Artículo Científico / Scientific Paper

UNIVERSIDAD POLITÉCNICA SALESIANA ECUADOR
INGENIUS
Revista de Ciencia y Tecnología
https://doi.org/10.17163/ings.n36.2026.02
pISSN: 1390-650X / eISSN: 1390-860X

Marco comparativo para el pronóstico de demanda eléctrica con machine learning y validación temporal rodante

Comparative framework for electricity demand forecasting using machine learning and rolling temporal validation

Juan Carlos Castillo[1,*], Jessica N. Castillo[1], Gabriel Pesántez[2], Wilian Guamán[2]

Recibido: 15-11-2025, Recibido tras revisión: 26-01-2026, Aceptado: 21-04-2026, Publicado: 01-07-2026

Resumen

La precisión en el pronóstico de la demanda eléctrica es un elemento central para la planificación y operación de los sistemas de potencia, en particular ante la variabilidad temporal de la carga y la presencia de deriva temporal. En este trabajo se desarrolla un marco comparativo reproducible de modelos de *machine learning* con validación temporal rodante (rolling-origin expanding), análisis multihorizonte y una métrica operativa de tolerancia relativa (%Tol). Se evalúan cuatro modelos representativos: EvoXGB (ensamble secuencial de XGBoost sobre residuales), XGB, TabNet y FT-Transformer, aplicados al pronóstico horario de potencia activa en subestaciones de distribución de un sistema eléctrico ecuatoriano. Para asegurar la comparabilidad cuando existen diferencias de cobertura o desalineación temporal entre predicciones, se incorpora una auditoría explícita basada en alineación y un conjunto común de evaluación (COMMONMASK), complementada con un bloque contiguo común para la figura de zoom. En la subestación representativa (con métricas sobre el conjunto común), XGB logra el mejor desempeño, con $R^2 = 0.993$ (corto) y 0.983 (mediano), y un RMSE de 21.16 y 30.84 kW, respectivamente. EvoXGB se mantiene competitivo, mientras que TabNet y FT-Transformer muestran mayor degradación en el horizonte mediano. En la verificación de *holdout* (90/10) se observa la caída esperada por deriva temporal, preservándose el orden comparativo. El marco propuesto entrega una base trazable para comparar modelos en series reales de subestaciones y para extender el análisis hacia esquemas híbridos y adaptativos.

Palabras clave: pronóstico de carga, *machine learning*, XGBoost, TabNet, FT-Transformer, validación temporal rodante.

Abstract

Accurate load forecasting is essential for power system planning and operation, particularly under pronounced temporal variability and temporal drift. This study presents a reproducible comparative framework for machine learning models based on rolling-origin expanding validation, multihorizon evaluation, and an operational relative tolerance metric denoted as %Tol. Four representative models are evaluated: EvoXGB, a sequential residual XGBoost ensemble; XGB; TabNet; and FT-Transformer. These models are applied to hourly active power forecasting in distribution substations within an Ecuadorian power system. To ensure a fair comparison when models exhibit differences in prediction coverage or temporal misalignment, the framework incorporates an explicit comparability audit based on temporal alignment and a common evaluation mask denoted as COMMONMASK, complemented the longest common contiguous block for the zoomed time-series visualization. For the representative substation, with metrics computed on the common set, XGB achieves the best performance, with $R^2 = 0.993$ for the short horizon and $R^2 = 0.983$ for the medium horizon, and RMSE values of 21.16 and 30.84 kW, respectively. EvoXGB remains competitive, whereas TabNet and FT-Transformer exhibit greater degradation in the medium horizon. The 90/10 holdout verification shows the expected performance decline associated with temporal drift while preserving the comparative ranking. Overall, the proposed framework provides a traceable benchmark for substation load forecasting and supports future extensions toward adaptive and hybrid forecasting approaches.

Keywords: load forecasting, machine learning, XGBoost, TabNet, FT-Transformer, rolling temporal validation

[1,*]Universidad Técnica de Cotopaxi, Facultad de Ciencias de la Ingeniería y Aplicadas, Ecuador.
Autor para correspondencia ✉: juan.castillo2321@utc.edu.ec.
[2]Escuela Superior Politécnica de Chimborazo (ESPOCH), GITEA, Riobamba, Ecuador.

Forma sugerida de citación: J. C. Castillo, J. N. Castillo, G. Pesántez y W. Guamán. "Marco comparativo para el pronóstico de demanda eléctrica con *machine learning* y validación temporal rodante," *Ingenius, Revista de Ciencia y Tecnología*, N.° 36, pp. 19-28, 2026. DOI: https://doi.org/10.17163/ings.n36.2026.02.

1. Introducción

El crecimiento sostenido de la demanda energética y la integración progresiva de fuentes renovables intermitentes han convertido el pronóstico de carga en un componente estratégico para la planificación y operación de los sistemas eléctricos. La precisión en distintos horizontes temporales resulta fundamental para la programación de subestaciones, la asignación de recursos y la gestión eficiente de la red bajo condiciones de estacionalidad, no linealidades y cambios de régimen [1, 2], [3–5].

