
Pronóstico de consumo de energía eléctrica residencial de corto plazo utilizando algoritmos de aprendizaje automático y profundo

Short-Term Residential Electric Power Consumption Forecast Using Machine and Deep Learning Algorithms

César Aristóteles Yajure Ramírez

<https://orcid.org/0000-0002-3813-7606>

cyajure@gmail.com

Universidad Central de Venezuela. Venezuela

RECIBIDO: 31/10/2022 - ACEPTADO: 23/11/2022 - PUBLICADO: 30/12/2022

RESUMEN

El objetivo de esta investigación es crear modelos de predicción de consumo de energía eléctrica para el corto plazo, utilizando el algoritmo de regresión lineal múltiple y una red neuronal artificial. Es una investigación de tipo descriptiva-explicativa, y se usa la metodología de la ciencia de datos para alcanzar el objetivo planteado. Los modelos se entrenan y evalúan utilizando los datos de energía eléctrica facturada mensual en Uruguay, durante el período 2000-2022, además de las mediciones de temperatura ambiente y humedad relativa, así como la variación porcentual del índice de protección al consumidor. A través de un análisis de correlación y de un análisis de componentes principales, se verifica la importancia de las variables explicativas seleccionadas, previo a la obtención de los modelos. El modelo de regresión lineal tuvo un R^2 de 0,782 y el modelo de la red neuronal de 0,797, los residuos se distribuyeron normalmente para ambos modelos, y las métricas RMSE, MAE, y MAPE tuvieron valores similares también en ambos modelos. El modelo de la red neuronal tuvo un mejor desempeño, en comparación con el modelo de regresión lineal, en lo que respecta a la predicción de la energía eléctrica para los meses de julio y agosto del año 2022.

Palabras clave: Análisis de componentes principales, análisis exploratorio de datos, aprendizaje automático, aprendizaje profundo, red neuronal artificial, regresión lineal.

ABSTRACT

The objective of this research is to create prediction models for electric power consumption, in the short term, using the multiple linear regression algorithm and an artificial neural network. It is a descriptive-explanatory type of research, and the methodology of data science is used to achieve the stated objective. The models are trained and evaluated using monthly invoiced electricity data in Uruguay, during the period 2000-2022, in addition to measurements of ambient temperature and relative humidity, as well as the percentage variation of the consumer protection index. Through correlation analysis and principal component analysis, the importance of the selected explanatory variables is verified, prior to obtaining the models. The linear regression model had an R^2 of 0.782 and the neural network model of 0.797, the residuals were normally distributed for both models, and the RMSE, MAE, and MAPE metrics had similar values in both models as well. The neural network model had a better performance, compared to the linear regression model, regarding the prediction of electrical energy for the months of July and August of the year 2022.

Keywords: Principal component analysis, exploratory data analysis, machine learning, deep learning, artificial neural network, linear regression.

I. INTRODUCCIÓN

La demanda de energía eléctrica está asociada a la carga eléctrica en un punto determinado del sistema, durante un período de tiempo específico. Esta carga está relacionada con los aparatos que se conectan a la red, y que requieren de energía eléctrica para su funcionamiento. El tipo de aparato o carga depende a su vez del tipo de cliente, es decir, residencial, comercial, industrial, oficial, entre otros. En cualquier caso, para un buen desempeño del sistema eléctrico completo, se requiere del monitoreo de la demanda eléctrica, así como de su estimación en el tiempo, para efectos de operación y/o planificación.

En ese sentido, el objetivo de esta investigación es crear modelos para predecir el consumo residencial de energía eléctrica, en el corto plazo, utilizando algoritmos de “Machine Learning” y “Deep Learning”. Se comparan los resultados obtenidos por cada uno de los modelos, y se utiliza la energía eléctrica residencial facturada mensual como proxy del consumo residencial de energía eléctrica. Específicamente, se utilizan por separado, el algoritmo de regresión lineal múltiple, y una red neuronal artificial para crear los modelos que permitirán hacer la predicción del consumo de energía.

Durante la revisión de investigaciones previas relacionadas con la aquí presentada, se pudo notar el uso del algoritmo de regresión lineal múltiple y/o de la red neuronal artificial, en combinación con otros algoritmos, para efectos de comparar resultados de los modelos de predicción de energía eléctrica, pero por lo general con datos horarios, y utilizando datos de distintos tipos de usuarios. Por ejemplo, en Parhizkar et al. (2021) se generan modelos de predicción de consumo de energía eléctrica, utilizando el algoritmo de regresión lineal múltiple, además de otros cuatro algoritmos, para predecir el consumo de energía eléctrica a partir de datos horarios. Las métricas consideradas en el proceso de comparación fueron: R^2 , MSE, análisis de los residuos, y tiempo de ejecución. De igual forma, en su trabajo Rambabu et al. (2022) utilizan distintos algoritmos de machine learning, incluido el de regresión lineal múltiple, para predecir el consumo residencial de energía eléctrica. Las métricas usadas para comparar los distintos modelos son el R^2 , y el tiempo de ejecución en el entrenamiento, y en la evaluación del modelo. En su estudio Peña Acción (2019) desarrollan modelos basados en el algoritmo K-NN y red neuronal artificial, para pronosticar el consumo eléctrico horario de un prosumidor residencial ubicado en Alemania, definiéndose prosumidor como

un usuario productor de energía eléctrica a través de energías renovables. Las métricas consideradas para comparar los resultados fueron: RMSE, MAE, MAPE, y el coeficiente de correlación. Asimismo, Peña et al. (2019) plantean una metodología de predicción de consumo en el horario de máxima demanda, desarrollando un modelo con el uso del algoritmo de regresión lineal múltiple, y otro modelo basado en una red neuronal artificial. Utilizan los datos provenientes de los medidores de una de las sedes de la Universidad de Cienfuegos, es decir, utilizan los datos de sólo un usuario. Para realizar la comparación de los métodos consideran las métricas R^2 y MAPE. Por otra parte, en su investigación, Lozada et al. (2022) realizan la proyección de la demanda eléctrica mediante un modelo basado sólo en redes neuronales artificiales, y utilizando datos históricos de demanda máxima mensual. Consideran el error porcentual absoluto medio MAPE para la evaluación del modelo. Asimismo, Ruiz et al. (2016) desarrollan modelos basados en una red neuronal artificial tipo perceptrón, en el algoritmo M5Rules, y en el algoritmo de conjunto de árboles, para pronosticar el consumo de energía eléctrica en un hospital, tomando los datos de los medidores inteligentes de consumo diario de energía. Para comparar los modelos se utilizaron las métricas R^2 , MAE, MSD, MSE, y RMSE. Finalmente, Yajure Ramírez (2022) aplicó la metodología de la ciencia de datos para analizar la facturación de energía eléctrica en Uruguay durante el período 2000-2022. Entre los algoritmos que utilizó, se encuentra el de Regresión Lineal Múltiple, considerando sólo variables explicativas del sector eléctrico.

