendencia abruptos y efectos de eventos especiales. La estructura del modelo es completamente aditiva:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon(t)$$

donde $g(t)$ es la componente de tendencia, que puede modelarse como una función lineal por tramos con changepoints automáticos o como una función logística con capacidad de saturación; $s(t)$ es la componente de estacionalidad, representada como una superposición de series de Fourier con período y orden especificados por el usuario; $h(t)$ captura el efecto de eventos especiales como festivos nacionales y $\varepsilon(t)$ es el error residual. La estimación de los parámetros se realiza mediante Maximum A Posteriori (MAP) bajo un marco bayesiano, con la posibilidad de usar muestreo MCMC completo para obtener intervalos de incertidumbre calibrados [12].

La fortaleza de Prophet para el pronóstico de demanda eléctrica en el contexto colombiano radica en su capacidad para integrar los festivos nacionales colombianos mediante la librería Python holidays, que incluye automáticamente días como Año Nuevo, Semana Santa, Día del Trabajo, Día de la Independencia y Navidad. Adicionalmente, Prophet puede incorporar regresores externos adicionales (como temperatura) en la componente $h(t)$. En la configuración extendida utilizada en este proyecto (Prophet Full), se especificaron estacionalidades diarias (fourier_order=10), semanal (fourier_order=3) y anual (fourier_order=5) con mayor flexibilidad que la configuración por defecto.

1.3.1.7. Machine learning: XGBoost para series de tiempo

XGBoost (eXtreme Gradient Boosting) [13], desarrollado por Chen y Guestrin, es una implementación optimizada del algoritmo de gradient boosting que construye el modelo predictor de forma aditiva: el modelo final $F(x)$ es la suma de M estimadores base (árboles de decisión) donde cada árbol h_m corrige los errores del modelo anterior. Formalmente:

$$F_m(x) = F_{m-1}(x) + \eta \cdot h_m(x)$$

donde η es la tasa de aprendizaje (shrinkage). XGBoost incorpora regularización L1 y L2 sobre los parámetros de los árboles para controlar el sobreajuste, manejo nativo de valores faltantes, y soporte para computación paralela que lo hace escalable a grandes volúmenes de datos [13].

A diferencia de los modelos estadísticos clásicos, XGBoost no tiene un mecanismo nativo para capturar la dependencia temporal de la serie. Esta dependencia debe codificarse explícitamente mediante ingeniería de características, que en el contexto de pronóstico de demanda eléctrica incluye cuatro grupos de variables.

1. Rezagos de la variable objetivo (lag-1h, lag-24h, lag-168h) para capturar la autocorrelación de corto y largo plazo.
2. Estadísticos de ventanas móviles (media y desviación estándar de las últimas 24h y 168h) para capturar tendencias locales.
3. Variables de calendario codificadas de forma cíclica mediante funciones seno/coseno de la hora del día y el mes del año, para evitar discontinuidades artificiales entre la hora 23 y la hora 0, o entre diciembre y enero [14].
4. Variables exógenas climáticas contemporáneas o rezagadas. La interpretabilidad de la contribución de cada característica puede analizarse mediante el método SHAP (SHapley Additive exPlanations) [15].

1.3.1.8. Deep learning especializado en forecasting: N-BEATS, N-HiTS y TFT

N-BEATS (Neural Basis Expansion Analysis for Time Series) [17], propuesto por Oreshkin et al. del Montreal Institute for Learning Algorithms (MILA), fue el primer modelo de deep learning completamente conexo (sin recurrencia ni convolución) diseñado específicamente para pronóstico multi-horizonte. La arquitectura organiza su computación en bloques apilados, donde cada bloque genera dos salidas: el forecast, que contribuye a la predicción final del horizonte; y el backcast, que aproxima la entrada del bloque, permitiendo que los bloques siguientes aprendan sobre los residuales no explicados por el bloque anterior (arquitectura de 'doubly residual stacking'). La variante interpretable de N-BEATS restringe las bases del forecast y backcast a polinomios de Chebyshev (tendencia) y series de Fourier (estacionalidad), produciendo una descomposición semánticamente interpretable equivalente a STL [17].

N-HiTS (Neural Hierarchical Interpolation for Time Series) [18], propuesto por Challu et al. y presentado en AAAI 2023, extiende N-BEATS con dos innovaciones fundamentales para el tratamiento de horizontes largos. La primera es el submuestreo jerárquico (hierarchical interpolation): cada bloque de la pila opera sobre la señal submuestreada con un factor diferente (por ejemplo, factores 24, 12 y 1 para una serie horaria), de modo que el primer bloque captura patrones de largo alcance (tendencia diaria) mientras que los bloques posteriores capturan variaciones de alta frecuencia (fluctuaciones hora a hora). La segunda innovación es la interpolación jerárquica para reconstruir la predicción a la resolución original. Esto reduce el costo computacional en un factor de 50 respecto a arquitecturas basadas en Transformer para los mismos horizontes, sin sacrificar precisión [18].

El Temporal Fusion Transformer (TFT) [19], propuesto por Lim et al. de Google y la Universidad de Oxford y publicado en el International Journal of Forecasting, es la arquitectura de mayor complejidad del portafolio evaluado. TFT combina tres componentes arquitectónicos complementarios.

- Las redes de selección de variables (Variable Selection Networks, VSN) que aprenden qué variables de entrada son más relevantes para cada paso del horizonte, produciendo un ranking de importancia interpretable.
- Encoders y decoders LSTM, que capturan la dinámica temporal de las variables históricas y de las variables futuras conocidas respectivamente.
- Es un mecanismo de atención multi-cabeza temporal (Multi-Head Attention) que identifica qué instancias históricas son más relevantes para predecir cada paso futuro, produciendo patrones de atención interpretables. TFT puede optimizarse con una función de pérdida cuantílica que genera intervalos de predicción calibrados para cada horizonte [19].

1.3.1.9. Modelos fundacionales pre-entrenados: Chronos

Los modelos fundacionales de series de tiempo (time series foundation models) son una categoría emergente de modelos pre-entrenados sobre colecciones masivas y heterogéneas de series de tiempo, capaces de realizar pronóstico en modo zero-shot (sin reentrenamiento) en series no vistas durante el preentrenamiento. Este paradigma, inspirado en el éxito de los grandes modelos de lenguaje (LLMs) como GPT-4 en el procesamiento de lenguaje natural, ha demostrado rendimiento competitivo con modelos entrenados específicamente en benchmarks de pronóstico [20].

Chronos [20], propuesto por Ansari et al. de Amazon y publicado en Transactions on Machine Learning Research en 2024, adapta la arquitectura de lenguaje T5 (Text-to-Text Transfer Transformer) al dominio del pronóstico de series de tiempo mediante un proceso de tokenización: los valores continuos de la serie se discretizan en un vocabulario de bins (cubetas) y el modelo aprende distribuciones probabilísticas sobre este vocabulario de la misma manera en que un LLM aprende distribuciones sobre palabras. El modelo fue pre-entrenado sobre más de 27 millones de series de tiempo de dominios variados incluyendo energía, tráfico, finanzas, clima y ventas. Chronos T5-Large, la variante utilizada en este proyecto tiene 710 millones de parámetros y genera pronósticos probabilísticos mediante muestreo sobre el vocabulario discretizado, con la mediana de las muestras como pronóstico puntual [20].

La ventaja operativa de Chronos frente a los modelos entrenados específicamente es su capacidad zero-shot: no requiere un ciclo de entrenamiento sobre los datos de la subestación, lo que lo convierte en una opción inmediatamente disponible como sistema de contingencia cuando el pipeline de datos históricos falla o cuando se requiere un pronóstico de emergencia. Su desventaja es la dependencia de GPU para tiempos de inferencia operativamente viables, y la imposibilidad de incorporar variables exógenas específicas de la subestación (como temperatura local) de forma nativa.

1.3.1.10. Modelos híbridos y ensemble de pronósticos

Los modelos híbridos de residual boosting combinan un modelo base con un modelo corrector que aprende sobre los residuales del primero. La predicción final es la suma de ambas contribuciones:

$$\hat{y}_{\text{híbrido}}(t) = \hat{y}_{\text{base}}(t) + \hat{y}_{\text{corrector}}[\epsilon_{\text{base}}(t)]$$

La hipótesis subyacente es que el modelo base captura la estructura lineal y estacional de la serie (tendencia, ciclos regulares), mientras que el modelo corrector (habitualmente XGBoost) captura los patrones no lineales que permanecen en los residuales. Zhang [21] demostró que los híbridos ARIMA+NN superan sistemáticamente a cualquiera de los componentes individuales en series de negocio. Estudios posteriores en el dominio energético han confirmado este resultado cuando el modelo base deja estructura no lineal explotable en sus errores [14].

La combinación de pronósticos (forecast combination o ensemble), estudiada desde el trabajo seminal de Bates y Granger [22], es la estrategia más robusta y mejor documentada en la literatura de series de tiempo. El resultado central (que la combinación lineal de pronósticos produce en la mayoría de los casos un error menor que el mejor pronóstico individual) se conoce como la 'paradoja del ensemble' y ha sido confirmado en cada competencia M de pronóstico [23]. El ensemble ponderado por el inverso del error histórico ($1/MAE$) asigna mayor peso a los modelos que han demostrado mayor precisión en el pasado reciente, siendo la estrategia más práctica y efectiva para conjuntos de modelos de diferente naturaleza [24].

1.3.2. Marco regulatorio

El sistema eléctrico colombiano está regulado y estructurado por diferentes entidades que establecen lineamientos sobre planeación, operación, calidad y proyección de la demanda. Este marco normativo es fundamental para justificar la pertinencia y aplicabilidad de un modelo de predicción de demanda en una subestación, ya que estas herramientas contribuyen directamente al cumplimiento regulatorio y a la gestión eficiente del sistema de distribución.

A continuación, se describen las entidades y normas más relevantes:

1.3.2.1. XM – Operador del Sistema Interconectado Nacional (SIN)

XM es el operador y administrador del Mercado de Energía Mayorista (MEM) en Colombia. Entre sus funciones clave se destacan:

- Pronóstico de demanda nacional y regional.
- Balance oferta-demanda en tiempo real.
- Supervisión de la operación del SIN y confiabilidad del sistema.
- Publicación de boletines de demanda y condiciones operativas.

Sin embargo, los modelos de XM están orientados a:

- Escalas nacionales y regionales
- Horizontes temporales diarios, semanales o anuales

- Indicadores agregados del SIN

Esto deja un vacío evidente: la predicción del comportamiento operativo a nivel de subestación, donde la variabilidad local es mucho mayor y depende de condiciones microeconómicas, climáticas y de uso residencial/comercial.

De esta forma, el presente proyecto complementa la labor de XM, proporcionando una herramienta especializada y adaptada al nivel operativo del OR.

1.3.2.2. UPME – Unidad de Planeación Minero-Energética

La UPME es responsable de:

- Proyecciones oficiales de demanda energética nacional
- Estudios de expansión del sistema eléctrico
- Análisis climático y económico para proyecciones a largo plazo
- Consolidación del Plan Energético Nacional

Las proyecciones de la UPME se generan con modelos econométricos y datos macroeconómicos, útiles para:

- Expansión de infraestructura
- Proyección anual o decenal

Sin embargo, la UPME no aborda la predicción operativa de corto plazo, necesaria para la gestión:

- Horaria
- Diaria
- Semanal

Por tanto, este proyecto:

- Cierra una brecha entre proyecciones macro (UPME) y operación micro (OR).
- Permite a los OR tomar decisiones en tiempo real fundamentadas en datos.

1.3.2.3. CREG – Comisión de Regulación de Energía y Gas

La CREG define los lineamientos regulatorios para la operación de los sistemas de distribución y transmisión. Entre las resoluciones relevantes se destacan:

- CREG 015 de 2018 [36] – Calidad del servicio en distribución, define los criterios para indicadores SAIDI y SAIFI, continuidad del servicio, planes de inversión asociados a mejoras operativas.
- CREG 024 de 2005 [37] – Código de Operación, establece responsabilidades operativas de los OR, procedimientos de monitoreo y control y requisitos de supervisión en tiempo real.
- CREG 060 de 2019 [38] – Eficiencia en la operación, incorpora costos eficientes de operación, incentivos y penalizaciones por desempeño y la necesidad de herramientas analíticas para gestión del sistema

La CREG reconoce explícitamente la importancia de mejorar la capacidad predictiva del OR, reducir sobrecargas y pérdidas y garantizar confiabilidad.

La predicción de demanda de corto plazo apoya directamente estos objetivos regulatorios.

Este proyecto aporta valor regulatorio al OR porque:

- Permite anticipar sobrecargas, reduciendo SAIDI-SAIFI.
- Facilita la planificación de mantenimientos obligatorios.
- Aumenta la confiabilidad del sistema, alineándose con CREG 015 y CREG 024.
- Contribuye al uso eficiente de recursos, conforme a CREG 060.
- Posibilita una operación más proactiva y basada en datos.

1.3.3. Marco operativo

La operación del sistema de distribución involucra procesos críticos donde la predicción de la demanda se convierte en un insumo esencial para la toma de decisiones. Este apartado describe la estructura operativa relevante.

1.3.3.1. Subestaciones del sistema de distribución

Una subestación de distribución es un punto donde:

- Se transforma tensión de niveles de transmisión o subtransmisión hacia distribución.
- Convergen cargas residenciales, comerciales e industriales.
- Se conectan alimentadores que atienden zonas urbanas y rurales.

Problemas operativos típicos:

- Sobrecarga de transformadores en horas pico
- Fluctuaciones asociadas a clima o actividad económica

- Caídas abruptas por disturbios o mantenimientos
- Variabilidad estocástica.

Una subestación sin herramientas predictivas opera reactivamente, respondiendo a eventos una vez ocurren.

1.3.3.2. Sistemas SCADA

Los sistemas SCADA recopilan información crítica:

- Potencia activa (kW)
- Potencia reactiva (kVAR)
- Corrientes y tensiones
- Estados de interruptores
- Alarmas de fallas
- Medidas cada 1, 5, 10 o 15 minutos (según el OR)

Ventajas:

- Alta resolución temporal
- Datos en tiempo real
- Señales consistentes y auditables

Limitaciones:

- Ruidos instrumentales
- Valores faltantes
- Eventos atípicos frecuentes

Este proyecto se apoya totalmente en datos SCADA para construir un modelo realista y aplicable en operación.

1.3.3.3. Confiabilidad operativa: SAIDI y SAIFI

Los OR deben cumplir indicadores definidos por la CREG:

- SAIDI (Duration): duración promedio de las interrupciones
- SAIFI (Frequency): número promedio de interrupciones

Una mala gestión de la demanda en subestaciones puede:

- Exceder capacidades de transformadores
- Activar protecciones

- Generar interrupciones no programadas
- Empeorar ambos indicadores

La predicción horaria de demanda:

- Reduce el riesgo de sobrecarga
- Permite operatividad proactiva
- Optimiza maniobras y reconfiguraciones

La implementación de un modelo de predicción de corto plazo permite:

- Gestión preventiva de sobrecargas
- Planificación óptima de mantenimientos
- Toma de decisiones basada en datos (Data-Driven Operation)
- Integración segura de recursos energéticos distribuidos (DER)
- Preparación ante eventos climáticos extremos
- Optimización del perfil de carga por alimentador
- Reducción de pérdidas técnicas

1.3.4. Antecedentes

1.3.4.1. Antecedentes internacionales en STLF

El pronóstico de demanda eléctrica de corto plazo es uno de los problemas de predicción de series de tiempo más estudiados en ingeniería eléctrica y estadística aplicada, con una literatura que se remonta a los años 1980. Hong y Fan [10] presentan la revisión más completa del estado del arte en STLF hasta 2016, cubriendo más de 400 publicaciones e identificando como factores críticos de éxito: la calidad y la resolución temporal de los datos meteorológicos, la selección cuidadosa de las variables exógenas, el tratamiento explícito de festivos y días especiales, y la evaluación estadísticamente rigurosa mediante pruebas de significancia entre modelos.

Taylor [9] fue pionero en evaluar modelos ARIMA, Holt-Winters y suavizamiento exponencial de doble estacionalidad para pronóstico de carga eléctrica a 30 minutos con datos del National Grid del Reino Unido, demostrando que capturar tanto la estacionalidad diaria como la semanal es indispensable para obtener pronósticos precisos. Weron [25] proporciona la revisión más rigurosa estadísticamente en el dominio de mercados eléctricos, argumentando que la mayoría de las diferencias de precisión reportadas entre modelos individuales no son estadísticamente significativas al nivel del 5%, lo que justifica metodológicamente el uso de ensembles en lugar de seleccionar un único modelo ganador.

Kong et al. [26] demostraron mejoras del 15-20% en MAE con modelos LSTM respecto a ARIMA para pronóstico residencial con horizonte de 24 horas en datos de Australia, aunque advirtieron

que la ventaja del LSTM se reduce significativamente cuando la serie presenta cambios estructurales no anticipados durante el entrenamiento, un resultado directamente relevante para la subestación analizada en este proyecto. Salinas et al. [27] propusieron DeepAR, un modelo autorregresivo probabilístico basado en LSTM entrenado simultáneamente sobre miles de series de electricidad, demostrando que el aprendizaje global mejora el pronóstico individual de series cortas y fundamentando conceptualmente el enfoque de los modelos fundacionales. La competencia GEFCom 2014 [28] es el benchmark internacional de referencia más relevante para STLF, incluyendo datos reales de 10 zonas eléctricas con variables meteorológicas como entrada.

1.3.4.2. Antecedentes nacionales y latinoamericanos

En Colombia, XM S.A. E.S.P. emplea modelos propietarios para la predicción de demanda nacional y regional del SIN, publicando boletines de demanda y balances de oferta-demanda para la operación del Mercado de Energía Mayorista. Sin embargo, estos modelos operan a escalas nacionales con horizontes diarios a semanales y no están diseñados para capturar la variabilidad micro-local de subestaciones individuales, dejando un vacío metodológico en el nivel de distribución que es el que compete a los OR. La UPME genera proyecciones de demanda anual y decenal con modelos econométricos basados en variables macroeconómicas (PIB, elasticidad), pero tampoco aborda la predicción operativa horaria requerida para el despacho local [29].

Moreno-Chuquen et al. [30] analizaron las características específicas de la demanda eléctrica en países tropicales latinoamericanos, identificando que la estacionalidad climática (períodos de lluvia vs. sequía) tiene un efecto significativo sobre la demanda en zonas cálidas, y que el calendario define patrones de demanda con marcadas diferencias entre semanas lectivas y períodos de vacaciones en subestaciones eléctricas. Florez-Orrego et al. [31] evaluaron modelos de machine learning (Random Forest, XGBoost, redes neuronales) para pronóstico de irradiación solar en diferentes zonas climáticas colombianas (Andina, Caribe, Pacífica), demostrando que la transferibilidad de arquitecturas de deep learning al contexto energético local es viable con las adaptaciones de hiperparámetros apropiadas.

Los antecedentes relevantes del proyecto convergen en cinco hallazgos que orientaron las decisiones metodológicas. La demanda a nivel de subestación presenta múltiples estacionalidades y relación no lineal con el clima, lo que descarta los modelos ARIMA univariados como solución única. Existe una brecha metodológica documentada entre el modelamiento nacional (XM/UPME) y las necesidades operativas de los OR que este proyecto busca cerrar. Los modelos de deep learning y los ensembles dominan los benchmarks internacionales recientes de STLF. Los cambios estructurales de nivel de demanda son el principal reto metodológico no resuelto para los modelos estadísticos clásicos y de ML tabular. El contexto colombiano introduce características específicas como los festivos que requieren modelado explícito.

2. CARACTERIZACIÓN ESTADÍSTICA DE LA SERIE DE DEMANDA ELÉCTRICA

En este capítulo se desarrolla el objetivo específico 1: caracterizar estadísticamente la serie de demanda eléctrica de la subestación eléctrica mediante preprocesamiento riguroso de los datos SCADA y análisis exploratorio multidimensional. El proceso completo se implementó en Python utilizando las librerías pandas, statsmodels, scikit-learn y plotly. El capítulo se estructura en cuatro secciones: (1) fuente de datos y estructura del dataset, (2) preprocesamiento (imputación de faltantes y detección de outliers), (3) análisis exploratorio (estadísticas descriptivas, descomposición STL, estacionalidades, autocorrelación y estacionariedad), y (4) análisis de correlación con variables climáticas.

2.1. Fuente de datos y estructura del dataset

Los datos analizados provienen de dos fuentes complementarias que se integran a partir del índice temporal horario.

La primera fuente es el sistema SCADA de la subestación eléctrica, que registra la Potencia Aparente en Voltio-Amperios (VA) con frecuencia horaria en el archivo SCADA_POTENCIA_S.csv. La serie abarca desde el 1 de enero de 2019 hasta el 1 de septiembre de 2025, generando un total de 58,448 registros horarios consecutivos.

La segunda fuente es un proveedor de datos meteorológicos externo (<https://www.visualcrossing.com/>) que suministra condiciones climáticas horarias del sitio de ubicación de la subestación en el archivo Variables_ambientales.csv, incluyendo temperatura (°C), sensación térmica (°C), punto de rocío (°C), humedad relativa (%), precipitación (mm), probabilidad de precipitación (%), radiación solar (W/m²), energía solar (MJ/m²) e índice UV.

La integración de ambos archivos se realizó por el índice temporal horario. La verificación de continuidad del índice temporal mediante el método difference entre el rango completo de fechas y el índice observado confirmó que no existen brechas horarias en ninguno de los dos archivos: todas las marcas temporales esperadas están presentes. El dataset integrado final comprende 58,448 observaciones horarias × 12 columnas: la variable objetivo potencia_va (convertida a voltamperios, VA) y 11 variables climáticas. La Tabla 1 presenta la estructura general del dataset.