Entre los enfoques contemporáneos, los algoritmos de impulso de gradiente, como XGBoost, se han consolidado por su robustez y su capacidad para modelar relaciones no lineales en datos tabulares de alta dimensionalidad [6–9]. No obstante, su desempeño depende de una calibración cuidadosa de hiperparámetros, que suele ser costosa y sensible a la configuración de los datos. Para reducir esta sensibilidad, se han propuesto variantes optimizadas mediante algoritmos evolutivos, por ejemplo, combinando XGBoost con *Differential Evolution o Genetic Algorithms*, que han mostrado mejoras en estabilidad y reducción de sobreajuste [10–12].

En paralelo, han surgido modelos de aprendizaje profundo específicos para datos tabulares. TabNet emplea mecanismos de atención secuencial con interpretabilidad inherente [13], mientras que FT-Transformer adapta la arquitectura Transformer mediante *embeddings* lineales de características y atención multicabeza [14, 15]. Sin embargo, diversos estudios señalan que la superioridad de las redes profundas sobre los métodos basados en árboles no es universal y depende del tamaño y la estructura del conjunto de datos [16, 17].

Una limitación frecuente en la literatura de pronóstico de carga es el uso de divisiones estáticas de entrenamiento y prueba que ignoran la no estacionariedad temporal. Las guías metodológicas actuales recomiendan esquemas de validación temporal con origen rodante (*rolling-origin expanding*) para evaluar el desempeño de los modelos en el tiempo y detectar degradación operacional [18–20]. Además de métricas globales como MAE, RMSE y R^2, resulta útil incorporar indicadores operativos que reflejen tolerancias aceptables de error desde la perspectiva de la planificación, como la métrica %Tol [21, 22].

Contraste explícito con enfoques estándar. A diferencia de la validación cruzada aleatoria (o de las particiones sin respetar el orden temporal), que puede producir estimaciones optimistas al mezclar pasado y futuro, la validación con origen rodante evalúa el desempeño en escenarios realistas donde el modelo predice sobre periodos posteriores.
Asimismo, el marco propuesto integra métricas clásicas y la métrica %Tol bajo una auditoría explícita de comparabilidad cuando la cobertura de predicciones difiere entre modelos.

En el contexto ecuatoriano, se han desarrollado trabajos sobre proyección de consumo a largo plazo mediante modelos de *machine learning* [23] y revisiones sobre planificación de sistemas eléctricos con aprendizaje por refuerzo [24]. Este trabajo desarrolla un marco comparativo para el pronóstico horario de potencia activa en subestaciones, empleando validación temporal con origen rodante, %Tol y una auditoría metodológica para comparaciones justas.

Las contribuciones específicas de este estudio son:

(i) Proponer un marco de evaluación reproducible basado en validación temporal rodante y agregación de métricas con criterio micro.

(ii) Comparar *boosting* y arquitecturas profundas tabulares en horizontes de corto y mediano plazo, manteniendo consistencia metodológica.

(iii) Incorporar una auditoría explícita de comparabilidad (alineación temporal + COMMONMASK) y una lectura operativa mediante %Tol.

2. Materiales y métodos

2.1. Datos y variables

El propósito de esta metodología es evaluar la estabilidad y precisión de modelos predictivos de demanda eléctrica aplicables a subestaciones ecuatorianas.

Se utilizaron series horarias de potencia activa registradas en nueve subestaciones de una red de distribución del sistema eléctrico nacional, con aproximadamente 40 000 observaciones horarias por estación, correspondientes a un periodo continuo de alrededor de 4.5 años (entre 2020 y 2024). Los datos provienen de registros históricos internos del operador eléctrico nacional.

La variable objetivo corresponde a la potencia activa (kW), mientras que las variables predictoras incluyen:

(i) atributos de calendario (hora, día de la semana y mes),

(ii) rezagos de la potencia entre 4 h y 24 h, y

(iii) medias móviles de 3, 6 y 24 h.

Por motivos de confidencialidad, los identificadores de las subestaciones fueron anonimizados. Cada modelo se entrena de forma independiente por subestación. Para mantener una presentación compacta, las tablas y figuras detallan los resultados de una subestación representativa, mientras que el resto de las estaciones se utiliza para confirmar la consistencia de los hallazgos. En particular, se considera “representativa” la

subestación cuyo perfil de desempeño (RMSE y %Tol en ambos horizontes) se ubica más cerca de la mediana del conjunto (según la suma de rangos), evitando seleccionar casos extremos.

2.2. Preprocesamiento de los datos

Las series se organizaron cronológicamente y se realizó una limpieza inicial de valores extremos o infinitos. Los valores faltantes se trataron mediante propagación hacia delante (*forward fill*) dentro de cada serie; los casos residuales se descartaron para evitar interpolaciones que pudieran introducir sesgos temporales. Posteriormente, se generaron rezagos y medias móviles, eliminando las primeras 24 observaciones para evitar inconsistencias de borde.

En cada iteración del proceso de validación, las variables explicativas fueron escaladas mediante *RobustScaler*, ajustado exclusivamente con los datos de entrenamiento de esa ventana y aplicado al conjunto de prueba, para evitar la fuga de información (*data leakage*). En TabNet se mantuvo la misma selección de atributos, con normalización interna del modelo.

2.3. Modelos de predicción analizados

Se evaluaron cuatro modelos representativos:

- **A_EvoXGB:** ensamble secuencial de cuatro etapas basado en XGBoost, donde cada componente entrena árboles sobre los residuales del modelo anterior. La predicción final corresponde a la suma de las salidas parciales, buscando reducir el error sistemático.
- **B_XGB:** implementación estándar de XGBoost, utilizada como línea base robusta.
- **C_TabNet:** red tabular con atención secuencial [13], configurada con $n_d = n_a = 32$, cuatro pasos de decisión y parada temprana.
- **D_FT-Transformer (D_FTT):** Transformer tabular con embeddings y atención multicabeza [14,15], con tres bloques codificadores, dimensión de token 192 y cuatro cabezas.