Ahora bien, el resto del artículo se organiza de la siguiente manera. En la sección 2 se presenta la metodología a utilizar, así como la descripción de los datos a utilizar. En la sección 3 se presentan y discuten los resultados obtenidos luego de la aplicación de los algoritmos, y se comparan los modelos generados por los mencionados algoritmos. Finalmente, en la sección 4 se presentan las conclusiones que se derivan de la investigación realizada.

II. METODOLOGÍA

En esta investigación se aplica la metodología de la ciencia de datos para predecir el consumo de energía eléctrica residencial en Uruguay, utilizando los datos históricos mensuales del período 2000 – 2022. Esta metodología consta de una serie de pasos, los cuales se ilustran en la Figura 1, la cual se basa en el trabajo desarrollado por Cielen (2016).

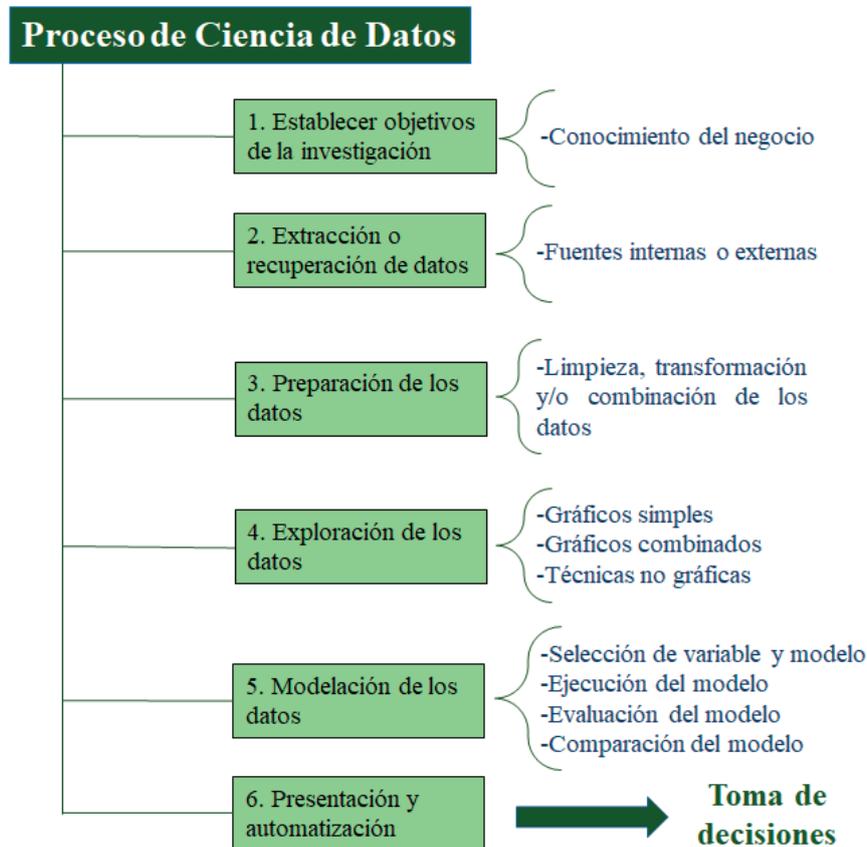
Tal como se plantea en la Figura 1, en primer lugar, debe tenerse al menos un conocimiento básico del negocio para así establecer el objetivo de la investigación. Posteriormente, deben obtenerse los datos que se utilizarán para llevar a cabo la investigación, pudiéndose obtener estos tanto de fuentes externas, como de fuentes internas. Luego de obtener los datos, debe hacerse un preprocesamiento de estos, lo cual incluye la detección e imputación de datos duplicados, datos faltantes, datos atípicos, entre otros. Adicionalmente, pudieran transformarse o combinarse los datos, de acuerdo con lo que sea requerido para un caso particular. El análisis exploratorio de los datos consiste principalmente en un análisis estadístico descriptivo de los mismos, ya sea por medio de gráficos y/o técnicas no gráficas, lo que permitirá tener una idea de la relación entre las variables, la estructura completa del conjunto, entre otras. En la etapa de la modelación, se aplican algoritmos de aprendizaje, automáticos y/o profundo, para obtener modelos que permitan, por ejemplo, hacer una predicción, determinar patrones, predecir clases, u otra necesidad. Finalmente, con los resultados obtenidos de la modelación, se

pueden tomar decisiones dentro del negocio. Aunque este proceso parece puramente secuencial, en la práctica pudiera haber realimentaciones desde cualquier etapa hacia el inicio del proceso.

En esta investigación, la etapa de modelación consistirá en la aplicación de un algoritmo de aprendizaje automático (Machine Learning) y otro de aprendizaje profundo (Deep Learning), para predecir el consumo de energía eléctrica residencial, y comparar sus resultados.

Ahora bien, debido a las etapas previamente mencionadas, se puede decir que el presente estudio tiene una parte descriptiva relacionada con el análisis exploratorio de los datos, y una parte explicativa relacionada con la etapa de modelación para la predicción del consumo de energía. Lo anterior concuerda con lo planteado por Arias (2012, p.23), quien indica que una investigación pudiera ubicarse en más de un tipo, que una investigación descriptiva busca caracterizar un hecho o fenómeno para definir su estructura, y que una investigación explicativa busca la razón de los hechos mediante las relaciones del tipo causa-efecto.

