Revista de investigación de sistemas e informática
15(1): 127-138 (2022)
DOI: https://doi.org/10.15381/risi.v15i1.23544

ISSN: 1815-0268 (Impreso) / ISSN: 1816-3823 (Electrónico)
Facultad de Ingeniería de Sistemas e Informática - UNMSM

Aplicación de la metodología de Ciencia de Datos para analizar datos de facturación de energía eléctrica. Caso de estudio: Uruguay 2000-2022

Application of the Data Science methodology to analyze electricity billing data. Case study: Uruguay 2000-2022

César Aristóteles Yajure Ramírez
https://orcid.org/0000-0002-3813-7606
cyajure@gmail.com
Universidad Central de Venezuela, Venezuela

Recibido: 20/09/2022 - Aceptado: 15/09/2022 - Publicado: 20/09/2022

RESUMEN

La presente investigación tiene como objetivo analizar los datos de facturación de energía eléctrica en Uruguay durante el período 2000-2022, utilizando algoritmos de aprendizaje automático. Es una investigación de tipo descriptiva-explicativa, y se utiliza la metodología de la Ciencia de Datos para alcanzar el objetivo planteado. Se maneja el algoritmo K-Means para agrupar los datos de acuerdo con los tipos de clientes, el algoritmo K-NN para generar un modelo que permite predecir el tipo de cliente de los nuevos registros, la técnica PCA para reducir la dimensionalidad de los datos previo a la aplicación del algoritmo de Regresión Lineal para obtener un modelo para predecir la energía eléctrica total facturada de los nuevos registros. El modelo obtenido con K-Means generó un clúster para cada tipo de clientes, agrupando perfectamente los datos. El modelo obtenido a través de K-NN, permitió predecir si el cliente era de tipo residencial o no residencial, con una exactitud del 100%. Combinando el análisis de correlación con el análisis PCA, se redujo la dimensionalidad hasta obtener sólo tres variables explicativas. El modelo de regresión lineal tuvo un alto coeficiente de determinación R^2 de 0,981, y los residuos se distribuyeron normalmente.

Palabras clave: Análisis de componentes principales; análisis exploratorio de datos; algoritmo K-Means; algoritmo K-NN; aprendizaje automático; regresión lineal.

ABSTRACT

The objective of this research is to analyze the electricity billing data in Uruguay during the period 2000-2022, using machine learning algorithms. It is a descriptive-explanatory type of research, and the Data Science methodology is used to achieve the stated objective. The K-Means algorithm is used to group the data according to the types of clients, the K-NN algorithm to generate a model that allows predicting the type of client of the new records, the PCA technique to reduce the dimensionality of the data, prior to the application of the Linear Regression algorithm to obtain a model to predict the total electric energy billed from the new records. The model obtained with K-Means generated a cluster for each type of customer, perfectly grouping the data. The model obtained through K-NN made it possible to predict whether the client was residential or non-residential, with 100% accuracy. Combining the correlation analysis with the PCA analysis, the dimensionality was reduced until only three explanatory variables were obtained. The linear regression model had a high coefficient of determination R^2 of 0.981, and the residuals were normally distributed.

Keywords: Principal Components Analysis; Exploratory Data Analysis; K-Means Algorithm; K-NN Algorithm; Machine Learning; Linear Regression.

I. INTRODUCCIÓN

Las políticas energéticas definidas por los estados a través de sus entes reguladores deben ser ejecutadas por los organismos operativos, los cuales deberán recoger la data necesaria para verificar el buen desempeño de dichas políticas. Una de las variables importantes a tomar en cuenta para este seguimiento es el de la facturación de la energía eléctrica por tipo de cliente, el cual se puede tomar como variable proxy del consumo de energía eléctrica. Esta variable nos permitirá observar las curvas de carga por tipo de cliente, los puntos de mayor consumo, la tendencia del consumo por tipo de cliente, entre otras características de utilidad.

En ese sentido, el objetivo de la presente investigación es aplicar la metodología de la ciencia de datos a la facturación de energía eléctrica por tipo de cliente de Uruguay para el período 2000-2022, haciendo uso de un análisis exploratorio y de algoritmos de aprendizaje automático, tanto supervisado como no supervisado, con el fin de determinar sus características subyacentes, agrupar los datos en clústers, predecir la clase de nuevos registros, y predecir la energía total, a través de modelos generados para ello. En concreto, se hace uso del algoritmo K-Means para generar grupos de datos con características similares, el algoritmo de K vecinos más cercanos (K-NN) para predecir la clase de los registros nuevos, el análisis de componentes principales (PCA) para reducir la dimensionalidad de los datos previo a la aplicación del algoritmo de Regresión Lineal para predecir la energía total facturada.

Haciendo una búsqueda sobre trabajos similares al aquí presentado, se encontró que la mayoría de éstos estaban enfocados a casos con datos de consumo de energía de tipo horario, al tipo de cliente residencial, y al uso del algoritmo K-Means para definir perfiles de usuario. Es así como en Lester et al. (2021) utilizan el algoritmo de aprendizaje no supervisado K-Means, para clasificar perfiles de clientes residenciales del sur de Chile, utilizando los datos provenientes de los medidores inteligentes instalados en los hogares de estos clientes. Del conjunto de 1179 medidores inteligentes, obtuvieron dos clústers, uno de alto consumo, y otro de consumo menor. Asimismo, en Nepal et al. (2019) los autores analizaron el patrón de consumo de energía eléctrica en un edificio utilizando el algoritmo K-Means. Para la selección del número de clústers hacen uso del método del percentil y de la función de distribución acumulada de los centroides aleatorios que se generan en cada corrida del algoritmo. Identificaron seis clústers con una exactitud del 89,3%. De igual manera, en Walker et al. (2019) aplican una serie de algoritmos de machine learning, así como algoritmos de redes neuronales, para predecir el consumo horario de energía en edificios comerciales. Concluyen que los algoritmos: bosques aleatorios, árbol repotenciado, y las redes neuronales, proporcionan los mejores resultados en la predicción. Igualmente, en Cheol et al. (2020) se determinan perfiles de carga de consumo de energía utilizando el algoritmo K-Means, para usuarios residenciales. Obtienen seis tipos de usuarios de acuerdo con su consumo de energía eléctrica.

Además, en Parhizkar et al. (2021) los autores utilizan los algoritmos de regresión lineal, regresión con soporte vectorial, árboles de regresión, bosques aleatorios, y K vecinos más cercanos, para predecir el consumo de energía eléctrica a cuatro casos de estudio. Previamente hicieron un preprocesamiento de los datos utilizando el análisis de componentes principales. Compararon los resultados basados en los criterios de tiempo de ejecución, R^2, y comportamiento de los residuos. Finalmente, en Hosseini et al. (2021) realizaron un estudio para predecir los factores que afectan el consumo de energía eléctrica en edificios utilizando algoritmos de Machine Learning. Específicamente, usaron algoritmos de árboles de decisión, bosques aleatorios, y K-vecinos más cercanos.