2.4. Ajuste de hiperparámetros

Los hiperparámetros más influyentes se ajustaron mediante experimentos piloto acotados alrededor de configuraciones recomendadas para datos tabulares de tamaño medio. La selección final priorizó la estabilidad temporal frente a mejoras marginales locales, manteniendo un presupuesto comparable de entrenamiento.

2.5. Esquema de validación temporal con origen rodante

Se aplicó validación con origen rodante (*rolling-origin expanding*). Se consideraron dos horizontes de pronóstico: corto y mediano:

- **Horizonte corto:** prueba 168 h, paso 24 h y purga 24 h.
- **Horizonte mediano:** prueba 720 h, paso 72 h y purga 72 h.

En cada iteración, el entrenamiento incluye todas las observaciones anteriores al inicio de la prueba (excluyendo la purga), con un mínimo de 1500 observaciones. Las métricas se agregaron con criterio micro, considerando conjuntamente todas las predicciones generadas por combinación modelo-horizonte.

2.6. Alineación temporal y conjunto común de evaluación (COMMONMASK)

En la práctica, distintos modelos pueden producir predicciones con cobertura desigual (p. ej., ejecución parcial de ventanas) o con desalineación temporal (desfase por construcción de lags o por el punto de anclaje del pronóstico). Para evitar comparaciones sesgadas, se aplicó un procedimiento de auditoría:

- **Alineación:** se aplicó un desfase de -24 h a las predicciones de B_XGB para homogeneizar el anclaje temporal del índice de prueba respecto a los demás modelos (corrección verificada por coincidencia del inicio del tramo de prueba).
- **COMMON ALL:** se definió la intersección de índices, donde los cuatro modelos presentan predicción finita; las métricas (MAE, RMSE, R^2, y %Tol) se calcularon sobre este conjunto.
- **COMMON CONTIG:** para la figura de zoom se seleccionó el bloque contiguo común más largo; de este bloque se presenta un segmento representativo de 168 h para facilitar la comparación visual sin discontinuidades.

Nota sobre cobertura y no imputación. Cuando un modelo presenta cobertura parcial (p. ej., por restricciones computacionales o ejecución incompleta en validación rodante), el conjunto común puede reducirse sustancialmente. Para evitar sesgos, no se imputaron ni interpolaron predicciones faltantes: en su lugar, se reporta explícitamente la cobertura (Tabla 2) y el tamaño del conjunto común (Tabla 3). La interpretación de desempeño se acota al tramo comparable; adicionalmente, se incluye una verificación de *hold-out* 90/10 y un análisis agregado de %Tol en nueve subestaciones para reforzar las conclusiones más allá del análisis de una sola serie.

2.7. Resumen del pipeline experimental (esquema)

La Figura 1 sintetiza la secuencia completa del marco experimental empleado, desde la preparación de datos y la ingeniería de atributos hasta la validación temporal, el entrenamiento de modelos y la agregación de métricas.

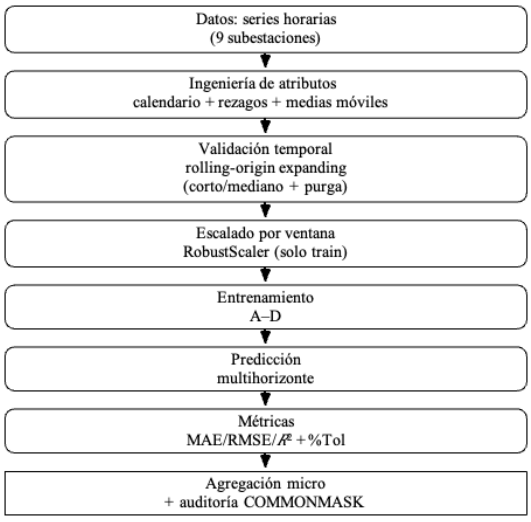


Figura 1. Esquema del flujo experimental: extracción de atributos, validación temporal, entrenamiento, predicción, métricas y auditoría de comparabilidad.

2.8. Validación independiente (*holdout* 90/10)

Como verificación complementaria, cada modelo se reentrenó con el 90 % inicial y se evaluó en el 10 % final, contrastando su desempeño fuera del entorno de recalibración temporal.

2.9. Métricas de evaluación

Se calcularon MAE, RMSE y R^2, junto con %Tol:

$$\%Tol_\delta = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(\frac{|y_i - \hat{y}_i|}{\max\{|y_i|, \varepsilon\}} \leq \delta\right) \times 100, \quad (1)$$

donde δ es el umbral de tolerancia relativa y $\varepsilon = 1$ kW evita cocientes inestables.

2.10. Reproducibilidad y recursos computacionales

Los experimentos se realizaron en Python 3.11 con XGBoost 2.0, PyTorch 2.3 y PyTorch-TabNet 4.1, fijando una semilla global. El entrenamiento se efectuó con una GPU NVIDIA RTX 3060 (12 GB) y 32 GB de RAM. El repositorio con scripts y configuraciones está disponible en GitHub [25].