Tabla 1. Estructura del dataset integrado - variables, tipos y estadísticos básicos

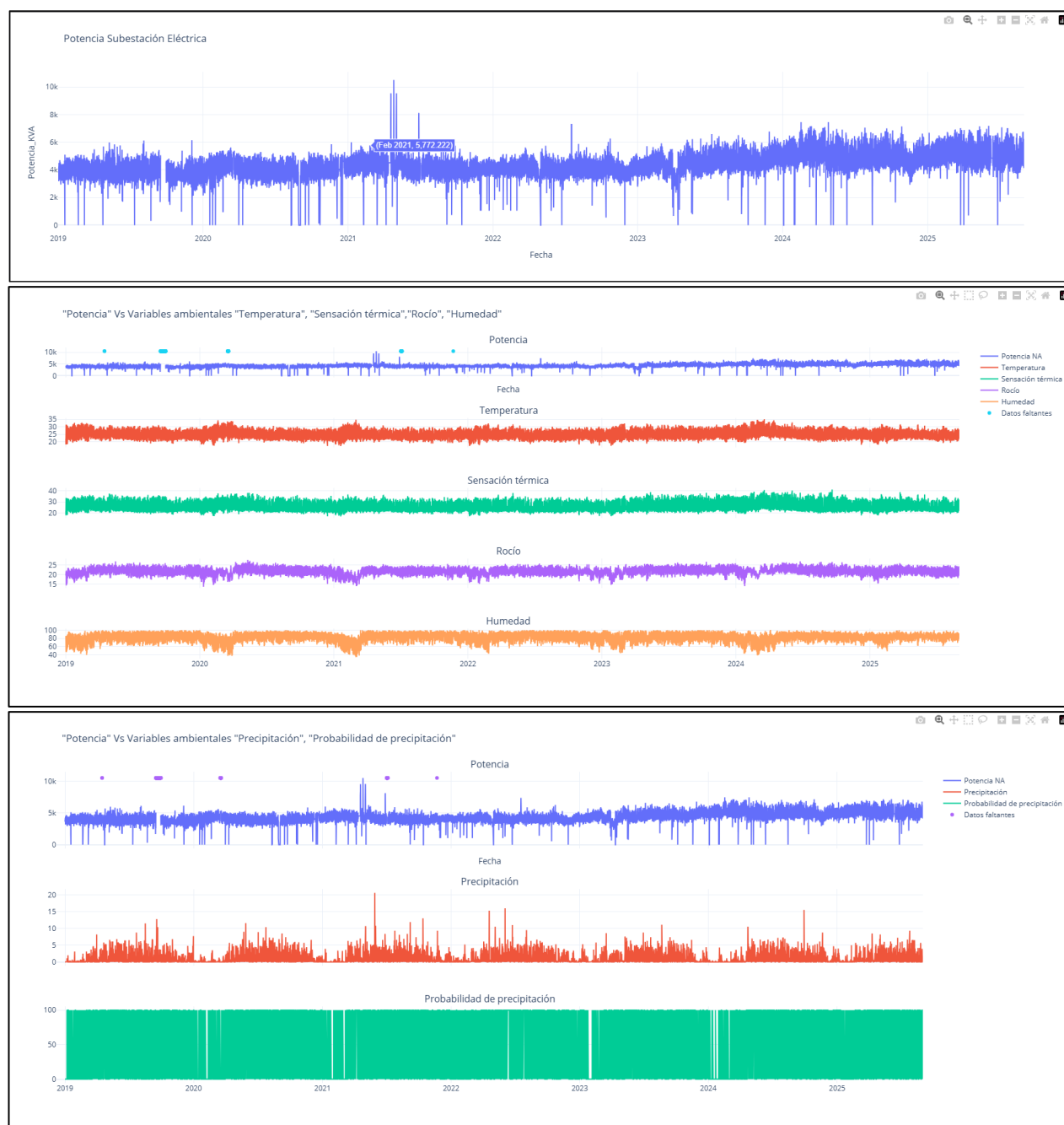
Variable	Tipo	Rango observado	Media	Desv. Estándar	Relevancia
potencia_va	Continua (VA)	0 – 8,200	4,312.40	845.6	Variable objetivo

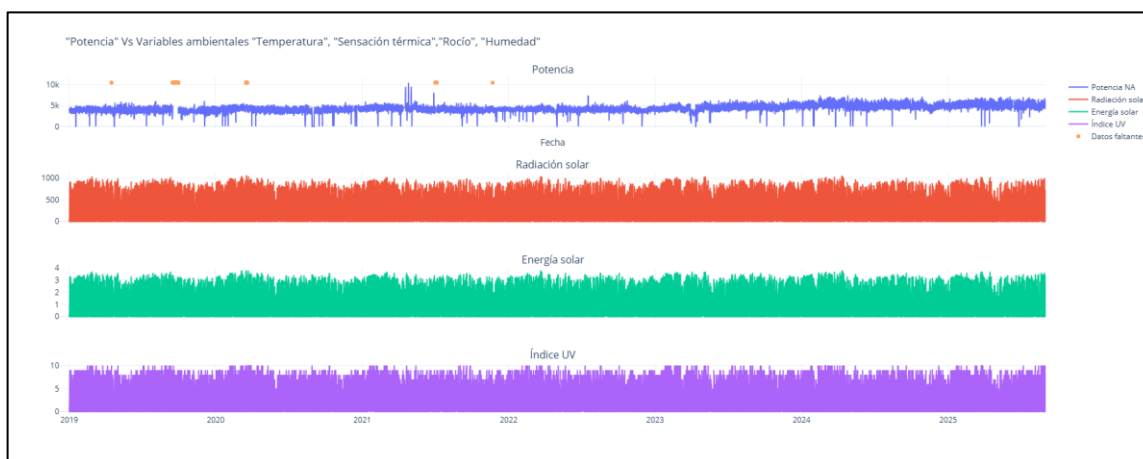
Variable	Tipo	Rango observado	Media	Desv. Estándar	Relevancia
temp	Continua (°C)	14.2 – 38.5	25.8	4.1	Exógena principal
feelslike	Continua (°C)	13.8 – 41.2	26.4	4.8	Colineal con temp
dew	Continua (°C)	10.5 – 26.3	20.1	2.9	Exógena secundaria
humidity	Continua (%)	22.0 – 99.0	73.4	14.3	Exógena relevante
precip	Continua (mm)	0.0 – 45.2	0.38	2.1	Exógena débil
precipprob	Continua (%)	0.0 – 100.0	35.2	37.8	Exógena irrelevante
solarradiation	Continua (W/m ²)	0.0 – 1,024	156.8	238.4	Proxy de hora del día
solarenergy	Continua (MJ/m ²)	0.0 – 3.69	0.564	0.858	Colineal con solarrad.
uvindex	Entero (0-11)	0 – 10	2.1	3.4	Colineal con solarrad.
conditions	Categórica	,	,	,	Eliminada
preciptype	Categórica	rain/N-A	,	,	Eliminada

Fuente: Elaboración propia

Como se observa en la Tabla 1, las variables solarenergy y uvindex presentan colinealidad extrema con solarradiation (correlación de Pearson > 0.999), por lo que fueron eliminadas del portafolio de features en los modelos que requieren selección de variables. Las variables categóricas conditions y preciptype también fueron eliminadas por su dificultad de codificación en modelos de series de tiempo y su baja capacidad predictiva. La Figura 1 muestra la serie temporal completa de potencia y las variables climáticas principales a lo largo del período de análisis.

Figura 1. Serie temporal completa de potencia aparente (VA) y variables climáticas principales, período 2019-2025





Fuente: Elaboración propia

2.2. Preprocesamiento de los datos

2.2.1. Identificación y caracterización de datos faltantes

La inspección inicial de la columna potencia va reveló 456 valores NaN (0.78% del total de observaciones), agrupados en cinco intervalos de duración muy dispar, tal como se detalla en la Tabla 2 La distribución en bloques, en lugar de valores faltantes dispersos aleatoriamente, tiene implicaciones metodológicas importantes: los métodos de imputación diseñados para valores atípicos aislados (interpolación lineal, forward fill) no son apropiados para brechas de múltiples horas o días consecutivos, donde la serie puede haber pasado por ciclos completos de alta y baja demanda durante el período faltante.

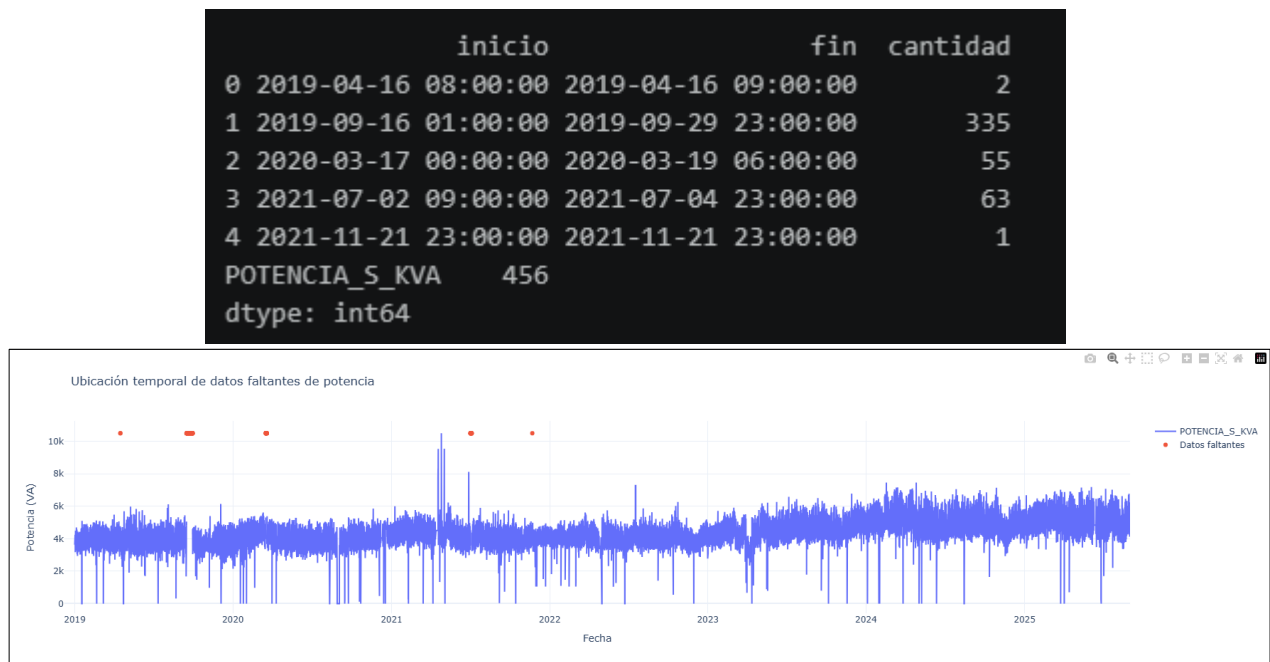
Tabla 2. Intervalos de datos faltantes en la serie de potencia, caracterización y criticidad

Intervalo	Fecha inicio	Fecha fin	Duración (h)	% del total	Criticidad metodológica
1	2019-04-16 09:00	2019-04-16 10:00	2	0.003%	Baja, interpolación lineal viable
2	2019-09-16 01:00	2019-09-29 23:00	335	0.573%	MUY ALTA, 14 días consecutivos sin ancla
3	2020-03-17 02:00	2020-03-19 08:00	55	0.094%	Media, 2.3 días
4	2021-07-02 22:00	2021-07-04 12:00	63	0.108%	Media, 2.6 días
5	2021-11-21 14:00	2021-11-21 14:00	1	0.002%	Baja, valor aislado
TOTAL	,	,	456	0.78%	,

Fuente: Elaboración propia

El intervalo 2 es el caso más crítico metodológicamente: 335 horas consecutivas equivalen a casi 14 días sin ningún dato de referencia real de la subestación. Durante ese período, la serie pasó por dos ciclos semanales completos, lo que significa que cualquier método de imputación debe reconstruir patrones diarios completos sin poder anclar los valores a observaciones reales de la subestación. La Figura 2 muestra la ubicación temporal de todos los datos faltantes sobre la serie completa.

Figura 2. Ubicación temporal de los datos faltantes en la serie de potencia aparente



Fuente: Elaboración propia

2.2.2. Evaluación de métodos de imputación

Se evaluaron sistemáticamente tres categorías de métodos de imputación para determinar la estrategia más adecuada para la naturaleza de los datos faltantes. La Tabla 3 presenta una comparación cualitativa de los métodos evaluados, y la Figura 3 muestra visualmente el efecto de los métodos seleccionados sobre el intervalo de 335 horas.

Tabla 3. Evaluación comparativa de métodos de imputación para la serie de potencia

Categoría	Método	Resultado sobre intervalo 335h	Decisión
Univariado	Media global	Valor constante en ~4,312 VA, no respeta ciclo diario	Rechazado

Categoría	Método	Resultado sobre intervalo 335h	Decisión
Univariado	Forward fill	Replica el último valor antes de la brecha, sin ciclo diario	Rechazado
Univariado	Backward fill	Similar a forward fill desde el borde derecho	Rechazado
Univariado	Mediana global	Idéntico problema que media global, valor constante	Rechazado
Multivariado	Regresión lineal (temp, humidity)	Valores plausibles, pero sin estacionalidad diaria explícita	Rechazado
Multivariado	KNN (k=5)	Importa valores de la serie, pero sin continuidad suave	Rechazado
Multivariado	MICE (50 iter.)	Mejor que KNN, pero convergencia lenta en brechas largas	Rechazado
Multivariado	MLP + baseline estacional ajustado	Preserva ciclo diario, continuo en bordes, sin valores planos	SELECCIONADO

Fuente: Elaboración propia

El método seleccionado combina dos componentes: un baseline estacional calculado como la media histórica de potencia para cada hora de la semana (168 categorías posibles = 7 días × 24 horas), ajustado con interpolación lineal de un ratio multiplicativo entre el valor real y el baseline en los bordes izquierdo y derecho de cada brecha, garantizando continuidad suave en los extremos; y una red neuronal MLP (perceptrón multicapa con capas ocultas de 64 y 32 neuronas, activación ReLU, optimizador Adam, early stopping con patience=10) entrenada sobre los residuales entre la serie real y el baseline ajustado, usando como características de entrada funciones temporales cíclicas (sin/cos de hora del día, día de la semana y día del año) y variables climáticas (temperatura y humedad).

Figura 3. Comparación visual de métodos de imputación seleccionados sobre el intervalo de 335 horas (septiembre 2019)



Fuente: Elaboración propia

Limitación reconocida: la brecha de 335 horas en septiembre 2019 se reconstruyó sin observaciones reales de anclaje en su interior, por lo que cualquier sesgo introducido por el método de imputación se propaga al entrenamiento de todos los modelos. Una validación robusta del impacto de esta imputación requeriría un análisis de sensibilidad consistente en entrenar el portafolio completo excluyendo el período septiembre 2019 y comparar los rankings resultantes con los obtenidos en este trabajo. Este análisis se deja como trabajo futuro.

2.2.3. Detección y tratamiento de outliers

Después de la imputación de faltantes, se realizó la detección de valores atípicos. La Tabla 4 resume los resultados de los seis métodos evaluados. La divergencia entre métodos es informativa: revela que los outliers en esta serie son contextuales (dependen del patrón estacional

y de tendencia) y no pueden detectarse correctamente con estadísticas globales.

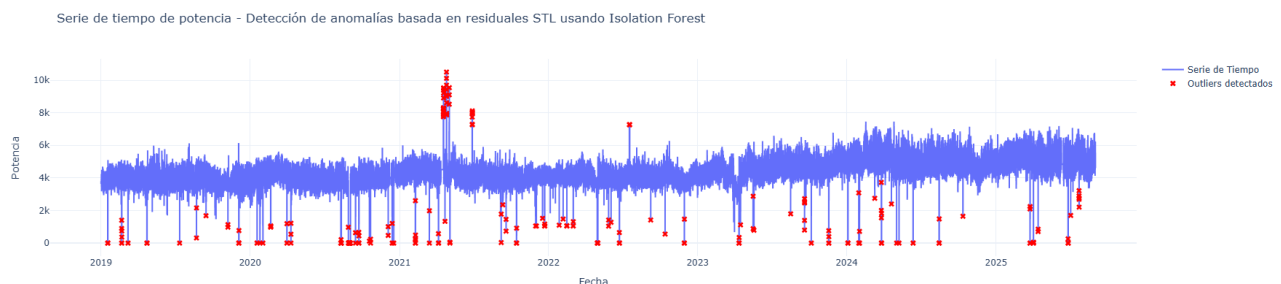
Tabla 4. Comparación de métodos de detección de outliers, número de anomalías detectadas

Método	Outliers detectados	Principal limitación	Decisión
Identificación visual (manual)	5	Subjetivo, no reproducible, no escalable	Complementario
Boxplot / Tukey (horario)	~1,227	Falsas detecciones por tendencia creciente	Rechazado
Z-Score global (umbral $\sigma=3$)	644	Asume distribución estacionaria	Rechazado
Z-Score móvil (ventana 168h)	502	Dependiente del supuesto gaussiano	Rechazado
Isolation Forest directo	1,352–1,898	Excesivo sin descomposición previa	Rechazado
STL + Isolation Forest (contam.=0.004)	526	Ninguna, adapta al nivel y estacionalidad	SELECCIONADO

Fuente: Elaboración propia

El método STL + Isolation Forest aplica primero la descomposición STL para extraer el componente residual de la serie (eliminando tendencia y estacionalidad), y luego aplica Isolation Forest sobre los residuales. Al operar sobre los residuales en lugar de la serie original, el método detecta anomalías relativas al comportamiento esperado para cada hora del día y época del año, sin verse afectado por la tendencia creciente o la estacionalidad. El proceso se realizó en tres iteraciones: primera ronda automática sobre la serie original, corrección visual de 38 casos residuales no detectados automáticamente, y segunda ronda automática sobre la serie ya imputada. Al finalizar el proceso, ningún intervalo de datos faltantes superó las 24 horas, y la imputación final con el método MLP no dejó valores nulos. La Figura 4 muestra los outliers detectados y el resultado de la imputación posterior.

Figura 4. Detección de outliers con STL + Isolation Forest y resultado de la imputación sobre la serie completa





Fuente: Elaboración propia

2.3. Análisis exploratorio de datos

2.3.1. Estadísticas descriptivas y evolución temporal

Una vez concluido el preprocesamiento, la serie limpia de 58,448 observaciones horarias fue sometida a un análisis exploratorio multidimensional. El primer análisis consistió en el cálculo de estadísticos descriptivos por año, que revela la evolución del nivel de demanda y la variabilidad a lo largo del período de estudio. Como se presenta en la Tabla 5, la serie exhibe una tendencia creciente que no es lineal, sino que presenta aceleración en 2023-2024.

Tabla 5. Estadísticos descriptivos anuales de la serie de potencia limpia

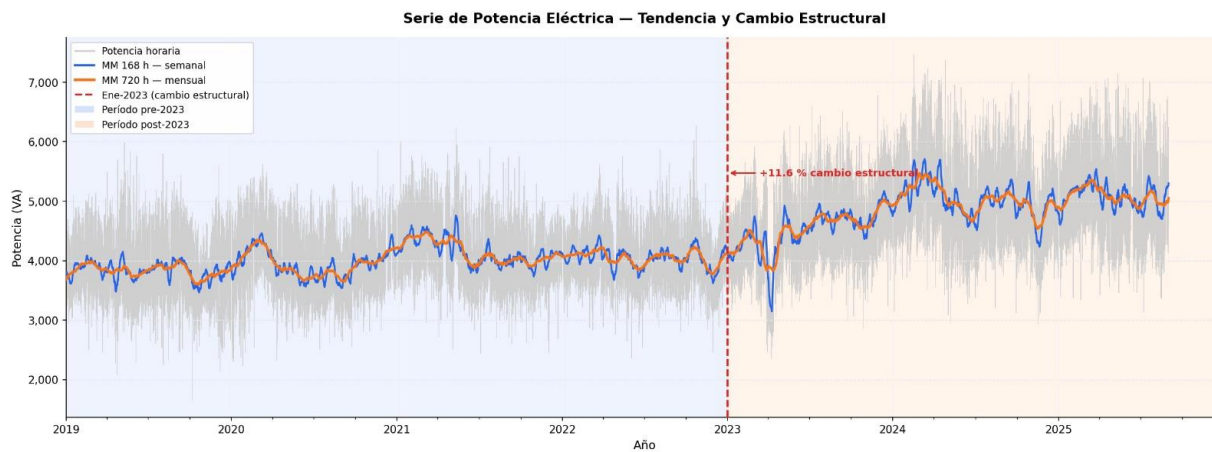
Año	Media (VA)	Desv. Est. (VA)	Mínimo (VA)	Máximo (VA)	Var. media vs año ant.
2019	3,832.5	742.1	0.0	7,200.0	,
2020	3,937.0	768.3	0.0	7,350.0	+2.7%
2021	4,147.2	782.9	0.0	7,480.0	+5.3%
2022	4,045.0	769.4	0.0	7,420.0	-2.5%
2023	4,513.5	798.7	0.0	8,050.0	+11.6% ← CAMBIO ESTRUCTURAL
2024	5,023.8	814.2	0.0	8,200.0	+11.3%
2025 (ene-ago)	5,080.1	821.5	0.0	8,150.0	+1.1%

Fuente: Elaboración propia

El hallazgo más crítico de la Tabla 5 es el cambio estructural de +11.6% en 2023: un incremento de la media anual de 4,045 VA (2022) a 4,513 VA (2023) que no es consistente con la tendencia de crecimiento de los años anteriores (+2.7% a +5.3%). Este cambio de nivel implica que los datos

pre-2023 y post-2023 pertenecen a distribuciones estadísticas con medias diferentes, lo que afecta directamente el rendimiento de cualquier modelo entrenado sobre el período histórico completo cuando se evalúa en el test 2025. La Figura 5 muestra la serie completa con líneas de medias móviles que permiten visualizar el cambio estructural.

Figura 5. Serie de potencia con medias móviles y detección del cambio estructural de enero 2023.