En la sección 2 se presenta la metodología de Ciencia de Datos utilizada en esta investigación. Posteriormente, en la sección 3 se discuten los resultados obtenidos luego de aplicar la metodología propuesta. Finalmente, en la sección 4 se presentan las conclusiones que se derivaron de la investigación realizada.

II. METODOLOGÍA

Esta investigación es de tipo descriptiva-explicativa, pues es descriptiva en cuanto a la etapa del análisis exploratorio de los datos para determinar su estructura y características, y es explicativa en cuanto a la etapa de aplicación de los algoritmos de aprendizaje automático para predecir clases y/o los valores de la variable objetivo. Esto concuerda por lo indicado en (Arias, 2012) en el sentido que una investigación se puede ubicar en más de un tipo, y además plantea que una investigación descriptiva se realiza con el fin de establecer la estructura o comportamiento de un hecho o fenómeno, mientras que la investigación explicativa se encarga de encontrar la razón de las cosas mediante el establecimiento de relaciones causa-efecto.

Ahora bien, el proceso de analizar datos con el fin de obtener información útil de ellos para coadyuvar en la toma de decisiones engloba en líneas generales lo que es la ciencia de datos. Según Cielen et al. (2016), la Ciencia de Datos está relacionada con la utilización de técnicas para estudiar grandes sumas de datos y sacar de éstos el conocimiento que pueda sernos útil para la toma de decisiones. Este proceso está compuesto por una serie de etapas con un propósito definido, y resaltan las etapas del análisis exploratorio de los datos y la de generación de los modelos. La primera comúnmente se lleva a cabo utilizando gráficos, figuras, y estadística descriptiva, mientras que la etapa de modelación se lleva a cabo aplicando algoritmos de machine learning tanto supervisados como no supervisados. En ese sentido, el primer paso de un proceso de ciencia de datos consiste en establecer los objetivos de la investigación, para lo cual es fundamental poseer un conocimiento mínimo del negocio del que surgen los datos. Seguidamente, deberán obtenerse los datos a analizar ya sea de una fuente oficial interna o externa, a estos datos se les conoce como datos "crudos", y por lo general deben someterse a una etapa de preparación, que puede incluir su limpieza, transformación, y/o combinación. Una vez estén preparados los datos, se realiza un análisis exploratorio de los mismos, a través de gráficos y técnicas no visuales. Este análisis exploratorio es la base para luego obtener los modelos necesarios que permitan cumplir con los objetivos de la investigación. Finalmente, con los resultados obtenidos se procede a la toma de decisiones en el área de negocios correspondiente. Los objetivos de la investigación se presentan en la introducción, la etapa de extracción de datos se presenta en esta sección, mientras que en la siguiente sección se aplican los algoritmos de aprendizaje automático y se discuten los resultados obtenidos.

Construcción del conjunto de datos

Los datos utilizados fueron extraídos de la página web del Ministerio de Industria, Minería y Energía de Uruguay (2022) el día 18 de agosto del 2022. Estos datos corresponden a la facturación mensual de energía eléctrica por sector, así como también el número de clientes mensuales por sector, ambos ítems para el período 2000-2022.

El conjunto de datos tiene 270 filas, correspondientes a cada uno de los registros de cada uno de los meses desde enero 2000 hasta junio 2022. Además, tiene 18 columnas, las cuales se corresponden con las siguientes 18 variables: año en que se factura la energía (ANHO), mes en el que factura la energía (MES), la energía total facturada en el año y mes correspondiente (E_TOTAL), la energía facturada al sector residencial en el año y mes correspondiente (E_RESID), la energía facturada al sector primario (agro, pesca, y minería) en el año y mes correspondiente (E_PRIM), la energía facturada al sector industrial en el año y mes correspondiente (E_IND), la energía facturada al sector energético (electricidad, agua, y gas) en el año y mes correspondiente (E_SE), la energía facturada al sector construcción en el año y mes correspondiente (E_SC), la energía facturada al sector del comercio y servicios en el año y mes correspondiente (E_CYS), la energía facturada al sector del alumbrado público en el año y mes correspondiente (E_AP), así como el número de clientes de los distintos sectores mencionados previamente, en el mes y año correspondiente. Toda la energía eléctrica facturada viene dada en unidades de megavatios-hora (MWh).

III. RESULTADOS Y DISCUSIÓN

Preparación de los datos

En esta etapa se aplicaron varias técnicas al conjunto de datos con el fin que tuvieran las características adecuadas previo al desarrollo del análisis exploratorio. Se verificó que cada columna de datos tuviera el formato adecuado: el año, el mes y el número de clientes, en el formato de número entero (int), mientras que el resto de los datos en el formato de número real (float). Asimismo, se comprobó que no había datos faltantes, ni registros duplicados.

Por otra parte, se inspeccionaron los datos para chequear la existencia de datos atípicos (outliers). Se determinó que en las columnas E_PRIM, E_SE, y E_AP, hay datos atípicos que según (Moreno Castellanos, 2012) son considerados como leves, por lo que no se tomaron acciones al respecto. De acuerdo con lo indicado por (McKinney, 2018), otra de las técnicas a aplicar es el de la transformación utilizando alguna función o mapeo, en ese sentido, se creó una columna con datos "string" que indica el tipo de cliente de cada registro (TIPO_CLIENTE). Adicionalmente, al combinar la energía facturada con el número de clientes, se pudo obtener una columna adicional (CONS_UNIT), con el consumo unitario por cada tipo de cliente, en el mes y año respectivo.