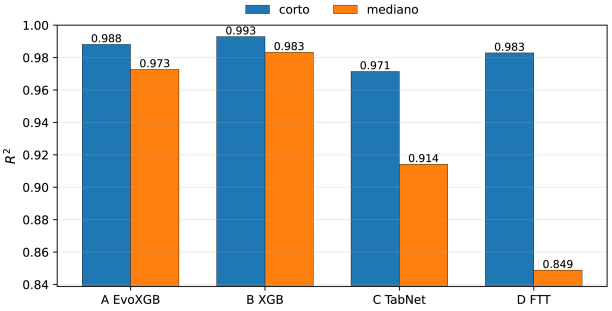
3. Resultados y discusión

3.1. Comportamiento de la demanda en las subestaciones

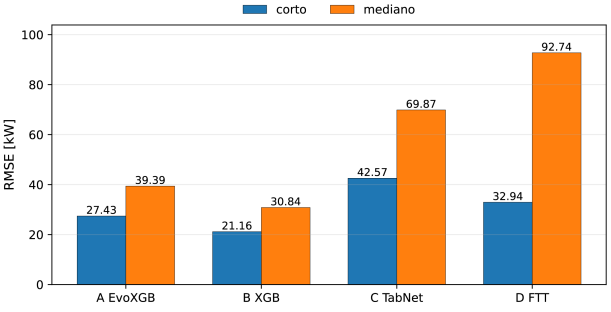
Las nueve subestaciones analizadas presentan patrones horarios típicos de redes de distribución: mínimos nocturnos, aumento matutino, mesetas diurnas y máximos al anochecer, además de estacionalidad semanal. En la subestación representativa, la potencia activa se mantiene dentro de un rango operativo estable y los cambios de régimen observados se asocian principalmente a variaciones semanales y estacionales, lo que justifica la necesidad de una validación temporal explícita.

3.2. Desempeño global por modelo y horizonte

La Figura 2 resume el desempeño por modelo y horizonte bajo validación con origen rodante. Las métricas se calcularon sobre COMMON ALL para garantizar que todos los modelos se evalúen sobre los mismos instantes. En esta estación, B_XGB logra el mejor equilibrio entre precisión y estabilidad temporal en ambos horizontes, mientras que A_EvoXGB se mantiene cercano. TabNet y FT-Transformer muestran una mayor degradación en el horizonte mediano.



(a) R^2 por modelo y horizonte (sobre COMMON ALL)



(b) RMSE [kW] por modelo y horizonte (sobre COMMON ALL)

Figura 2. Desempeño por modelo y horizonte en validación con origen rodante.

Tabla 1. Desempeño (agregación micro) en subestación representativa

Modelo	Hor.	MAE [kW]	RMSE [kW]	R^2	%Tol@5 %
A_EvoXGB	Corto	20.89	27.43	0.988	94.17
B_XGB	Corto	15.84	21.16	0.993	96.67
C_TabNet	Corto	29.63	42.57	0.972	82.50
D_FTT	Corto	24.67	32.94	0.983	87.92
A_EvoXGB	Mediano	25.61	39.39	0.973	87.92
B_XGB	Mediano	20.04	30.84	0.983	93.33
C_TabNet	Mediano	44.51	69.87	0.914	70.90
D_FTT	Mediano	54.38	92.74	0.849	68.26

3.3. Auditoría de comparabilidad: cobertura y COMMONMASK

La Tabla 2 reporta la cobertura de predicciones por modelo y horizonte (longitud total, número de predicciones finitas y porcentaje de NaN). La Tabla 3 resume el tamaño del conjunto común (COMMON ALL) y el bloque contiguo común (COMMON CONTIG) utilizado para la figura de zoom. En esta estación, el tamaño de COMMON ALL queda determinado por el modelo con menor cobertura (D_FTT), por lo que la auditoría se incluye explícitamente para garantizar la transparencia.

Tabla 2. Cobertura de predicción por modelo y horizonte (subestación representativa)

Hor.	Modelo	shift [h]	N total	N finito	%NaN	idx_{min}–idx_{max}
Corto	A_EvoXGB	0	37686	37686	0.00	0–37685
Corto	B_XGB	-24	37686	37662	0.06	0–37661
Corto	C_TabNet	0	37686	37686	0.00	0–37685
Corto	D_FTT	0	37686	1200	96.81	1620–4571
Mediano	A_EvoXGB	0	37686	37686	0.00	0–37685
Mediano	B_XGB	-24	37686	37662	0.06	0–37661
Mediano	C_TabNet	0	37686	37686	0.00	0–37685
Mediano	D_FTT	0	37686	1440	96.18	1572–2291

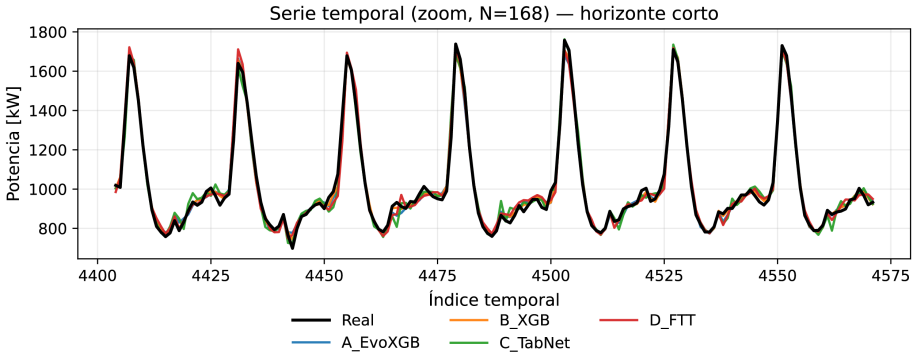
Tabla 3. Resumen del conjunto común (COMMONMASK) por horizonte (subestación representativa).