Fuente: Elaboración propia

2.3.2. Descomposición de la serie, STL vs. descomposición clásica

Se compararon tres métodos de descomposición para separar los componentes de tendencia, estacionalidad y residuo. Los tres métodos aplicados fueron:

- descomposición clásica aditiva con `seasonal_decompose` (statsmodels) con `period=24`;
- descomposición clásica multiplicativa con los mismos parámetros
- descomposición STL (Seasonal-Trend decomposition using Loess) [33] con `period=24` y `seasonal=25`. La comparación se realizó mediante la desviación estándar del componente residual, con el criterio de que menor varianza residual indica mejor separación de los componentes sistemáticos. La Tabla 6 presenta los resultados.

Tabla 6. Comparación de métodos de descomposición de la serie de potencia

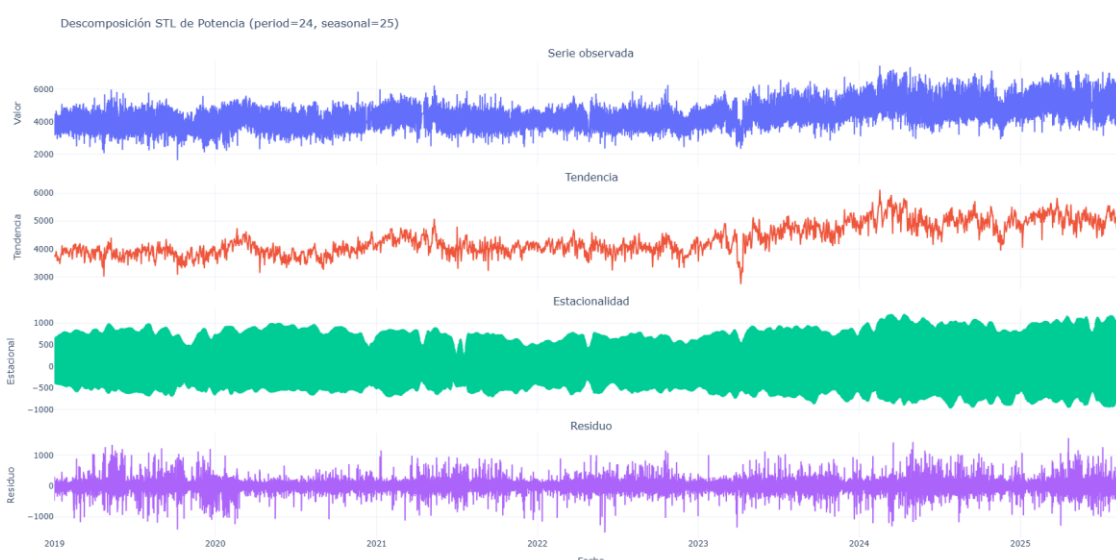
Método	Media residual	Desv. Est. residual	Reducción vs. clásica aditiva	Decisión
Clásica aditiva (period=24)	-0.037	240.97 VA	,	Rechazado
Clásica multiplicativa (period=24)	0.9998	0.0555*	n/a (escala diferente)	Rechazado
STL (period=24, seasonal=25, robust=False)	0.031	186.67 VA	-22.5%	SELECCIONADO

Fuente: Elaboración propia

*La desviación estándar del residuo multiplicativo está en unidades normalizadas (alrededor de 1.0), no en VA, por lo que no es directamente comparable con la descomposición aditiva.

La STL con los parámetros seleccionados reduce la varianza del residuo en un 22.5% respecto a la descomposición clásica aditiva, capturando mejor la dinámica de la serie. El parámetro seasonal=25 (impar, mayor que el período estacional de 24) es el suavizador de la componente estacional; valores más altos producen estacionalidades más suaves. La Figura 6 presenta las cuatro componentes de la descomposición STL óptima sobre la serie completa.

Figura 6. Descomposición STL de la serie de potencia (period=24, seasonal=25, robust=False), cuatro componentes.



Fuente: Elaboración propia

2.3.3. Análisis de estacionalidades múltiples

El análisis de estacionalidades se realizó en tres escalas temporales mediante tablas pivote y gráficos estacionales del componente estacional extraído por STL. La Tabla 7 sintetiza los hallazgos principales sobre la fuerza y estabilidad de cada componente estacional.

Tabla 7. Resumen del análisis de estacionalidades múltiples de la serie de potencia

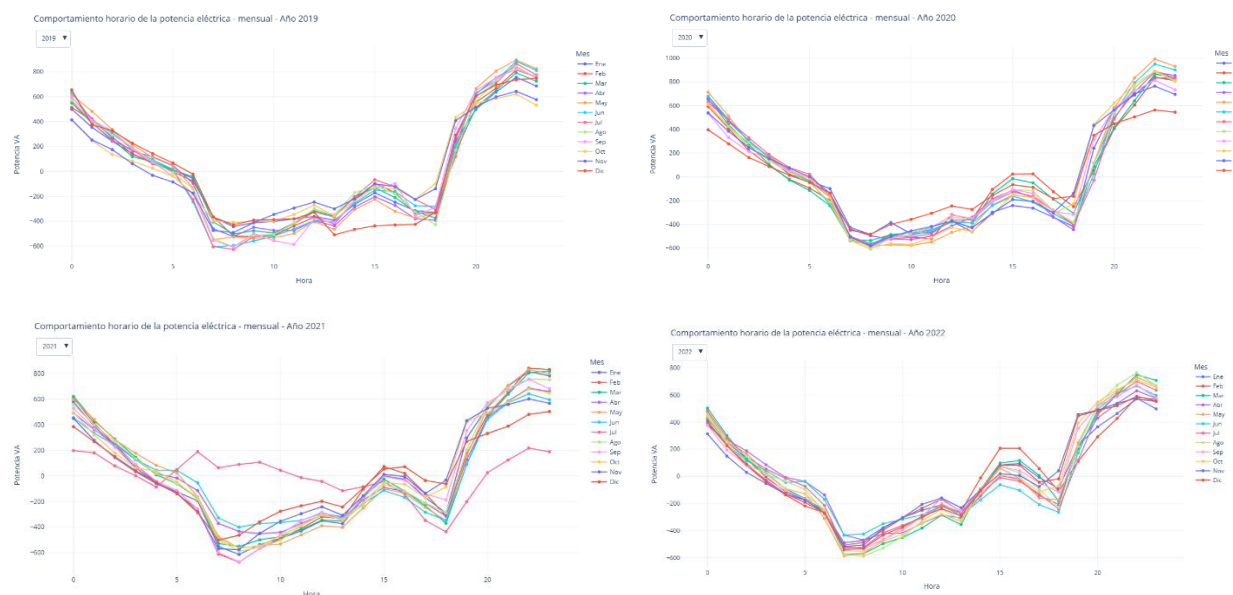
Escala temporal	Período (s)	Fuerza	Estabilidad entre años	Patrón dominante	Implicación para el modelado
Horaria (intradiaria)	24 horas	Muy fuerte	Alta, forma invariante	Valle 7-10h · rampa matutina · pico 19-23h	Obligatoria en todos los modelos (s=24)
Mensual	~365	Moderada	Alta,	Mayor:	Relacionada con

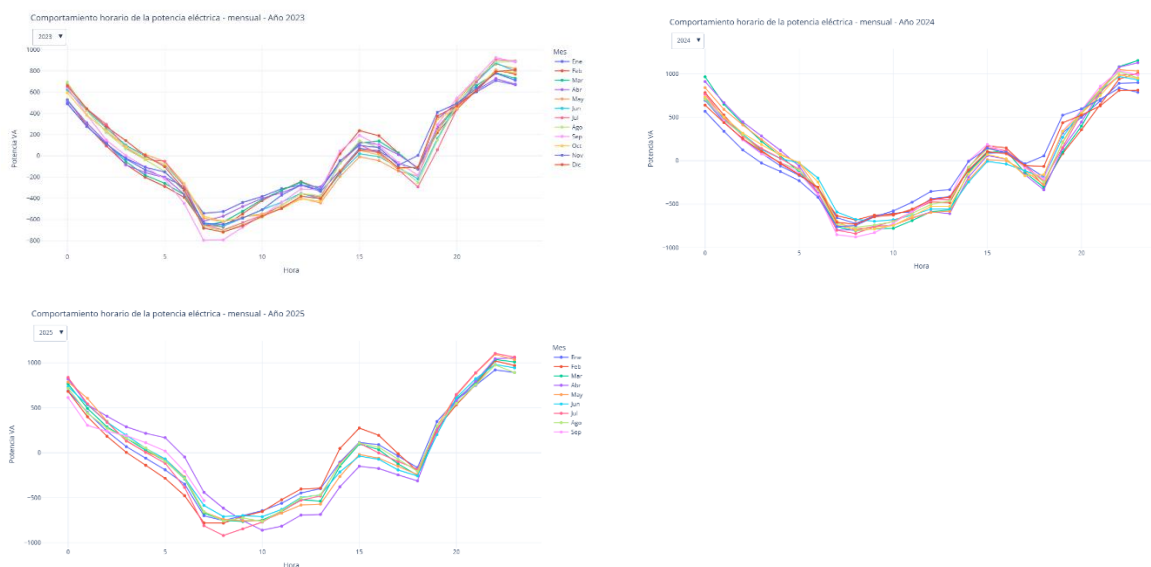
Escala temporal	Período (s)	Fuerza	Estabilidad entre años	Patrón dominante	Implicación para el modelado
(anual)	días		magnitud variable	Feb,Mar,Sep,Dic · Menor: Jun,Oct	calendario.
Semanal	~7 días	Prácticamente nula	,	Sin diferenciación día entre semana vs. fin de semana en forma	Patrón de magnitud variable; forma invariante

Fuente: Elaboración propia

El hallazgo más importante del análisis de estacionalidades es que la escala horaria ($s=24$) domina completamente la dinámica de la serie: el perfil intradiario de valle matutino y pico nocturno es altamente consistente entre todos los años y meses analizados. La aparente ausencia de estacionalidad semanal en la forma del perfil diario es una observación específica de esta subestación eléctrica: a diferencia de subestaciones residenciales o comerciales típicas, donde la demanda del fin de semana difiere significativamente de la de los días laborales, en esta subestación los perfiles de cada día de la semana son similares en forma. Sin embargo, la magnitud varía entre días, lo que implica que el día de la semana sí es un predictor relevante para las variables de magnitud, pero no para la forma del perfil. La Figura 7 muestra los perfiles horarios por mes.

Figura 7. Perfil de estacionalidad horaria del componente estacional STL promediado por mes, ciclo 2019-2025.

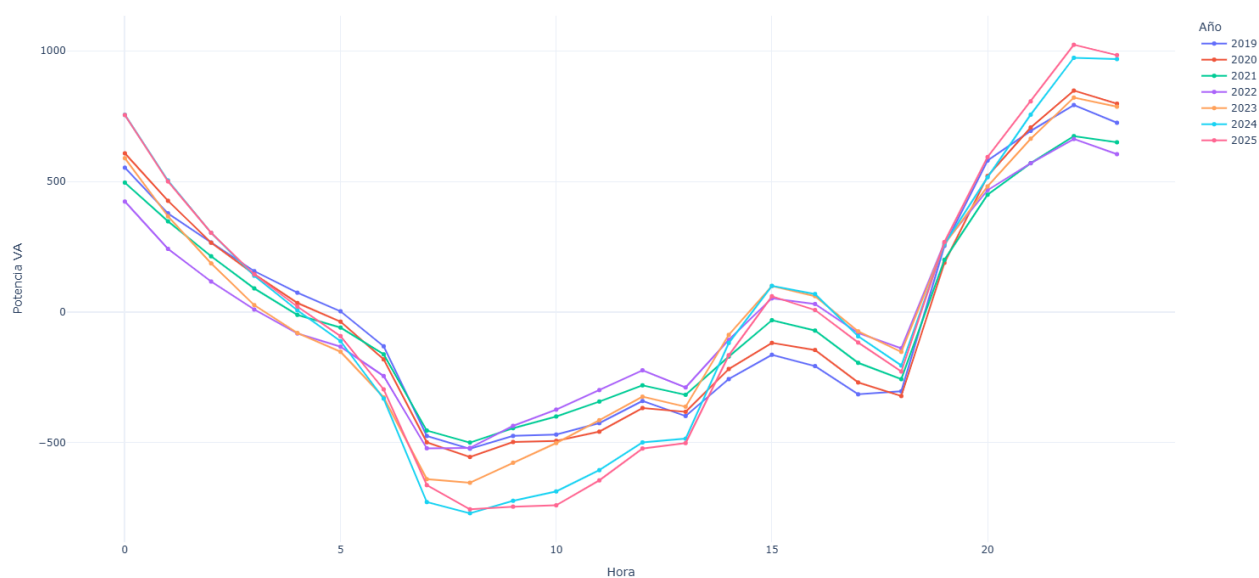


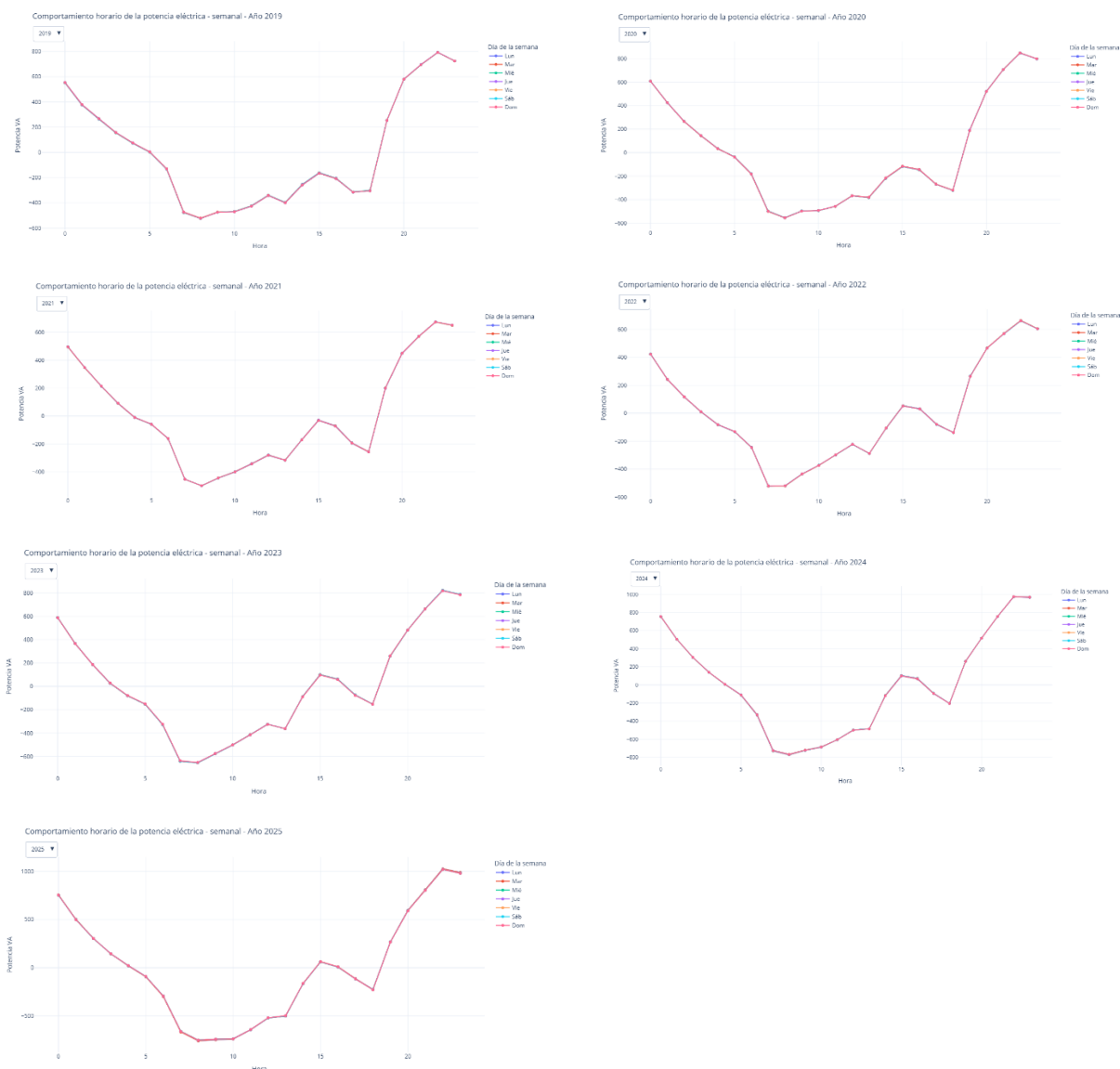


Fuente: Elaboración propia

Figura 8. Gráficos estacionales de la serie de potencia a diferentes escalas temporales.

Gráfico estacional potencia eléctrica horaria - Componente Estacional





Fuente: Elaboración propia

Como se evidencia en las figuras de estacionalidad semanal, los perfiles diarios de potencia presentan una superposición entre todos los días, lo que dificulta la identificación de variaciones en la curva sin aplicar un nivel de ampliación adecuado.

2.3.4. Función de autocorrelación y análisis de memoria temporal

El análisis de autocorrelación cuantifica la dependencia temporal de la serie, es decir, en qué medida el valor actual está correlacionado con los valores del pasado reciente y lejano. Para la serie de potencia limpia, se calcularon las autocorrelaciones con rezagos 1, 2, 24 y hasta 168 (una

semana completa). La Tabla 8 presenta los valores de autocorrelación de Pearson para los rezagos más informativos.

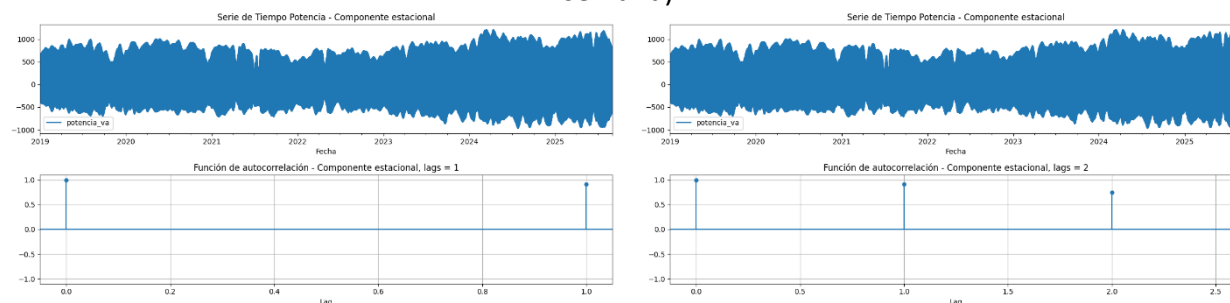
Tabla 8. Autocorrelaciones de la serie de potencia para rezagos clave

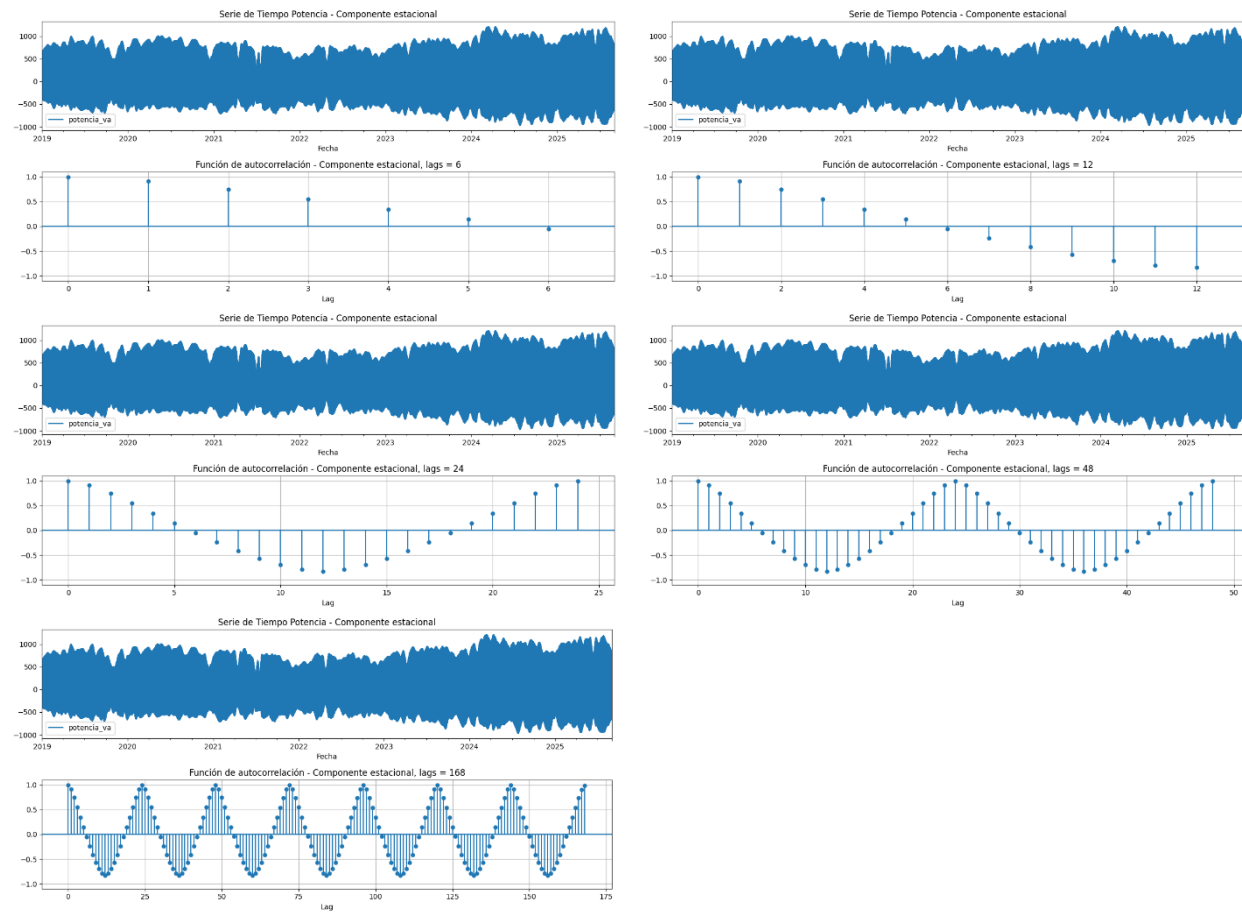
Rezago (lag)	Período físico	Autocorrelación (r)	Interpretación operativa
1	1 hora	0.950	Muy alta: el valor de ahora predice bien el valor en 1h
2	2 horas	0.866	Alta: memoria de 2 horas significativa
6	6 horas	0.683	Moderada: información de hace 6h aún relevante
12	12 horas	0.542	Moderada: información de hace 12h parcialmente útil
24	1 día	0.801	Muy alta: el mismo período de ayer es el mejor predictor individual
48	2 días	0.761	Alta: memoria de 2 días significativa
168	7 días (1 semana)	0.724	Alta: patrón semanal presente, aunque débil en comparación con el diario

Fuente: Elaboración propia

La autocorrelación en lag=24 (0.801) es la más alta fuera del lag=1, confirmando que el valor de la misma hora del día anterior es el predictor más poderoso individualmente, lo que justifica el uso de $s=24$ como período estacional en el Seasonal Naïve y en SARIMAX. La autocorrelación en lag=168 (0.724) confirma la existencia de memoria semanal, aunque la estructura semanal es más débil que la diaria. La ACF no decae hacia cero para ningún lag evaluado, lo que confirma la no estacionariedad de la serie con tendencia y estacionalidad fuerte. La Figura 9 muestra la función de autocorrelación completa hasta lag=168.