Análisis exploratorio de los datos

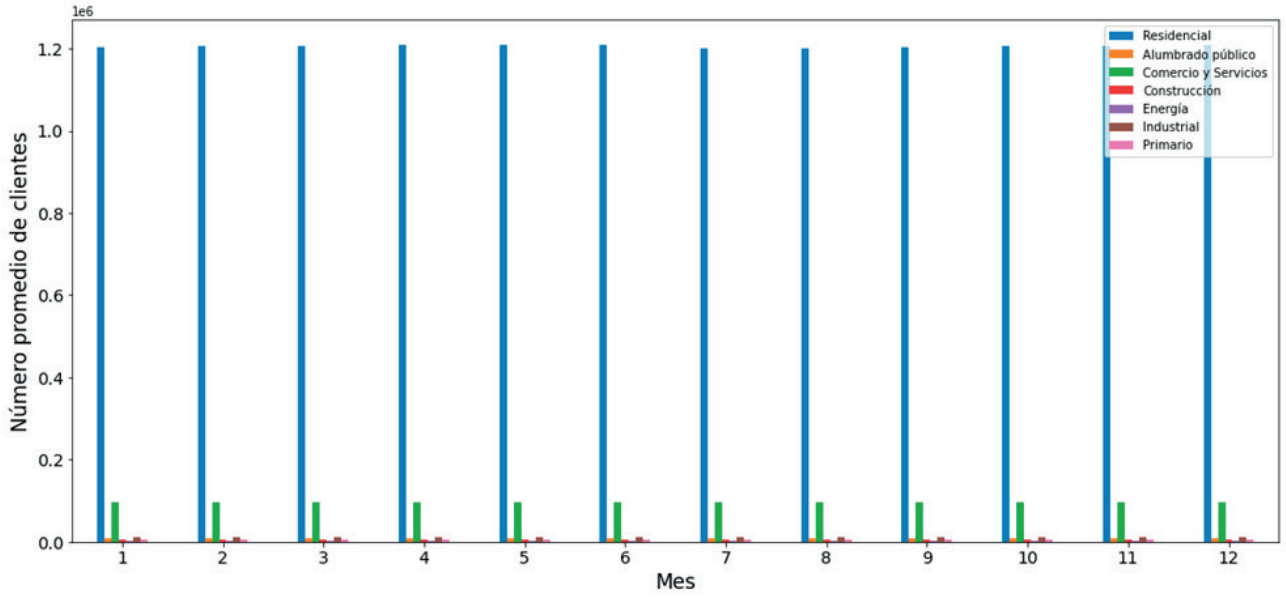
El número total de clientes a los que se les facturó energía está definido principalmente por el número

de clientes residenciales. En la figura 1 se puede observar que el número promedio de clientes residenciales por mes es de alrededor de 1.200.000, mientras que el siguiente sector con más clientes es el de comercio y servicios con menos de 100.000 clientes.

Por otra parte, el sector residencial es el de mayor promedio de facturación de energía durante el período de estudio, seguido nuevamente por el sector de comercio y servicios. En la figura 2 se presenta la facturación promedio para todos los tipos de clientes, de la cual se puede ver que los meses de julio y agosto son los de mayor facturación para los clientes residenciales, pero para los clientes comerciales los meses que en promedio tienen mayor facturación son enero y febrero.

Figura 1

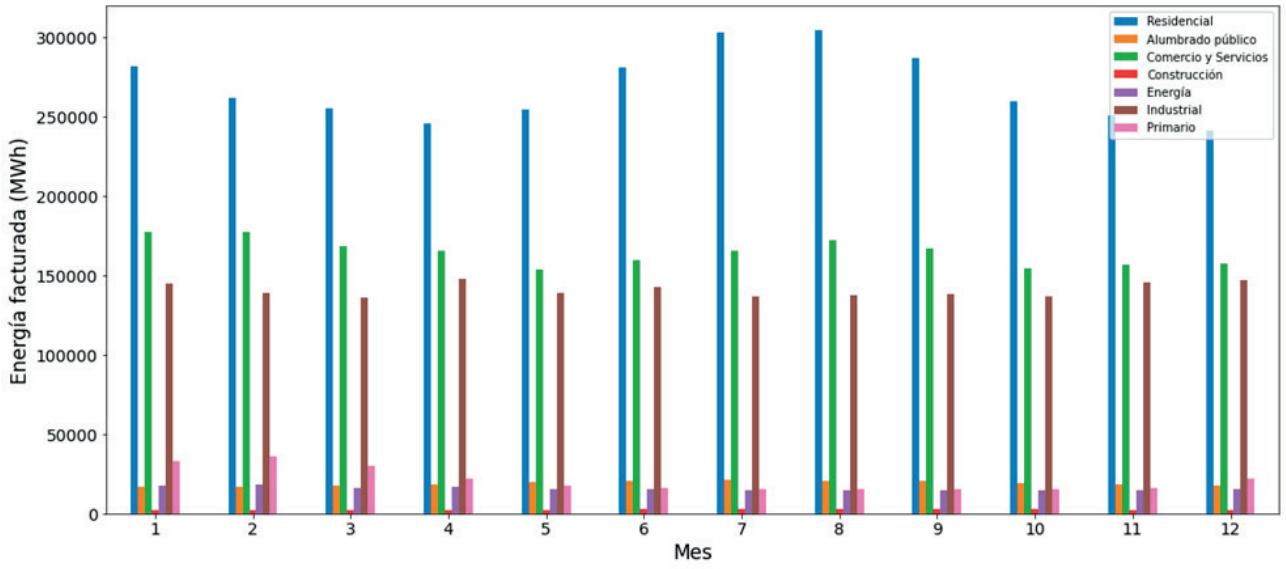
Número promedio mensual de clientes.



Fuente: Elaboración propia.

Figura 2

Energía facturada promedio mensual.



Fuente: Elaboración propia.

En lo que respecta al consumo unitario, se tiene que, a excepción del mes de febrero, el sector industrial es el que tiene el mayor valor promedio, seguido del sector de energía, y el sector primario. En la figura 3 se presenta la información para todos los tipos de cliente, de la que se observa como era de esperarse, que los clientes residenciales tienen el menor valor promedio del consumo unitario de energía eléctrica.

