Residential Energy Consumption Forecasting in Electric Utilities: An Approach Based on Random Forests and Time Series

José Luis Hernández, David De Yong, Fernando Magnago, Sergio Bragagnolo, and Juan Amaya

***Abstract*—Forecasting the monthly electricity consumption of residential users is a critical task for improving energy planning, demand management, and the efficient integration of renewable energy sources into the electrical system. This study predicts the consumption of a single residential user based on historical data, including monthly consumption and average temperature records from November 2018 to December 2024. Five forecasting approaches are compared: moving averages, ARIMA, standard Random Forest, Random Forest with lag variables, and Random Forest with hyperparameter optimization using RandomizedSearchCV. The models' performance is evaluated using MAE, MSE, and RMSE metrics over the last 12 months of the analyzed period, with 95% confidence intervals calculated via bootstrapping for both the validation phase and the estimation for January 2025. The results show that Random Forest models with lag variables and hyperparameter optimization outperform traditional methods such as moving averages and ARIMA in terms of accuracy. Additionally, the use of confidence intervals provides a more robust assessment of prediction reliability. It is concluded that the combined use of machine learning techniques, selection of relevant historical variables, and uncertainty quantification methods offers an effective tool for anticipating residential electricity consumption behavior. This approach can be valuable for electric utilities and policymakers seeking data-driven, reliable, and reproducible decisions.**

Link to graphical and video abstracts, and to code:
https://latamt.ieeer9.org/index.php/transactions/article/view/10246

***Index Terms*—Forecasting, time series, ARIMA, Random Forests, residential users**

I. Introducción

La predicción del consumo eléctrico residencial constituye una herramienta fundamental para la planificación energética, la gestión de la demanda y la operación eficiente de los sistemas de distribución eléctrica. Más allá de su uso en estudios de expansión de redes, las estimaciones de consumo son empleadas cotidianamente en tareas operativas tales como la validación de facturación, la detección de anomalías y la gestión de reclamos por parte de los usuarios. Estas aplicaciones adquieren especial relevancia en contextos donde los datos de consumo se recolectan con baja resolución temporal y bajo condiciones operativas heterogéneas.

En la provincia de Córdoba, Argentina, el servicio de distribución de energía eléctrica es brindado por la Empresa Provincial de Energía de Córdoba (EPEC) y por más de doscientas cooperativas eléctricas. A diferencia de otros sistemas eléctricos que cuentan con infraestructura de medición avanzada basada en medidores inteligentes, una proporción significativa de los consumos residenciales en estas redes aún se registra mediante lecturas manuales mensuales. Este proceso implica elevados costos operativos, demanda una considerable cantidad de recursos humanos y es susceptible a errores, particularmente en zonas rurales o geográficamente dispersas [1]. En consecuencia, las distribuidoras recurren con frecuencia a estimaciones de consumo cuando no se dispone de lecturas confiables [2].

Las estimaciones inexactas de consumo derivan en facturaciones incorrectas, generando insatisfacción en los usuarios y un aumento significativo en la cantidad de reclamos. Tradicionalmente, muchas distribuidoras han utilizado enfoques heurísticos simples para estimar o validar consumos, tales como la comparación con el período inmediato anterior o el cálculo de promedios móviles sobre ventanas temporales recientes [3]. Si bien estos métodos son fáciles de implementar, presentan limitaciones importantes: no capturan adecuadamente la estacionalidad, ignoran relaciones no lineales y no proporcionan una medida explícita de la incertidumbre asociada a la predicción [4-7].

Desde el punto de vista metodológico, los modelos de series temporales, como ARIMA (Autoregressive Integrated Moving Average), ofrecen un marco estadístico más formal para modelar la dependencia temporal del consumo eléctrico. Sin embargo, estos modelos suelen requerir un ajuste manual de parámetros, asumen relaciones esencialmente lineales y presentan dificultades para incorporar variables exógenas de manera flexible. En los últimos años, los avances en aprendizaje automático han impulsado el uso de modelos más complejos capaces de capturar patrones no lineales y comportamientos heterogéneos en los datos de consumo [8].

Gran parte de la literatura reciente en predicción de demanda residencial centra su análisis en el uso de datos de alta frecuencia provenientes de medidores inteligentes y en arquitecturas avanzadas de aprendizaje profundo, tales como redes neuronales recurrentes, modelos LSTM o Transformers. Si bien estos enfoques alcanzan altos niveles de precisión, su aplicación práctica requiere grandes volúmenes de datos

The associate editor coordinating the review of this manuscript and approving it for publication was Giner Alor-Hernández (*Corresponding author: Sergio Bragagnolo*).

J. L. Hernández, D. D. Yong, and F. Magnago are with Facultad de Ingenieria, Universidad Nacional de Rio Cuarto, Rio Cuarto, Argentina (e-mail: jhl@ing.unrc.edu.ar, ddeyong@ing.unrc.edu.ar, and fmagnago@ing.unrc.edu.ar).

Sergio Bragagnolo, and J. Amaya are with CIDTIEE, Facultad Regional Córdoba, Universidad Tecnológica Nacional, Córdoba, Argentina (e-mails: sbragagnolo@frc.utn.edu.ar, and jamaya@frc.utn.edu.ar).

históricos, alta capacidad computacional y una infraestructura de medición que no está disponible en muchas distribuidoras, especialmente en países en desarrollo [9-11].

En este trabajo se aborda un problema complementario y de alta relevancia práctica: la predicción del consumo eléctrico residencial bajo restricciones reales de datos, caracterizadas por registros mensuales, profundidad histórica limitada y escasa información auxiliar. En lugar de proponer un nuevo algoritmo de predicción, el aporte principal del trabajo radica en la evaluación sistemática y comparativa de métodos estadísticos tradicionales y modelos de aprendizaje automático en este contexto operativo, priorizando la robustez, la interpretabilidad y la posibilidad de implementación a gran escala.