Hor.	N total	N_{common}	%common	idx_{min}–idx_{max}	inicio–fin contig	L_{contig}
Corto	37686	240	0.64	1620–4571	4404–4571	168
Mediano	37686	1440	3.82	1572–2291	1572–2291	720

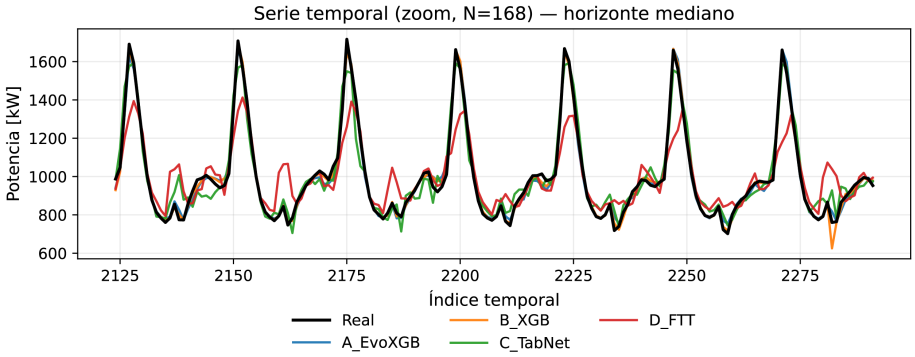
En el horizonte corto, el conjunto común queda reducido (240 h) porque D_FTT produjo predicciones en un bloque parcial y la intersección simultánea con los demás modelos (tras el ajuste de alineación de B_XGB) limita el solapamiento. Estas métricas describen el desempeño comparativo en ese tramo común (sin imputación). Para sostener conclusiones operativas, se enfatiza además que (i) el horizonte mediano, donde el solapamiento es mayor (COMMON CONTIG = 720 h en esta estación), (ii) la verificación *holdout* 90/10 y (iii) la sensibilidad de %Tol agregada en nueve subestaciones.

3.4. Reproducción temporal de la señal de carga

La Figura 3 compara la serie real y las predicciones en un segmento representativo (168 h) extraído del bloque COMMON CONTIG para ambos horizontes. En los modelos basados en *boosting* se observa una mejor reproducción de picos y valles. En el horizonte mediano, TabNet y FT-Transformer tienden a degradar su ajuste, consistente con el incremento de RMSE y la caída de R^2.



(a) Horizonte corto (zoom en COMMON CONTIG)



(b) Horizonte mediano (zoom en COMMON CONTIG)

Figura 3. Series reales y predichas en un segmento representativo.

3.5. Relación entre valores observados y predichos

Las Figuras 4 y 5 muestran la relación entre valores reales y predichos en validación *rolling-origin* (en ambos horizontes, evaluados sobre COMMON ALL). En los modelos basados en *boosting*, especialmente en **B_XGB**, se observa una mayor concentración alrededor de la diagonal $y = x$ en ambos horizontes. En **C_TabNet** y, en particular, en **D_FTT**, la dispersión aumenta en el horizonte mediano, lo cual es consistente con el incremento de RMSE y la reducción de R^2.

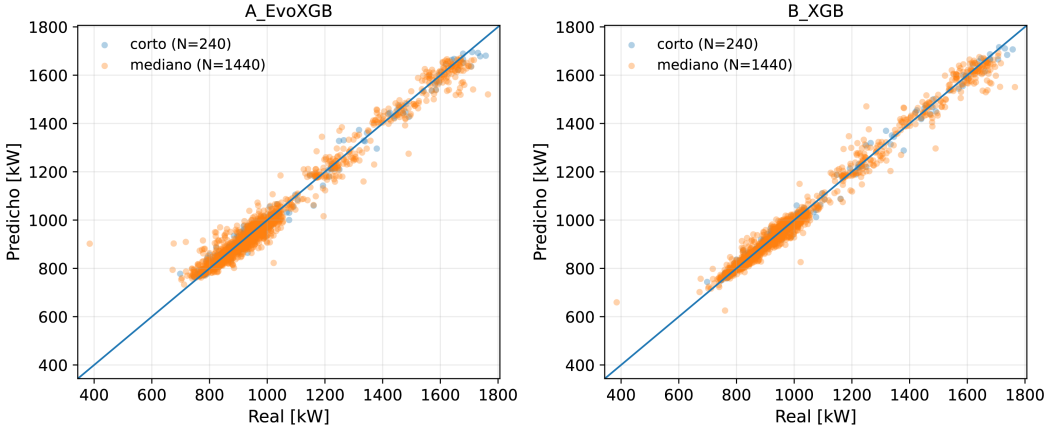


Figura 4. Relación entre valores reales y predichos para A_EvoXGB y B_XGB en ambos horizontes (*rolling-origin*; COMMON ALL). La línea representa y=x.

