Revista de +D TECNOLÓGICO
Portal de revistas: http://revistas.utp.ac.pa
Vol. 18, No. 2: 17-31
Julio - diciembre 2022
ISSN 1680-8894
ISSN-L 2219-6714

Uso de algoritmos de aprendizaje automático para analizar datos de energía eléctrica facturada. Caso: Chile 2015 – 2021

Use of machine learning algorithms to analyze billed electricity data. Case: Chile 2015 – 2021

César Yajure Ramírez [1]
[1] Universidad Central de Venezuela, Escuela de Ingeniería Básica, Venezuela

cyajure@gmail.com
Fecha de recepción:08 de agosto de 2022; Fecha de aprobación: 4 de noviembre de 2022.
***Autor de correspondencia:** César Yajure Ramírez (cyajure@gmail.com)

RESUMEN. En el mercado eléctrico chileno, los usuarios finales se clasifican en clientes libres y clientes regulados. El análisis del consumo energético de los clientes regulados es importante para efectos del diseño y aplicación de las políticas públicas del sector. En esta investigación se hace el análisis de los datos de energía eléctrica facturada mensual de los clientes regulados de Chile, durante el período 2015-2021, con el fin de detectar patrones y predecir la categoría a la que pertenecen. Se utilizan los algoritmos K-Means para la detección de patrones, K-NN para la predicción de la categoría de los clientes, y PCA para determinar las variables más significativas dentro del conjunto de datos. Con K-Means se encontró que los datos se agrupan de acuerdo con el tipo de cliente, con K-NN se obtuvo un modelo que permite predecir a qué tipo de clientes pertenecen los datos, y con PCA se encontró que las variables tipo de cliente, el año y el mes, son las más importantes en el conjunto de datos. Más del 96% de los clientes analizados corresponde al tipo residencial, quienes consumieron el 50% de la energía facturada durante el período de estudio, y además imponen la estacionalidad mensual de los datos. Los resultados obtenidos son de ayuda para el establecimiento y revisión de las políticas aplicadas a los clientes regulados, en cuanto a tarifas, límites de consumo en invierno, y eficiencia energética. Se recomienda continuar la investigación orientándola hacia la predicción del consumo de energía eléctrica.

Palabras clave. *Agrupamiento, análisis de componentes principales, aprendizaje automático, energía facturada.*

ABSTRACT. In the Chilean electricity market, end users are classified as free customers and regulated customers. The analysis of its behavior is important for the design and application of public policies in the sector. In this research, the monthly billed electricity data of Chilean regulated customers is studied during the 2015-2021 period, to detect patterns and predict the category to which they belong. K-Means algorithms are used for pattern detection, K-NN for customer category prediction, and principal component analysis to determine the most significant variables within the data set. With K-Means it was found that the data is grouped according to the type of client, with K-NN a model was obtained that allows predicting to which type of clients the data belongs, and with the analysis of principal components it was found that the variables customer type, year, and month, are the most important in the data set. More than 96% of the customers analyzed correspond to the residential type, who consumed 50% of the energy invoiced during the study period and imposed the monthly seasonality of the data. The machine learning algorithms applied to the data made it possible to generate models to group them, to predict their category, and to establish the most significant variables in terms of their variance.

Keywords. *Clustering, principal component analysis, machine learning, billed energy.*

1. Introducción

En Chile los sectores de generación, transmisión y distribución eléctrica son manejados por actores del sector privado, y debido a sus características de monopolio, la transmisión y la distribución funcionan dentro de un esquema de regulación, mientras que la generación eléctrica funciona bajo las reglas de la libre competencia. Por otra parte, desde un punto de vista geográfico se tienen tres sistemas eléctricos independientes, el sistema eléctrico nacional (SEN) compuesto por las instalaciones de generación eléctrica, transmisión y consumo que abarcan el territorio desde las regiones de Arica Parinacota, hasta la Isla Grande de Chiloé, en la región de Los Lagos. El Sistema de Aysén (SEA) en la región del mismo nombre, y el Sistema de

Magallanes (SEM) que abarca la región de Magallanes y la Antártica Chilena [1].