Las principales contribuciones de este trabajo son las siguientes:

- La comparación de métodos estadísticos clásicos (medias móviles y ARIMA) con enfoques de aprendizaje automático basados en Random Forest para la predicción mensual del consumo residencial.
- La incorporación de variables rezagadas como estrategia de ingeniería de características para capturar la autocorrelación de corto plazo y la estacionalidad anual del consumo.
- La cuantificación de la incertidumbre de las predicciones mediante técnicas de remuestreo (bootstrapping), permitiendo construir intervalos de confianza útiles para la validación de facturación.
- El desarrollo de un marco de predicción reproducible, aplicable a distribuidoras que carecen de infraestructura de medición avanzada, ni una alta capacidad de computación, o los recursos necesarios para metodos más sofisticados de medición y predicción.

El resto del artículo se organiza de la siguiente manera. En la Sección II se revisan trabajos relacionados sobre predicción del consumo eléctrico residencial. La Sección III describe los métodos de predicción y las métricas de evaluación utilizadas. En la Sección IV se presentan los experimentos y resultados obtenidos, mientras que la Sección V discute los principales hallazgos. Finalmente, la Sección VI resume las conclusiones y plantea líneas de trabajo futuro.

II. Trabajos Relacionados

La predicción del consumo eléctrico residencial ha sido ampliamente estudiada utilizando tanto modelos estadísticos de series temporales como técnicas de aprendizaje automático. Los enfoques tradicionales, basados en medias móviles y modelos ARIMA se utilizan históricamente debido a su simplicidad, interpretabilidad y bajos requerimientos computacionales. No obstante, su capacidad para representar relaciones no lineales, cambios estructurales y patrones complejos es limitada, especialmente en series con alta variabilidad y estacionalidad marcada [8].

En los últimos años, los modelos de aprendizaje automático han ganado protagonismo en aplicaciones de predicción de demanda eléctrica. Algoritmos como Support Vector Regression, Random Forest, Gradient Boosting y redes neuronales artificiales han demostrado mejoras significativas en la precisión frente a los métodos estadísticos clásicos. Estudios comparativos recientes señalan que estos modelos resultan particularmente efectivos cuando los datos presentan comportamientos no lineales y dependencias temporales complejas.

El desarrollo de arquitecturas de aprendizaje profundo, como redes LSTM y modelos basados en Transformers, ha impulsado avances importantes en la predicción de demanda a partir de datos de alta frecuencia provenientes de medidores inteligentes. Estos enfoques modelan dependencias de largo plazo y capturan patrones complejos de consumo. Sin embargo, su aplicación práctica suele estar condicionada por la disponibilidad de grandes volúmenes de datos, elevados costos computacionales y una infraestructura de medición avanzada [9-11].

Dentro de los modelos de aprendizaje automático, Random Forest se ha consolidado como una alternativa robusta y eficiente para problemas de regresión. Su naturaleza basada en ensambles captura relaciones no lineales manteniendo una buena capacidad de generalización. Diversos trabajos aplican Random Forest a la predicción del consumo eléctrico residencial, combinándolo con técnicas de ingeniería de características como variables rezagadas, información climática y atributos calendarios. Asimismo, se han propuesto modelos híbridos que integran Random Forest con redes neuronales profundas, logrando mejoras adicionales en la precisión a costa de una mayor complejidad [12].

A pesar de estos avances, una parte significativa de la literatura se enfoca en escenarios con datos de alta resolución temporal y no considera explícitamente las restricciones operativas de muchas distribuidoras eléctricas. En particular, son escasos los estudios que analizan el desempeño de estos modelos cuando solo se dispone de registros mensuales de consumo y que, además, incorporan una cuantificación explícita de la incertidumbre asociada a las predicciones.

En contraste con los trabajos existentes, el presente estudio se centra en un escenario operativo realista para numerosas distribuidoras, evaluando métodos estadísticos y de aprendizaje automático bajo condiciones de datos limitadas. El énfasis se pone en la interpretabilidad, la robustez y la utilidad práctica de las predicciones, especialmente en aplicaciones relacionadas con la validación de consumos y la reducción de reclamos por facturación.

III. Metodos Utilizados

III-A. Series Temporales

Una serie temporal es una secuencia de observaciones de una variable que se registran regularmente a lo largo de un período de tiempo. En esencia, es una variable estadística cuyas mediciones están organizadas en función del tiempo. Estas observaciones dependen del momento en que se recopilan, lo que las convierte en una colección ordenada de datos. Algunos ejemplos comunes de series temporales incluyen las ventas anuales en un supermercado, las temperaturas por hora a lo largo de un día, los precios de las acciones a lo largo de varios meses o los consumos de energía eléctrica residencial en los últimos años [13]. El estudio de las series temporales es diferente al del resto de las variables estadísticas porque el

interés reside habitualmente en la evaluación de sus cambios a lo largo del tiempo.

Una serie temporal se descompone en tres componentes: tendencia, estacionalidad y residuo. La tendencia da idea de cómo evoluciona una variable a lo largo del tiempo. La estacionalidad consiste en variaciones periódicas que se presentan en forma regular en la serie de tiempo. Finalmente el residuo, también llamado ruido son alteraciones sin periodicidad ni tendencia reconocible [7].

El análisis de series temporales consiste en examinar los datos disponibles para identificar patrones o tendencias. Este proceso extrae y modela las relaciones entre los datos a lo largo del tiempo, con el fin de predecir valores futuros basados en los registros históricos. Las predicciones se realizan considerando únicamente el comportamiento pasado de la serie (predicción autorregresiva) o incorporando variables externas adicionales [7].

III-B. Predicción

Una predicción consiste en estimar, de la manera más precisa posible, el comportamiento futuro de una variable de interés. Cuando esta variable está determinada por procesos físicos conocidos y modelables matemáticamente, la predicción se reduce a resolver las ecuaciones del modelo asociado [14]. Sin embargo, en muchos casos reales, como en la previsión de demanda eléctrica, el sistema presenta una elevada complejidad: intervienen numerosas variables, algunas de ellas desconocidas o difíciles de cuantificar, y existe una componente estocástica que introduce incertidumbre [15].