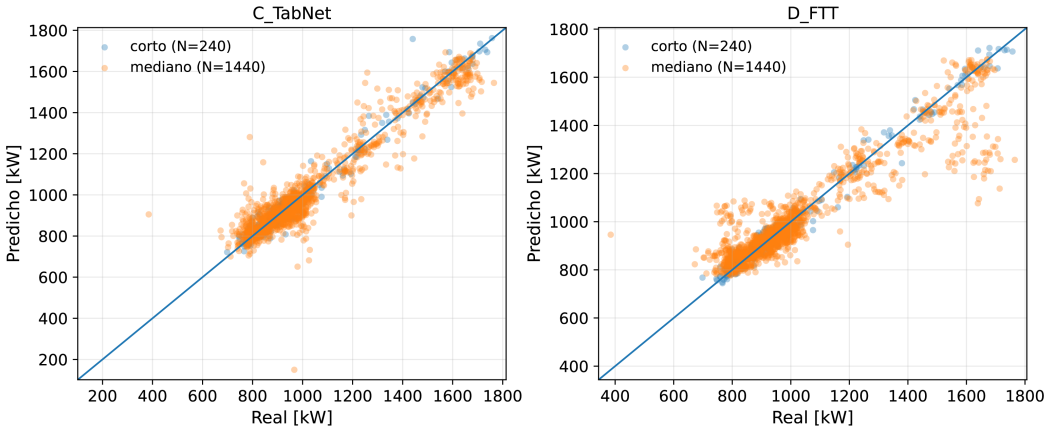


Figura 5. Relación entre valores reales y predichos para C_TabNet y D_FTT en ambos horizontes (*rolling-origin*; COMMON ALL). La línea representa y=x.

3.6. Distribución de errores relativos

Para complementar las métricas agregadas y caracterizar la variabilidad, la Figura 6 presenta la distribución del error absoluto relativo (%). Los modelos basados en *boosting* concentran el error en rangos bajos y con colas más cortas, mientras que TabNet y D_FTT exhiben una mayor dispersión, particularmente en el horizonte mediano.

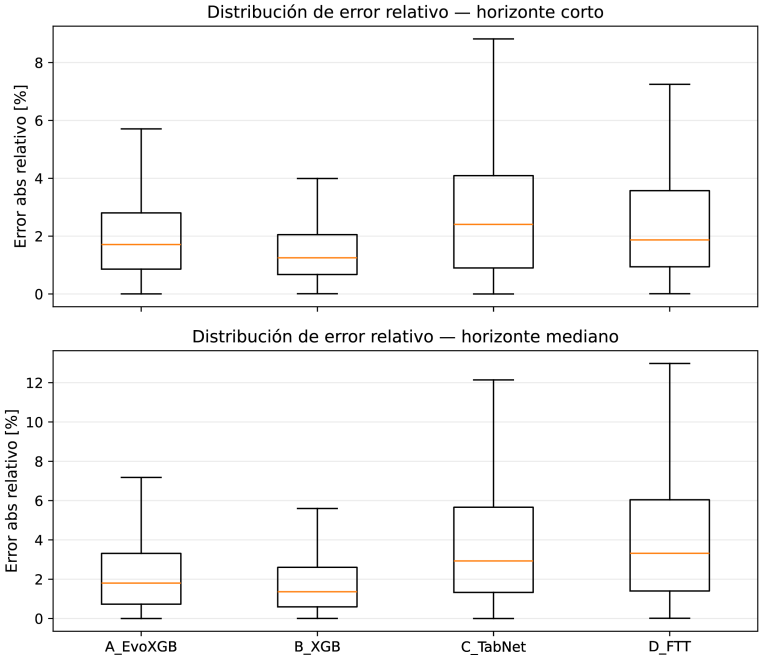


Figura 6. Distribución del error absoluto relativo (%) por modelo y horizonte (*rolling-origin*).

3.7. Métrica operativa: sensibilidad de %Tol al umbral δ

La Figura 7 resume la sensibilidad de %Tol ante distintos umbrales de tolerancia. Se evaluó δ en 2 %, 5 %, 10 %, 15 % y 20 %, agregando predicciones en nueve subestaciones. Para $\delta = 5$ % la métrica separa con claridad el desempeño entre familias de modelos; para $\delta \geq 10$ % la mayoría de los métodos se aproxima a 100 %, lo que reduce la capacidad discriminativa del indicador.

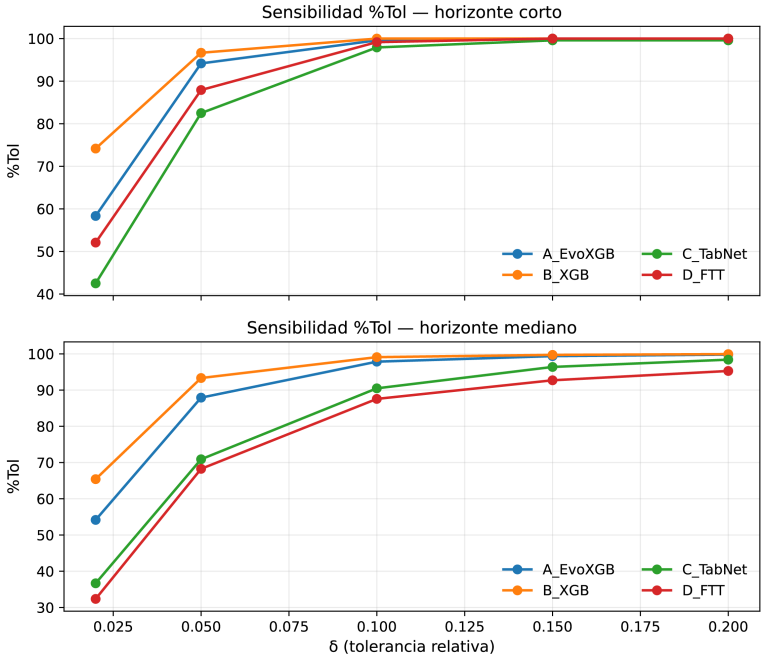


Figura 7. Sensibilidad de %Tol frente al umbral δ por modelo y horizonte (agregado en nueve subestaciones).