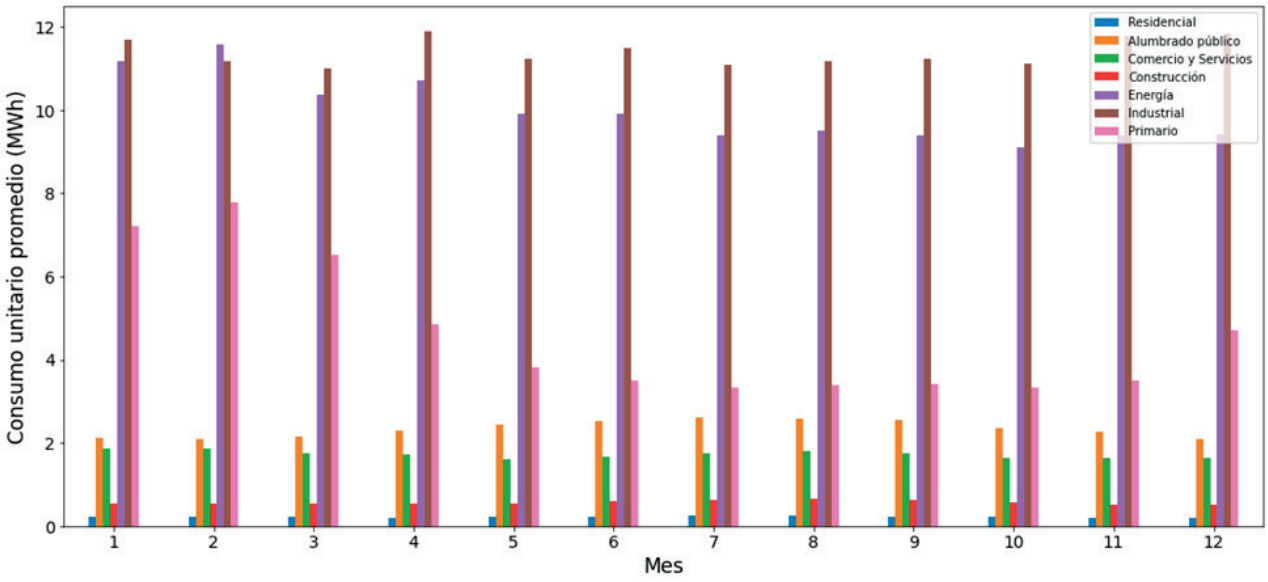
Análisis de correlación

Se cálculo el coeficiente de correlación entre cada par de variables del conjunto de datos, tanto de energía facturada como de número de clientes. Para ambos casos se utilizó el método de Pearson, el de Spearman, y el de Kendall, obteniéndose resultados similares para cada uno de ellos. En la tabla 1 se presentan la matriz de correlación para los datos de energía facturada, obtenidos con el método de Pearson. Las celdas resaltadas corresponden a los coeficientes de correlación mayor a 0,5 entre el par de variables respectivo.

De la tabla 1 se puede ver que la variable E_TOTAL tiene un alto coeficiente de correlación con la mayoría de las variables de energía facturada, a excepción de E_PRIM, E_SE, y E_AP. Aunque en general una alta correlación no implica causalidad, en este caso si se cumple pues efectivamente E_TOTAL

Figura 3

Consumo unitario promedio mensual.



Fuente: Elaboración propia.

Tabla 1

Matriz de correlación – Variables de energía facturada.

	E_TOTAL	E_RESID	E_PRIM	E_IND	E_SE	E_SC	E_CYS	E_AP
E_TOTAL	1,00	0,91	0,37	0,85	-0,08	0,92	0,96	0,25
E_RESID	0,91	1,00	0,09	0,62	-0,15	0,90	0,81	0,29
E_PRIM	0,37	0,09	1,00	0,34	0,21	0,14	0,46	-0,34
E_IND	0,85	0,62	0,34	1,00	-0,07	0,77	0,83	0,17
E_SE	-0,08	-0,15	0,21	-0,07	1,00	-0,19	-0,18	-0,27
E_SC	0,92	0,90	0,14	0,77	-0,19	1,00	0,88	0,29
E_CYS	0,96	0,81	0,46	0,83	-0,18	0,88	1,00	0,26
E_AP	0,25	0,29	-0,34	0,17	-0,27	0,29	0,26	1,00

Fuente: Elaboración propia.

representa la sumatoria de la energía facturada a cada uno de los clientes en el mes y año respectivo.

En cuanto a las variables de número de clientes, la mayoría tiene una alta correlación con todas las otras variables, resaltando el coeficiente de correlación casi unitario entre el número de clientes residenciales con el número de clientes total. La única excepción es la variable C_AP, la cual tiene una correlación baja con las otras variables. Es interesante resaltar que las variables asociadas al alumbrado público (E_AP, C_AP) tienen un bajo coeficiente de correlación con el resto de las variables.

Modelación de los datos

A continuación, se obtienen los modelos a utilizar para extraer la información pertinente de los datos, a través de la aplicación de los algoritmos de aprendizaje automático, y se discuten los resultados correspondientes. En otras palabras, se obtuvo un modelo que permitió agrupar los datos de acuerdo con sus características principales, aplicando el algoritmo K-Means. De igual manera, se generó un modelo de predicción de la categoría de nuevas instancias a través de la aplicación del de predicción K-NN. Finalmente, se aplicó el algoritmo de Regresión Lineal Múltiple para predecir la energía total en los nuevos registros, utilizando el algoritmo de análisis de componentes principales para reducir la dimensionalidad de los datos.

Aplicación del algoritmo K-Means

Este es un algoritmo que se utiliza usualmente para detectar patrones dentro de los datos, al agruparlos de acuerdo con la similitud entre ellos. Es decir, el algoritmo detecta cuales datos tienen características similares entre sí y los agrupa en un mismo clúster, mientras que los datos con características diferentes los envía a otros clústers. Se trabaja sólo con datos numéricos. Según Mailund (2017, p. 182) "en agrupamiento K-Means se intenta separar los datos en K clústers, donde tu determinas el número K. Los datos usualmente están en la forma de vectores numéricos. Estrictamente hablando, el método trabajará siempre que se tenga la manera de calcular la media de los puntos del conjunto de datos y la distancia entre pares de puntos de datos".

La mayoría de los algoritmos de aprendizaje automático tienen hiperparámetros que deben ser definidos por el usuario. Para el caso de K-Means el hiperparámetro es el número de clústers, para lo cual hay técnicas que permiten obtener su valor óptimo. Una de estas técnicas es el método del codo, que según lo indicado por Shi et al. (2021) es uno de los métodos más comúnmente utilizados para obtener el número óptimo de clústers. En su investigación, Cui (2020) propone utilizar el método del codo para obtener el número de clústers, utilizando la inercia como métrica, que no es más que la suma de los errores al cuadrado, siendo el error la distancia entre los puntos del clúster y su centroide. Por otra parte, en su investigación, Tambunan et al. (2021) utilizan la métrica Silhouette para obtener el valor óptimo del número de clústers.

En el presente trabajo se utilizaron tanto el método del codo como la maximización de la métrica Silhouette. En el primer caso se obtuvo un número óptimo de clústers igual a 8, mientras que en el segundo método se consiguió un K óptimo igual a 7. En la figura 4 se gráfica la relación Silhouette vs. K, de la cual se observa que el valor máximo se consigue cuando K vale 7.

Con el valor óptimo de K, se obtiene el modelo aplicando el algoritmo de K-Means. En la figura 5 se presenta el número de registros por clústers y por de cliente, de la cual se puede decir que el modelo agrupa perfectamente los datos de acuerdo con el tipo de cliente. Es importante mencionar que, con el valor de K igual a ocho, obtenido con el método del codo, el modelo divide los clientes residenciales en dos clústers, por lo que se seleccionó el K óptimo obtenido maximizando la métrica Silhouette.

Aplicación del algoritmo K-NN

K-NN es una de las técnicas de clasificación más simples, pero es de las más utilizadas. Clasifica los registros de acuerdo con el número de vecinos más cercanos a ellos según su distancia euclidiana. Un registro particular se asigna a la clase a la que pertenezca la mayoría de sus vecinos más cercanos. Como lo plantean en Fan et al. (2019), el algoritmo K-NN se utiliza para encontrar K muestras de entrenamiento que estén lo más cerca posible de la muestra objetivo en el conjunto de entrenamiento, y luego se asigna la categoría dominante a la muestra objetivo, donde K es el número de muestras de entrenamiento.