Ante este escenario, donde las relaciones causales no son explícitas pero se dispone de datos históricos, el modelado estadístico y las técnicas de aprendizaje automático emergen como herramientas válidas. Estos métodos identifican patrones, capturan dependencias no lineales y generan predicciones robustas a partir de los datos observados [15]. Los modelos estadísticos se clasifican como lineales o no lineales, univariantes o multivariantes, paramétricos o no paramétricos, estacionarios o no estacionarios, y la estimación de sus parámetros se realiza utilizando los datos provenientes de las observaciones disponibles [8]. Además, el avance de la tecnología, el abaratamiento de los dispositivos de almacenamiento y la velocidad de cálculo de los actuales procesadores han facilitado el resurgimiento de algoritmos provenientes del campo de la inteligencia artificial (IA) tales como el aprendizaje supervisado [16].

Cuando la predicción contiene un conjunto de probabilidades asociadas con todos los posibles resultados futuros, se está en presencia de una predicción probabilística. No se obtiene un resultado en particular sino la distribución esperada del resultado. Proporciona mayor información y permite construir intervalos de confianza dentro de los cuales se espera que caiga el valor predicho [17].

III-C. Métricas para Evaluación de Predicciones

Para poder comparar los diferentes modelos de predicción es necesario definir como evaluar la calidad de las mismas. Una métrica de predicción es una herramienta utilizada para evaluar la precisión y efectividad de un modelo predictivo. Estas métricas cuantifican la calidad de las predicciones realizadas por el modelo, permitiendo comparar diferentes modelos y seleccionar el más adecuado para una tarea específica.

Algunas métricas de predicción más comunes incluyen:

Error Medio (ME): Promedio de todos los errores de un conjunto de observaciones.

Error Absoluto Medio (MAE): Promedio de los valores absolutos de los errores.

Error Cuadrático Medio (MSE): Promedio de los cuadrados de los errores.

Raíz del Error Cuadrático Medio (RMSE): Raíz cuadrada del MSE, útil para interpretar el error en las mismas unidades que los datos originales.

Error Porcentual Absoluto Medio (MAPE): Promedio de los errores porcentuales absolutos, útil para comparar errores relativos

En este trabajo se han seleccionado tres de las métricas listadas: MAE, MSE y RMSE.

Error Absoluto Medio (MAE). Calcular el error medio conduce a la cancelación de los errores. La solución más directa es tomar el valor absoluto de la diferencia entre el valor real y_i y el valor predicho f_i con $\mathrm{i} = 1, \ldots, n$ [7]. La expresión matemática está dada por la ecuación 1:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|(y_i - f_i)| \tag{1}$$

Error Cuadrático Medio (MSE). Esta métrica no solo evita la cancelación de errores, sino que también penaliza más los errores de predicción. La expresión matemática está dada por la ecuación 2 [7]:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2 \tag{2}$$

Raíz Cuadrada del Error Cuadrático Medio (RMSE). Esta métrica es la desviación estándar del error y se expresa mediante:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f_i)^2} \tag{3}$$

El RMSE comparte las ventajas del MSE y se utiliza comúnmente en la predicción y análisis de las regresiones para verificar resultados experimentales. Además, tiene la ventaja de tener las mismas unidades que la variable predicha, por lo que es más fácil de interpretar directamente [7].

III-D. Modelos de Predicción

Los modelos de predicción tienen como objetivo estimar valores futuros de una variable de interés a partir de observaciones históricas y, cuando se dispone de ellas, de variables explicativas adicionales. En el contexto de la predicción del consumo eléctrico residencial, estos modelos suelen formularse como problemas de regresión, en los cuales el consumo se expresa como una función de valores pasados, variables exógenas, o una combinación de ambas.

Desde un punto de vista metodológico, los modelos de regresión utilizados para predicción se clasifican, de manera general, en modelos lineales y modelos no lineales, cada uno de los cuales presenta diferentes supuestos, capacidades de modelado y requerimientos computacionales.

Los modelos de regresión lineales asumen una relación lineal entre la variable dependiente y las variables explicativas. Estos modelos han sido ampliamente utilizados debido a su simplicidad, interpretabilidad y bajo costo computacional. Dentro de este grupo se incluyen métodos clásicos de series temporales, como las medias móviles y los modelos ARIMA, que se expresan mediante formulaciones lineales. Si bien estos enfoques resultan adecuados para capturar tendencias y dependencias temporales de corto plazo, presentan limitaciones para representar relaciones no lineales complejas y efectos de interacción, característicos del consumo eléctrico residencial.

Por otro lado, los modelos de regresión no lineales permiten una representación más flexible de la relación entre las variables. Esta categoría se subdivide en dos grandes grupos. El primero corresponde a modelos no lineales tradicionales y basados en datos, como los métodos basados en árboles de decisión y técnicas de ensamble, que permiten capturar dependencias no lineales y relaciones complejas manteniendo un adecuado equilibrio entre capacidad predictiva e interpretabilidad. El segundo grupo incluye los modelos basados en inteligencia artificial (IA), que abarcan técnicas de aprendizaje automático y aprendizaje profundo, tales como redes neuronales artificiales y modelos recurrentes. Estos enfoques son capaces de modelar patrones altamente complejos y dependencias temporales de largo plazo, aunque su aplicación práctica suele requerir grandes volúmenes de datos históricos, mayor capacidad computacional y procesos de ajuste más costosos.

En aplicaciones reales, la selección del modelo de predicción implica un compromiso entre precisión, robustez, interpretabilidad, disponibilidad de datos y complejidad computacional. En particular, para distribuidoras eléctricas que operan con registros mensuales de consumo y con información auxiliar limitada, resulta especialmente relevante el uso de modelos que ofrezcan un balance adecuado entre desempeño predictivo y factibilidad de implementación.

Sobre la base de este marco teórico, en la Sección IV-B se describen y evalúan los modelos de predicción específicos considerados en este trabajo, abarcando desde enfoques lineales clásicos hasta modelos no lineales basados en aprendizaje automático, y analizando su desempeño bajo restricciones reales de disponibilidad de datos.