Figura 1
Etapas de la Ciencia de Datos



Construcción del conjunto de datos

El conjunto de datos a utilizar consta de 8 columnas y 270 filas. Las columnas representan las siguientes variables: energía eléctrica facturada mensual en megavatios-hora, en Uruguay, para los 270 meses que van desde enero del año 2000 hasta junio del año 2022 (E_RESID), año en que se factura la energía respectiva (ANHO), mes en que se factura la energía respectiva (MES), número de clientes en el mes y año respectivo (C_RESID), variación porcentual del índice de precios al consumidor en el año y mes respectivo (IPC), humedad relativa media en el año y mes respectivo (HUM_REL), temperatura máxima registrada en el año y mes respectivo (TEMP_MAX), y temperatura mínima registrada en el año y mes respectivo (TEMP_MIN).

Como se puede observar, para la investigación se consideran tanto variables económicas como climáticas, puesto que típicamente las variables explicativas del consumo de energía eléctrica residencial son de ese tipo. Es así como Fumo, & Rafe Biswas (2015) utilizan la temperatura exterior como una de las variables explicativas del consumo de energía residencial. De la misma forma, Peña-Guzmán, & Rey (2020) consideran como variables explicativas del consumo de energía eléctrica residencial al número de usuarios, ingresos por persona, precio del gas, producto interno bruto, y temperatura superficial.

Los datos de energía facturada residencial y número de clientes residenciales fueron tomados de la página web del Ministerio de Industria, Energía y Minería de Uruguay el día 04/09/2022 (Ministerio de Industria, Energía y Minería, 2022). Los datos de humedad relativa media, temperatura máxima, y temperatura mínima, fueron extraídos de la página web del Instituto Nacional de Investigación Agropecuaria de Uruguay el día 04/09/2022 (INIA, 2022). Es importante destacar, que los datos de temperatura y humedad corresponden a la estación ubicada en la ciudad de Montevideo, en la que se concentra la mayor cantidad de población en Uruguay. Por otra parte, los datos de IPC fueron extraídos de la página web del Instituto Nacional de Estadística de Uruguay el día 04/09/2022 (INE, 2022).

III. RESULTADOS Y DISCUSIÓN

Preparación de los datos

La limpieza y preparación de los datos consistió en aplicar las técnicas planteadas por McKinney (2018), utilizando el lenguaje de programación Python. Concretamente, se verificó la posible exis-

tencia de datos faltantes, y la posibilidad de duplicación de muestras (filas), resultando que en el conjunto de datos no se presenta ninguno de esas dos anomalías. Adicionalmente, se comprobó que las variables de cada una de las columnas tuvieran el formato adecuado. En ese sentido, las variables del año, mes, y número de clientes tienen el formato de número entero, mientras que las variables de energía eléctrica, IPC, humedad relativa y temperatura tienen el formato de número real.

Aunado a lo anterior, se procedió a detectar datos atípicos (outliers) aplicando la prueba de Tukey a cada una de las variables del conjunto de datos, pues según Moreno (2012) es uno de los métodos más utilizados para detección de outliers. En esta prueba se clasifican los outliers como “leves” y “graves”, los cuales se detectan a través del Diagrama de Caja y Bigotes (Box-Plot). Un outlier “grave” es aquel dato que está alejado una distancia de tres veces el rango intercuartil, del primer cuartil (hacia abajo) o del tercer cuartil (hacia arriba). Para este caso se detectaron outliers sólo para la variable IPC, tres outliers graves para el año 2002, tal como se muestra en la [Figura 2](#).

Específicamente, se detectaron para los meses de julio, agosto, y septiembre del año 2002. Es importante destacar que, durante el año 2002, Uruguay sufrió una grave crisis económica-financiera, que afectó el desempeño de todas las variables económicas, tal como lo menciona Antía (2003) en su investigación. Dado que fue una situación coyuntural, para efectos de esta investigación, se imputaron estos outliers con los valores promedios de IPC de los meses correspondientes, dentro del período de estudio. Los outliers que se observan en los otros años, cayeron dentro de la categoría de “leves”, y por lo tanto no fueron considerados.

Por último, al combinar las variables de temperatura máxima y temperatura mínima, se generó una nueva columna con la temperatura promedio y se agregó al conjunto de datos. De igual manera, la columna de la energía residencial se combinó con la columna de los clientes residenciales para obtener el consumo unitario en kilovatios-hora (kWh).

Análisis exploratorio de los datos

Primeramente, se obtuvo un resumen estadístico descriptivo del conjunto de datos, el cual se muestra en la Tabla 1. Se observa que en la mayoría de las variables se cumple que su valor promedio es cercano a su mediana. En la última fila de la Tabla 1 se observa la variación porcentual de la mediana con respecto a la media.

Figura 2

Box-Plot del IPC filtrado por año.