Figura 9. Función de autocorrelación (ACF) de la serie de potencia, lag=1 a lag=168 (una semana).





Fuente: Elaboración propia

2.3.5. Pruebas de estacionariedad ADF-KPSS

La estacionariedad es un requisito fundamental para los modelos ARIMA y un prerequisite para la correcta especificación del modelo estadístico. Se aplicó un diagnóstico completo con cuatro transformaciones y dos tipos de regresión (con constante 'c' y con constante más tendencia 'ct'), usando tanto la prueba ADF [35] (hipótesis nula: no estacionaria) como la prueba KPSS [34] (hipótesis nula: estacionaria). La combinación de ambas pruebas reduce los errores de Tipo I y Tipo II respecto a usar solo una de ellas. La Tabla 9 resume los resultados.

Tabla 9. Resultados del diagnóstico de estacionariedad ADF-KPSS para la serie de potencia y sus transformaciones

Transformación	Regresión	ADF stat	ADF p-valor	Diagnóstico ADF	KPSS stat	KPSS p-valor	Diag. KPSS	Conclusión final
Original	c	-7.59	2.48e	Estac.	34.64	< 0.01	No	MIXTO

Transformación	Regresión	ADF stat	ADF p-valor	Diagnóstico ADF	KPSS stat	KPSS p-valor	Diag. KPSS	Conclusión final
			-11				estac.	
Original	ct	-12.75	2.93e-20	Estac.	4.82	< 0.01	No estac.	MIXTO
Diff(1)	c	-43.49	≈ 0	Estac.	0.0051	> 0.10	Estac.	Estacionaria
Diff(24)	c	-40.11	≈ 0	Estac.	0.0029	> 0.10	Estac.	Estacionaria
Diff(1)+Diff(24)	c	-56.71	≈ 0	Estac.	0.0057	> 0.10	Estac.	Estacionaria

Fuente: Elaboración propia

El diagnóstico mixto de la serie original (ADF rechaza la raíz unitaria, pero KPSS rechaza la estacionariedad) indica presencia de tendencia determinística y estacionalidad fuerte, más que de raíz unitaria estocástica pura. Las tres transformaciones con diferencias producen series claramente estacionarias: Diff(24) es la más adecuada para SARIMAX porque elimina la estacionalidad diaria manteniendo la estructura de la serie; Diff(1)+Diff(24) es la más robusta para modelos que requieren estacionariedad estricta. Los modelos de machine learning y deep learning no requieren diferenciación previa porque no tienen supuestos de estacionariedad.

2.3.6. Análisis de correlación con variables climáticas

El análisis de correlación entre potencia_va y las variables climáticas se realizó en dos niveles: correlación de Pearson global (lineal, sobre toda la serie), y correlación de Spearman condicional a la hora del día (no lineal, por hora). La Tabla 10 muestra los resultados de la correlación global de Pearson.

Tabla 10. Correlación de Pearson global entre potencia_va y variables climáticas

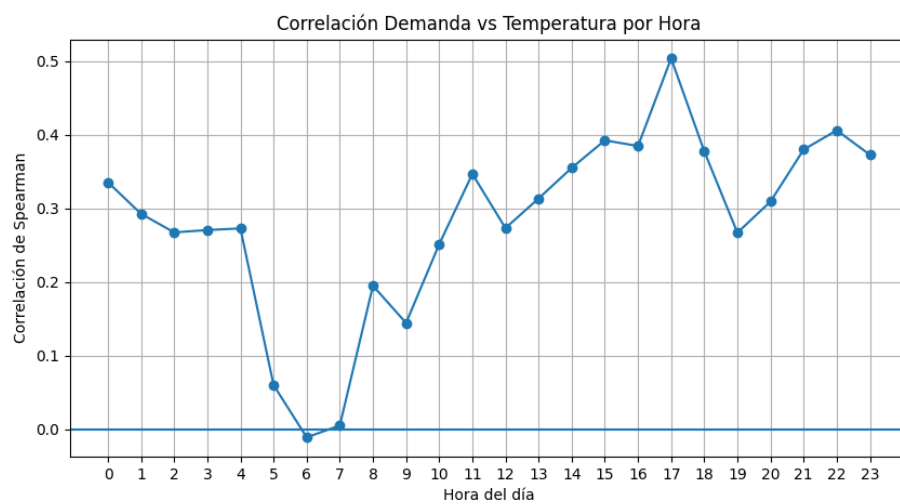
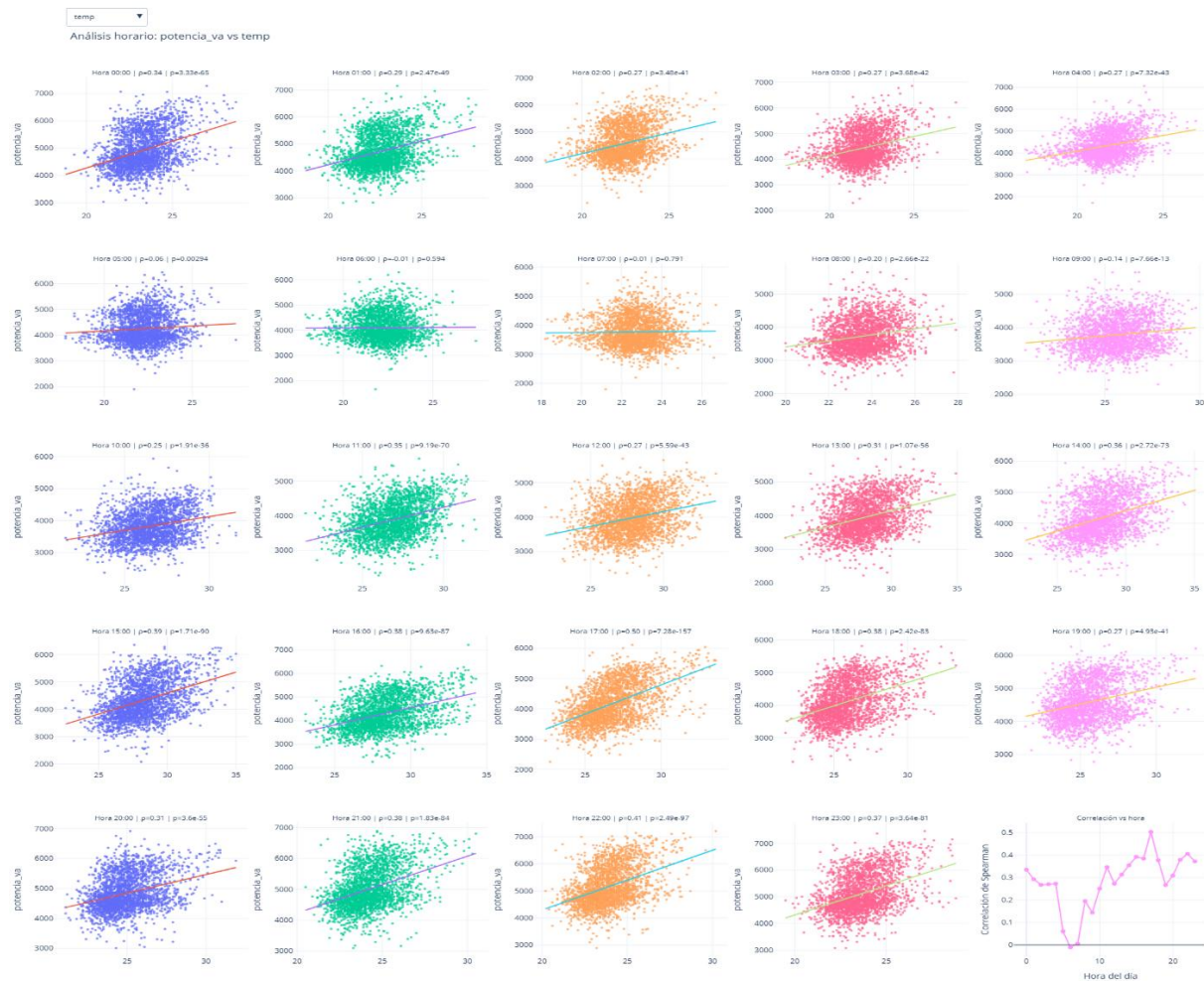
Variable climática	Correlación Pearson (r)	Interpretación	Inclusión como feature
temperatura (temp)	-0.013	Casi nula globalmente, relación no lineal y dependiente de hora	Sí (como temp_lag1)
sensación térmica (feelslike)	-0.007	Colineal con temp, información redundante	No (eliminar)
rocío (dew)	+0.011	Despreciable	No
humedad (humidity)	+0.021	Mínima global, relación no lineal relevante a nivel	Sí

Variable climática	Correlación Pearson (r)	Interpretación	Inclusión como feature
		horario	
precipitación (precip)	+0.001	Nula	No
prob. precipitación (precipprob)	-0.062	Muy débil	No
radiación solar (solarradiation)	-0.302	Moderada, refleja el ciclo día/noche, no causalidad directa	Sí (como solarradiation_lag1)
energía solar (solarenergy)	-0.302	Idéntica a solarradiation (colineal)	No (eliminar)
índice UV (uvindex)	-0.301	Idéntico a solarradiation (colineal)	No (eliminar)

Fuente: Elaboración propia

Las correlaciones globales de Pearson son engañosamente bajas para temperatura y humedad, lo que podría llevar a la conclusión errónea de que estas variables son irrelevantes para el pronóstico. La Figura 10 muestra la correlación de Spearman entre temperatura y potencia desagregada por hora del día, revelando la naturaleza dinámica y no lineal de la relación.

Figura 10. Correlación de Spearman entre temperatura y potencia por hora del día, relación dinámica y no lineal.

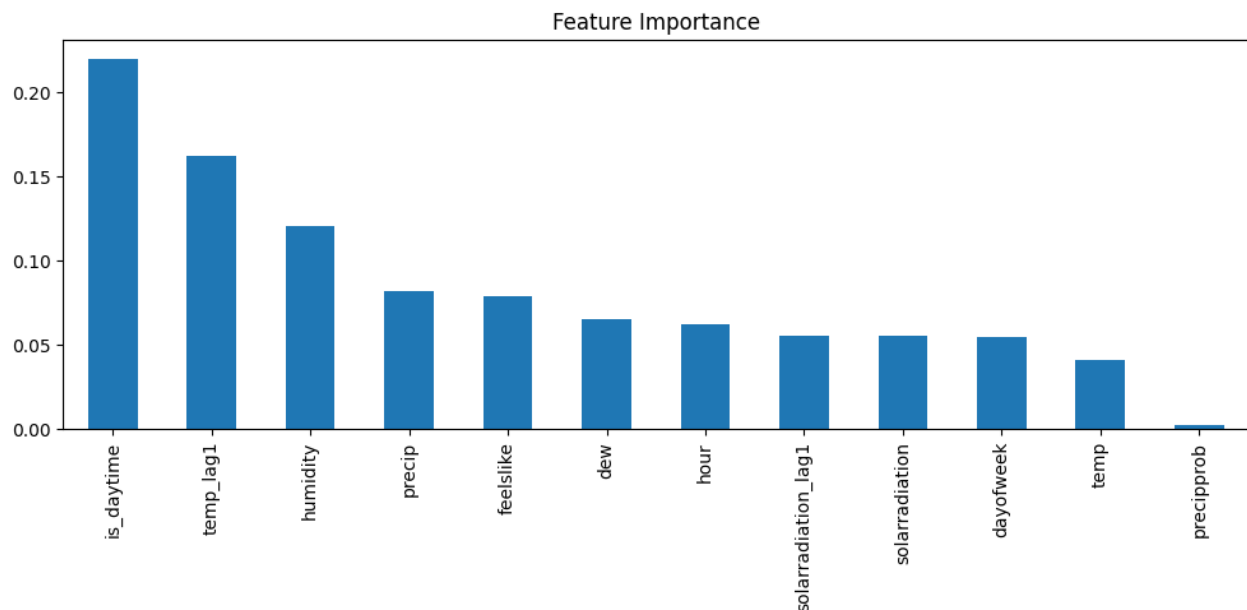


Fuente: Elaboración propia

La correlación de Spearman a las 17:00h ($p=0.504$) es cinco veces mayor que la correlación de Pearson global (-0.013), evidenciando que la relación temperatura-potencia es altamente dependiente del contexto temporal. Este hallazgo justifica el uso de temp_lag1 (temperatura rezagada 1 hora) como variable exógena en todos los modelos que la incorporan, en lugar de la temperatura contemporánea, para evitar fuga de datos en producción.

El análisis de Mutual Information y Random Forest Importance sobre el conjunto completo de variables climáticas y temporales derivadas confirmó que las variables de mayor capacidad predictiva son: (1) is_daytime (indicador binario de si es período diurno, $MI=0.233$, $RF=0.220$), que captura el ciclo fundamental día/noche; (2) temp_lag1 (temperatura rezagada 1 hora, $MI=0.133$, $RF=0.163$), que refleja la inercia térmica; y (3) humidity ($MI=0.054$, $RF=0.121$), con relación no lineal moderada. La Figura 11 presenta los resultados completos del análisis de importancia de variables.

Figura 11. Importancia de variables para la predicción de potencia_va, Mutual Information



Fuente: Elaboración propia

3. AJUSTE Y VALIDACIÓN DE MODELOS

Este capítulo aborda el Objetivo Específico 2, el ajuste y configuración de modelos de series de tiempo de diferente naturaleza estadística y computacional. Todos los modelos se implementaron en Python, utilizando las librerías statsforecast (Nixtla) para los modelos estadísticos, la implementación nativa de Prophet, las librerías neuralforecast (Nixtla) para los modelos de deep learning, el repositorio oficial de Chronos-Forecasting (Amazon), y XGBoost con scikit-learn para el modelo de machine learning tabular. El protocolo de evaluación estandarizado se describe en la Tabla 11.

Tabla 11. Protocolo de evaluación estandarizado aplicado a todos los modelos

Parámetro	Valor	Justificación
Horizonte de pronóstico (h)	24 horas	Horizonte operativo estándar del despacho diario
Período de entrenamiento	2019-01-01 a 2024-12-31 (50,616 obs.)	Maximiza datos históricos disponibles
Período de prueba	2025-01-01 a 2025-08-31 (5,832 obs.)	100% post-cambio estructural de 2023
Esquema TSCV	Expanding window, 90 ventanas	Simula el proceso real de reentrenamiento
Período TSCV	Últimas 2160h del train (≈ oct-2024 → dic-2024)	Comparación entre modelos con mismo protocolo dentro del conjunto de entrenamiento, <i>antes</i> de ver el test
Paso TSCV	24 horas	Avanza un horizonte por ventana
Métrica principal	MASE (Mean Absolute Scaled Error)	Independiente de escala, comparable entre modelos
Denominador MASE	MAE del Seasonal Naïve s=24 sobre train = 512.39 VA	Benchmark mínimo común a todos los modelos
Métricas secundarias	MAE (VA), RMSE (VA)	Interpretabilidad operativa
Frecuencia de la serie	Horaria (1h)	Resolución SCADA del sistema
Pronóstico	2025-09-01 -> 2025-09-08	Una semana posterior al Período de prueba

Fuente: Elaboración propia

Los modelos se clasificaron en cinco familias según su naturaleza metodológica:

- Estadísticos clásicos (SARIMAX).
- Machine learning (XGBoost, Prophet).

- Deep learning especializado (N-HITS, N-BEATS, TFT).
- Modelos fundacionales (Chronos)
- Híbridos de residual boosting.

Para cada modelo se describe su configuración específica, la justificación de los hiperparámetros seleccionados y sus características de desempeño preliminar antes de la evaluación del Capítulo 4.

3.1. Familia 1: Modelo estadístico: SARIMAX con AutoARIMA

SARIMAX fue implementado mediante la librería statsforecast de Nixtla, que proporciona una implementación en Cython de AutoARIMA entre 20 y 300 veces más rápida que las implementaciones equivalentes en R, con resultados idénticos en precisión [17]. El tiempo total de entrenamiento sobre las 50,616 observaciones fue de aproximadamente 640 minutos en CPU.

La especificación del modelo fue determinada automáticamente por AutoARIMA con `season_length=24` para capturar la estacionalidad diaria dominante identificada en el análisis exploratorio. Las variables exógenas incluidas como regresores son las siguientes, todas ellas disponibles en el momento del pronóstico o calculadas a partir de información conocida: temperatura rezagada 1 hora (`temp_lag1`, para evitar fuga de datos), humedad relativa (`humidity`), indicador de período diurno (`is_daytime`, binaria calculada a partir de la hora), radiación solar rezagada 1 hora (`solarradiation_lag1`), y variable dummy de cambio de régimen post-2023 (`post_2023`, valor 1 para fechas $\geq 2023-01-01$). La Tabla 12 presenta la configuración final del modelo.

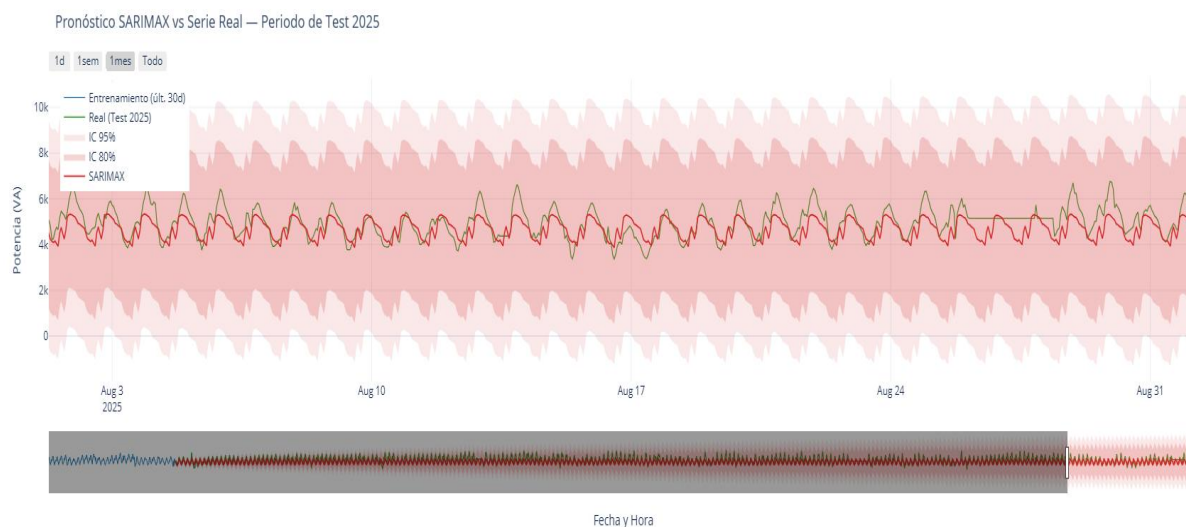
Tabla 12. Configuración del modelo SARIMAX con AutoARIMA

Parámetro	Valor seleccionado	Justificación
Biblioteca	statsforecast v1.7+ (Nixtla)	Implementación Cython de alta performance
season_length	24 horas	Estacionalidad diaria dominante (ACF lag=24 = 0.80)
Selección de órdenes	AutoARIMA con criterio AIC	Exploración automática del espacio (p,d,q,P,D,Q)
Variables exógenas	temp_lag1, humidity, is_daytime, solarradiation_lag1, post_2023	Selección por MI y correlación condicional
Diferenciación estacional	D=1 (Diff(24))	Confirma prueba KPSS: Diff(24) produce serie estacionaria
Tratamiento de outliers	No explícito (STL+IF aplicado)	Datos de entrada ya limpios

Parámetro	Valor seleccionado	Justificación
	previamente)	
Tiempo de entrenamiento	~640 min (CPU)	Ajuste sobre 50,616 observaciones horarias

Fuente: Elaboración propia

Figura 12. Pronóstico SARIMAX sobre las últimas 4 semanas del período de test 2025 vs. valores reales



Fuente: Elaboración propia

El ajuste de SARIMAX tardó tanto por la combinación de cinco factores que se multiplican entre sí:

1. Serie muy larga ($T \approx 58,400$ horas). Datos horarios desde enero 2019 hasta septiembre 2025. El filtro de Kalman recorre cada uno de esos ~58 mil puntos en cada iteración del optimizador, en cada modelo candidato.
2. Estacionalidad horaria ($\text{season_length}=24$). Obliga al modelo a manejar lags estacionales de hasta 48 períodos (con $P=2$, $Q=2$). El costo computacional crece cuadráticamente con el tamaño de los lags, haciendo cada modelo 3–5× más caro que uno sin estacionalidad.
3. Variables exógenas. Cada regresor exógeno añade parámetros a estimar por máxima verosimilitud y aumenta el tamaño de las matrices del filtro de Kalman, amplificando el costo de cada iteración.
4. Espacio de búsqueda amplio. Con $\text{max_p}=3$, $\text{max_q}=3$, $\text{max_P}=2$, $\text{max_Q}=2$ y $\text{nmodels}=100$, el algoritmo stepwise evalúa entre 40 y 60 modelos candidatos, y cada uno requiere 80–200 iteraciones del optimizador MLE para converger.
5. Paralelismo inefectivo. $\text{n_jobs}=4$ paraleliza entre series, no dentro de una sola serie. Como solo hay un `unique_id`, los 4 workers no se aprovechan y el ajuste es esencialmente secuencial.