IV. Implementación

La implementación de los modelos de predicción se llevó a cabo siguiendo un enfoque sistemático y reproducible, que abarca desde la preparación de los datos hasta la evaluación del desempeño y la cuantificación de la incertidumbre. Todas las etapas fueron desarrolladas en el lenguaje *Python*, utilizando bibliotecas ampliamente adoptadas en la comunidad científica, tales como *pandas*, *numpy*, *matplotlib*, *scikit-learn* y *statsmodels*.

IV-A. Preparación de los Datos

Los datos históricos de consumo eléctrico residencial y temperatura media mensual se almacenan en archivos en formato CSV y son cargados en estructuras *DataFrame* para facilitar su manipulación. Se realizó una inspección inicial con el objetivo de identificar valores faltantes, inconsistencias temporales y posibles errores de registro. Posteriormente, se construyó una variable de fecha a partir del año y el mes, asegurando la correcta ordenación cronológica de la serie temporal.

Para cada usuario residencial se extrajeron sus registros individuales y se definieron dos subconjuntos de datos: un conjunto de entrenamiento, que incluye los datos desde noviembre de 2018 hasta diciembre de 2023, y un conjunto de validación correspondiente al año calendario 2024. Esta partición temporal evalúa la capacidad predictiva de los modelos en datos no utilizados durante el entrenamiento, respetando la naturaleza secuencial de la serie.

IV-B. Modelos de Predicción Evaluados

Con el objetivo de comparar enfoques de distinta complejidad y fundamento teórico, se implementaron cinco modelos de predicción:

- **Medias móviles (MA):** se calcularon promedios móviles con ventanas de 3, 6, 12 y 18 meses. Cada predicción se obtuvo como el promedio de los consumos previos definidos por la ventana correspondiente.
- **ARIMA:** se ajustó un modelo ARIMA(1,1,1) utilizando exclusivamente la serie histórica de consumo. Esta configuración permite capturar dependencias autorregresivas de corto plazo, eliminar tendencias mediante diferenciación e incorporar un componente de media móvil.
- **Random Forest simple:** se entrenó un modelo de Random Forest Regressor utilizando únicamente la temperatura media mensual como variable explicativa. Este modelo se emplea como línea base para evaluar el aporte predictivo de las variables exógenas.
- **Random Forest con variables rezagadas:** se incorporaron como características adicionales el consumo del mes anterior (lag1) y el consumo del mismo mes del año anterior (lag12), junto con la temperatura. Estas variables capturan la autocorrelación temporal y la estacionalidad anual del consumo residencial.
- **Random Forest optimizado:** se aplicó un proceso de optimización de hiperparámetros mediante *RandomizedSearchCV*, manteniendo las mismas variables de entrada que en el modelo anterior.

IV-C. Optimización de Hiperparámetros

La optimización del modelo Random Forest se realizó explorando distintas combinaciones de hiperparámetros clave: número de árboles, profundidad máxima del árbol y número mínimo de muestras por hoja. Se utilizó *RandomizedSearchCV* con un total de diez combinaciones aleatorias, lo cual permite un balance adecuado entre exploración del espacio de búsqueda y costo computacional.

Se empleó validación cruzada con cinco pliegues. Esta elección se justifica por el tamaño relativamente reducido del conjunto de entrenamiento y por la necesidad de preservar un número suficiente de observaciones en cada partición para garantizar estimaciones estables del error. El uso de un mayor número de pliegues, como diez, reduciría el tamaño de los subconjuntos de entrenamiento, afectando negativamente la confiabilidad de la validación en series temporales cortas.

La métrica utilizada para seleccionar la mejor configuración fue el error cuadrático medio negativo. Esta elección se debe a que el MSE penaliza fuertemente los errores grandes, lo cual resulta particularmente relevante en aplicaciones de facturación eléctrica, donde desviaciones significativas generan reclamos. Además, el MSE es la métrica estándar empleada internamente por *scikit-learn* para problemas de regresión, y su versión negativa permite compatibilidad con procedimientos de maximización durante la búsqueda de hiperparámetros.

IV-D. Evaluación del Desempeño

El desempeño de todos los modelos se evaluó utilizando las métricas MAE, MSE y RMSE sobre el período de validación correspondiente al año 2024. Estas métricas permiten una comparación integral entre enfoques, considerando tanto la magnitud promedio del error como la penalización de errores extremos.

IV-E. Cuantificación de la Incertidumbre

Con el objetivo de incorporar una medida explícita de incertidumbre en las predicciones, se aplicó una técnica de remuestreo basada en *bootstrapping*. Para cada modelo se realizaron 1000 iteraciones de remuestreo sobre el conjunto de entrenamiento, generando distribuciones empíricas de las predicciones y de las métricas de error. A partir de estas distribuciones se estimaron intervalos de confianza del 95 %, tanto para los errores de validación como para la predicción del consumo correspondiente a enero de 2025. Este enfoque no solo obtiene un valor puntual estimado, sino también evalúa la confiabilidad de la predicción, aspecto fundamental para su uso como herramienta de apoyo en la validación de consumos y la toma de decisiones operativas. El flujo completo del procedimiento de implementación, entrenamiento, evaluación y estimación de incertidumbre se resume en la Fig. 1, garantizando la reproducibilidad del enfoque propuesto y facilitando su extensión a múltiples usuarios residenciales.

V. Experimentos y Resultados

En esta sección se presentan los experimentos realizados y los resultados obtenidos a partir de la aplicación de los distintos modelos de predicción descritos en las secciones anteriores. El objetivo principal es evaluar y comparar el desempeño de cada enfoque en un escenario operativo realista, caracterizado por datos de consumo mensual y disponibilidad limitada de información auxiliar.

V-A. Descripción del Conjunto de Datos

Los experimentos se realizaron utilizando datos correspondientes a un usuario residencial, con registros mensuales de consumo eléctrico desde noviembre de 2018 hasta diciembre de 2024. Adicionalmente, se dispone de la temperatura media mensual asociada a cada período. Los datos se almacenan en un archivo CSV que incluye, para cada registro, el identificador del usuario, el año, el mes, el consumo eléctrico, la temperatura y la fecha asociada.