Ahora, para efectos de definir y/o hacer seguimiento a las políticas públicas en el área energética y/o mejorar la gestión del servicio que se presta desde las empresas distribuidoras de electricidad, es conveniente conocer el comportamiento del consumo de energía eléctrica, a través del análisis de datos de consumo o de facturación de la energía eléctrica de los clientes del servicio. En ese sentido, la normativa chilena establece dos segmentos principales en el área de consumo de energía eléctrica: clientes regulados y clientes libres. De acuerdo con [1], el segmento de clientes regulados lo conforman consumidores con una potencia conectada igual o inferior a 5MW, pero aquellos con una potencia conectada entre 500kW y 5MW, y que están ubicados en el área de concesión de una empresa distribuidora, pueden optar a ser clientes libres. También plantean que el segmento de clientes libres está compuesto por consumidores cuya potencia conectada es superior a 5MW, y que pueden pactar libremente los precios y condiciones con sus suministradores. Siguiendo con [1], aquellos con potencia superior a 500kW que opten a ser cliente libre, deben permanecer al menos 4 años en esta categoría.

Por su definición, los clientes regulados se relacionan únicamente con la empresa de distribución eléctrica. Ésta deberá contratar el suministro de energía y potencia y traspasar estos costos, además de los cargos de transmisión, al cliente. Además, debe recaudar el valor agregado de distribución, es decir, los costos de generación, transmisión y distribución se traspasan al cliente final. Según lo indican en [2], el costo de la energía asociado al segmento de generación se calcula a través del precio de nudo promedio. En cuanto a la transmisión, el costo debe considerar el uso de las instalaciones a nivel nacional y zonal, además de los sistemas de interconexión internacional. Por último, las empresas de distribución reciben sus ingresos a través del llamado valor agregado de distribución. La Comisión Nacional de Energía (CNE), es el ente encargado de fijar las tarifas que pueden cobrar las empresas por la distribución de electricidad, esto lo realiza cada cuatro años. Como lo indican en [2], para los clientes residenciales todos los costos mencionados se establecen de manera regulada, a través de decretos.

La normativa vigente chilena establece distintas opciones tarifarias para los clientes regulados, quienes podrán elegir libremente una de estas opciones tarifarias, siempre que cumplan con las limitaciones y condiciones de aplicación establecidas en cada caso y dentro del nivel de tensión que les corresponda. Según [3] la normativa chilena indica que los clientes en alta tensión son aquellos que se conectan a la red con un voltaje superior a los 400 voltios, mientras que los clientes en baja tensión se conectan a la red con un voltaje igual o inferior a los 400 voltios. Para los clientes residenciales se tienen las tarifas: BT1a, BT1b, TRBT2, TRBT3, TRAT1, TRAT2, TRAT3, y para los clientes no residenciales se tienen las tarifas: BT2, BT3, BT4.1, BT4.2, BT4.3, BT5, AT2, AT3, AT4.1, AT4.2, AT4.3, AT5. En cuanto a las tarifas para clientes no residenciales, en la consultoría desarrollada por [4] se indica que las tarifas BT2 y AT2 son utilizadas principalmente por clientes comerciales, y las tarifas BT3, BT4, AT3 y AT4, son utilizadas principalmente por usuarios industriales.

En la presente investigación, tomando en cuenta los datos estadísticos oficiales de la CNE, se realizó el análisis de los datos de energía eléctrica facturada mensual por tipo de cliente, tipo de tarifa, y ubicación geográfica de los usuarios, durante el período 2015-2021 en la república de Chile. Los objetivos fueron describir, a partir de los resultados cuantitativos obtenidos, sus características principales, descubrir patrones en la energía eléctrica facturada, y predecir categorías en los datos nuevos. La obtención de esta información podría ser útil para la revisión y/o evaluación de las políticas implementadas para los clientes regulados. Para lograrlo, se hizo uso de algoritmos de aprendizaje automático, tanto de aprendizaje supervisado como de no supervisado. Específicamente, se utilizó el algoritmo K-Means para encontrar patrones en los datos de energía eléctrica facturada, ya que es un algoritmo muy utilizado para analizar datos de consumo de energía eléctrica, el algoritmo K-NN para predecir las categorías de nuevos datos, puesto que es un algoritmo sencillo de utilizar y con pocos hiperparámetros por definir. Asimismo, se utilizó el análisis de componentes principales para determinar las variables con mayor impacto en el conjunto de los datos , pues es una técnica muy utilizada

para la reducción de dimensionalidad, y para detectar variables importantes.