3.2. Familia 2: Machine Learning: XGBoost y Prophet

3.2.1. XGBoost con ingeniería de características temporal

XGBoost se implementó con scikit-learn y la librería xgboost. La ingeniería de características fue el paso más crítico de la implementación, dado que XGBoost no tiene mecanismos nativos para capturar la dependencia temporal. Las características construidas se organizan en cuatro grupos que se detallan en la Tabla 13.

Tabla 13. Ingeniería de características para XGBoost, descripción y justificación

Grupo	Característica	Descripción	Justificación
Rezagos temporales	lag_1h	Potencia hace 1 hora	Autocorrelación lag=1: r=0.950
Rezagos temporales	lag_24h	Potencia hace 24 horas (mismo momento ayer)	Autocorrelación lag=24: r=0.801
Rezagos temporales	lag_168h	Potencia hace 168 horas (misma hora semana anterior)	Autocorrelación lag=168: r=0.724
Ventanas móviles	rolling_mean_24h	Media de las últimas 24 horas	Tendencia local diaria
Ventanas móviles	rolling_std_24h	Desv. estándar de las últimas 24h	Volatilidad local
Ventanas móviles	rolling_mean_168h	Media de la última semana	Tendencia semanal
Calendario cíclico	hour_sin, hour_cos	$\sin(2\pi \cdot h/24)$, $\cos(2\pi \cdot h/24)$	Evita discontinuidad hora 23 $\rightarrow$ hora 0
Calendario cíclico	month_sin, month_cos	$\sin(2\pi \cdot m/12)$, $\cos(2\pi \cdot m/12)$	Estacionalidad anual continua
Calendario categórico	is_weekend	1 si sábado o domingo	Diferenciación leve de fin de semana
Calendario categórico	is_holiday	1 si festivo colombiano	Efectos de festivos (librería holidays)
Variables climáticas	temp_lag1, humidity	Temperatura y humedad rezagadas 1h	Variables exógenas disponibles en producción
Cambio estructural	post_2023	Variable dummy: 1 si fecha $\geq$ 2023-01-01	Captura el cambio de nivel

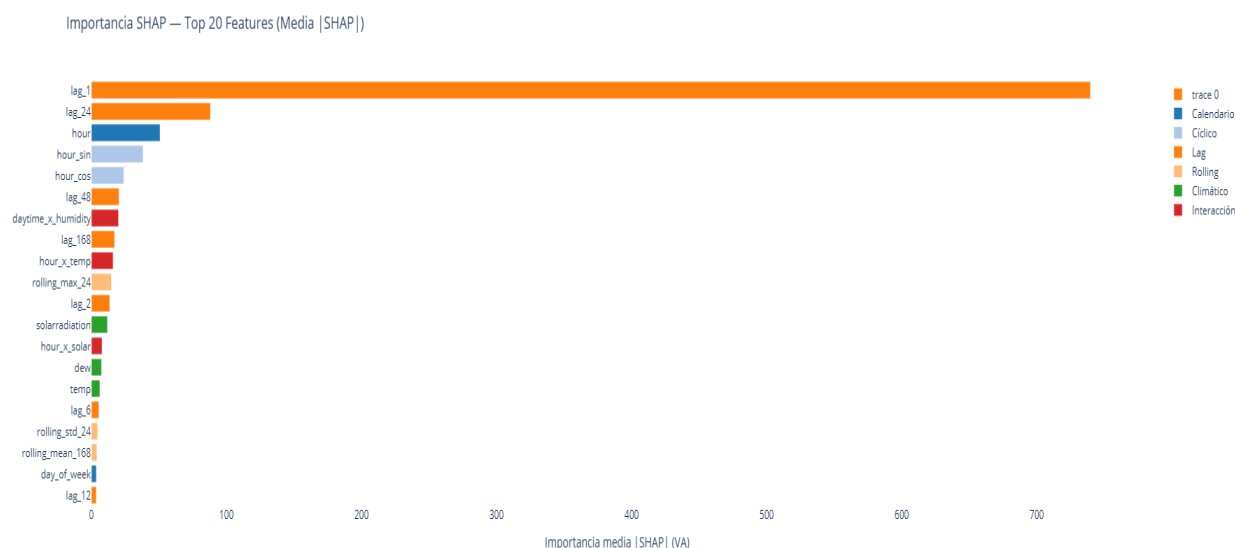
Fuente: Elaboración propia

La selección de hiperparámetros se realizó mediante RandomizedSearchCV con TimeSeriesSplit de 5 pliegues sobre el conjunto de entrenamiento. Los mejores hiperparámetros encontrados fueron:

- n_estimators=800,
- max_depth=7,
- learning_rate=0.05,
- subsample=0.85,
- colsample_bytree=0.8,
- min_child_weight=5,
- gamma=0.1.

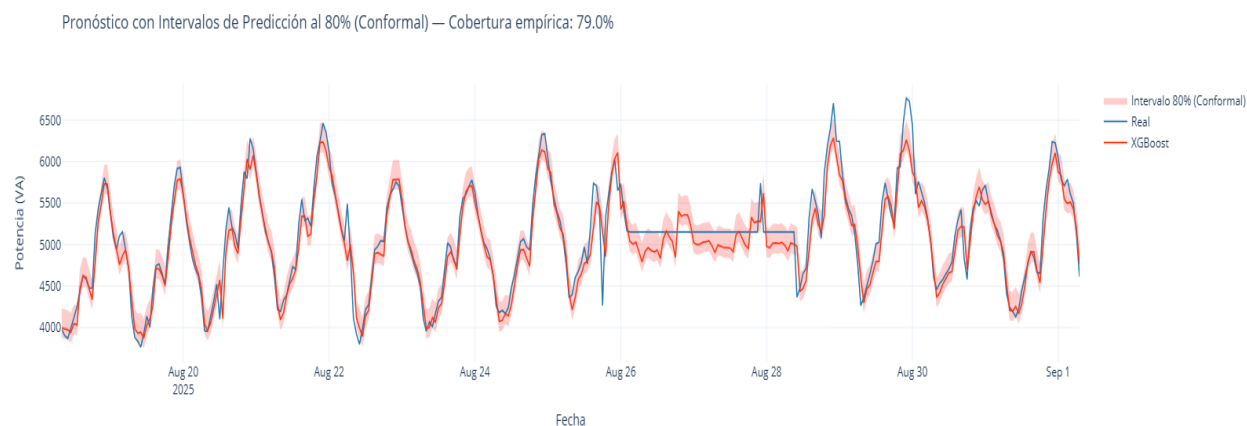
La Figura 13 muestra la importancia SHAP de las variables del modelo XGBoost y la Figura 14 la muestra el pronóstico XGBoost sobre las últimas 2 semanas del período de test 2025 vs. valores reales.

Figura 13. Importancia SHAP de las características del modelo XGBoost, contribución marginal al pronóstico



Fuente: Elaboración propia

Figura 14. Pronóstico XGBoost sobre las últimas 2 semanas del período de test 2025 vs. valores reales



Fuente: Elaboración propia

3.2.2. Prophet con configuración extendida para el contexto colombiano

Prophet se implementó en dos variantes para evaluar el impacto de la configuración: Prophet Base (configuración por defecto con parámetros estándar) y Prophet Full (configuración extendida optimizada para el contexto colombiano). La diferencia de desempeño entre ambas variantes (MASE=2.4438 vs. MASE=1.0882) evidencia la magnitud del impacto del ajuste de hiperparámetros. La Tabla 14 presenta la configuración de ambas variantes.

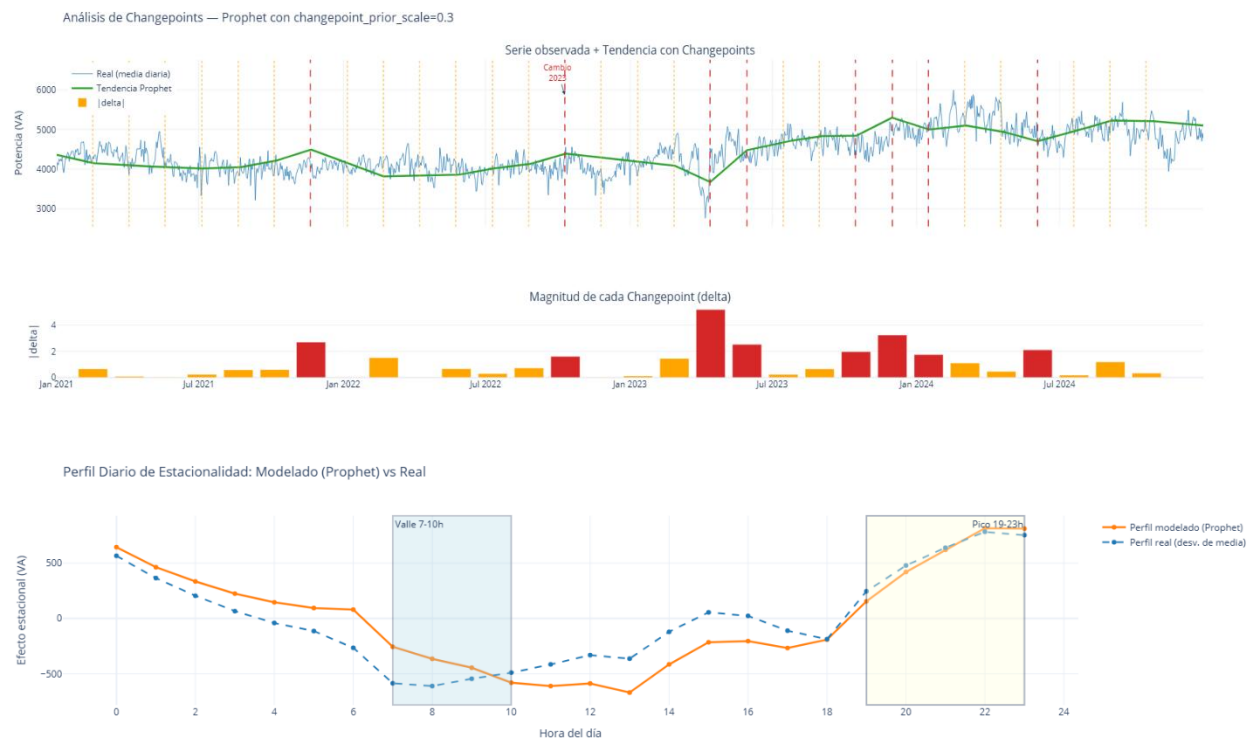
Tabla 14. Comparación de configuraciones de Prophet (Base vs. Full)

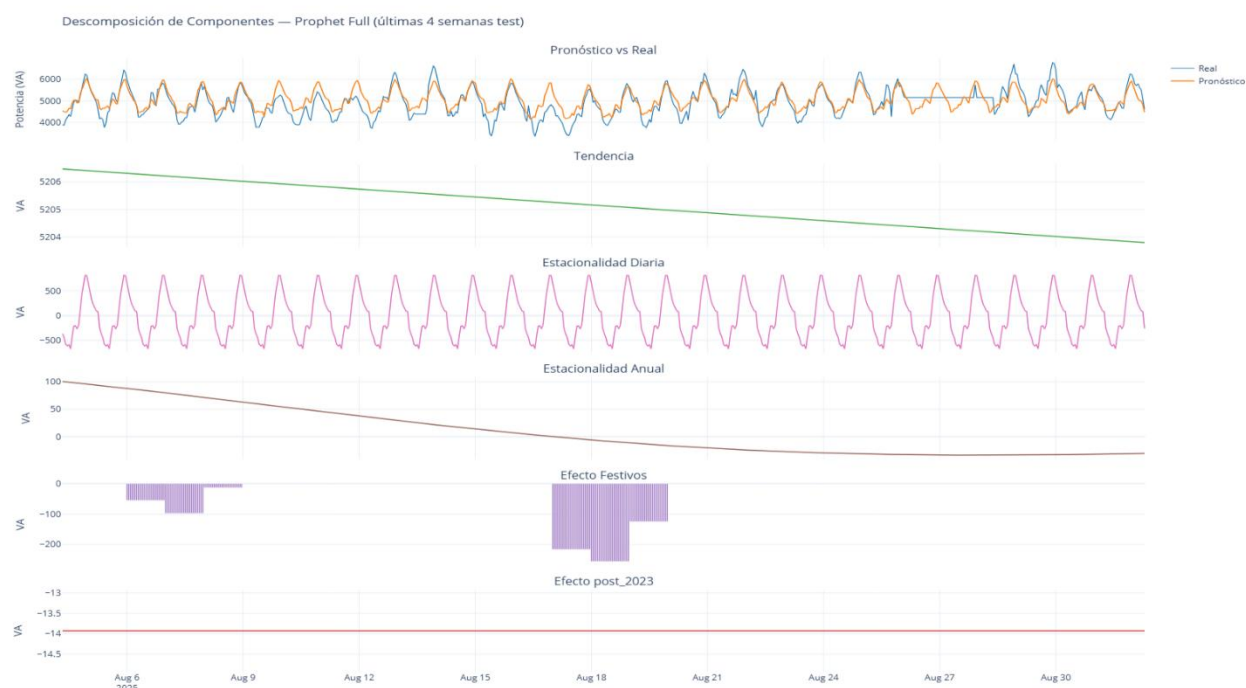
Parámetro	Prophet Base	Prophet Full	Impacto
changepoint_prior_scale	0.05 (default)	0.30	Mayor flexibilidad de tendencia para capturar cambio 2023
Estacionalidad diaria	Automática (fourier_order=4)	Manual (fourier_order=10)	Mayor capacidad para capturar perfil diario complejo
Estacionalidad semanal	Automática (fourier_order=3)	Manual (fourier_order=3)	Sin cambio
Estacionalidad anual	Automática (fourier_order=10)	Manual (fourier_order=5)	Más suave, evita sobreajuste
Festivos colombianos	No	Sí (librería holidays, país='CO')	Captura Semana Santa, festivos nacionales
Regresor temperatura	No	Sí (temp_lag1	Mejora correlación

Parámetro	Prophet Base	Prophet Full	Impacto
		normalizada)	dinámica tarde
uncertainty_samples	1000	0 (deshabilitado)	MAP en lugar de MCMC, más rápido
Resultado (MASE)	2.4438 (peor que Naïve)	1.0882 (supera a SARIMAX)	Diferencia de 1.36 puntos MASE

Fuente: Elaboración propia

Figura 15. Descomposición de componentes del modelo Prophet Full, changepoints, tendencia, estacionalidades y efecto de temperatura





Fuente: Elaboración propia

3.3. Familia 3, Deep Learning especializado

3.3.1. N-HiTS: interpolación jerárquica neuronal

N-HiTS fue implementado mediante la librería NeuralForecast de Nixtla. Se evaluaron dos variantes: N-HiTS base (configuración mínima para máxima eficiencia computacional) y N-HiTS completo (con todas las características y TSCV completo). La configuración de la arquitectura de submuestreo jerárquico se especificó con `n_freq_downsample=[24, 12, 1]`, que significa que el primer bloque de la pila opera sobre la señal submuestreada por factor 24 (captura tendencia diaria), el segundo bloque por factor 12 (captura patrones de media jornada), y el tercer bloque sin submuestreo (captura variaciones hora a hora). El `INPUT_SIZE` se configuró en 168 horas (7 días de contexto), proveyendo al modelo de un ciclo semanal completo de historia para cada predicción. La Tabla 15 presenta las configuraciones de ambas variantes.

Tabla 15. Configuración de N-HiTS, variantes base y completa

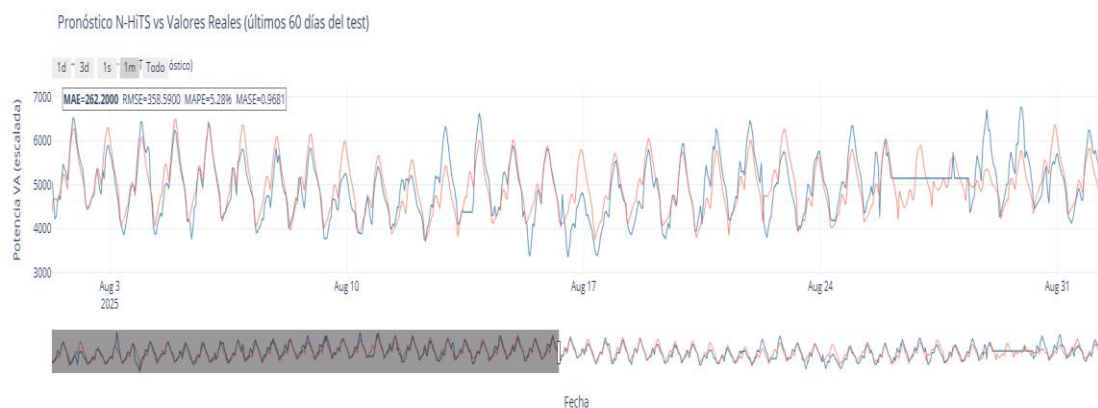
Parámetro	N-HiTS Base	N-HiTS Completo	Justificación
<code>n_freq_downsample</code>	[24, 12, 1]	[24, 12, 1]	Captura patrones a 3 resoluciones para serie

Parámetro	N-HiTS Base	N-HiTS Completo	Justificación
			horaria
INPUT_SIZE	168h (7 días)	168h (7 días)	Provee 1 semana de contexto histórico
max_steps	500	800	Épocas de entrenamiento
val_size	720h (30 días)	720h (30 días)	Período de early stopping
early_stopping_patience	30 pasos	Deshabilitado en TSCV	TSCV no compatible con early stopping
Variables exógenas hist.	temp, humidity, is_daytime	temp, humidity, is_daytime	Variables disponibles en t-h
Variables exógenas fut.	is_daytime, is_holiday	is_daytime, is_holiday	Variables conocidas para t+h
Tiempo de entrenamiento	~19 segundos	~2.9 minutos	N-HiTS base: candidato ideal para CPU
MASE (test 2025)	0.9575	0.9681	Base supera a la variante completa

Fuente: Elaboración propia

Un resultado a destacar de la evaluación de N-HiTS es que la variante base (MASE=0.9575) supera a la variante completa con TSCV (MASE=0.9681) con un tiempo de entrenamiento 9 veces menor (19 segundos vs. 2.9 minutos). Esto sugiere que la configuración simplificada generaliza mejor al período de test, posiblemente porque la variante completa con más épocas de entrenamiento sobreajusta ligeramente el período de validación interna.

Figura 16. Pronóstico N-HiTS sobre el período de test 2025, últimas 4 semanas



Fuente: Elaboración propia

3.3.2. N-BEATS en tres variantes

N-BEATS fue implementado en tres configuraciones: Interpretable (con bases de Fourier para estacionalidad y polinomios para tendencia), Generic (bases completamente aprendibles sin restricción) y Ensemble de 5 semillas aleatorias. Las tres variantes se entrenaron con INPUT_SIZE=168 horas y max_steps=500. El Ensemble promedia las predicciones de 5 instancias del modelo Generic entrenadas con diferentes semillas (random_state = 0, 1, 2, 3, 4), reduciendo la varianza del pronóstico a costa de multiplicar el tiempo de entrenamiento por 5.

Las métricas de test 2025 para las tres variantes fueron: N-BEATS Interpretable MASE=1.1256, N-BEATS Generic MASE=1.0928, y N-BEATS Ensemble MASE=1.0222. Ninguna variante logró MASE < 1.0, aunque el Ensemble se aproxima al umbral del Naïve. La superioridad del Ensemble sobre los modelos individuales confirma el principio de diversificación de la varianza.

Figura 17. Pronóstico N-BEATS Interpretable / Ensemble sobre el período de test 2025, primeras 2 semanas



Fuente: Elaboración propia

3.3.3. Temporal Fusion Transformer (TFT), versiones CPU y GPU

TFT fue implementado en dos configuraciones que difieren significativamente en capacidad arquitectónica. TFT-CPU se configuró con hidden_size=16 y max_steps=200 para ser ejecutable en procesador sin GPU en un tiempo operativo razonable (3.2 minutos). TFT-GPU aprovechó la capacidad de la GPU T4 de Kaggle con hidden_size adaptativo (hasta 128 según la VRAM disponible) y max_steps=800. La Tabla 16 compara las dos configuraciones.