La estructura del conjunto de datos se resume en la Tabla I. Si bien el análisis experimental presentado en este trabajo se centra en un único usuario, la metodología desarrollada es directamente extensible a un conjunto mayor de usuarios residenciales, como los registrados por las distribuidoras eléctricas de la provincia de Córdoba.

TABLA I
Estructura del Conjunto de Datos Históricos de Consumo Eléctrico y Temperatura Media Mensual para un Usuario Residencial (Período Noviembre 2018 – Diciembre 2024)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 148 entries, 0 to 147
Data columns (total 6 columns):
# Column Non-Null Count Dtype
--- ------ -------------------
0 usuario 148 non-null int64
1 año 148 non-null int64
2 mes 148 non-null int64
3 consumo 148 non-null int64
4 temperatura 148 non-null float64
5 fecha 148 non-null datetime64[ns]
dtypes: datetime64[ns](1), float64(1), int64(4)
memory usage: 7.1 KB datos.info()
```

V-B. Preparación y Validación de los Datos

Previo a la aplicación de los modelos de predicción, se realiza una inspección detallada del conjunto de datos con el fin de identificar valores faltantes, inconsistencias temporales y posibles errores de registro. Este proceso incluye tareas de limpieza, filtrado y corrección, garantizando la coherencia temporal de la serie y la calidad de los datos utilizados en los experimentos.

La validación de los modelos se lleva a cabo utilizando una partición temporal, reservando el año 2024 como conjunto de validación y emplea los datos anteriores como conjunto de entrenamiento. Esta estrategia evalúa el desempeño predictivo en datos no observados durante el entrenamiento, respetando la naturaleza secuencial de la serie temporal.

V-C. Resultados de los Modelos de Predicción

En primer lugar, se aplicaron modelos basados en medias móviles, utilizando ventanas de 3, 6, 12 y 18 meses. Las predicciones para el año 2024 se obtuvieron a partir del promedio de los consumos correspondientes a cada ventana. La Fig. 2 presenta la comparación entre los valores observados y los estimados para cada una de estas configuraciones. Se

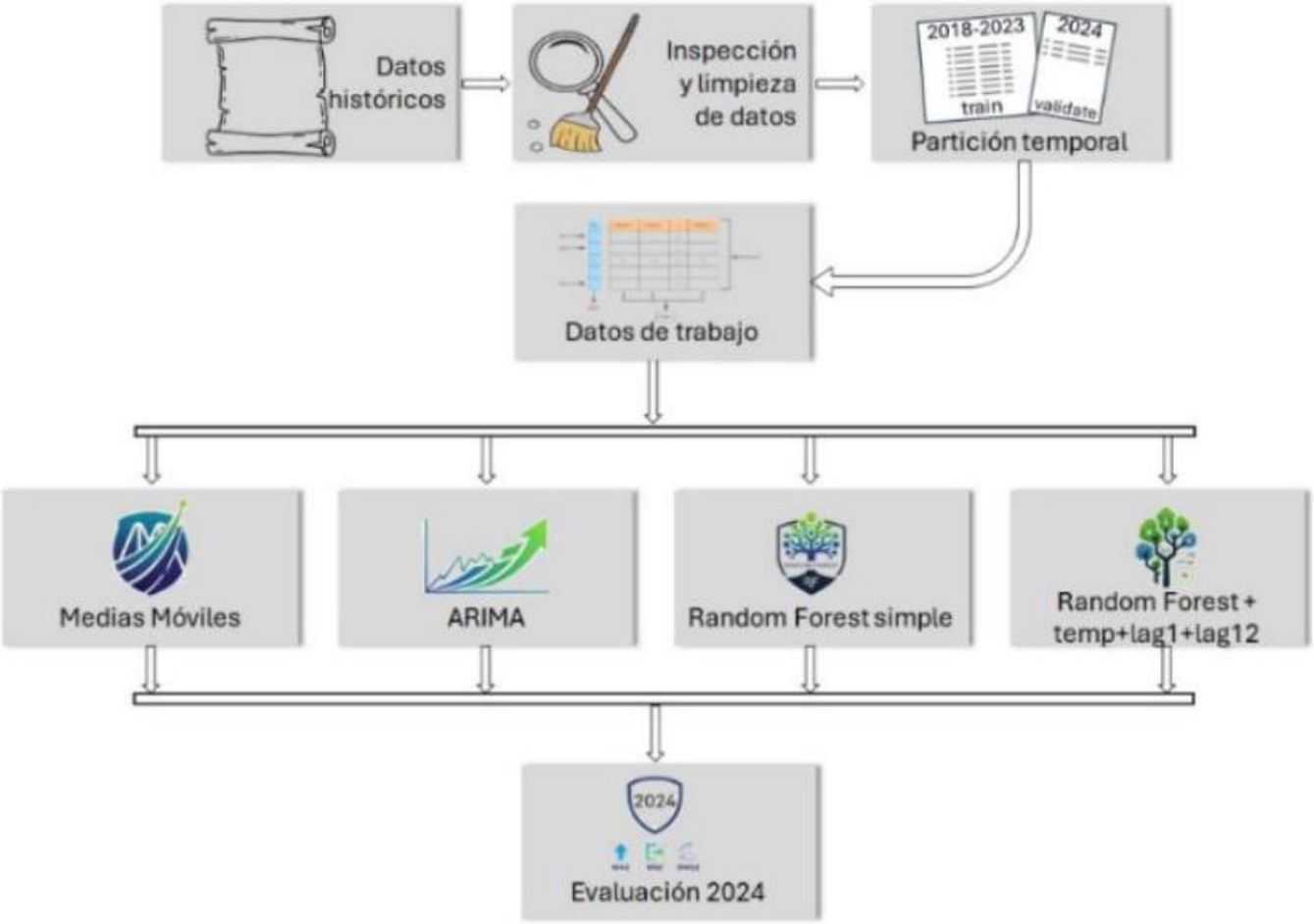


Fig. 1. Procedimientos aplicado para la elaboración y evaluación de predicciones de consumo.

observa que el modelo MA(12) logra capturar de manera más efectiva la estacionalidad anual del consumo, mientras que ventanas más cortas presentan mayor variabilidad y menor estabilidad predictiva.

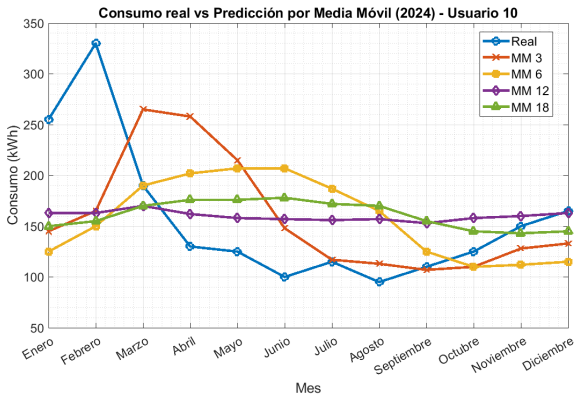


Fig. 2. Comparación de valores observados y estimados por MA(3), MA(6), MA(12), MA(18).

Posteriormente, se ajustó un modelo ARIMA(1,1,1) utilizando exclusivamente la serie histórica de consumo eléctrico. Este modelo permite capturar dependencias autorregresivas de corto plazo y eliminar tendencias mediante diferenciación. La Fig. 3 muestra los valores observados y estimados durante el período de validación. Si bien el modelo ARIMA mejora el desempeño respecto de las medias móviles, su capacidad predictiva se ve limitada por la ausencia de variables exógenas.

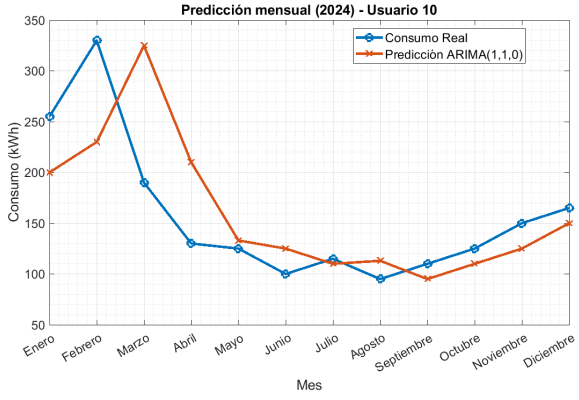


Fig. 3. Valores observados y estimados ARIMA(1,1,1).

Como línea base dentro de los modelos de aprendizaje automático, se implementó un modelo de Random Forest simple utilizando únicamente la temperatura media mensual como variable explicativa. Este experimento permite evaluar el poder explicativo de la temperatura por sí sola. Los resultados,

ilustrados en la Fig. 4, muestran que este enfoque no resulta adecuado para aplicaciones operativas, dado que el consumo eléctrico residencial presenta una fuerte autocorrelación temporal y patrones estacionales que no se obtienen únicamente a partir de la temperatura.

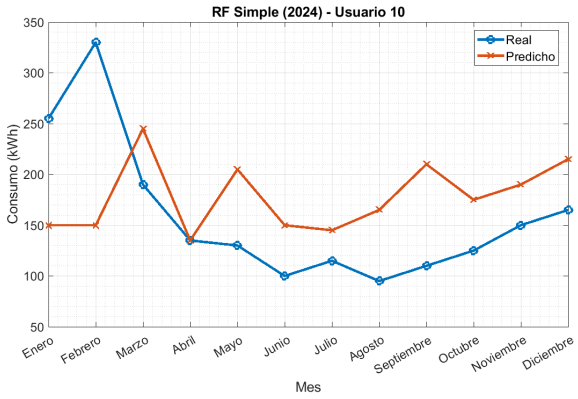


Fig. 4. Valores observados y estimados por RF simple.

Con el objetivo de incorporar información histórica relevante, se entrenó un modelo de Random Forest que incluye como variables de entrada la temperatura, el consumo del mes anterior (lag1) y el consumo del mismo mes del año anterior (lag12). Esta combinación captura tanto la dependencia temporal de corto plazo como la estacionalidad anual. La Fig. 5 muestra una mejora significativa en el ajuste del modelo, evidenciando el aporte de las variables rezagadas en la predicción del consumo. Finalmente, se evaluó un modelo

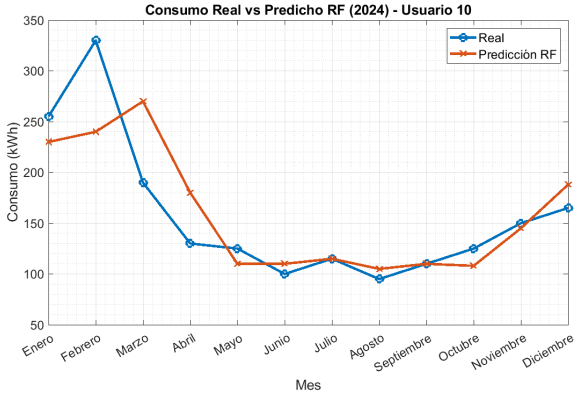