Para aplicar este algoritmo, se reagruparon los datos del conjunto original para poder crear la variable TIPO_CLIENTE y así seleccionarla como variable objetivo, es decir, el modelo obtenido debe predecir si el registro de prueba debe asignarse al conjunto de clientes residenciales o algún otro de los tipos de cliente que no entran en la categoría residencial. Previo a la aplicación del algoritmo, se debe definir el número K de vecinos más cercanos, esto lo pue-

Figura 4

Obtención del número óptimo de clústers.

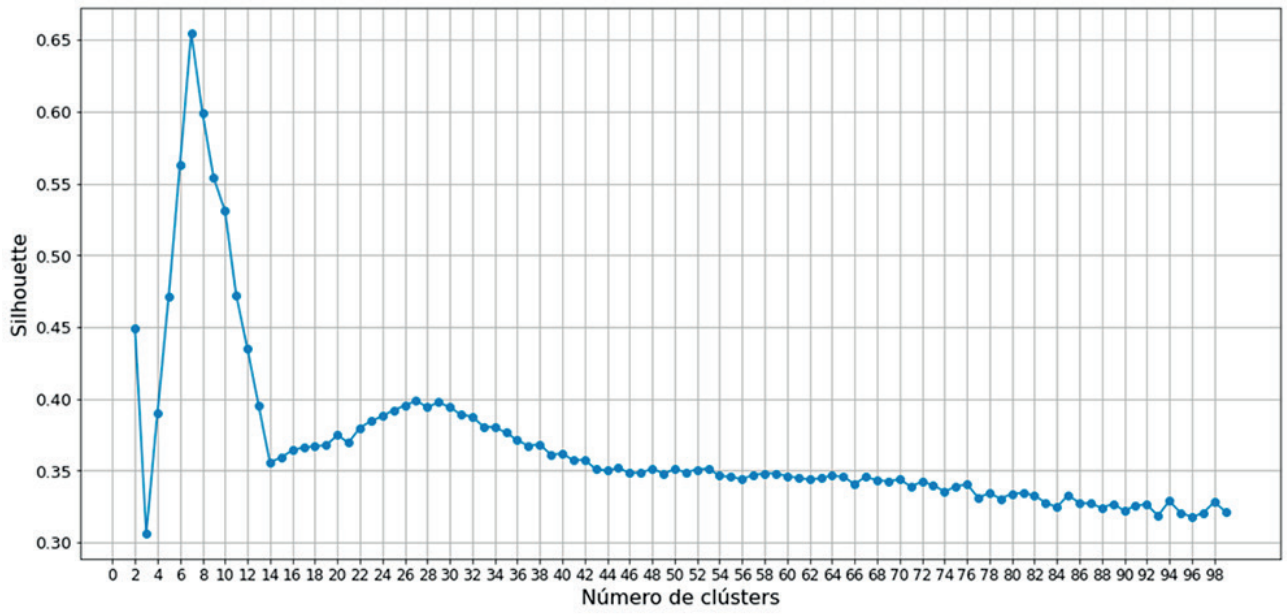
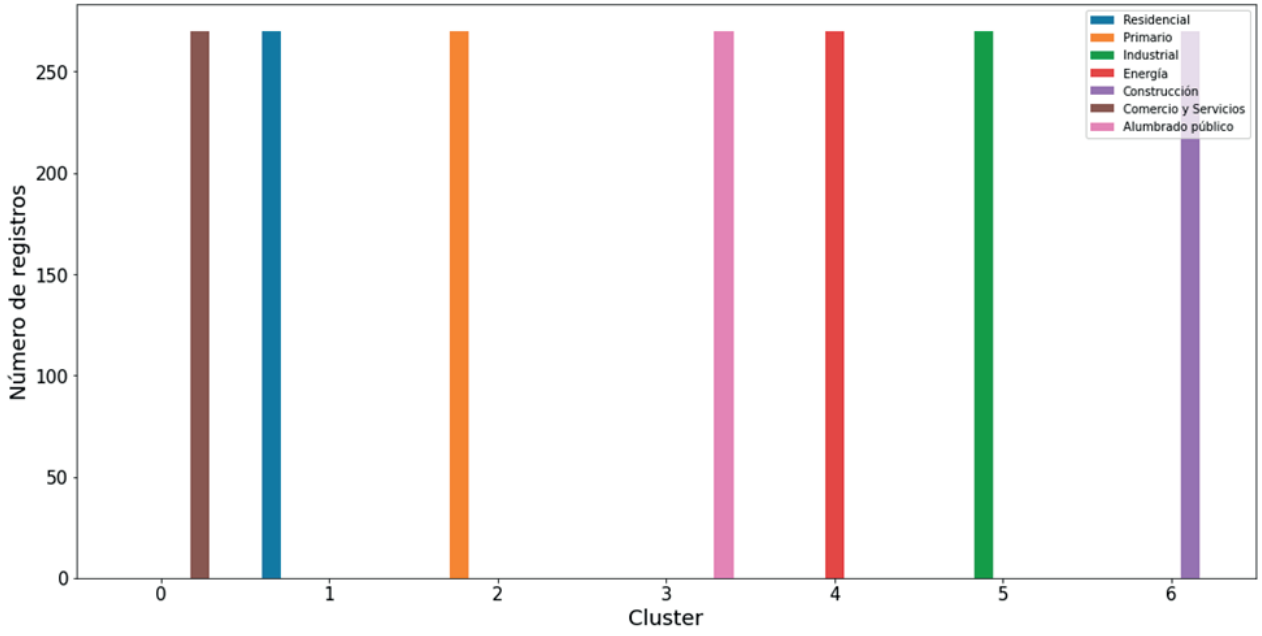


Figura 5

Número de registros de datos por clúster.



Fuente: Elaboración propia.