Tabla 4. Sensibilidad de %Tol(δ) por modelo y horizonte (agregado en nueve subestaciones)

Horizonte corto					
Modelo	δ=2%	5%	10%	15%	20%
A_EvoXGB	58.3	94.2	99.6	100.0	100.0
B_XGB	74.2	96.7	100.0	100.0	100.0
C_TabNet	42.5	82.5	97.9	99.6	99.6
D_FTT	52.1	87.9	99.2	100.0	100.0
Horizonte mediano					
Modelo	δ=2%	5%	10%	15%	20%
A_EvoXGB	54.2	87.9	97.8	99.4	99.8
B_XGB	65.4	93.3	99.1	99.7	99.9
C_TabNet	36.7	70.9	90.5	96.4	98.4
D_FTT	32.4	68.3	87.6	92.7	95.3

3.8. Validación independiente (*holdout* 90/10)

La Tabla 5 muestra la disminución esperada frente a *rolling-origin*, reflejando la deriva temporal entre periodos. Se conserva el orden comparativo, con B_XGB por encima del resto en esta estación.

Tabla 5. Desempeño en validación *holdout* 90/10 (subestación representativa): MAE, RMSE, R^2 y %Tol@5 %

Modelo	MAE [kW]	RMSE [kW]	R^2	%Tol@5 %
A_EvoXGB	36.95	96.25	0.872	68.00
B_XGB	29.67	70.04	0.932	79.13
C_TabNet	55.48	125.29	0.783	58.26
D_FTT	47.84	105.94	0.845	65.17

3.9. Discusión general

En la estación representativa, los métodos basados en árboles mantienen el mejor equilibrio entre precisión e interpretación operativa. La auditoría de alineación y COMMONMASK evita comparaciones sesgadas cuando existen diferencias de cobertura o desalineación temporal. TabNet y D_FTT pueden presentar un buen ajuste en el horizonte corto, pero en el horizonte mediano muestran degradación y una mayor dispersión.

Limitaciones. Los datos provienen de un sistema específico y se utilizaron variables de calendario y rezagos/medias móviles; en otros contextos o con variables exógenas podrían cambiar los patrones. La selección de una estación representativa se usa solo para dar claridad; las conclusiones operativas se refuerzan con la sensibilidad de %Tol agregada en nueve subestaciones. Finalmente, cuando un modelo presenta cobertura parcial, el conjunto común de evaluación puede reducirse considerablemente (p. ej., 240 h en el horizonte corto de esta estación); por ello, se reporta la auditoría de cobertura y COMMONMASK de forma explícita y se evita extrapolar más allá del tramo común evaluado.

Como líneas futuras, se propone incorporar variables meteorológicas y de generación renovable, y explorar esquemas de recalibración adaptativa e híbridos.

4. Conclusiones

Este estudio presentó un marco comparativo de modelos de *machine learning* para el pronóstico horario de

demanda eléctrica en subestaciones, basado en validación temporal con origen rodante, análisis multihorizonte y una métrica operativa de tolerancia relativa. Para garantizar la comparabilidad bajo diferencias de cobertura y desalineación temporal, se incorporó una auditoría explícita basada en alineación y COMMONMASK (máscara común de evaluación), reportando además la cobertura para contextualizar.

En la validación con origen rodante (*rolling-origin*) sobre COMMON ALL, **B_XGB** logró el mejor desempeño en la subestación representativa, seguido por **A_EvoXGB.** TabNet y D_FTT mostraron una degradación más marcada en el horizonte mediano. En la verificación de *holdout* 90/10 se observó la disminución esperada del desempeño asociada con la deriva temporal, manteniéndose el orden relativo.

En particular, cuando la intersección COMMON ALL es pequeña por cobertura parcial (p. ej., en el horizonte corto de la estación representativa), las métricas se interpretan como desempeño en el tramo estrictamente comparable, y las conclusiones se respaldan principalmente con el horizonte mediano, la verificación de *holdout* y el análisis agregado de %Tol.

El marco propuesto constituye una base trazable para comparar enfoques y orientar decisiones de planificación y operación. Como trabajo futuro, se plantea incorporar variables exógenas, explorar la recalibración adaptativa y extender la auditoría por ventana a todos los modelos para caracterizar la estabilidad temporal de forma uniforme.

Rol de los autores

- **Juan Carlos Castillo:** conceptualización; metodología; curación de datos; software; análisis formal; investigación; validación; visualización; administración del proyecto.
- **Jessica N. Castillo:** metodología; supervisión; validación; escritura – borrador original; escritura – revisión y edición.
- **Gabriel Pesántez:** metodología; supervisión; validación; escritura – borrador original; escritura – revisión y edición.
- **Wilian Guamán:** metodología; supervisión; validación; escritura – borrador original; escritura – revisión y edición.