Fig. 5. Valores observados y estimados con RF con temperatura, consumo del mes anterior y consumo del mismo mes en el año anterior.

de Random Forest optimizado mediante búsqueda aleatoria de hiperparámetros. En este caso, se mantuvieron las mismas variables de entrada que en el modelo anterior, optimizando el número de árboles, la profundidad máxima y el número mínimo de muestras por hoja. La Fig. 6 presenta los resultados obtenidos, donde se observa el mejor ajuste entre los valores observados y estimados durante el período de validación.

V-D. Análisis de Errores

La Tabla II resume las métricas de error obtenidas para cada uno de los modelos evaluados. Los resultados muestran

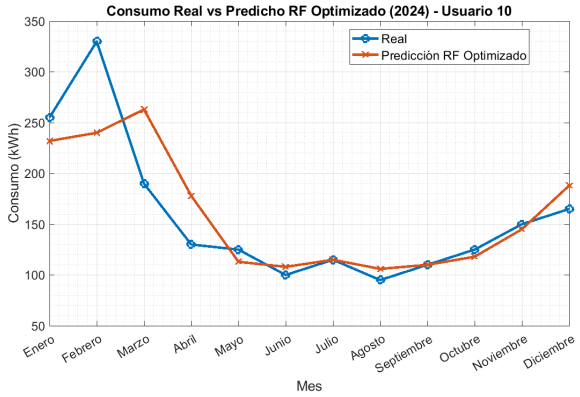


Fig. 6. Valores observados y estimados por RF con variable exógena, lags y optimización de hiperparámetros.

que, dentro de los modelos de medias móviles, la ventana de 12 meses presenta el menor error, lo cual es consistente con la periodicidad anual del consumo residencial. El modelo ARIMA reduce el error respecto de las medias móviles, evidenciando su mayor capacidad para modelar la dinámica temporal.

TABLA II
ERRORES DE PREDICCIÓN OBTENIDOS EN EL CONJUNTO DE VALIDACIÓN (AÑO 2024)

Método	MAE	MSE	RMSE
MA(3)	58.5	5409.5	73.6
MA(6)	68.7	7084.8	84.2
MA(12)	49.1	4122.9	64.2
MA(18)	56.4	4985.8	70.6
ARIMA	39.2	2869.0	53.6
RF simple	64.8	5623.8	74.5
RF c/LAGS	27.5	1515.0	38.9
RF/OPT	25.2	1364.0	36.9

El modelo de Random Forest simple presenta errores elevados, confirmando que la temperatura, por sí sola, no resulta un predictor suficiente del consumo eléctrico mensual. En contraste, la incorporación de variables rezagadas en el modelo de Random Forest produce una reducción significativa de los errores. El mejor desempeño se alcanza con el modelo de Random Forest optimizado, que obtiene el menor error absoluto medio (MAE = 25.2), confirmando la importancia del ajuste de hiperparámetros para mejorar la capacidad predictiva. Si bien el MAE es el usado para optimizar también es el que tiene menor MSE y RMSE

V-E. Comparación de Métodos

Desde una perspectiva comparativa, los resultados obtenidos indican que los modelos estadísticos tradicionales, si bien son simples y de bajo costo computacional, presentan limitaciones para capturar la complejidad del consumo eléctrico residencial. El modelo ARIMA representa una mejora respecto de las medias móviles, aunque su desempeño sigue siendo inferior al de los enfoques basados en aprendizaje automático.

Los modelos de Random Forest muestran un desempeño claramente superior, especialmente cuando se combinan

variables exógenas con información histórica relevante. La versión optimizada del modelo logra un balance adecuado entre precisión, robustez y costo computacional, lo que la convierte en una alternativa viable para su implementación en entornos operativos reales sin inversiones costosas en AMI y capacidad de cómputo.

Adicionalmente, la incorporación de técnicas de *bootstrapping* para la estimación de intervalos de confianza permite complementar las predicciones puntuales con una medida explícita de incertidumbre. Este aspecto resulta clave para aplicaciones como la validación de facturación, donde no solo interesa el valor estimado sino también el rango de variabilidad esperable.

VI. Conclusiones

En este trabajo se abordó el problema de la predicción del consumo eléctrico residencial bajo un escenario operativo realista, caracterizado por datos de consumo mensual, profundidad histórica limitada y disponibilidad reducida de información auxiliar. Este contexto es representativo de numerosas distribuidoras eléctricas que aún no cuentan con infraestructura de medición avanzada basada en medidores inteligentes.

Se realizó una comparación sistemática entre métodos estadísticos tradicionales, como medias móviles y modelos ARIMA, y enfoques de aprendizaje automático basados en Random Forest. Los resultados obtenidos muestran que, si bien los modelos simples presentan ventajas en términos de facilidad de implementación y bajo costo computacional, su capacidad predictiva es limitada cuando se enfrentan a patrones de consumo con estacionalidad y autocorrelación pronunciadas.