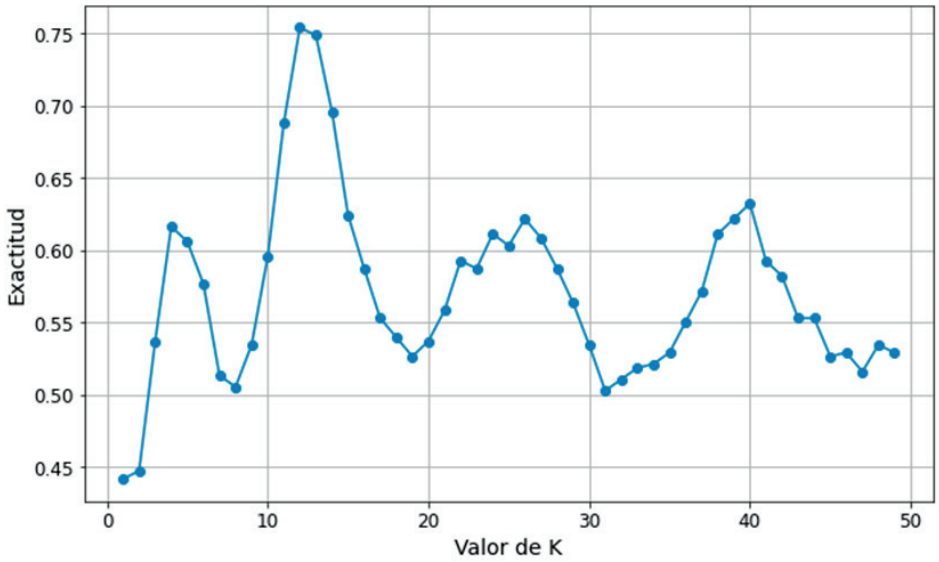
de establecer el usuario de manera arbitraria, pero usualmente se alcanza al maximizar alguna métrica de interés. En este artículo se hace uso de la métrica exactitud (accuracy), la cual de acuerdo con Fenner (2020, p.163) “es la métrica que tenemos para evaluar que tan bien nuestra conjetura o predicción coincide con la realidad”. Para cada valor de K se computa el valor de la exactitud, y los pares de valores K vs exactitud se grafican para obtener la figura 6, que al observarla se puede decir que el valor máximo de la exactitud (75,39%) se obtiene cuando el valor de K es igual a doce.

Con el valor óptimo de K vecinos más cercano, se aplica entonces el algoritmo para obtener el modelo que me permita predecir el tipo de clientes de los registros nuevos. El procedimiento implica dividir el conjunto de datos en dos partes. Una parte, representando el 80% de los datos originales, se utiliza para entrenar el modelo, a esta parte se le llama el set de entrenamiento. La otra parte, representando el 20% de los datos originales se le llama set de prueba, se utiliza para evaluar el modelo obtenido y así verificar que cumpla con niveles mínimos de calidad. Del proceso de evaluación se produce la llamada "matriz de confusión", que no es más que una matriz cuadrada, con los verdaderos negativos y los verdaderos positivos en la diagonal principal, y los falsos negativos y los falsos positivos en las otras celdas. Como tenemos siete categorías de clientes, esta matriz tiene una dimensión de siete filas y siete columnas, y se presenta en la tabla 2.

De la tabla 2 se puede decir que el set de prueba estuvo conformado por 378 filas del conjunto de datos (sumatoria de los valores de todas las celdas), el cuál es el 20% del total de datos. De las filas del set de prueba, el modelo clasificó correctamente 285 filas, correspondientes al 75,39% de exactitud. Adicionalmente, 43 registros originalmente pertenecientes a los clientes residenciales fueron clasificados de esa misma forma por parte del modelo. De igual manera, 64 filas eran realmente de alumbrado público y el modelo clasificó correctamente 52

Figura 6

Obtención del valor óptimo de vecinos más cercanos.



Fuente: Elaboración propia.

Tabla 2

Matriz de confusión.

TIPO DE CLIENTE	Residencial	Alumbrado Público	Comercio y Servicio	Construcción	Energía	Industrial	Primario
Residencial	43	0	0	0	0	0	0
Alumbrado Público	0	52	0	1	2	0	9
Comercio y Servicio	0	0	58	0	0	7	0
Construcción	0	10	0	33	1	0	0
Energía	0	19	0	2	24	0	8
Industrial	0	0	5	0	0	55	0
Primario	0	13	0	0	16	0	20

Fuente: Elaboración propia.

de ellas, 9 las clasificó como del sector primario, 2 como del sector de energía, y la fila restante como del sector construcción. Al hacer el mismo análisis para los otros tipos de clientes, se nota que sólo se clasificó correctamente a los clientes residenciales.

Para efectos de comparación, se generó otro modelo que clasifica los registros como cliente residencial o cliente no residencial, es decir, se creó una nueva columna en los datos para indicar si la fila correspondiente pertenece a clientes residenciales o a clientes no residenciales. Para este segundo modelo, la exactitud fue del 100% para cualquier valor de K, manteniendo el tamaño del set de prueba en 20% de los datos y el set de entrenamiento en 80% de los datos, y la matriz de confusión fue de sólo 2x2. De las 378 filas del set de prueba, 335 eran de clientes no residenciales, y el modelo las catalogó correctamente, mientras que 43 filas eran de clientes residenciales y el modelo las catalogó de forma correcta.

La utilidad de este modelo está en el hecho que puede comprobar el tipo de cliente, y por consiguiente el tipo de tarifa a aplicar, cuando debido a alguna causa, se requiera verificar este tópico en los registros nuevos del conjunto de datos.

Aplicación del algoritmo de Regresión Lineal Múltiple

Con la aplicación de este algoritmo se busca generar un modelo de regresión en el que se quiere predecir una variable objetivo a partir de un conjunto de otras variables (explicativas). Para este caso como variable objetivo se selecciona facturación total de energía eléctrica en Uruguay, y como variables explicativas se tienen inicialmente las restantes variables del conjunto de datos. Sin embargo, se hace un análisis previo para determinar cuáles de esas posibles variables explicativas se podrían descartar del modelo.

En primer lugar, se hace uso del análisis de correlación entre las variables del número de clientes. De esos resultados se decide mantener dentro de las variables explicativas sólo el número de clientes totales y el número de clientes de alumbrado público. Posteriormente, se obtienen los coeficientes de correlación entre la variable objetivo y cada una de las variables explicativas, encontrándose que el número de clientes de alumbrado público (C_AP) tiene un coeficiente casi nulo con la variable objetivo, por lo cual también se descarta, al igual que las variables E_PRIM, E_AP, y E_SE.

Seguidamente, se lleva a cabo un análisis de componentes principales (PCA) para continuar con el proceso de reducción de dimensionalidad. PCA transforma los datos originales en un nuevo conjunto con una dimensionalidad menor, este nuevo conjunto lo conforman las componentes principales, que se obtienen como una combinación lineal de las variables originales. La idea es que el nuevo conjunto de componentes tenga la mayor cantidad posible de información de los datos originales, y eso se mide con la proporción de varianza explicada por las componentes principales. En Núñez et al. (2020) se aplica PCA para reducir la dimensionalidad del sistema y visualizar el desempeño de la técnica de clusterización, mientras que en Dana et al. (2021) se utiliza PCA para eliminar variables explicativas irrelevantes y ruido.