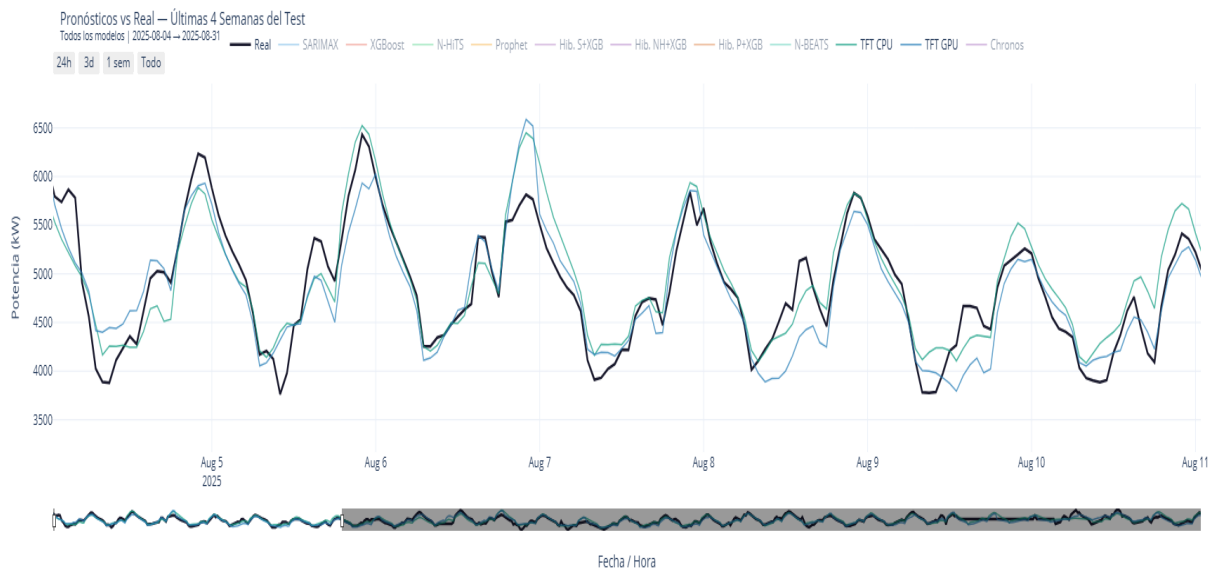
Tabla 16. Configuración de TFT, variantes CPU (arquitectura mínima) y GPU (arquitectura completa)

Parámetro	TFT-CPU	TFT-GPU	Impacto en capacidad
hidden_size	16	32-128 (adaptativo)	Capacidad de representación interna del Transformer
n_head	2	4-8 (adaptativo)	Número de cabezas del mecanismo de atención
max_steps	200	800	Épocas de entrenamiento
INPUT_SIZE	48h	168h	Ventana de contexto histórico
dropout	0.1	0.1	Regularización
Variables exógenas	temp_lag1, humidity, is_daytime	temp_lag1, humidity, is_daytime, solarradiation_lag1	Más contexto meteorológico en GPU
Tiempo de entrenamiento	3.2 minutos (CPU)	11.1 minutos (GPU T4)	4x más rápido en CPU, pero menor capacidad
MASE (test 2025)	0.9811 (TFT CPU min) / 0.9545 (con Prob)	0.9179	TFT-GPU supera a TFT-CPU en 0.0363 puntos
Sesgo sistemático	+30.03 VA	+3.95 VA	TFT-GPU tiene menor sesgo: mejor captura del nivel post-2023

Fuente: Elaboración propia

La diferencia más relevante entre TFT-CPU y TFT-GPU no es solo la diferencia en MASE (0.0363 puntos), sino el sesgo sistemático: TFT-GPU sobreestima la potencia en solo +3.95 VA/hora en promedio, mientras que TFT-CPU lo hace en +30.03 VA/hora. Un sesgo de +3.95 VA es operativamente despreciable, mientras que +30.03 VA puede acumularse significativamente a lo largo del día. La Figura 18 muestra la comparación de pronósticos TFT-CPU vs. TFT-GPU sobre una semana del test.

Figura 18. Comparación de pronósticos TFT-CPU vs. TFT-GPU sobre una semana del test 2025



Fuente: Elaboración propia

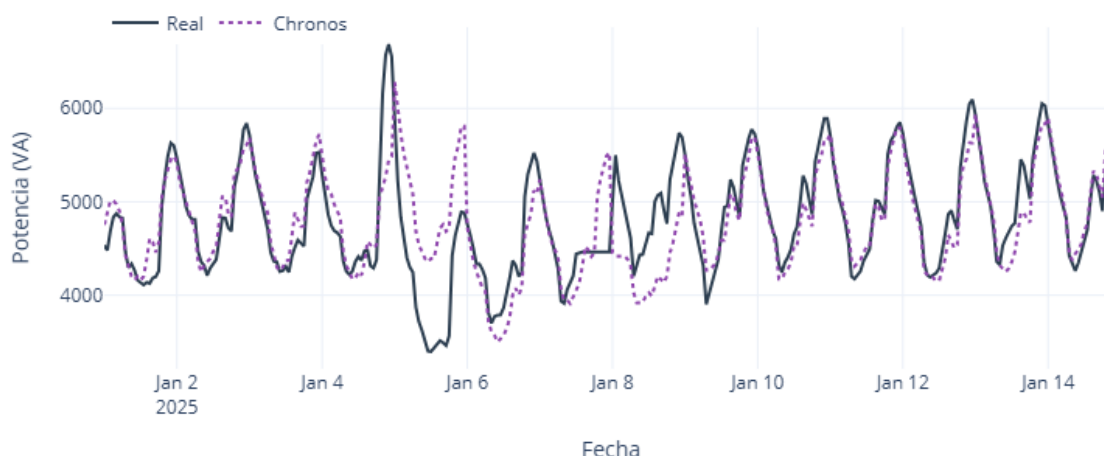
3.4. Familia 4, Modelo fundacional pre-entrenado: Chronos

Chronos T5-Large (710 millones de parámetros) fue implementado mediante el repositorio oficial chronos-forecasting de Amazon, ejecutado en la GPU T4 de Kaggle. A diferencia de todos los demás modelos, Chronos opera en modo zero-shot: no se realizó ningún ajuste fino (fine-tuning) sobre los datos de la subestación, lo que significa que el modelo no vio los datos de potencia durante el proceso de ajuste.

Se evaluaron dos variantes de contexto: Chronos estándar (contexto de 512 horas previas al punto de predicción, correspondiente a aproximadamente 21 días) y Chronos con contexto post-2023 (datos de contexto restringidos al período posterior al cambio estructural, es decir, solo datos desde enero 2023). La variante con contexto post-2023 logró $MASE=0.8645$ vs. $MASE=0.8723$ para el contexto estándar, evidenciando que restringir el contexto al nuevo régimen mejora marginalmente la precisión porque el modelo no recibe señales confusas del régimen anterior de menor demanda.

Figura 19. Chronos sobre el período de test 2025 – primeras 2 semanas del test

Chronos vs Real — primeras 2 semanas del test



Fuente: Elaboración propia

3.5. Familia 5, Modelos híbridos de residual boosting

Los tres modelos híbridos implementados siguen la misma arquitectura de residual boosting: un modelo base genera el pronóstico inicial, se calculan sus residuales sobre el conjunto de entrenamiento mediante validación cruzada interna, y XGBoost se entrena sobre esos residuales para aprender a corregirlos. La predicción final es la suma del pronóstico base más la corrección de XGBoost. La Tabla 17 describe los tres modelos híbridos evaluados.

Tabla 17. Descripción de los modelos híbridos de residual boosting evaluados

Modelo híbrido	Base	Corrector	Residuales calculados sobre	MASE test 2025	vs. modelo base
Híbrido SARIMAX + XGBoost	SARIMAX	XGBoost	TSCV interno de SARIMAX (90 ventanas)	1.4188	Mejora +24% (base=1.8598)
Híbrido N-HiTS + XGBoost	N-HiTS	XGBoost	Cross-validation interna de N-HiTS (refit=False)	0.9700	Empeora -0.2% (base=0.9681)
Híbrido Prophet + XGBoost	Prophet Full	XGBoost	In-sample de Prophet desde 2023-01-01	1.1352	Mejora trivial (base=1.1364)

Fuente: Elaboración propia

El análisis de los modelos híbridos muestra que el único híbrido con mejora apreciable respecto a su modelo base es el que tiene el peor modelo base (SARIMAX), con una reducción del 24% en MASE. La reducción del 24% en MASE del Híbrido SARIMAX + XGBoost respecto a SARIMAX solo no es suficiente para superar el umbral del Naïve ($MASE=1.4188 > 1.0$).

El Híbrido N-HiTS + XGBoost no solo no mejora a N-HiTS, sino que lo empeora marginalmente, lo que sugiere que los residuales de N-HiTS son prácticamente ruido blanco sin estructura explotable por XGBoost.

El Híbrido Prophet+XGBoost aporta una mejora estadísticamente trivial (0.0012 puntos MASE).

Estos resultados tienen una implicación metodológica importante: la estrategia de residual boosting solo aporta valor cuando el modelo base deja estructura no lineal explotable en sus residuales, lo que requiere verificación previa mediante análisis ACF de los residuales del modelo base.

3.6. Ensemble ponderado 1/MAE

El ensemble ponderado combina los pronósticos de los 11 modelos base mediante pesos inversamente proporcionales a su MAE en el período de evaluación. En la implementación ejecutada, los pesos se calculan a partir del MAE almacenado en la matriz de métricas (rama directa o recalculado sobre serie_test_eval), lo que para 10 de los 11 modelos corresponde al período de test 2025 y para el Híbrido NH+XGB al promedio de TSCV. ***Esta dependencia del MAE de test introduce una fuga de datos por construcción: el ensemble conoce el desempeño real del período que pretende predecir, por lo que su $MASE=0.4359$ (ver Tabla 19) debe interpretarse como una referencia de máximo teórico bajo conocimiento perfecto del MAE de test, no como desempeño esperado en operación.*** Una formulación libre de fuga (pesos derivados íntegramente de TSCV o de un período de validación reservado) se identifica como trabajo futuro. El peso w_i de cada modelo i se calcula como:

$$w_i = (1/MAE_i) / \sum_j (1/MAE_j)$$

Garantizando que los pesos sumen 1.0. La predicción del ensemble para cada hora t es:

$$\hat{y}_{ensemble}(t) = \sum_i w_i \cdot \hat{y}_i(t)$$

La Tabla 18 muestra los pesos asignados a cada modelo.

Tabla 18. Pesos del ensemble ponderado 1/MAE, distribución de contribución entre modelos

Modelo	MAE para cálculo de pesos (VA)	Peso 1/MAE	Participación (%)
Chronos (ctx post-2023)	224.23	0.1075	10.75%
Chronos	226.25	0.1061	10.61%
TFT GPU T4	238.08	0.1009	10.09%
TFT CPU min Prob	247.57	0.0970	9.70%
N-HiTS (base)	259.35	0.0928	9.28%
N-HiTS	262.20	0.0916	9.16%
TFT CPU min	254.46	0.0944	9.44%
N-BEATS Ensemble	265.12	0.0906	9.06%
Híbrido NH+XGB	262.48	0.0916	9.16%
SARIMAX	482.37	0.0500	5.00%
XGBoost	326.11	0.0739	7.39%
TOTAL	,	1.0000	100%

Fuente: Elaboración propia

Nota 1: para 10 de los 11 modelos del ensemble, el MAE corresponde al período de test 2025; para el Híbrido NH+XGB corresponde al promedio de validación cruzada con ventana deslizante (TSCV) por la ruta de cálculo del código fuente. Esta heterogeneidad introduce una fuente adicional de inconsistencia metodológica que se documenta para transparencia y no se corrige para preservar trazabilidad con la implementación ejecutada.

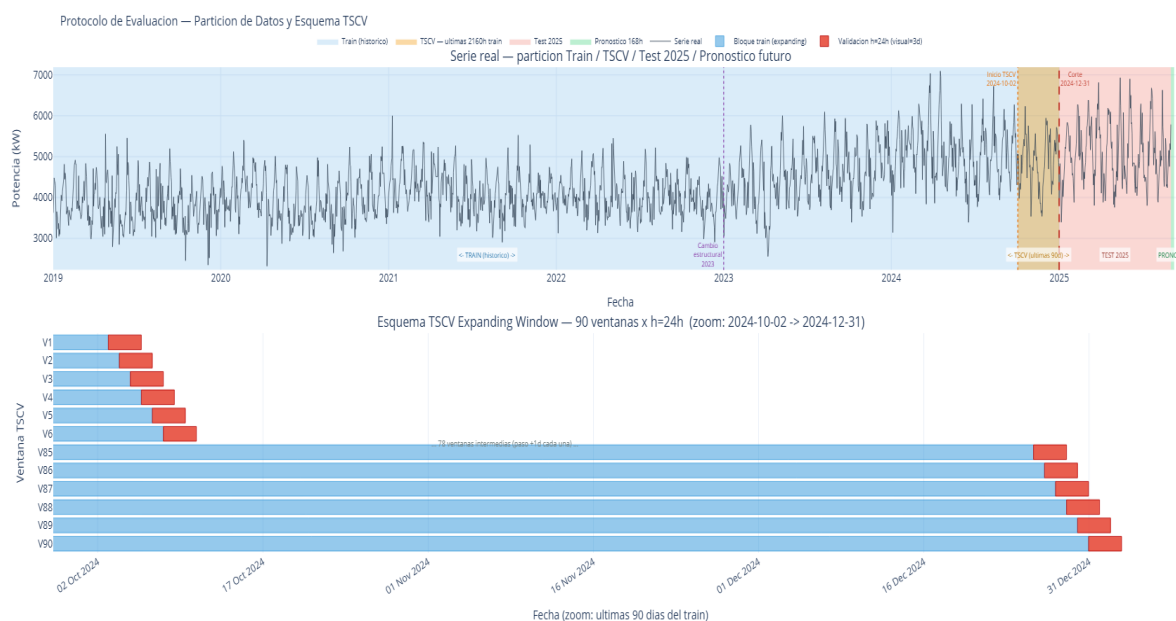
Nota 2: el portafolio total de modelos son 17 configuraciones individuales; el ensemble se construye sobre 11 de ellas (las que pasaron filtro de calidad)

La distribución de pesos en la Tabla 18 muestra que los tres modelos con mayor participación son Chronos post-2023 (10.75%), Chronos estándar (10.61%) y TFT-GPU (10.09%), los que mejor se adaptaron al régimen post-2023. SARIMAX recibe el menor peso (5.00%) por su mayor MAE, pero sigue contribuyendo al ensemble dado que su comportamiento en ciertos períodos (como los primeros meses del año donde el nivel de demanda es más estable) puede complementar a los modelos de deep learning. El ensemble alcanzó MAE=223.36 VA y MASE=0.4359 sobre el período de test 2025. Como se advirtió al inicio de esta sección, este valor refleja el máximo teórico alcanzable cuando los pesos se ajustan con el MAE del propio período de prueba (fuga de datos por construcción) y por tanto no se compara directamente con los modelos del ranking principal en 4.1.

4. EVALUACIÓN DEL DESEMPEÑO Y COMPARACIÓN DE MODELOS

Este capítulo desarrolla el Objetivo Específico 3: la evaluación comparativa del desempeño de los modelos mediante el protocolo estandarizado descrito en la Tabla 11. Adicionalmente, en la Figura 20 también se presentan los diferentes periodos de train, test, pronóstico y TSCV. La evaluación se presenta en seis dimensiones complementarias: (1) ranking global por MASE en el período de test 2025; (2) clasificación por cuadrantes eficiencia-precisión; (3) análisis de sesgo sistemático post-cambio estructural; (4) curvas de degradación del error por paso de horizonte; (5) análisis temporal del error por hora del día y por mes; y (6) comparación de consistencia entre TSCV y test para detección de sobreajuste.

Figura 20. Protocolo de Evaluación, Partición de Datos y Esquema TSCV



Fuente: Elaboración propia

4.1. Ranking global de desempeño (período de test 2025)

La Tabla 19 presenta el ranking completo de las 17 configuraciones de modelos individuales evaluadas, junto con el Ensemble ponderado y el Seasonal Naïve de referencia (19 filas en total), ordenadas por MASE ascendente. Los modelos con $MASE < 1.0$ (resaltados en verde en el ranking) superan el benchmark del Seasonal Naïve y son candidatos viables para producción. Los modelos con $MASE > 1.0$ no justifican su complejidad computacional frente al simple pronóstico ingenuo. Se observa que solo 7 modelos individuales superan el benchmark Seasonal Naïve, evidenciando la alta dificultad del período de prueba post-cambio estructural. El mejor desempeño individual lo obtiene Chronos (ctx post-2023) con $MASE=0.8645$, una mejora del 13.55% sobre el Seasonal Naïve.

Tabla 19. Ranking completo de desempeño en el período de test 2025, ordenado por MASE ascendente (menor MASE = mejor)

Rank	Modelo	MAE (VA)	RMSE (VA)	MASE	Mejora vs Naïve	T (min)	Familia	Super a Naïve
	Ensemble ponderado	223.36	,	0.4359	56.3%	,	Ensemble	Sí
1	Chronos (ctx post-2023)	224.23	331.65	0.8645	56.2%	16.9	Fundacional	Sí
2	Chronos	226.25	332.41	0.8723	55.8%	16.9	Fundacional	Sí
3	TFT GPU T4	238.08	333.72	0.9179	53.5%	11.1	Deep Learning	Sí
4	TFT CPU min Prob	247.57	341.16	0.9545	51.7%	3.2	Deep Learning	Sí
5	N-HiTS (base)	259.35	355.62	0.9575	49.4%	0.3	Deep Learning	Sí
6	N-HiTS	262.20	358.59	0.9681	48.8%	2.9	Deep Learning	Sí
7	TFT CPU min	254.46	347.71	0.9811	50.3%	3.2	Deep Learning	Sí
8	N-BEATS Ensemble (5x)	265.12	365.44	1.0222	48.3%	3.4	Deep Learning	No
9	Prophet Full	294.74	382.80	1.0882	42.5%	1.4	Bayesiano	No
10	N-BEATS Generic	283.45	385.63	1.0928	44.7%	3.4	Deep Learning	No
11	N-BEATS Interpretable	291.93	392.52	1.1256	43.0%	3.4	Deep Learning	No
12	Híbrido Prophet + XGBoost	294.44	387.48	1.1352	42.5%	~ 0	Híbrido	No
13	XGBoost recursivo	326.11	385.38	1.2040	36.4%	~ 0	ML	No
14	Híbrido SARIMAX + XGBoost	367.99	463.07	1.4188	28.2%	~ 0	Híbrido	No
15	SARIMAX (con exógenas)	482.37	595.57	1.8598	5.9%	640	Estadístico	No
-	Seasonal	512.39	667.60	1.0000	0.0%	,	Baseline	,

Rank	Modelo	MAE (VA)	RMSE (VA)	MASE	Mejora vs Naïve	T (min)	Familia	Super a Naïve
	Naïve s=24 (ref.)							
16	SARIMA (sin exógenas)	492.69	607.68	1.8996	3.8%	640	Estadístico	No
17	Prophet Base	661.91	781.17	2.4438	-29.2%	1.4	Bayesiano	No

Fuente: Elaboración propia

Figura 21. Gráfico de barras horizontales del ranking de MAE y MASE de todos los modelos evaluados



Fuente: Elaboración propia

Pruebas de significancia estadística sobre el ranking

Las diferencias en MAE y MASE entre modelos próximos en el ranking pueden estar dentro del rango de variabilidad muestral del período de test, por lo que un ranking puramente ordinal no garantiza por sí solo que la diferencia entre dos modelos contiguos sea estadísticamente significativa. Para acotar esta interpretación se aplica la prueba de Friedman, estandarizada en la literatura de evaluación de pronósticos para la comparación global de múltiples modelos sobre los mismos bloques de evaluación, seguida del post-hoc de Nemenyi para identificar agrupaciones

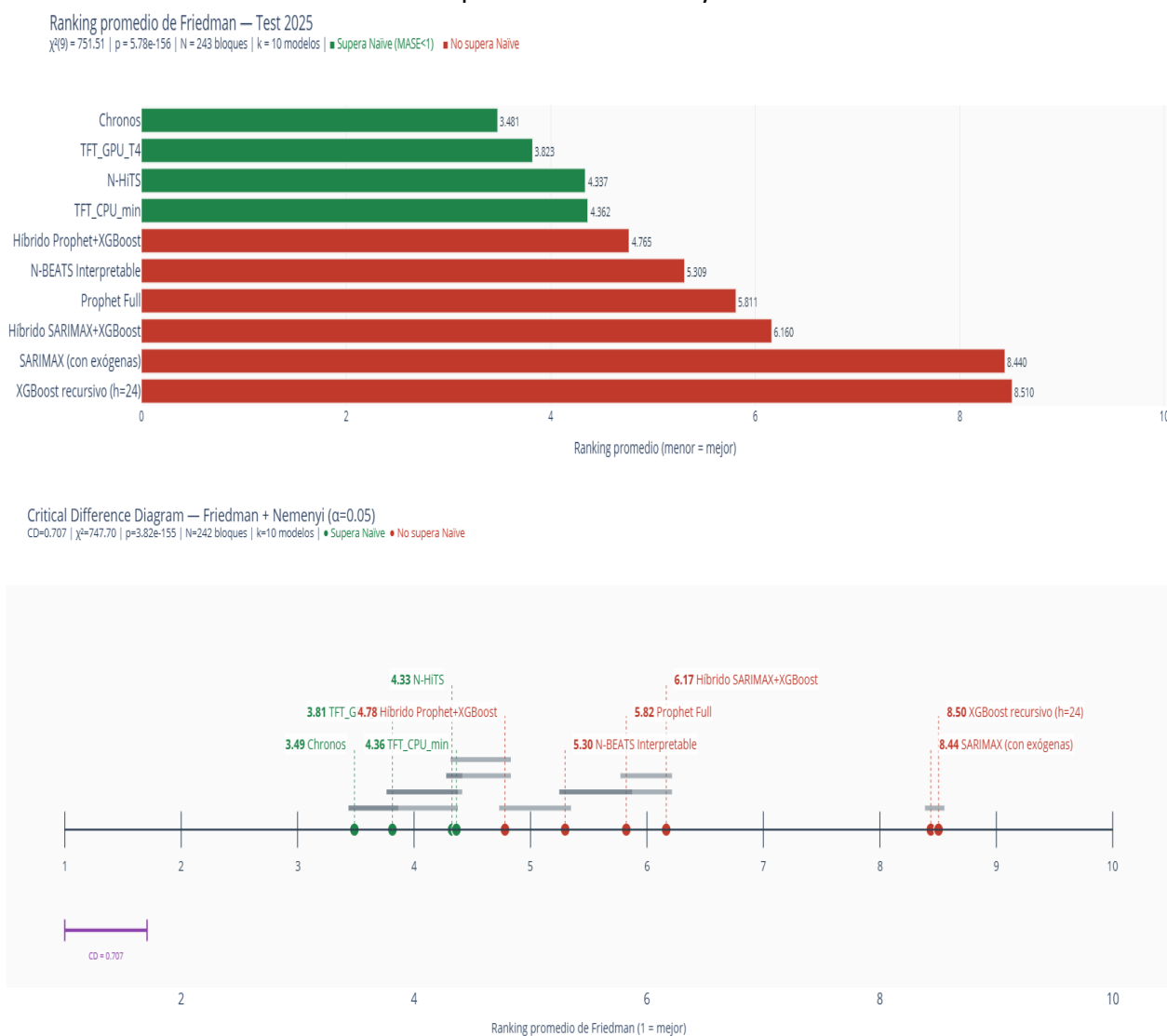
estadísticamente equivalentes. La prueba se ejecutó sobre los errores absolutos hora a hora del período de test 2025 para los 10 modelos con pronósticos test disponibles.