Se encontró una gran variedad de investigaciones sobre uso de algoritmos de aprendizaje automático para detectar patrones y/o hacer predicciones a partir de datos de consumo de energía eléctrica. La mayoría de ellas está orientada al consumo eléctrico residencial y/o al uso de algoritmo K-Means para definir perfiles de usuarios, principalmente con datos de consumo horario. Por ejemplo, en [5] desarrollan un estudio comparativo de técnicas de agrupamiento para patrones de segmentación de carga eléctrica, utilizando datos de consumo diario de energía eléctrica, y haciendo uso de distintas métricas para comparar los distintos algoritmos empleados, siendo K-Means el algoritmo de mejor desempeño con respecto a las métricas MSE y tiempo de procesamiento. Por otra parte, en [6] realizaron una clasificación de clientes residenciales a partir de los datos de consumo obtenidos desde medidores inteligentes, utilizando una metodología basada en el algoritmo K-Means. Los resultados obtenidos muestran que dos de los clústeres son los que mejor representan el conjunto de clientes analizados. En su estudio, [7] proponen una metodología para desarrollar modelos de predicción de consumo de energía eléctrica utilizando el algoritmo de K vecinos más cercanos, entre otros. El modelo que mejor desempeño tuvo fue el obtenido del algoritmo del árbol de regresión. De igual forma, [8] se utilizan algoritmos de aprendizaje automático para predecir el consumo de energía en edificios inteligentes. Aplicaron los algoritmos Máquina de Soporte Vectorial y K vecinos más cercanos, junto a redes neuronales, utilizando la plataforma Azure. El algoritmo Máquina de Soporte Vectorial, tuvo el mejor desempeño en términos de las métricas NRMSE, RMSE, y MAPE. Por otra parte, [9] desarrollaron un análisis comparativo de patrones de uso de la electricidad, utilizando técnicas de minería de datos. Más específicamente, utilizaron el algoritmo de agrupamiento K-Means sobre un conjunto de datos de mil edificios en Suiza, obteniendo patrones de uso de la electricidad significativamente diferentes entre sí. De igual manera, [10] aplicaron algoritmos de aprendizaje automático para modelar el consumo de energía eléctrica en un hospital, utilizando los datos de consumo diarios, además de otras trece variables sobre la actividad rutinaria en el hospital. Obtuvieron que los valores de temperatura son buenos predictores del consumo de energía eléctrica en el hospital. Además, en [11] utilizan algoritmos de aprendizaje automático para predecir el consumo de electricidad en edificios. Específicamente utilizan árboles de decisión, bosques aleatorios, K vecinos más cercanos, y regresión univariada. Obtienen los mejores resultados con el algoritmo de árboles de decisión. En su investigación, [12] analizaron los datos de consumo de electricidad en las municipalidades de la provincia de Pichincha en Ecuador, utilizando el algoritmo de agrupamiento K-Means. Lograron identificar tres grupos de municipalidades, de acuerdo con su nivel de consumo. Adicionalmente, [13] realizaron una investigación para predecir la demanda de energía eléctrica de edificios, utilizando un predictor basado en el algoritmo de K vecinos más cercanos, que resultó significativamente más preciso que otros modelos utilizados previamente. En su estudio, [14] diseñaron una estructura para seleccionar el mejor algoritmo de clasificación para detectar fraudes en el consumo de energía eléctrica. Utilizaron distintas métricas para comparar los algoritmos, entre los cuales estuvo K vecinos más cercanos.

El resto del artículo se organiza de la siguiente manera. En la sección 2 se presenta la metodología utilizada en la investigación. Seguidamente, en la sección 3 se presenta el desarrollo de la metodología aplicada y la discusión de los resultados obtenidos. En la sección 4 se presentan las conclusiones que se derivaron de la investigación realizada.