Dentro de los métodos estadísticos evaluados, el modelo de medias móviles con ventana de 12 meses fue el que obtuvo el mejor desempeño, lo cual resulta coherente con la periodicidad anual típica del consumo residencial. El modelo ARIMA logró mejorar los resultados respecto de las medias móviles, evidenciando una mayor capacidad para capturar la dinámica temporal de la serie, aunque su desempeño sigue siendo inferior al de los modelos basados en aprendizaje automático.

Los modelos de Random Forest demostraron un desempeño significativamente superior, especialmente cuando se incorporaron variables rezagadas que representan el consumo reciente y la estacionalidad anual. En particular, el modelo de Random Forest con optimización de hiperparámetros alcanzó el menor error de predicción, confirmando que el ajuste fino de estos parámetros mejora sustancialmente la capacidad predictiva sin requerir cambios en la estructura de los datos ni un aumento significativo de la complejidad computacional.

Un aporte relevante de este trabajo es la incorporación de técnicas de remuestreo mediante *bootstrapping* para la estimación de intervalos de confianza. Este enfoque complementa las predicciones puntuales con una medida explícita de incertidumbre, aspecto fundamental para aplicaciones operativas como la validación de facturación y la detección de consumos atípicos, donde no solo interesa el valor estimado sino también su confiabilidad.

Desde una perspectiva práctica, los resultados sugieren que el enfoque propuesto constituye una herramienta robusta, interpretable y escalable para distribuidoras eléctricas que operan con datos de baja resolución temporal. La metodología se extende fácilmente a un gran número de usuarios residenciales, contribuyendo a la reducción de errores de estimación y, potencialmente, a la disminución de reclamos por facturación incorrecta.

Como líneas de trabajo futuro, se propone incorporar variables exógenas adicionales, tales como características socioeconómicas del hogar o información sobre equipamiento eléctrico, que mejoran aún más la precisión de los modelos. Asimismo, se plantea la exploración de técnicas más avanzadas de aprendizaje automático, como Gradient Boosting y modelos de aprendizaje profundo, así como la validación del enfoque propuesto mediante su implementación en entornos reales de operación dentro de distribuidoras eléctricas, evaluando tanto el impacto técnico como la aceptación por parte de usuarios y empresas.

Referencias

[1] International Energy Agency, "Unlocking Smart Grid Opportunities in Emerging Markets and Developing Economies," IEA, Paris, 2023. [Online]. Available: https://www.iea.org/reports/unlocking-smart-grid-opportunities-in-emerging-markets-and-developing-economies

[2] World Bank, "Deployment of smart meters: Benefits and barriers," World Bank, Washington, DC, 2020. [Online]. Available: https://openknowledge.worldbank.org/handle/10986/33969

[3] H. Yu and C. Yao, "A low cost design of the rural intelligent meter reading system," in *Proc. 2014 Int. Conf. Future Computer and Communication Engineering (ICFCCE)*, Mar. 2014, pp. 123–126, doi: 10.2991/icfcce-14.2014.30.

[4] Y. L. Guan, B. Q. Chen, and X. Y. Liu, "Smart meter data analytics for accurate billing and fraud detection," *IEEE Trans. Smart Grid*, vol. 12, no. 4, pp. 3364–3373, Jul. 2021, doi: 10.1109/TSG.2020.3026512.

[5] L. M. Zeger, "Information reliability essential for use of smart grid DER behind the meter," *IEEE Smart Grid eBulletin*, Jul. 2022.

[6] P. Jokar, N. Arianpoo, and V. C. M. Leung, "Electricity theft detection in AMI using customers' consumption patterns," *IEEE Trans. Smart Grid*, vol. 7, no. 1, pp. 216–226, Jan. 2016, doi: 10.1109/TSG.2015.2425222.

[7] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 3rd ed. Melbourne, Australia: OTexts, 2021. [Online]. Available: https://otexts.com/fpp3/

[8] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ: Wiley, 2015.

[9] H. Hewamalage, C. Bergmeir, and K. Bandara, "Recurrent neural networks for time series forecasting: Current status and future directions," *Int. J. Forecasting*, vol. 37, no. 1, pp. 388–427, Jan.–Mar. 2021, doi: 10.1016/j.ijforecast.2020.06.008.

[10] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," *PLOS ONE*, vol. 13, no. 3, p. e0194889, Mar. 2018, doi: 10.1371/journal.pone.0194889.

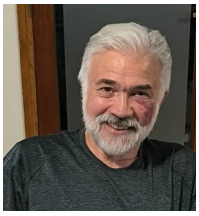
[11] J. Lago, F. De Ridder, and B. De Schutter, "Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms," *Appl. Energy*, vol. 221, pp. 386–405, Jul. 2018, doi: 10.1016/j.apenergy.2018.02.069.