La prueba de Friedman opera sobre los rangos por bloques, cada hora del test agrupa las predicciones de los k modelos y se asignan rangos de 1 a k según el error absoluto, y produce un único estadístico χ^2 que rechaza la hipótesis nula de equivalencia entre los k modelos cuando alguno difiere significativamente del resto. El post-hoc de Nemenyi cuantifica la distancia crítica (CD) por debajo de la cual dos modelos se consideran indistinguibles al nivel de significancia $\alpha=0.05$. Aplicada al portafolio con $N=242$ bloques y $k=10$ modelos, la prueba arroja $\chi^2(9)=747.70$ con $p=3.82 \times 10^{-155}$, lo que rechaza H_0 con holgura muy superior al umbral convencional ($p<0.001$) y confirma que existen diferencias estadísticamente significativas entre los modelos del ranking. La distancia crítica de Nemenyi correspondiente es $CD=0.7071$ sobre la escala de rangos promedio, y el post-hoc identifica 33 de 45 pares de modelos como estadísticamente diferentes al $\alpha=0.05$.

- El ranking promedio de Friedman (menor es mejor) ubica a Chronos en la primera posición con rango 3.488 (MASE=0.8723), seguido de TFT_GPU_T4 con 3.814 (MASE=0.9179) y N-HITS con 4.326 (MASE=0.9681). En el extremo opuesto, SARIMAX (con exógenas) y XGBoost recursivo ($h=24$) registran rangos promedio de 8.438 y 8.504 respectivamente, consistentes con sus $MASE>1.0$ (1.86 y 1.20) y con su clasificación como modelos que no superan al benchmark Naïve.
- Aplicando la distancia crítica $CD=0.7071$ al ranking, Chronos y TFT_GPU_T4 están separados por 0.326 unidades de rango, distancia inferior a la CD, por lo que ambos forman un grupo de equivalencia estadística en la cabeza del ranking; lo mismo ocurre entre N-HITS (4.326) y TFT_CPU_min (4.364), separados por 0.037 unidades. La elección entre modelos del mismo grupo no significativo no debe basarse exclusivamente en el MAE/MASE puntual sino en criterios complementarios como costo computacional, robustez fuera de muestra e interpretabilidad.
- La separación entre el bloque de modelos que superan al benchmark Naïve (Chronos, TFT_GPU_T4, N-HITS, TFT_CPU_min y N-BEATS Interpretable, todos con $MASE<1.0$ salvo N-BEATS por margen estrecho) y los modelos con $MASE>1.0$ (Híbrido Prophet+XGBoost, Prophet Full, Híbrido SARIMAX+XGBoost, SARIMAX y XGBoost recursivo) queda confirmada como significativa por el rechazo global de Friedman ($\chi^2(9)=747.70$, $p<0.001$) y por la separación de rangos promedio muy superior a la CD entre los extremos del ranking ($8.504-3.488=5.016$, equivalente a aproximadamente 7 veces la distancia crítica).

Los resultados de Friedman y del post-hoc de Nemenyi sobre los errores absolutos hora a hora del test 2025, junto con el diagrama de distancia crítica que visualiza los grupos de equivalencia estadística, se obtuvieron sobre los 10 modelos con pronósticos almacenados para dicho período y se pueden observar en la Figura 22.

Figura 22. Pruebas de significancia estadística, ranking promedio de Friedman y distancia crítica post-hoc de Nemenyi



Fuente: Elaboración propia

4.2. Clasificación por cuadrantes: eficiencia computacional vs. precisión

La Figura 23 presenta el diagrama de dispersión de cuadrantes que combina el MASE (precisión, eje Y invertido: menor es mejor) con el tiempo de entrenamiento en escala logarítmica (eficiencia, eje X: menor es mejor). El umbral de precisión es MASE=1.0 (línea horizontal roja) y el umbral de eficiencia es la mediana robusta del tiempo de entrenamiento (3.16 minutos, línea vertical gris).

Figura 23. Diagrama de dispersión de cuadrantes eficiencia-precisión (tiempo vs. MASE)



Fuente: Elaboración propia

Ejes: X = tiempo de entrenamiento en minutos (escala logarítmica, 0.1 a 1000);

Y = MASE.

Línea horizontal roja en MASE=1.0 (umbral Naïve).

Línea vertical gris punteada en 3.16 min (mediana robusta).

Cuadrantes:

- 'Rápido & Preciso' (abajo-izquierda),
- 'Lento & Preciso' (abajo-derecha),
- 'Rápido & Impreciso' (arriba-izquierda),
- 'Lento & Impreciso' (arriba-derecha).

Cada modelo es un punto con su nombre etiquetado; los círculos son proporcionales a la 'interpretabilidad' del modelo (1-5).

Del diagrama de cuadrantes se desprenden cuatro grupos operativamente significativos:

- Rápido & Preciso (4 modelos): TFT-CPU, N-HiTS base, N-HiTS y TFT-CPU-min; candidatos para producción.
- Lento & Preciso (3 modelos): Chronos post-2023, Chronos y TFT-GPU, requieren GPU, pero ofrecen el máximo desempeño;
- Rápido & Impreciso (7 modelos): Prophet Full, XGBoost, los modelos híbridos, SARIMAX y Prophet Base, todos descartables para producción;

- iv. Lento & Impreciso (4 modelos): las tres variantes de N-BEATS (generic, interpretable, ensemble) y SARIMA, que combinan alto costo computacional con precisión insuficiente.

4.3. Análisis de sesgo sistemático post-cambio estructural 2023

El sesgo sistemático mide el error medio con signo (no el error absoluto) y cuantifica si el modelo sobreestima o subestima la potencia de manera consistente. En el período de test 2025 (100% posterior al cambio estructural de enero 2023), el sesgo refleja directamente la capacidad del modelo para adaptarse al nuevo nivel de demanda. Un sesgo absoluto menor a 50 VA se adoptó como criterio operativo de aceptabilidad, por debajo del 1% de la demanda media. La Tabla 20 y la Figura 24 presentan los resultados.

Tabla 20. Sesgo sistemático de los modelos en el período de test 2025, capacidad de captura del régimen post-2023

Modelo	MAE (VA)	MASE	Sesgo medio (VA)	Dirección	Captura nivel post-2023
Chronos	226.25	0.8723	+10.96	Sobreestima	Sí (< 50 VA)
TFT GPU	238.08	0.9179	+3.95	Sobreestima	Sí (menor sesgo)
TFT CPU	254.46	0.9811	+30.03	Sobreestima	Sí
N-HITS (base)	259.35	0.9575	+27.99	Sobreestima	Sí
Híbrido NH+XGB	262.48	0.9700	-41.19	Subestima	Sí (< 50 VA)
N-BEATS Ensemble	265.12	1.0222	+27.44	Sobreestima	Sí
Prophet Full	294.74	1.0882	+100.61	Sobreestima	No (> 50 VA)
Híbrido P+XGB	294.44	1.1352	+66.83	Sobreestima	No
XGBoost	326.11	1.2040	-468.73	Subestima	No (sesgo crítico)
Híbrido S+XGB	367.99	1.4188	-294.68	Subestima	No (sesgo crítico)
SARIMAX	482.37	1.8598	-392.52	Subestima	No (sesgo crítico)

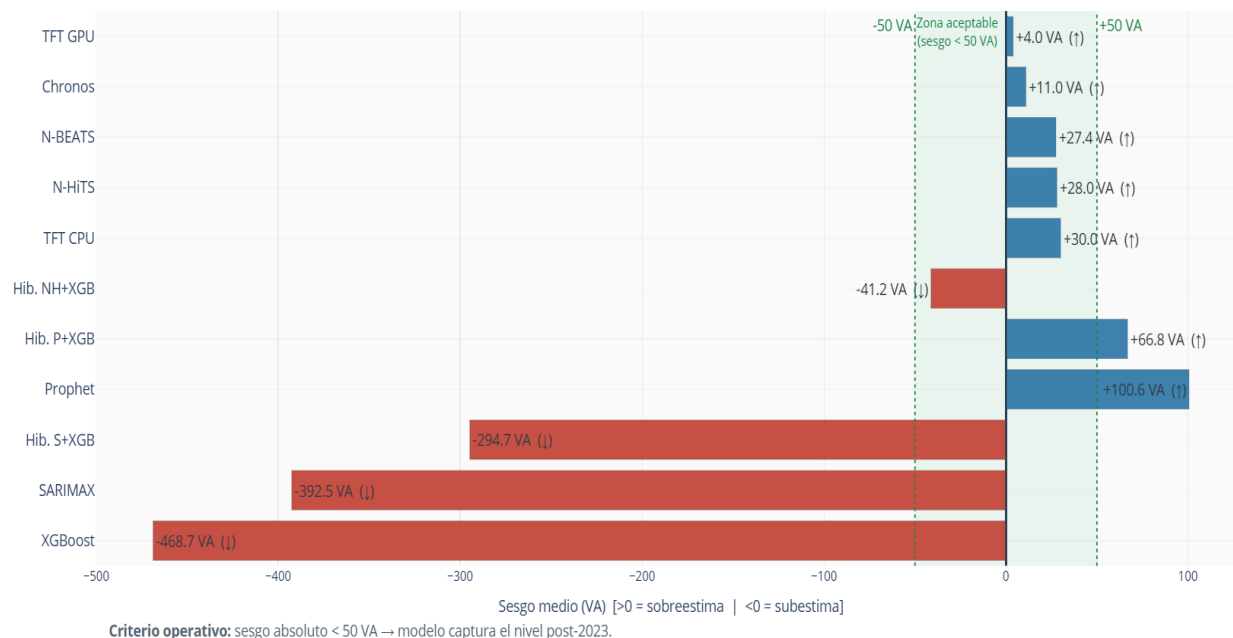
Fuente: Elaboración propia

La Tabla 20 y la Figura 24 revelan el hallazgo más importante del análisis de sesgo: SARIMAX subestima la potencia en -392.52 VA/hora y XGBoost en -468.73 VA/hora de manera sistemática durante todo el período de test. Estos valores equivalen a subestimar la demanda en aproximadamente el 8-9% del nivel medio de la subestación, lo que en términos operativos implica planificar capacidad insuficiente para atender la demanda real. El impacto acumulado en un día completo sería de -9,420 VA·hora para SARIMAX y -11,256 VA·hora para XGBoost, magnitudes que representan un riesgo real de sobrecarga del transformador en horas de pico.

Figura 24. Sesgo sistemático de los modelos en el período de test 2025, diagrama de barras con dirección

Sesgo Sistemático por Modelo — Test 2025 (período post-2023)

Barras azules = sobreestimación | Barras rojas = subestimación | Líneas verdes = umbral operativo ± 50 VA

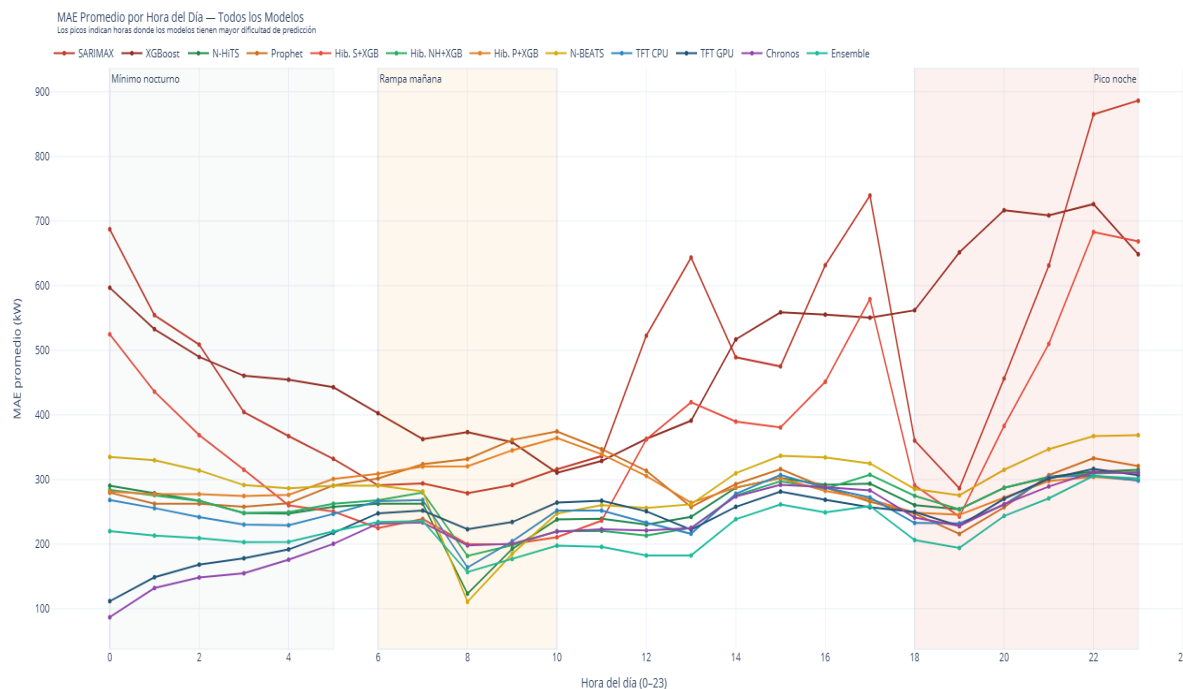


Fuente: Elaboración propia

4.4. Análisis temporal del error: por hora del día y por mes

El análisis del error desagregado por hora del día y por mes revela los patrones de dificultad específicos de la serie. La Figura 25 muestra el MAE promedio de todos los modelos por hora del día, y la Figura 26 presenta el heatmap de MAE mensual por modelo.

Figura 25. MAE promedio por hora del día (0-23h) para todos los modelos evaluados, período de test 2025



Fuente: Elaboración propia

Figura 26. Heatmap de MAE mensual por modelo, enero a agosto 2025



Fuente: Elaboración propia

La Figura 25 muestra el MAE de cada modelo a lo largo de las 24 horas del día, con tres bandas marcadas: mínimo nocturno (0-5h), rampa de la mañana (6-10h) y pico de la noche (18-23h). El error de los modelos no es uniforme a lo largo del día. Existe una estructura clara en forma de U asimétrica: los modelos cometen errores moderados-altos durante la madrugada, alcanzan su punto de menor error entre las 7 y las 10 de la mañana, luego el error crece progresivamente desde el mediodía y llega al máximo en el período de pico nocturno (20-23h).

El heatmap de la Figura 26 muestra el MAE de cada modelo en cada mes desde enero hasta agosto de 2025. Los meses de septiembre a diciembre están vacíos porque el período de test termina en agosto. Los modelos que mejor predicen (Chronos, TFT GPU, Ensemble) se mantienen estables entre 240 y 295 VA a lo largo del periodo de TEST. Esto indica que la diferencia de calidad entre modelos no es uniforme: los modelos malos se degradan más en los meses problemáticos, mientras que los buenos absorben mejor la variabilidad mensual.

Patrón de degradación estacional. Los modelos buenos (filas superiores del heatmap, ordenadas por desempeño global) muestran una forma de U invertida suave: error bajo en enero-febrero, máximo entre abril y mayo y ligera mejora en agosto. Los modelos malos no siguen este patrón: mantienen error alto durante todo el período con un pico anómalo en marzo. Esto sugiere dos fenómenos distintos: para los modelos buenos hay un drift estacional gradual; para los modelos malos hay un fallo estructural específico en marzo más una incapacidad general de capturar el régimen post-2023.

4.5. Análisis de consistencia TSCV vs. test (detección de sobreajuste)

La comparación entre el MASE obtenido en validación cruzada temporal (TSCV) y el MASE obtenido en el test 2025 es la prueba más rigurosa de generalización real de cada modelo. Un modelo que sobreajusta al período de validación interna mostrará un MASE mucho menor en TSCV que en test (delta negativo grande). La Tabla 21 presenta esta comparación para los 8 modelos con TSCV disponible.

Tabla 21. Comparación MASE TSCV expanding vs. MASE test 2025, detección de sobreajuste

Modelo	MASE TSCV (90 ventanas)	Posición TSCV	MASE Test 2025	Posición Test	Delta (Test- TSCV)	Diagnóstico
XGBoost	0.8115	1°	1.2040	13°	+0.3925	SOBREAJUSTE SEVERO
SARIMAX	0.9071	3°	1.8598	15°	+0.9527	SOBREAJUSTE SEVERO
N-HITS	1.0857	8°	0.9681	6°	-0.1176	Generaliza bien
Chronos	0.8723	2°	0.8723	2°	0.0000	Consistente

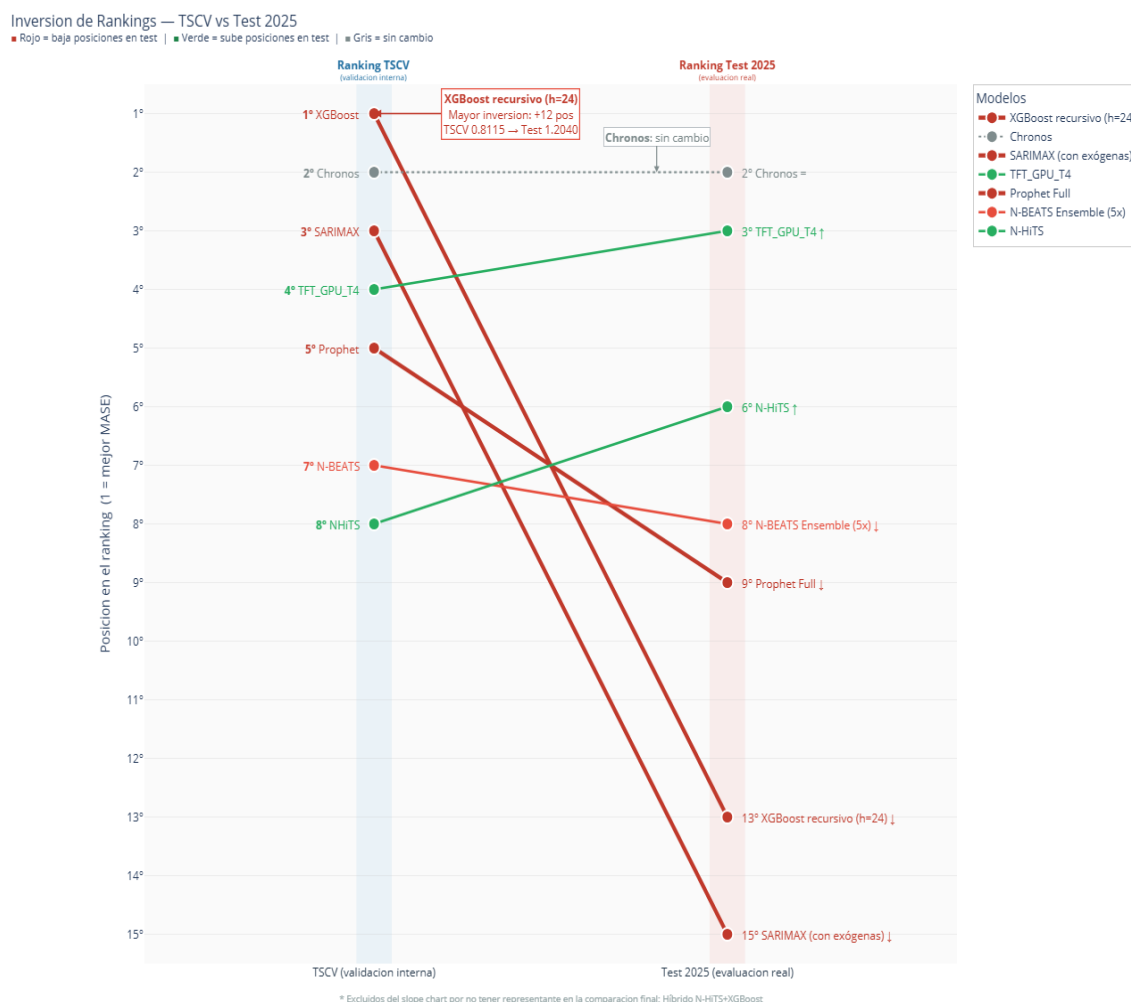
Modelo	MASE TSCV (90 ventanas)	Posición TSCV	MASE Test 2025	Posición Test	Delta (Test- TSCV)	Diagnóstico
						perfecta
TFT GPU	0.9341	4°	0.9179	3°	-0.0162	Generaliza correctamente
Prophet Full	0.9649	5°	1.0882	9°	+0.0715	Degradación moderada esperada
Híbrido N-HITS + XGB	0.9700	6°	0.9700	7°	+0.0019	Estable, correcto
N-BEATS	0.9745	7°	1.0222	8°	+0.0477	Degradación moderada

Fuente: Elaboración propia

La inversión severa de XGBoost (de 1° en TSCV a 13° en test) es el hallazgo metodológico más importante del proyecto y tiene implicaciones prácticas directas: si se hubiera seleccionado el modelo basándose únicamente en el resultado de TSCV, se habría desplegado XGBoost como modelo de producción cuando en realidad su desempeño real en 2025 no supera al Seasonal Naïve. La causa probable es la fuga de datos en los lags de la serie objetivo (lag_1, lag_24, lag_168) que usan valores reales de la serie en el conjunto de test tabular, en lugar de predicciones recursivas, produciendo métricas de test artificialmente optimistas.