Se aplica entonces PCA al conjunto de variables explicativas. Debido a que en este punto se tienen 10 variables explicativas, ese será el número máximo de componentes principales que genera el algoritmo. En la figura 7 se presenta la varianza explicada por cada una de las componentes principales. Se puede notar que en las primeras 6 componentes ya se explica el 98% de la varianza de los datos.

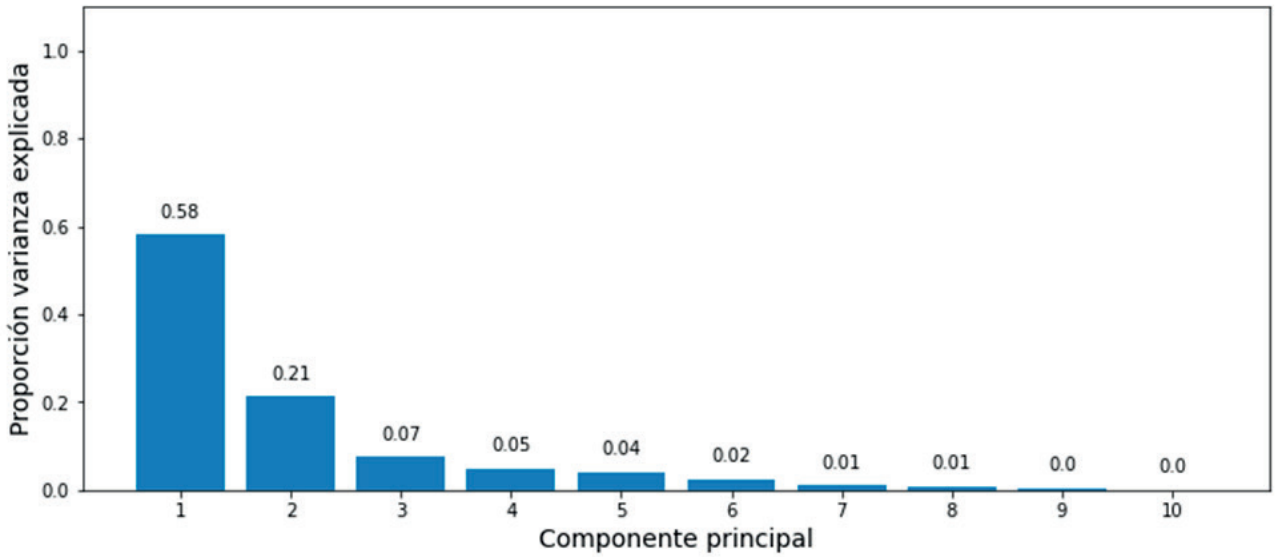
Para realizar la reducción de la dimensionalidad, se debe definir un criterio en cuanto a la varianza explicada por las componentes principales. Sí se toma como referencia el 98% de la varianza explicada, entonces se mantienen aquellas variables cuyos pesos en las seis primeras componentes sea relativamente alto, es decir, que sean significativas para esas componentes. En ese sentido, se descartan las variables E_IND y E_SC.

Finalmente, se aplica el algoritmo de Regresión Lineal teniendo a la energía facturada total (E_TOTAL) como la variable objetivo, y teniendo como variables explicativas a la energía facturada residencial (E_RESID), la energía facturada para el sector de comercio y servicios (E_CYS), y el número de clientes totales (C_TOTAL).

Se consideran los datos desde enero del 2000 hasta diciembre del 2021. El modelo se entrenó con el 75% de estos datos, y el restante 25% se utilizó para evaluar el modelo. Se obtienen los coeficientes: 0,678, 1,782, y 0,078, para las variables E_RESID, E_CYS, y C_TOTAL, respectivamente. Es decir, un aumento unitario en la variable E_RESID implica un aumento de 678 kWh de la energía facturada total, un aumento unitario en E_CYS implica un aumento de 1.782 kWh en E_TOTAL, y un

Figura 7

Proporción de varianza explicada por cada componente principal.



Fuente: Elaboración propia.

aumento unitario en C_TOTAL significa un aumento de 78 kWh en E_TOTAL.

Para la evaluación del modelo uno de los indicadores utilizados es el coeficiente de determinación R^2, que según Walpole et al. (2012, p.407) "es una medida de la calidad de ajuste del modelo, y muestra la proporción de la variabilidad explicada por el modelo ajustado". Es un indicador que varía entre 0 y 1, valdrá 1 cuando el ajuste del modelo es perfecto. Por el contrario, un valor cercano a 0, indica un ajuste deficiente del modelo. Por lo tanto, lo que se busca es que este indicador este lo más cercano posible a la unidad. Los otros dos indicadores utilizados son la raíz cuadrada de la media de los errores al cuadrado (RMSE), y la media del valor absoluto de los errores (MAE). En ambos casos, las unidades son las mismas de la variable objetivo, es decir mega watt-hora (MWh). El error o residuo, no es más que la diferencia entre el valor real y la predicción respectiva.

Por otra parte, se aplicó la prueba de Shapiro-Wilk para contrastar la normalidad de los residuos. La hipótesis nula es que los residuos provienen de una población normalmente distribuida, y el estadístico utilizado varía entre 0 y 1. Si el p-valor es mayor al nivel de significancia (por lo general 5%) se concluye que no se puede rechazar la hipótesis nula. En la tabla 3 se presentan los resultados de la evaluación.

Tabla 3

Parámetros resultantes del modelo de regresión.

Indicador de desempeño	Valor Obtenido
R^2	0,981
RMSE	12.458,21
MAE	10.173,93
Prueba de Shapiro-Wilk a los residuos	
Estadístico	0,973
p-valor	0,157

Fuente: Elaboración propia.

De los resultados de la evaluación del modelo se puede observar que el R^2 es cercano a uno, lo cual es indicativo de un buen ajuste del modelo. Adicionalmente, se puede notar que no se rechaza la hipótesis de que los residuos provienen de una población normalmente distribuida. De lo anterior se deduce que el modelo se puede utilizar para hacer las predicciones.

Finalmente se utiliza el modelo obtenido para predecir la energía total facturada, en MWh, durante los primeros seis meses del año 2022, y se comparan con los valores reales. Los resultados se muestran en la tabla 4, de la cual se puede ver que el mayor error cometido en la predicción fue del 4,26% para la predicción del mes de enero, y en promedio el error absoluto cometido fue de apenas 1,23%.

Tabla 4

Predicción energía total facturada 2022.

Mes	Real	Predicción	Residuos	Error
Enero	836.374	800.728	35.646	4,26%
Febrero	793.616	788.916	4.700	0,59%
Marzo	726.011	728.632	-2.621	-0,36%
Abril	714.636	715.351	-715	-0,10%
Mayo	695.098	701.079	-5.982	-0,86%
Junio	800.623	790.953	9.670	1,21%

Fuente: Elaboración propia.