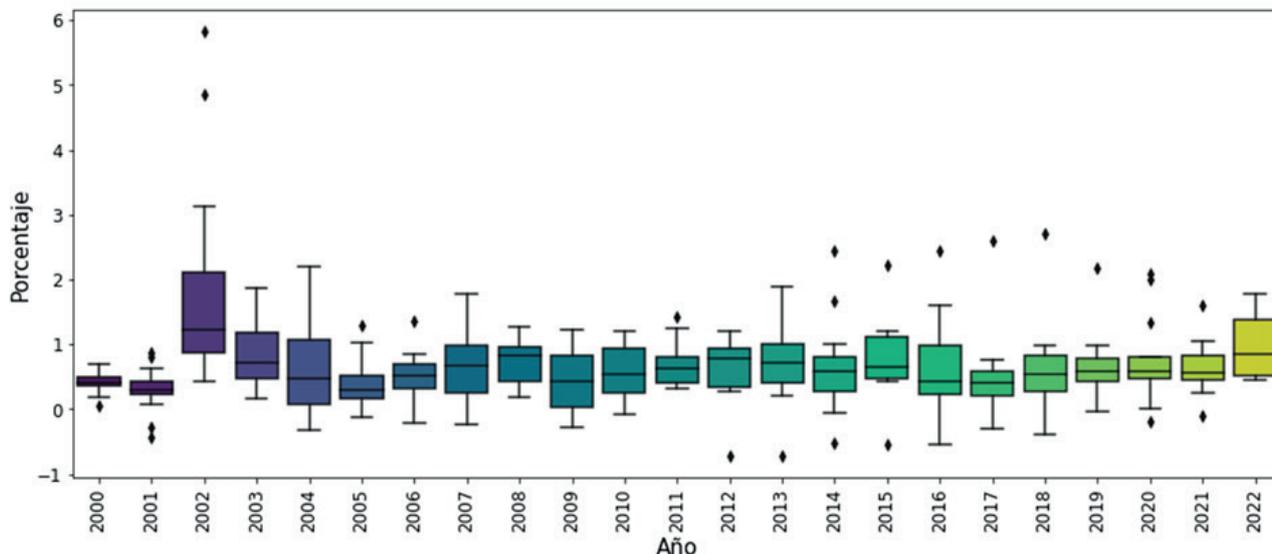


Tabla 1

Resumen estadístico descriptivo de las variables.

Parámetro	E_RESID	C_RESID	IPC	HUM_REL	TEMP_MAX	TEMP_MIN	TEMP_MED	CONS_UNIT
Media	268.663,81	1.205.927,96	0,64	77,40	22,25	11,34	16,79	222,06
DesvStd	43.391,29	113.166,66	0,56	8,32	5,00	4,19	4,55	22,67
Mínimo	189.283,00	1.050.359,00	-0,73	55,45	12,81	1,85	7,60	175,51
25%	233.449,25	1.104.407,25	0,32	71,28	17,78	7,62	12,85	205,68
50%	263.909,50	1.187.072,00	0,57	77,48	22,14	11,28	16,83	217,64
75%	296.735,75	1.297.668,00	0,91	83,25	26,68	15,22	21,16	237,17
Máximo	389.574,00	1.442.849,00	2,71	95,13	31,08	19,03	24,55	280,98
Var%	1,77%	1,56%	11,26%	-0,10%	0,49%	0,49%	-0,21%	1,99%

Destaca el IPC con una variación porcentual del 11,26%, mientras que para el resto de las variables es menor al 2% absoluto. También se puede observar de la Tabla 1 que la diferencia entre el valor mínimo del consumo de energía y su valor máximo, es mayor al valor mínimo, mientras que esa condición no se cumple para el número de clientes. Esto último pudiera indicar un aumento en el consumo de energía per cápita. En ese sentido, en la Figura 3 se presenta el consumo unitario promedio en kilovatios-hora, durante el período de estudio.

Se puede observar de la Figura 3 que el consumo unitario tuvo una tendencia descendente durante los cinco primeros años del período de estudio, y luego comenzó a recuperarse, tomando una tendencia ascendente desde el año 2005, con un valor máximo para el año 2016. Esto se confirma al

observar la Figura 4, la cual muestra el consumo residencial absoluto de energía eléctrica durante el período considerado, mostrando su tendencia ascendente continua a partir de alrededor del año 2005.

En cuanto a la estacionalidad mensual, se tiene que el mayor consumo unitario promedio ocurre durante los meses de junio a septiembre, tal como se observa en la Figura 5.

También se puede ver de la Figura 5 que el consumo unitario promedio durante el mes de enero es igualmente significativo. Estos resultados coinciden con el hecho que las menores temperaturas promedio ocurren precisamente durante los meses de junio a septiembre, mientras que el valor máximo de las temperaturas promedio ocurre durante el mes de enero. Es decir, los extremos en el consumo de

Figura 3

Consumo unitario de energía anual.

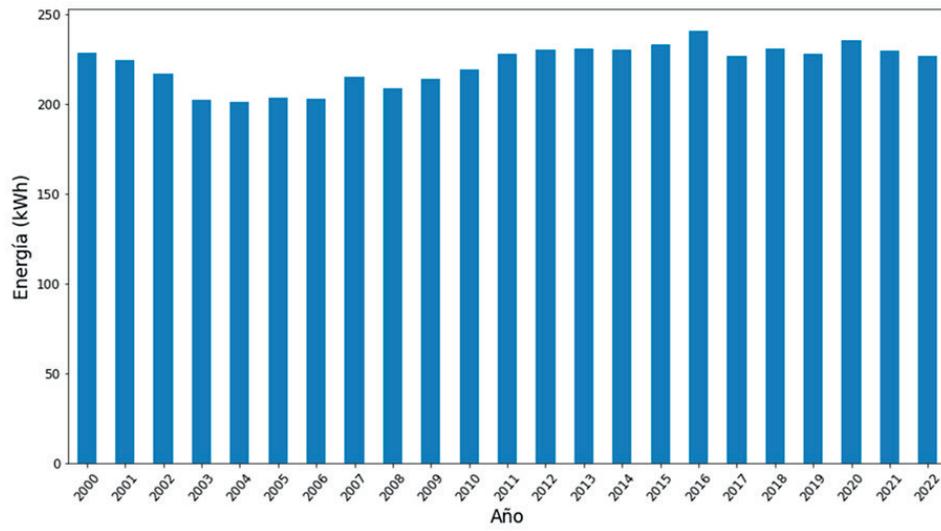


Figura 4

Consumo absoluto de energía.