2. Materiales y métodos

El análisis de grandes cantidades de datos con el fin de extraer de ellos la información pertinente para la toma de decisiones se conoce como Ciencia de Datos. De acuerdo con [15], la Ciencia de Datos involucra el uso de métodos para analizar cantidades masivas de datos y extraer el conocimiento que contienen. La extracción de información y/o conocimiento a partir de los datos se lleva a cabo a través de dos etapas claramente diferenciadas: el análisis exploratorio de los datos y la modelación de los datos. La primera se ejecuta, usualmente, utilizando medios visuales y estadística descriptiva, mientras que la etapa de modelación se lleva

a cabo aplicando algoritmos de aprendizaje automático para generar modelos que nos permitan detectar patrones en los datos, predecir categorías, predecir valores de una variable objetivo, entre otras funciones.

En ese sentido, en [15] se presentan las etapas que conforman un proceso de Ciencia de Datos. En la Figura 1 se ilustra dicho proceso, basado en su propuesta. En esta investigación se desarrollaron los primeros cinco pasos de la propuesta, siendo el paso 2 presentado en esta sección, y los pasos 3, 4, y 5 presentados en la siguiente sección.

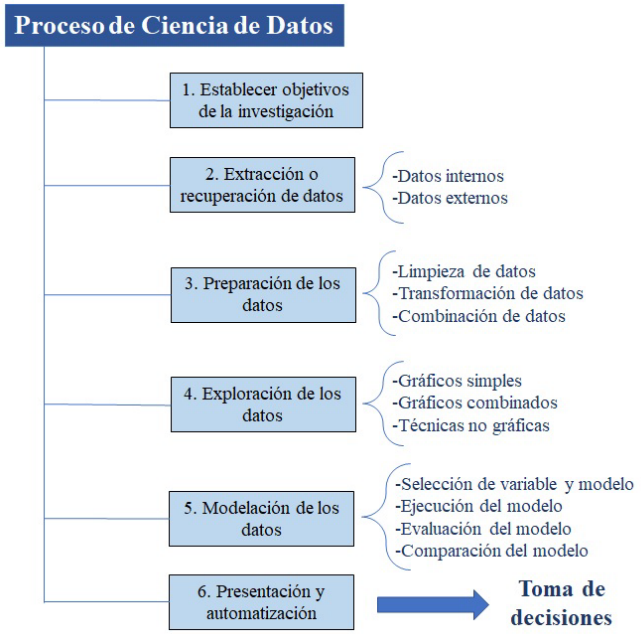

Figura 1. Proceso de Ciencia de Datos.
Fuente: Elaboración propia en base a [15].

En otro orden de ideas, éste trabajo tiene rasgos de una investigación descriptiva, asociado al análisis exploratorio de los datos, pero también rasgos de una investigación explicativa relacionados con la aplicación de los algoritmos de aprendizaje automático. Pues, tal como lo indica [16], en la investigación descriptiva se refieren las características del fenómeno objeto de estudio. Adicionalmente [16] plantea que en la investigación de tipo explicativa se analizan causas y efectos de la relación entre variables existentes.

Ahora bien, tal como se indica en la Figura 1, una de las etapas consiste en la extracción de los datos. En ese sentido, los datos utilizados se extrajeron el 16/07/2022 de la plataforma online "Energía Abierta" de la Comisión Nacional de Energía de Chile (CNE), la cual es el ente regulador del mercado energético chileno. Estos datos corresponden a la energía eléctrica facturada mensual para clientes regulados en Chile durante el período 2015-2021 [17].

El conjunto de datos tiene 338,652 filas y 10 columnas. Las columnas equivalen a las 10 variables existentes, las cuales son: el año en que se consume esta energía facturada ("Year"), el mes en que se consume la energía facturada (Mes), la región del país en la cual está la subestación eléctrica desde donde se abastece al grupo de clientes ("Region"), la comuna de esa región donde la empresa distribuidora hace el retiro de esta energía para los clientes regulados ("Comuna"), el tipo de clientes ya sean residenciales o no residenciales ("Tipo_clientes"), el tipo de tarifa correspondiente para los tipos de clientes ("Tarifa"), la cantidad de clientes que son abastecidos con la energía eléctrica retirada del punto de suministro ("Numero_Clientes"), la energía eléctrica base en kWh facturada a los clientes regulados durante el período informado ("E1_kwh"), la energía eléctrica adicional de invierno en kWh facturada a los clientes regulados ("E2_kwh"), y la energía eléctrica total en kWh facturada a los clientes regulados durante el período informado ("Energia_kwh").