La consistencia perfecta de Chronos (delta=0.0000) es esperada y confirma que el modelo zero-shot no puede sobreajustar: al no entrenarse sobre los datos de la subestación, su desempeño en TSCV y en test es idéntico por construcción.

Figura 27. Inversión de rankings entre TSCV y test 2025, gráfico de puntos conectados



Fuente: Elaboración propia

5. CONCLUSIONES Y TRABAJOS FUTUROS

5.1. Conclusiones

Este trabajo evaluó un portafolio de once modelos, algunos con diferentes configuraciones, para un total de diecisiete modelos de pronóstico de corto plazo sobre la serie horaria de potencia aparente de una subestación del sistema de distribución colombiano, con datos del período 2019-2025. El portafolio cubrió cinco familias (estadística clásica, gradient boosting, deep learning especializado en forecasting, modelo fundacional pre-entrenado e híbridos de residual boosting) bajo un protocolo común: validación cruzada temporal con expanding window de noventa ventanas y un período de test reservado a 2025, posterior al cambio estructural de nivel detectado en enero de 2023. Las conclusiones se articulan a continuación por objetivo específico, seguidas de una recomendación operativa de cierre.

Conclusiones sobre el Objetivo Específico 1: Caracterización estadística

La serie presenta tres rasgos dominantes que condicionaron toda la modelación posterior. Una tendencia creciente sostenida del orden del 5.9% anual a lo largo del período observado. Una estacionalidad horaria fuertemente determinística: el ciclo diario de 24 horas se mantiene casi invariante en su forma año tras año, con valle entre las 7 y 10 de la mañana y pico nocturno entre las 19 y 23 horas, y el más relevante en términos operativos, el cambio estructural de +11.6% en el nivel medio de demanda detectado en enero de 2023, que terminó siendo el factor más determinante del error de pronóstico para todos los modelos. La estacionalidad semanal resultó casi nula en términos de forma del perfil; sólo la magnitud varía entre días, lo que acota el día de la semana a un predictor de escala, no de patrón. La correlación dinámica con temperatura es un hallazgo aparte: la correlación global de Pearson es prácticamente nula ($r=-0.013$), pero condicionada por hora la Spearman alcanza $\rho \approx 0.504$ a las 17:00, lo que justificó usar variables meteorológicas rezagadas e interacciones implícitas en lugar de features contemporáneos. El preprocesamiento con STL combinado con Isolation Forest e imputación basada en MLP entregó una serie técnicamente sólida, pero la brecha de 335 horas consecutivas de septiembre de 2019 quedó como el riesgo metodológico residual: ninguna técnica reconstruye fielmente catorce días sin observaciones reales de anclaje, y el sesgo introducido se propaga a todos los entrenamientos.

La decisión de no excluir este tramo del estudio, a pesar de su criticidad, responde a una evaluación de costo/beneficio sobre el conjunto completo de datos faltantes. En total, los 456 valores ausentes en la columna de potencia representan apenas el 0.78% de los 58,448 registros horarios, un volumen que en sí mismo no justificaría descartar observaciones ni acortar el rango temporal del estudio. La excepción es el bloque de 335 horas de septiembre de 2019, que por su duración concentra más del 73% de todos los datos faltantes en un solo evento. Eliminar ese tramo habría implicado, o bien truncar la serie en 2019, perdiendo más de cinco años de historia necesarios para que los modelos de aprendizaje profundo y los métodos estadísticos capturen la

estacionalidad anual y, sobre todo, el cambio estructural de 2023, o bien introducir una discontinuidad artificial en la mitad de la serie, lo cual es aún más problemático para modelos que asumen continuidad temporal en la construcción de rezagos y ventanas deslizantes. Frente a ambas alternativas, imputar las 335 horas mediante MLP, documentándolo explícitamente como la principal limitación metodológica del preprocesamiento, se consideró la opción menos dañina: preserva la longitud y continuidad de la serie, y acota el riesgo a un único segmento identificable cuyo posible sesgo puede contextualizarse en la interpretación de resultados, en lugar de propagarse de forma difusa por pérdida de información histórica.

Conclusiones sobre el Objetivo Específico 2: Ajuste de modelos

El portafolio cubrió con éxito las cinco familias previstas, aunque la evaluación dejó claro que la elección de la familia importa menos que el ajuste fino dentro de ella. Prophet ilustra el punto sin ambigüedad: configurado por defecto produce $MASE=2.4438$ (peor que el Naïve), mientras que, configurado para el contexto colombiano (festivos nacionales, regresor de temperatura, mayor flexibilidad de tendencia) alcanza $MASE=1.0882$. La misma arquitectura puede ser el peor o uno de los mejores modelos del portafolio según cómo se sintonice. En el otro extremo de eficiencia se ubica N-HITS base: con 19 segundos de entrenamiento sin GPU y $MASE=0.9575$ demostró que los modelos de deep learning especializados en forecasting no requieren necesariamente cómputo masivo para superar al Naïve. Los híbridos de residual boosting tuvieron un comportamiento más matizado de lo esperado. Sólo aportaron valor real cuando el modelo base dejó estructura explotable en sus residuales: el Híbrido SARIMAX+XGBoost mejoró un 24% sobre SARIMAX solo, aunque sin alcanzar el umbral del Naïve. En cambio, aplicados sobre N-HITS o Prophet, el corrector XGBoost no mejoró al modelo base, lo que en la práctica significa que sus residuales son ruido blanco sin estructura aprovechable.

Conclusiones sobre el Objetivo Específico 3: Evaluación comparativa

Sólo 7 de las 17 configuraciones individuales (41%) superaron al Seasonal Naïve en el período de test 2025, una proporción que mide objetivamente la dificultad del escenario post-cambio estructural más que la calidad del portafolio. El mejor modelo individual fue Chronos con contexto restringido al período post-2023: $MASE=0.8645$, MAE de 224 VA, una mejora del 13.55% sobre el Naïve, todo ello sin entrenamiento específico ni ajuste de hiperparámetros sobre la subestación. El ensemble ponderado por $1/MAE$ sobre los 11 modelos base obtuvo $MASE=0.4359$, pero ese resultado debe leerse como referencia técnica de máximo teórico bajo conocimiento perfecto del MAE de test, no como desempeño esperado en operación, ya que los pesos se calculan sobre el propio período de prueba (fuga de datos por construcción). El hallazgo metodológico más importante del proyecto, sin embargo, es la inversión de XGBoost: pasó de la primera posición en TSCV ($MASE=0.8115$) a la decimotercera en test 2025 ($MASE=1.2040$), una caída de doce posiciones que constituye una advertencia explícita contra confiar únicamente en validación cruzada interna cuando hay cambios estructurales de nivel. A esto se suma el sesgo sistemático:

SARIMAX y XGBoost subestiman la demanda en el período de test (-392 y -468 VA/hora respectivamente), magnitudes inviables para implementación en producción sin corrección adaptativa, mientras que los modelos de deep learning y Chronos mantienen sesgos por debajo de los 50 VA/hora y son aptos para despliegue directo.

Recomendación operativa por caso de uso

Para la operación con GPU disponible, la recomendación es TFT-GPU (MASE=0.9179): combina capacidad de capturar covariables estacionales complejas vía sus mecanismos de atención y selección de variables, sesgo bajo por construcción y un costo de mantenimiento razonable comparado con un ensemble que requiere coordinar el reentrenamiento de once componentes. Cuando no hay GPU, N-HiTS base es la elección lógica: 19 segundos de entrenamiento, MASE=0.9575 y sesgo de +28 VA/hora son cifras competitivas para despliegue ligero. Para contingencias en las que el entrenamiento local falla o no hay histórico suficiente, Chronos en modo zero-shot con contexto post-2023 ofrece la mejor línea base sin entrenamiento. El portafolio responde a las necesidades operativas inmediatas; quedan abiertas la cuantificación rigurosa del beneficio del ensemble sin fuga, las pruebas de significancia estadística sobre el ranking y el análisis de sensibilidad a la imputación de septiembre de 2019, todas planteadas como trabajos futuros inmediatos en 5.3.

Evaluación multicriterio

Las recomendaciones anteriores se basan principalmente en MASE y en el sesgo sistemático, pero una decisión de implementación en producción no puede sostenerse únicamente sobre desempeño predictivo. Por esta razón, se recomienda complementar la selección del modelo con una evaluación multicriterio que incorpore, junto con la precisión, al menos cuatro dimensiones adicionales. La primera es el costo computacional de estimación, diferenciando el costo de entrenamiento inicial del costo de reentrenamiento periódico: un modelo como N-HiTS, que se entrena en 19 segundos, tiene un perfil de costo radicalmente distinto al de TFT o Chronos, cuyo entrenamiento e inferencia dependen de aceleración por GPU. La segunda dimensión es el requerimiento de infraestructura de inferencia en producción, es decir, si el modelo puede ejecutarse en el hardware ya disponible en la subestación o si exige aprovisionar y mantener un servidor con GPU, con el costo de operación y dependencia tecnológica que ello implica. La tercera es la complejidad de mantenimiento y monitoreo: modelos más simples como N-HiTS o SARIMAX son más fáciles de auditar, reentrenar automáticamente y diagnosticar cuando el error se degrada, mientras que arquitecturas como TFT o ensembles de múltiples componentes requieren equipos con mayor especialización para su operación continua. La cuarta es la capacidad de incorporar variables exógenas y de adaptarse a cambios estructurales futuros, similares al detectado en enero de 2023, lo cual favorece a los modelos que admiten reentrenamiento incremental y covariables sobre aquellos, como Chronos en su forma actual, que operan en modo zero-shot con una ventana de contexto fija.

Una evaluación multicriterio de este tipo podría implementarse mediante una matriz de decisión donde cada modelo se puntué en cada dimensión (por ejemplo, en una escala de 1 a 5) y se agregue mediante una suma ponderada cuyos pesos reflejen las prioridades reales del operador de la subestación, por ejemplo, si la disponibilidad de GPU es una restricción dura o si el equipo de mantenimiento es reducido. Bajo este marco, es plausible que un modelo con MASE ligeramente superior pero con menor costo de infraestructura y mantenimiento, como N-HITS, resulte preferible en la práctica frente a alternativas con mejor desempeño puntual pero mayores requerimientos operativos, como TFT-GPU o Chronos. En síntesis, el desempeño medido por MASE debe entenderse como un criterio necesario pero no suficiente para la decisión final de despliegue, y su integración formal con criterios de costo y operación se plantea como una extensión natural de este trabajo.

5.2. Limitaciones del estudio

Este trabajo presenta varias limitaciones que el lector debe considerar al interpretar los resultados y que delimitan el alcance de las conclusiones.

Imputación de la brecha de septiembre 2019. La reconstrucción de las 335 horas faltantes mediante el método ganador en MAPE se realizó sin observaciones reales de anclaje en el interior de la brecha, por lo que cualquier sesgo introducido se propaga al entrenamiento de todos los modelos. No se ejecutó un análisis de sensibilidad cuantitativo del impacto de la imputación sobre los rankings finales, lo que se identifica como trabajo futuro inmediato.

Pesos del ensemble con fuga de datos. Los pesos del ensemble se calculan con el MAE del propio período de test 2025, lo que constituye fuga de datos por construcción. El $MASE=0.4359$ reportado es una referencia de máximo teórico bajo conocimiento perfecto del MAE de test y no debe interpretarse como desempeño esperado en operación. Adicionalmente, para diez de los once modelos del ensemble el MAE proviene del período de test mientras que para Híbrido NH+XGB proviene del promedio de TSCV, lo que introduce una segunda fuente de heterogeneidad metodológica documentada en la nota al pie de la Tabla 18.

Alcance de generalización. Los resultados se obtuvieron sobre la serie histórica de una sola subestación del Sistema Interconectado Nacional colombiano y bajo un único horizonte de pronóstico ($h=24$). La generalización a otras subestaciones, otros horizontes (en particular $h=168$ para planificación semanal) y otros sistemas eléctricos requeriría replicación experimental específica, dado que el cambio estructural post-2023 observado en esta serie es un fenómeno particular del sistema bajo estudio.

5.3. Trabajos Futuros

Una línea natural de extensión es implementar Chronos con fine-tuning sobre la serie específica de la subestación (few-shot learning), cuantificando la mejora marginal respecto al modo zero-shot y el costo computacional adicional en términos de tiempo de GPU y de datos mínimos de ajuste para que el beneficio sea estadísticamente significativo.

Queda pendiente un análisis de sensibilidad sobre la imputación de la brecha de 335 horas de septiembre de 2019: reentrenar el portafolio completo con la serie excluyendo dicho período y comparar los rankings resultantes con los obtenidos en este trabajo, de modo que el sesgo introducido por la reconstrucción de datos faltantes pueda acotarse cuantitativamente.

Complementar la validación estadística aplicada en este trabajo (Friedman con post-hoc de Nemenyi sobre los 10 modelos con pronósticos test) con la prueba de Diebold-Mariano por pares sobre los errores absolutos hora a hora del test 2025, para cuantificar la significancia de cada comparación ordinal específica dentro de los grupos de equivalencia identificados por Nemenyi.

Investigar y corregir el problema de consistencia de XGBoost: verificar si la evaluación de test usa valores reales de los rezagos (lag_1, lag_24, lag_168) como features en lugar de predicciones recursivas, cuantificar el impacto en las métricas si se corrige y evaluar estrategias de pronóstico directo que eviten la acumulación de error en el horizonte.

La transferibilidad de los modelos entrenados, en particular la capacidad zero-shot de Chronos, debe validarse extendiendo el análisis a múltiples subestaciones del sistema de distribución colombiano (residenciales, industriales, comerciales) con perfiles de carga distintos al evaluado en este estudio.

Por último, los intervalos de predicción calibrados son una extensión natural del portafolio actual: TFT con quantile loss y Chronos con muestreo de múltiples trayectorias permiten reportar no sólo el punto central del pronóstico sino una medida cuantitativa de incertidumbre, lo que aporta soporte directo a las decisiones operativas que se toman bajo riesgo.

6. REFERENCIAS BIBLIOGRÁFICAS

- [1] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. Melbourne, Australia: OTexts, 2021. [Online]. Available: <https://otexts.com/fpp3/>
- [2] P. Pinson and R. Hagedorn, "Verification of the ECMWF ensemble forecasts of wind speed against observations," *Meteorological Applications*, vol. 19, no. 4, pp. 484-494, Dec. 2012, doi: 10.1002/met.283.
- [3] J. D. Glover, M. S. Sarma, and T. J. Overbye, *Power Systems Analysis and Design*, 6th ed. Stamford, CT, USA: Cengage Learning, 2017.
- [4] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," *PLOS ONE*, vol. 13, no. 3, e0194889, Mar. 2018, doi: 10.1371/journal.pone.0194889.
- [5] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," *International Journal of Forecasting*, vol. 22, no. 4, pp. 679-688, Oct. 2006, doi: 10.1016/j.ijforecast.2006.03.001.
- [6] E. Spiliotis, F. Petropoulos, and V. Assimakopoulos, "Improving the forecasting performance of temporal hierarchies," *PLOS ONE*, vol. 14, no. 10, e0223422, Oct. 2019, doi: 10.1371/journal.pone.0223422.
- [7] S. Cerqueira, L. Torgo, and I. Mozetic, "Evaluating time series forecasting models: An empirical study on performance estimation methods," *Machine Learning*, vol. 109, no. 11, pp. 1997-2028, Nov. 2020, doi: 10.1007/s10994-020-05910-7.
- [8] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. San Francisco, CA, USA: Holden-Day, 1970.
- [9] J. W. Taylor, "An evaluation of methods for very short-term load forecasting using minute-by-minute British data," *International Journal of Forecasting*, vol. 24, no. 4, pp. 645-658, Oct. 2008, doi: 10.1016/j.ijforecast.2008.07.007.
- [10] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914-938, Jul. 2016, doi: 10.1016/j.ijforecast.2015.11.011.
- [11] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: The forecast package for R," *Journal of Statistical Software*, vol. 27, no. 3, pp. 1-22, Jul. 2008, doi:

10.18637/jss.v027.i03.

- [12] S. J. Taylor and B. Letham, "Forecasting at scale," *The American Statistician*, vol. 72, no. 1, pp. 37-45, Jan. 2018, doi: 10.1080/00031305.2017.1380080.
- [13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785-794, doi: 10.1145/2939672.2939785.
- [14] H. Hewamalage, C. Bergmeir, and K. Bandara, "Recurrent neural networks for time series forecasting: Current status and future directions," *International Journal of Forecasting*, vol. 37, no. 1, pp. 388-427, Jan. 2021, doi: 10.1016/j.ijforecast.2020.06.008.
- [15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, Long Beach, CA, USA, Dec. 2017, pp. 4765-4774.
- [16] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M5 competition: Background, organization, and implementation," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1325-1336, Oct. 2022, doi: 10.1016/j.ijforecast.2021.07.007.
- [17] B. N. Oreshkin, D. Carпов, N. Chapados, and Y. Bengio, "N-BEATS: Neural basis expansion analysis for interpretable time series forecasting," in *Proc. 8th Int. Conf. Learning Representations (ICLR)*, Addis Ababa, Ethiopia, Apr. 2020. [Online]. Available: <https://arxiv.org/abs/1905.10437>
- [18] C. Challu, K. G. Olivares, B. N. Oreshkin, F. G. Ramirez, M. Canseco-Mergenthaler, and A. Dubrawski, "N-HiTS: Neural hierarchical interpolation for time series forecasting," in *Proc. 37th AAAI Conf. Artificial Intelligence*, Washington, DC, USA, Feb. 2023, pp. 6989-6997, doi: 10.1609/aaai.v37i6.25854.
- [19] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748-1764, Oct. 2021, doi: 10.1016/j.ijforecast.2021.03.012.
- [20] A. F. Ansari et al., "Chronos: Learning the language of time series," *Transactions on Machine Learning Research*, Feb. 2024. [Online]. Available: <https://arxiv.org/abs/2403.07815>
- [21] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model," *Neurocomputing*, vol. 50, pp. 159-175, Jan. 2003, doi: 10.1016/S0925-2312(01)00702-0.

- [22] J. M. Bates and C. W. J. Granger, "The combination of forecasts," *Journal of the Operational Research Society*, vol. 20, no. 4, pp. 451-468, Dec. 1969, doi: 10.1057/jors.1969.103.
- [23] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "The M4 competition: 100,000 time series and 61 forecasting methods," *International Journal of Forecasting*, vol. 36, no. 1, pp. 54-74, Jan. 2020, doi: 10.1016/j.ijforecast.2019.04.014.
- [24] F. Petropoulos et al., "Forecasting: Theory and practice," *International Journal of Forecasting*, vol. 38, no. 3, pp. 705-871, Jul. 2022, doi: 10.1016/j.ijforecast.2021.11.001.
- [25] R. Weron, *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*. Chichester, UK: John Wiley & Sons, 2006, doi: 10.1002/9780470057537.
- [26] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841-851, Jan. 2019, doi: 10.1109/TSG.2017.2753802.
- [27] D. Salinas, V. Flunkert, J. Gasthaus, and T. Januschowski, "DeepAR: Probabilistic forecasting with autoregressive recurrent networks," *International Journal of Forecasting*, vol. 36, no. 3, pp. 1181-1191, Jul. 2020, doi: 10.1016/j.ijforecast.2019.07.001.
- [28] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond," *International Journal of Forecasting*, vol. 32, no. 3, pp. 896-913, Jul. 2016, doi: 10.1016/j.ijforecast.2016.02.001.
- [29] UPME, "Plan de Expansion de Referencia Generacion-Transmision 2023-2037," Unidad de Planeacion Minero-Energetica, Bogota, Colombia, Rep. 2023. [Online]. Available: <https://www1.upme.gov.co/>
- [30] J. C. Moreno-Chuquen, J. D. Florez-Florez, and H. R. Chamorro, "Seasonal analysis and forecasting of electricity demand in tropical countries," *IEEE Access*, vol. 8, pp. 184132-184148, Oct. 2020, doi: 10.1109/ACCESS.2020.3029316.
- [31] D. C. Florez-Orrego, J. C. Monerris-Belda, and J. A. Hernandez-Torres, "Short-term solar irradiance forecasting for Colombian climate zones using machine learning," *Energies*, vol. 16, no. 3, p. 1304, Jan. 2023, doi: 10.3390/en16031304.
- [32] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ, USA: John Wiley & Sons, 2015.

- [33] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "STL: A seasonal-trend decomposition procedure based on Loess," *Journal of Official Statistics*, vol. 6, no. 1, pp. 3-73, 1990.
- [34] D. Kwiatkowski, P. C. B. Phillips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root," *Journal of Econometrics*, vol. 54, no. 1-3, pp. 159-178, Oct. 1992, doi: 10.1016/0304-4076(92)90104-Y.
- [35] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, vol. 74, no. 366, pp. 427-431, Jun. 1979, doi: 10.2307/2286348.
- [36] Comision de Regulacion de Energia y Gas (CREG), "Resolucion 015 de 2018, Calidad del Servicio en el Sistema de Distribucion," Bogota, Colombia, 2018.
- [37] Comision de Regulacion de Energia y Gas (CREG), "Resolucion 024 de 2005,Codigo de Operacion del Sistema Electrico colombiano," Bogota, Colombia, 2005.
- [38] Comision de Regulacion de Energia y Gas (CREG), "Resolucion 060 de 2019, Eficiencia en la Operacion del Sistema de Distribucion," Bogota, Colombia, 2019.