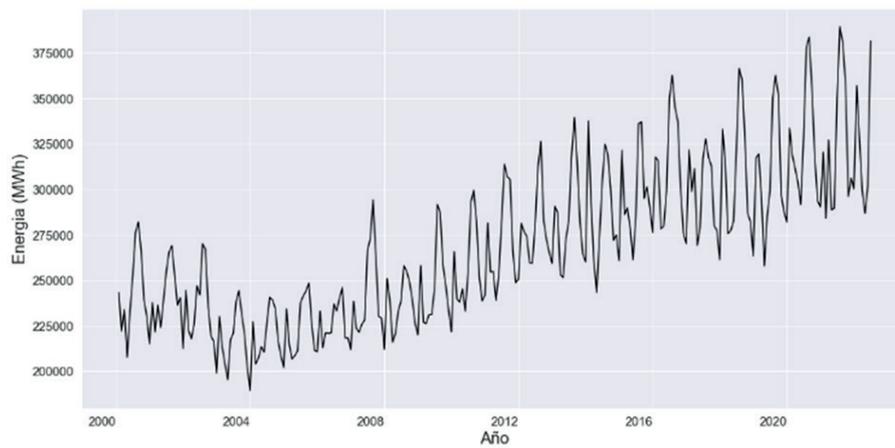
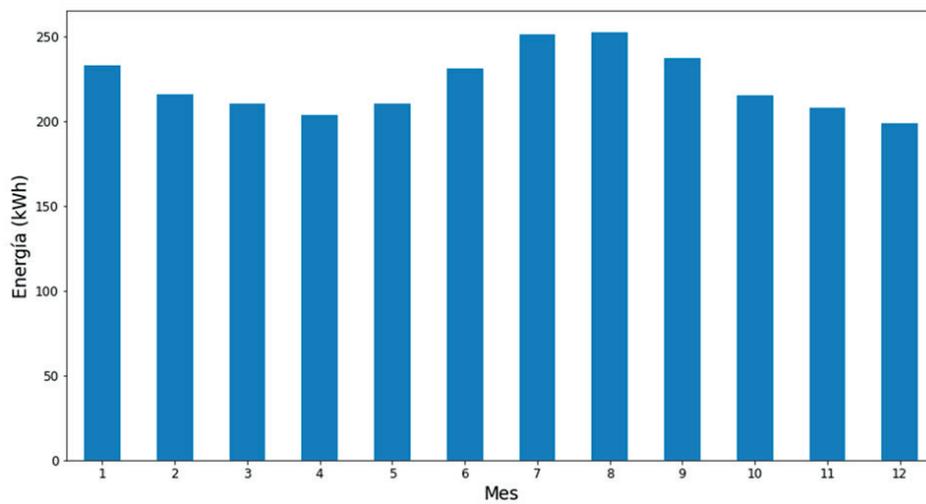


Figura 5

Consumo unitario de energía mensual.



energía coinciden con los valores extremos de la temperatura promedio.

Por otra parte, se hizo un análisis de correlación entre las variables del conjunto de datos, considerando los métodos de Pearson, Spearman, y Kendall. Según lo planteado por Amat (2022), el coeficiente de Pearson es adecuado para datos cuantitativos que están normalmente distribuidos, mientras que el estadístico Rho de Spearman o el estadístico Tau de Kendall son útiles cuando se requieren técnicas no paramétricas, es decir, cuando no se cumple la condición de normalidad de los datos. En todo caso, en esta investigación los resultados fueron similares al comparar los tres métodos, por lo que en la Figura 6 se presentan para el caso tradicional de Pearson.

De la Figura 6 se puede ver que las celdas con los colores más intensos corresponden a coeficientes más cercanos a 1 ó a -1. Es así como las variables de temperatura tienen una alta correlación directa entre ellas, como era de esperarse. También se observa una alta correlación directa del consumo de energía con la cantidad de clientes y con el consumo unitario. Asimismo, se observa una correlación negativa, relativamente alta, entre la humedad y las

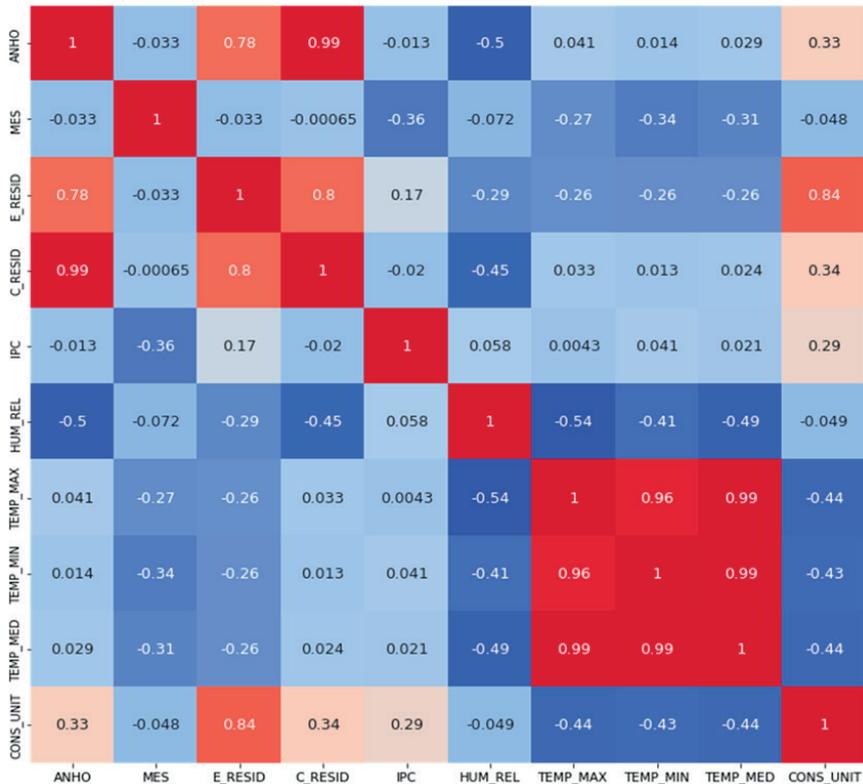
temperaturas, así como entre las temperaturas y el consumo unitario. Entonces, las variables de temperatura máxima y mínima se descartan por su alta correlación con la temperatura media, así como el consumo unitario y el año con la energía facturada.

Aplicación del algoritmo de Regresión Lineal Múltiple

El algoritmo de regresión lineal múltiple (RLM) es del tipo de aprendizaje automático supervisado. En este caso, el objetivo es generar un modelo de regresión que permita predecir el consumo de energía eléctrica residencial en Uruguay utilizando los datos mensuales disponibles para el período 2000 – 2022, teniendo como variables explicativas o regresoras al número de clientes residenciales, el índice de precios al consumidor, la humedad relativa, y la temperatura media. Todos los datos de las variables con una regularidad mensual.