IV. CONCLUSIONES

- La energía facturada promedio mensual a los clientes residenciales es mayor que la de cada uno de los otros tipos de clientes, para cada uno de los doce meses del año, siendo los clientes del sector comercio y servicios los que les siguen en cuanto a la energía facturada promedio mensual. En cuanto al número de clientes, el mayor valor promedio mensual corresponde a los clientes residenciales, siendo su más cercano seguidor los clientes del sector comercio y servicios con alrededor de un millón de clientes menos. En contraste, al hablar del consumo unitario promedio mensual por tipo de cliente, el sector industrial es el que tiene el mayor valor promedio, seguido del sector de energía, y el sector primario. Los clientes del sector residencial son los que tienen menor consumo unitario promedio mensual.
- El modelo obtenido a partir del algoritmo K-Means permitió agrupar los datos de manera perfecta de acuerdo con los tipos de clientes, siendo el número de clústers igual a la cantidad de tipos de clientes diferentes presentes en los datos. El número de clústers óptimo K se obtuvo comparando el resultado del método del codo (K igual a 8), con el método de maximización de la métrica Silhouette (K igual a 7). Con ocho clústers, el modelo tendía a dividir los clientes residenciales en dos clústers, mientras que con siete clústers el modelo agrupó los clientes residenciales en un solo clúster.
- Al aplicar la técnica K-NN, se logró un modelo que permitió predecir el tipo de cliente al cual pertenecen las nuevas instancias. Para ello, previamente se obtuvo el valor óptimo de K igual a doce, para el cual se maximizó la exactitud del modelo. En el conjunto de prueba el modelo clasificó correctamente a los clientes residenciales, pero presentó un error significativo al clasificar a los clientes de las distintas clases no residenciales, siendo la exactitud del modelo de sólo 75,39%. Se generó un modelo nuevo, que permitió predecir si la clase de los registros nuevos eran de clientes residenciales o de clientes no residenciales, es decir, sólo dos clases. Para este segundo modelo se obtuvo una exactitud del 100% para cualquier valor del parámetro K.
- Antes de aplicar el algoritmo de regresión lineal, se realizó un análisis de correlación, y un análisis de componentes principales, para reducir la dimensionalidad del conjunto de datos. La variable objetivo fue la energía facturada total, mientras que las variables explicativas fueron: energía facturada residencial, energía facturada del sector comercio y servicios, y el número total de clientes. El modelo obtenido tuvo un coeficiente de determinación R^2 de 0,981, y con la prueba de normalidad de los residuos se obtuvo estadístico de 0,973 y un p-valor de 0,157.

V. REFERENCIAS

[1] Arias, F. G. (2012). *El Proyecto de Investigación - Introducción a la metodología científica.* Caracas: Editorial EPISTEME, C.A.

[2] Cheol Jeong, H., Jang, M., Kim, T., & Joo, S.-K. (2020). Clustering of Load Profiles of Residential Customers Using Extreme Points and Demographic Characteristics. *Electronics*, MDPI. https://doi.org/10.3390/electronics10030290

[3] Cielen, D., Meysman, A., & Ali, M. (2016). *Introducing Data Science.* Shelter Island, NY: Manning Publications Co.

[4] Cui, M. (2020). Introduction to the K-Means Clustering Algorithm Based on the Elbow Method. *Accounting, Auditing and Finance (2020)*, 1-5. https://dx.doi.org/10.23977/accaf.2020.010102

[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] Fan, G.-F., Guo, Y.-H., Zheng, J.-M., & Hong, W.-C. (2019). Application of the Weighted K-Nearest Neighbor Algorithm for Short-Term Load Forecasting. *Energies*. https://doi.org/10.3390/en12050916

[7] Fenner, M. E. (2020). *Machine Learning with Python for Everyone.* Boston: Pearson Education, Inc.

[8] Hosseini, S., & Hafezi Fard, R. (2021). Machine Learning Algorithms for Predicting Electricity Consumption of Buildings. *Wireless Personal Communications, 121*, 3320-3341. https://doi.org/10.1007/s11277-021-08879-1.

[9] Lester, M., Carrizo, D., Ulloa-Vásquez, F., & García-Santander, L. (2021). Uso de algoritmo K-means para clasificar perfiles de clientes con datos de medidores inteligentes de consumo eléctrico: Un caso de estudio. *Ingeniare. Revista chilena de ingeniería, 29*(4), 778-787. http://dx.doi.org/10.4067/S0718-33052021000400778.

[10] Mailund, T. (2017). *Beginning Data Science in R - Data Analysis, Visualization, and Modelling for the Data Scientist.* New York: Apress.

[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] Nepal, B., Yamaha, M., Sahashi, H., & Yokoe, A. (2019). Analysis of Building Electricity Use Pattern Using KMeans Clustering Algorithm by Determination of Better Initial Centroids and Number of Clusters. *Energies*. https://doi.org/10.3390/en12122451

[15] Núñez-Barrionuevo, O., Llanes-Cedeño, E., Martinez-Gomez, J., Guachimboza-Davalos, J., & Lopez-Villada, J. (2020). Clustering Analysis of Electricity Consumption of Municipalities in the Province of Pichincha-Ecuador Using the K-Means Algorithm. *Proceedings of ICCIS 2020 Springer, 1273*, 187-195. https://doi.org/10.1007/978-3-030-59194-6_16.

[16] 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

[17] Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *Wireless Com Network*. https://doi.org/10.21203/rs.3.rs-58011/v1

[18] Tambunan, H., Barus, D., Hartono, J., Alam, A., Nugraha, D., & Usman, H. (2020). Electrical Peak Load Clustering Analysis Using K-Means Algorithm and Silhouette Coefficient. *2020 International Conference on Technology and Policy in Electric Power & Energy (ICT-PEP)*. IEEE. https://doi.org/10.1109/ICT-PEP50916.2020.9249773

[19] Walker, S., Khan, W., Katic, K., Maassen, W., & Zeiler, W. (2019). Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings. *Energy and Buildings 209*. https://doi.org/10.1016/j.enbuild.2019.109705

[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.

Fuentes de financiamiento:

Propia.

Conflictos de interés:

El autor declara no tener conflictos de interés.

Contribución del Autor

El autor desarrolló todas las etapas del proceso de Ciencia de Datos al conjunto de datos de energía eléctrica facturada en Uruguay durante el período 2000 – 2022. Es decir, extrajo el conjunto de datos utilizados, aplicó las técnicas de preparación a dichos datos, realizó el análisis exploratorio, generó los modelos de aprendizaje automático, y redactó las conclusiones. Asimismo, participó en la redacción y revisión final del artículo.