Referencias

[1] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914–938, 2016. [Online]. Available: https://doi.org/10.1016/j.ijforecast.2015.11.011

[2] IEA, "Renewables 2023: Analysis and forecast to 2028," International Energy Agency, Paris, France, Tech. Rep., 2023, accessed: May 15, 2026. [Online]. Available: https://upsalesiana.ec/ing36ar2r2

[3] M. G. Pinheiro, S. C. Madeira, and A. P. Francisco, "Short-term electricity load forecasting—a systematic approach from system level to secondary substations," *Applied Energy*, vol. 332, p. 120493, Feb. 2023. [Online]. Available: https://doi.org/10.1016/j.apenergy.2022.120493

[4] S. Akhtar, S. Shahzad, A. Zaheer, H. S. Ullah, H. Kilic, R. Gono, M. Jasiński, and Z. Leonowicz, "Short-term load forecasting models: A review of challenges, progress, and the road ahead," *Energies*, vol. 16, no. 10, p. 4060, May 2023. [Online]. Available: https://doi.org/10.3390/en16104060

[5] F. Rodrigues, C. Cardeira, J. M. F. Calado, and R. Melicio, "Short-term load forecasting of electricity demand for the residential sector based on modelling techniques: A systematic review," *Energies*, vol. 16, no. 10, p. 4098, May 2023. [Online]. Available: https://doi.org/10.3390/en16104098

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. ACM, Aug. 2016, pp. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[7] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: a highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 3149–3157. [Online]. Available: https://upsalesiana.ec/ing36ar2r4

[8] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," in *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*. arXiv, 2017, pp. 6639–6649. [Online]. Available: https://doi.org/10.48550/arXiv.1706.09516

[9] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, Aug. 2020. [Online]. Available: https://doi.org/10.1007/s10462-020-09896-5

[10] Z. Mustaffa and M. H. Sulaiman, "Advanced forecasting of building energy loads with XGBoost and metaheuristic algorithms integration,"

Energy Storage and Saving, vol. 4, no. 4, pp. 421–438, Dec. 2025. [Online]. Available: https://doi.org/10.1016/j.enss.2025.03.005

[11] T.-N. Tran and Q.-D. Nguyen, "Research on the influence of genetic algorithm parameters on XGBoost in load forecasting," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 18 849–18 854, Dec. 2024. [Online]. Available: https://doi.org/10.48084/etasr.8863

[12] B. Liang, W. Qin, and Z. Liao, "A differential evolutionary-based xgboost for solving classification of physical fitness test data of college students," *Mathematics*, vol. 13, no. 9, p. 1405, Apr. 2025. [Online]. Available: https://doi.org/10.3390/math13091405

[13] S. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 6679–6687, May 2021. [Online]. Available: https://doi.org/10.1609/aaai.v35i8.16826

[14] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems*. arXiv, 2021, pp. 18 932–18 943. [Online]. Available: https://doi.org/10.48550/arXiv.2106.11959

[15] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 6, pp. 7499–7519, 2024. [Online]. Available: https://doi.org/10.1109/TNNLS.2022.3229161

[16] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?" in *NIPS'22: Proceedings of the 36th International Conference on Neural Information Processing Systems*. arXiv, 2022, pp. 507–520. [Online]. Available: https://doi.org/10.48550/arXiv.2207.08815

[17] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, May 2022. [Online]. Available: https://doi.org/10.1016/j.inffus.2021.11.011

[18] V. Cerqueira, L. Torgo, and I. Mozetič, "Evaluating time series forecasting models: an empirical study on performance estimation methods," *Machine Learning*, vol. 109, no. 11, pp. 1997–2028, Oct. 2020. [Online]. Available: https://doi.org/10.1007/s10994-020-05910-7

[19] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. Melbourne, Australia: OTexts, 2021, accessed: May 15, 2026. [Online]. Available: https://upsalesiana.ec/ing36ar2r16

[20] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Information Sciences*, vol. 191, pp. 192–213, May 2012. [Online]. Available: https://doi.org/10.1016/j.ins.2011.12.028

[21] M. Q. Raza and A. Khosravi, "A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings," *Renewable and Sustainable Energy Reviews*, vol. 50, pp. 1352–1372, Oct. 2015. [Online]. Available: https://doi.org/10.1016/j.rser.2015.04.065

[22] C. Borges, Y. Penya, I. Fernández, J. Prieto, and O. Bretos, "Assessing tolerance-based robust short-term load forecasting in buildings," *Energies*, vol. 6, no. 4, pp. 2110–2129, Apr. 2013. [Online]. Available: https://doi.org/10.3390/en6042110

[23] W. Guamán, P. Benalcázar, J. Córdova-García, and M. Torres, *Machine Learning-Based Projections of Long-Term Electricity Consumption: The Case Study of Ecuador*. Springer Nature Switzerland, 2025, pp. 174–187. [Online]. Available: https://doi.org/10.1007/978-3-031-83432-5_12

[24] G. Pesántez, W. Guamán, J. Córdova, M. Torres, and P. Benalcázar, "Reinforcement learning for efficient power systems planning: A review of operational and expansion strategies," *Energies*, vol. 17, no. 9, p. 2167, May 2024. [Online]. Available: https://doi.org/10.3390/en17092167

[25] J. C. Castillo. (2026) Forecasting-rolling-energy. GitHub repository. [Online]. Available: https://upsalesiana.ec/ing36ar2r27