En primera instancia, se lleva a cabo un análisis de componentes principales (PCA) para verificar la importancia de las variables seleccionadas para el modelo de regresión. Por medio de esta técnica se genera un nuevo conjunto conformado por

Figura 6
Matriz de correlación.



las componentes principales, las cuales son una combinación lineal de las variables del conjunto de datos original. Se busca que el grupo componentes tenga la mayor cantidad posible de información del conjunto de datos originales, y eso se mide con la proporción de varianza explicada por cada una de las componentes principales. Por ejemplo, en Dana et al. (2021) se utiliza PCA para eliminar las variables explicativas que eran irrelevantes, además del ruido.

Se aplica entonces PCA al conjunto de variables explicativas de nuestro caso. En la Figura 7 se presenta la varianza explicada por cada una de las componentes principales, de la cual se deduce que las primeras componentes principales son las que explican la mayor proporción de varianza. Se puede notar que en las primeras cinco componentes ya se explica el 100% de la varianza de los datos, por lo que se debe verificar si nuestras variables explicativas son significativas para esas componentes.

Se comprueba que el número de clientes es significativo para la primera componente, la temperatura media es significativa para la segunda, tercera, y cuarta componente, la humedad relativa es significativa para la tercera y cuarta componente, mientras que el IPC es significativo sólo para la quinta componente. Al ser significativas todas las variables explicativas para al menos una de las primeras cinco componentes principales, se consideran todas para el modelo de regresión.

Se aplica entonces el algoritmo de regresión. Dada la cantidad de muestras disponibles (270 filas), se tomó el 65% de estas para entrenar el modelo (175

filas), y el restante 35% (95 filas) para evaluar el modelo obtenido.

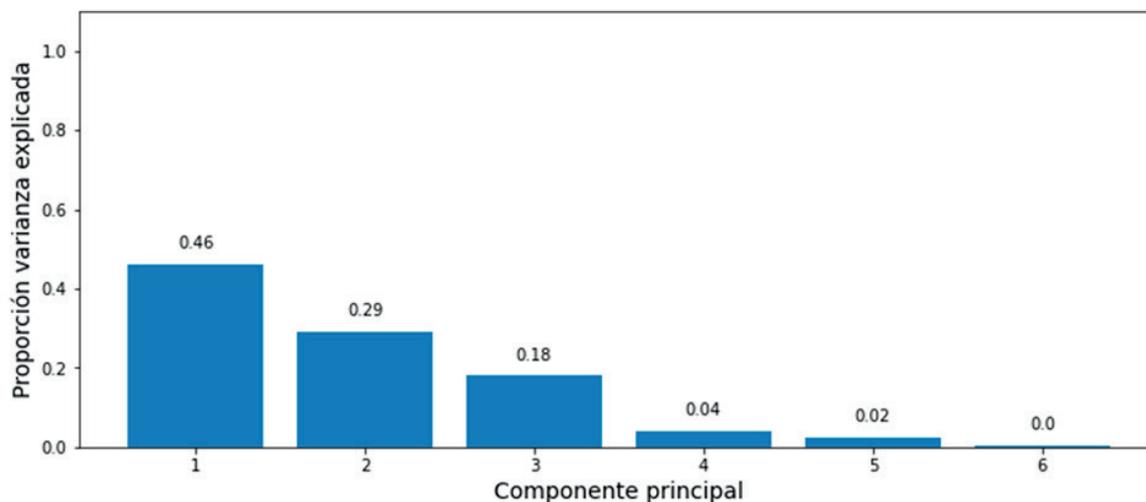
Se obtuvieron los coeficientes: 0,283, 10.791,7, -602,7, y -3.269, para las variables C_RESID, IPC, HUM_REL y TEMP_MED, respectivamente. De acuerdo con el p-valor obtenido para cada coeficiente, se puede decir que son estadísticamente significativos para un nivel del 5%.

Para la evaluación del modelo se utilizaron los indicadores: coeficiente de determinación R^2 , la raíz cuadrada de la media de los errores al cuadrado (RMSE), y la media del valor absoluto de los errores (MAE). Los dos últimos indicadores tienen las mismas unidades que la variable objetivo, es decir, MWh. El error se define como la diferencia entre el valor real de la variable objetivo, para una muestra determinada, y la predicción correspondiente. Por otra parte, Walpole et al. (2012) plantean que el R^2 es una medida de la calidad de ajuste del modelo, e indica la proporción de la variabilidad que es explicada por el modelo obtenido. Varía entre 0 y 1, un valor muy cercano a 1 indica que el ajuste del modelo es casi perfecto. Mientras que un valor cercano a 0 es indicativo de un ajuste deficiente del modelo.

De igual manera, se verificó la normalidad de los residuos aplicando la prueba de Shapiro-Wilk, puesto que según lo que plantean Flores Tapia & Flores Cevallos (2021) se ha convertido en la prueba preferida para detectar desviaciones de la normalidad, debido a su buen nivel de potencia. Se plantea la hipótesis nula de que los residuos provienen de una población normalmente distribuida, y el estadístico

Figura 7

Varianza explicada por componente.



utilizado varía entre 0 y 1. Si el p-valor es mayor al nivel de significancia del 5% (nivel considerado para este estudio) se concluye que no se puede rechazar la hipótesis nula. En la Tabla 2 se presentan los resultados de la evaluación.

Tabla 2
Indicadores modelo RLM.

Indicador	Valor Obtenido
R ²	0,782
RMSE	21.039,35
MAE	17.317,12
Prueba de Shapiro-Wilk a los residuos	
Estadístico	0,988
p-valor	0,540

De la Tabla 2 se puede ver que la calidad del ajuste del modelo es relativamente buena, con un 78,2%, y que el p-valor de la prueba de normalidad de los residuos es mayor al nivel de significancia del 5%, por lo que no se rechaza la hipótesis nula de que los residuos provienen de una distribución normal.

Aplicación de Red Neuronal Artificial

Una red neuronal artificial (RNA) se clasifica como algoritmo de aprendizaje profundo (Deep Learning), y según como se fijen sus parámetros, se puede utilizar para predecir el valor de una variable explicativa continua. En el presente estudio se utilizó una RNA para predecir el consumo de energía eléctrica residencial en Uruguay utilizando los datos mensuales disponibles para el período 2000 – 2022. Se utilizan las mismas variables explicativas del punto anterior.