[12] R. Kumar and A. Singh, "Residential electricity consumption prediction using hybrid CNN-BiLSTM and Random Forest model," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 1, pp. 55–63, 2024.

[13] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and Its Applications: With R Examples*, 4th ed. New York, NY: Springer, 2017.

[14] T. Hong, P. Pinson, and S. Fan, "Global energy forecasting competition 2012," *Int. J. Forecasting*, vol. 30, no. 2, pp. 357–363, Apr.–Jun. 2014, doi: 10.1016/j.ijforecast.2013.07.001.

[15] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond," *Int. J. Forecasting*, vol. 32, no. 3, pp. 896–913, Jul.–Sep. 2016, doi: 10.1016/j.ijforecast.2016.02.001.
[16] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks: The state of the art," *Int. J. Forecasting*, vol. 14, no. 1, pp. 35–62, Mar. 1998, doi: 10.1016/S0169-2070(97)00044-7.
[17] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aaa8415.
[18] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY: Springer, 2006.



Jose Luis Hernandez (Magíster en Ingeniería) received the Ingeniero Mecánico-Electricista degree from the Universidad Nacional de Río Cuarto (UNRC), Argentina, the M.Sc. degree in Data Networks from the Universidad Nacional de La Plata, Argentina, and is a Ph.D. candidate in Sciences from the Universidad Nacional de La Plata. He is currently a Professor and Director of the Master's Program in Engineering at the Facultad de Ingeniería, UNRC, with extensive experience in electrical and mechanical engineering education, technological innovation, and graduate thesis supervision. His research interests include data networks, engineering systems, technological needs in industry, and higher education methodologies. He has supervised multiple master's theses, contributed to academic panels on innovation, and participated in virtual defense committees during challenging times. Prof. Hernández is an active member of the UNRC academic community, fostering engineering development and student guidance initiatives.

David De Yong received the Ph.D. degree in Engineering from the Universidad Nacional de Río Cuarto (UNRC), Argentina. He was the Secretary of Postgraduate Studies at the Facultad de Ingeniería, UNRC, with over a decade of experience in academic administration, research supervision, and tutoring programs in engineering. His research interests include engineering education, student support systems, technological innovation in higher education, and collaborative academic projects. He has co-authored publications on tutoring trajectories and participated in institutional analyses of educational initiatives. Dr. de Yong is a dedicated contributor to the UNRC community, presiding over thesis defenses and promoting sustainable educational practices.

Fernando H. Magnago (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in Electrical Engineering from the University of Texas A&M, USA. He is currently a Principal Software Engineer at Resource Innovations (formerly Nexant) and a Professor at the Universidad Nacional de Río Cuarto (UNRC), Argentina, with more than 30 years of experience in power system analysis, optimization, and software development. His research interests include power system optimization, security assessment, fault analysis, state estimation, security-constrained unit commitment, and renewable energy integration. He has authored four books, multiple book chapters, over 30 journal papers, and more than 80 conference publications. Prof. Magnago is a Senior Member of IEEE, a former Chair of the IEEE PES Argentina Chapter, and an active contributor to international power system research and development initiatives.

Sergio Nicolás Bragagnolo has a PhD in Engineering Sciences at the FCEFyN of the UNC (2022). He is team research at the CIDTIEE of Cordoba Regional Faculty belonging to National Technological University. He has experience in the use of different software, in the design of transformer stations and electrical installations. His areas of interest are Smart Grids, Demand Management and Electrical Power Systems.

Juan Ignacio Amaya is an Electrical Engineer from the National Technological University. He is a researcher at the CIDTIEE of the Department of Electrical Engineering of the UTN-FRC. His areas of interest are modeling, control, and operation of electrical power systems.