Al ser relativamente bajo el número de muestras, se recomienda tener un número bajo de capas. La RNA utilizada tiene 3 capas, una de entrada, una oculta, y una de salida. La primera capa, que es de tipo densa, recibe los valores de entrada, a los que se les aplica una función de activación, para obtener la salida de esa capa. Se dice que es una capa densa porque está densamente conectada, es decir, completamente conectada. Para esta capa se aplica la función de activación lineal rectificadora (ReLU), la cual elimina los valores negativos. La capa oculta también es densa, y con el mismo tipo de función de activación. La capa salida si se diferencia de las dos primeras en el sentido que tiene una función de activación de tipo lineal, para no limitar los valores de la predicción. En la investigación de Sun et al. (2021), utilizaron una red de cuatro capas, teniendo una de entrada, una de salida, y dos ocultas.

Como función de pérdida se utiliza el error absoluto medio (MAE), cuya tarea es medir que tanto se aleja la salida de la RNA del valor esperado, es decir, compara la predicción con el valor real y calcula un “score” de su distancia. En el lazo de entrenamiento, esa comparación se repite varias veces y cada vez se van acercando los valores debido a que en cada iteración se ajustan los pesos asignados a cada una de las entradas. El ajuste de los pesos se realiza a través del optimizador. Para este caso se utilizó el optimizador de propagación de raíz cuadrática media (RMSProp), y como métricas se utilizan el MSE y el error absoluto medio (MAE).

Las 270 muestras se dividen en 65% (175 muestras) para el entrenamiento de la RNA, y el 35% restante (95 muestras) para la prueba de la RNA. Si embargo, el set de prueba se divide a su vez en dos partes iguales, una para la etapa de validación, y la otra parte para la prueba.

En la Tabla 3 se muestran los resultados de la evaluación de la RNA obtenida. Se puede observar una buena calidad de ajuste del 79,7%, así como también se observa del análisis de los residuos que el p-valor es mayor al nivel de significancia del 5%, lo que indica que no se puede rechazar la hipótesis nula de que provienen de una distribución normal.

Tabla 3
Indicadores RNA.

Indicador	Valor Obtenido
R ²	0,797
RMSE	19.840,29
MAE	15.667,11
Prueba de Shapiro-Wilk a los residuos	
Estadístico	0,985
p-valor	0,793

Comparación de resultados modelos RLM y RNA

Tanto en el caso del modelo RLM, como en el caso del modelo de la RNA, se obtuvieron valores aceptables del coeficiente R², con una mayor calidad del ajuste para la RNA. Por otra parte, los indicadores de error RMSE y MAE son similares para ambos casos, pero de nuevo la RNA tiene los valores más bajos. Se incorporó el error porcentual absoluto medio (MAPE), el cual como para los otros indicadores, mientras más bajo sea mejor los resultados. El modelo obtenido con RNA tuvo error porcentual de 5,92%, y para el modelo RLM el error porcentual fue de 6,25%.

En cuanto al análisis de los residuos, como se mencionó previamente, en ambos casos se aplicó la prueba de normalidad de Shapiro-Wilk. De acuerdo con esta prueba se puede decir que tanto para el modelo RLM como para el modelo RNA, los residuos provienen de una distribución normal, dado que el estadístico es cercano a uno, y además el p-valor es mayor al nivel de significancia del 5%.

Adicionalmente, se utilizaron los modelos para generar las predicciones de energía eléctrica residencial para los meses de julio y agosto del año 2022. En la Tabla 4 se muestran los resultados obtenidos, junto con los valores reales de energía eléctrica en MWh.

Tabla 4

Predicciones Julio – Agosto 2022.

Valores	Julio 2022	Agosto 2022
Valor Real Consumo (MWh)	404.066,80	391.854,90
Predicción (MWh) - RLM	350.919,29	355.635,89
Error porcentual	13,15%	9,24%
Predicción (MWh) - RNA	382.369,97	388.530,19
Error porcentual	5,37%	0,85%

De la tabla 4 se puede ver que para ambos modelos el error porcentual del mes de julio fue mayor que el correspondiente al mes de agosto. Adicionalmente, se puede decir que el modelo RNA arroja mejores predicciones que el modelo RLM, para los dos meses. Incluso, los errores porcentuales de predicción del modelo RLM están muy por encima que el error porcentual promedio obtenido de los valores de prueba (6,25%), mientras que para el modelo RNA los errores porcentuales de predicción están por debajo del error porcentual promedio obtenido de los valores de prueba (5,92%).

CONCLUSIONES

Tanto el consumo unitario, como el consumo absoluto de energía eléctrica de los clientes residenciales aumenta continuamente a partir del año 2005. El consumo unitario presenta una estacionalidad mensual, en la que los meses de mayor consumo corresponden a los meses de junio a septiembre durante el invierno, y el mes de enero durante el verano.

Desde el punto de vista del análisis de componentes principales, la variable explicativa menos significativa es el IPC, puesto que es importante sólo para la quinta componente principal, la cual es la que explica menos la varianza del conjunto de da-

tos completos. Siendo el número de usuarios residenciales, la variable explicativa más importante de todas.

El modelo obtenido al aplicar el algoritmo de regresión lineal múltiple tiene una calidad de ajuste del 78,2%, con un error porcentual absoluto del 6,25%, y cuyos residuos provienen de una población distribuida normalmente. El modelo se entrenó con el 65% de los datos, y se evaluó con el restante 35%.

En cuanto al modelo obtenido al aplicar la red neuronal artificial, este tiene una calidad de ajuste del 79,7%, con un error absoluto porcentual de 5,92%. Los residuos provienen de una población normalmente distribuida. El modelo se entrenó con el 65% de los datos, y se evaluó con el restante 35%.

Se puede decir que el modelo obtenido de la RNA tiene un mejor desempeño que el modelo obtenido del algoritmo de RLM, en cuanto a la calidad del ajuste, y de los indicadores de error RMSE. MAE y MAPE. Adicionalmente, la predicción del consumo de energía para los meses de julio y agosto del año 2022, obtenida con el modelo RNA, es más cercana a los valores reales, al compararla con la predicción obtenida con el modelo RLM.

REFERENCIAS

- [1] Amat Rodrigo, J. (15 de 07 de 2022). *Ciencia de Datos, Estadística, Machine Learning y Programación*. (Joaquin Amat Rodrigo) Recuperado el 16 de Julio de 2022, de <https://www.cienciadedatos.net/documentos/pystats05-correlacion-lineal-python.html>
- [2] Antía, F. (2003). Uruguay 2002: contagio, crisis bancaria y perspectivas. *Revista ICONOS*, 145-153.
- [3] Arias, F. G. (2012). *El Proyecto de Investigación - Introducción a la metodología científica*. Caracas: Editorial EPISTEME, C.A.
- [4] Cielen, D., Meysman, A., & Ali, M. (2016). *Introducing Data Science*. Shelter Island, NY: Manning Publications Co.
- [5] Dana, R. D., Soilihudin, D., Silalahi, R. H., Kurnia, D., & Hayati, U. (2021). Competency test clustering through the application of Principal Component Analysis (PCA) and the K-Means algorithm. *IOP Conf. Series: Materials Science and Engineering 1088 (2021) 012038*. Cirebon. doi:10.1088/1757-899X/1088/1/012038.
- [6] Flores Tapia, C. E., & Flores Cevallos, K. L. (2021). Pruebas para comprobar la

- normalidad de datos en procesos productivos: Anderson Darling, Ryan-Joiner, Shapiro-Wilk y Kolmogórov-Smirnov. *Societas. Revista de Ciencias Sociales y Humanísticas*, 23(2).
- [7] Fumo, N., & Rafe Biswas, M. (2015). Regression analysis for prediction of residential energy consumption. *Renewable and Sustainable Energy Reviews*, 332-343. <https://doi.org/10.1016/j.rser.2015.03.035>.
- [8] Instituto Nacional de Estadística del Uruguay. (04 de 09 de 2022). Obtenido de INE Uruguay: <https://www.ine.gub.uy/web/guest/indicadores?indicadorCategoryId=11421>
- [9] Instituto Nacional de Investigación Agropecuaria de Uruguay. (04 de 09 de 2022). Obtenido de INIA Uruguay: <http://www.inia.uy/#>
- [10] Lozada Cabrera, C., Acosta Apolo, N., Paredes Moran, D., & Vique Salazar, G. (2022). Predicción de demanda eléctrica utilizando redes neuronales artificiales para un sistema de distribución de energía eléctrica. *Ciencia Latina Revista Científica Multidisciplinar*, 3871-3894. https://doi.org/10.37811/cl_rcm.v6i3.2502.
- [11] McKinney, W. (2018). *Python for Data Analysis*. Sebastopol, CA: O'Reilly Media, Inc.
- [12] Ministerio de Industria, Minería y Energía de Uruguay. (2022). Recuperado el 18 de agosto de 2022, de <https://www.gub.uy/ministerio-industria-energia-mineria/datos-y-estadisticas/datos/series-estadisticas-energia-electrica>
- [13] Moreno Castellanos, J. G. (2012). *Método de detección temprana de outliers*. Bogotá: Trabajo de Grado, Pontificia Universidad Javeriana.
- [14] Parhizkar, T., Rafieipour, E., & Parhizkar, A. (2021). Evaluation and improvement of energy consumption prediction models using principal component analysis based feature reduction. *Journal of Cleaner Production*. <https://doi.org/10.1016/j.jclepro.2020.123866>.
- [15] Peña Acción, J. A., Viego Felipe, P. R., Gómez Sarduy, J. R., & Padrón Padrón, A. (2019). Pronóstico del consumo pico para la gestión energética de la Universidad de Cienfuegos. *UNIVERSIDAD Y SOCIEDAD - Revista Científica de la Universidad de Cienfuegos*, 220-228.
- [16] Peña-Guzmán, C., & Rey, J. (2020). Forecasting residential electric power consumption for Bogotá Colombia using regression models. *Energy Reports*, 561-566. <https://doi.org/10.1016/j.egy.2019.09.026>.
- [17] Rambabu, M., Ramakrishna, N., & Polamarasetty, P. (2022). Prediction and Analysis of Household Energy Consumption by Machine Learning Algorithms in Energy Management. *International Conference on Environment, Renewable Energy and Green Chemical Engineering (EREGCE 2022)*. E3S Web of Conferences. <https://doi.org/10.1051/e3sconf/202235002002>.
- [18] Ruiz, E., Pacheco-Torres, R., & Casillas, J. (2017). Energy Consumption Modeling by Machine Learning from Daily Activity Metering in a Hospital. *2017 22nd IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*. Limassol, Cyprus: IEEE. DOI: 10.1109/ETFA.2017.8247667.
- [19] Sun, J., Dong, H., Gao, Y., Fang, Y., & Kong, Y. (2021). The Short-Term Load Forecasting Using an Artificial Neural Network Approach with Periodic and Nonperiodic Factors: A Case Study of Tai'an, Shandong Province, China. *Computational Intelligence and Neuroscience*, 1-8. <https://doi.org/10.1155/2021/1502932>.
- [20] Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). *Probabilidad y estadística para ingeniería y ciencias*. Ciudad de México: Pearson Educación de México, S.A. de C.V.
- [21] Yajure Ramírez, C. A. (2022). Aplicación de la metodología de Ciencia de Datos para analizar datos de facturación de energía eléctrica. Caso de estudio: Uruguay 2000-2022. *Revista de investigación de sistemas e informática*, 127-138. DOI: <https://doi.org/10.15381/risi.v15i1.23544>.

Fuentes de financiamiento:

Propia.

Conflictos de interés:

El autor declara no tener conflictos de interés.

Contribución de los autores

El autor desarrolló el preprocesamiento de los datos, el análisis exploratorio, así como los modelos de regresión lineal múltiple, y el de la red neuronal artificial. Asimismo, utilizó los modelos para realizar las predicciones de energía eléctrica residencial de los meses de julio y agosto del año 2022, y redactó las conclusiones. De igual manera, participó en la redacción y revisión final del artículo.