Cada una de las 338,652 filas corresponden a un lote de energía eléctrica retirado del punto de suministro por parte de la empresa distribuidora durante el período informado, para abastecer a un determinado número de clientes, que tienen un mismo tipo de tarifa, y que están ubicados en la misma región y comuna del país.

3. Resultados y discusión

La siguiente sección presenta los resultados obtenidos luego de aplicar la metodología propuesta. Se explica la preparación inicial de los datos, luego el análisis exploratorio de éstos, para finalmente presentar los modelos de aprendizaje automático obtenidos.

3.1 Preparación de los datos

La limpieza y preparación de los datos se hizo aplicando las técnicas mencionadas en [18], utilizando el software Python. Consiste en verificar, y corregir de ser

necesario, que el formato de los datos sea correcto, que no haya datos faltantes, que no haya datos duplicados, entre otras acciones.

Los datos numéricos y los categóricos deben tener el formato correcto, de acuerdo con su naturaleza. Para los datos categóricos se utiliza el formato "object", y para los datos numéricos se utilizan los formatos "int" (entero) o "float" (decimal). En esta investigación, se ajustó el formato del número de clientes de decimal a entero.

Adicionalmente, se verifica la posibilidad de la existencia de datos faltantes. Se detectó un total de 25 datos faltantes, uno en la variable "Numero_Clientes", doce en la variable "E1_kwh", y 12 en la variable "E2_kwh". Estos 25 datos corresponden a 13 filas del conjunto de datos, las cuales fueron alrededor del 0.004% del total de filas, por lo que fueron eliminadas. Por otra parte, se comprobó la posible existencia de filas duplicadas, de las cuales sólo se encontró una de ellas, y fue eliminada, quedando 338,638 filas sin datos faltantes, y sin duplicación.

Ahora, haciendo una revisión más relacionada con el área de negocios de los datos analizados, se detectaron filas que no tenían clientes asociados, es decir, el número de clientes era nulo. Las filas con esta característica de número de clientes nulos no tenían sentido, puesto que el conjunto de datos está referido a la energía eléctrica facturada a un número determinado de clientes regulados. El número de filas con esta situación fueron 4,468, representando sólo el 1.32% del total de filas, por lo que fueron eliminadas del conjunto de datos. Finalmente, quedaron 334,170 filas, para desarrollar el análisis de los datos.

3.2 Análisis exploratorio de los datos

Consistió en un análisis descriptivo de los datos, utilizando tanto herramientas visuales como no visuales, con el fin de obtener un mayor entendimiento de los datos y de la interacción entre las variables. En primer lugar, se determina que en los datos hay presentes tarifas para clientes residenciales y para clientes no residenciales. Para clientes residenciales se tienen: BT1a, BT1b, TRBT2, TRBT3, TRAT1, TRAT2, TRAT3. Para clientes no residenciales se tienen: BT2, BT3, BT4.1, BT4.2, BT4.3, BT5 AT2, AT3, AT4.1, AT4.2, AT4.3, AT5.

Del conjunto de datos analizados se puede establecer que durante el período de estudio se abastecieron un total de 538,384,986 clientes regulados, de los cuales el 96.94% corresponden a clientes con el tipo de tarifa BT1a, mientras que sólo el 0.69% correspondió a clientes regulados con tarifa BT2. En la Tabla 1 se muestran los datos completos.

Tabla 1. Número de clientes por tipo de tarifa

Tarifa	Numero_Clientes	Porcentaje
BT1a	521,915,775	96.94%
BT2	3,691,788	0.69%
BT3	4,105,437	0.76%
BT43	2,580,918	0.48%
AT43	2,087,771	0.39%
BT1b	1,889,195	0.35%
AT2	1,103,566	0.21%
AT3	673,798	0.13%
BT41	160,273	0.03%
AT41	71,807	0.01%
AT42	44,744	0.01%
TRAT1	34,285	0.006%
BT42	24,112	0.005%
TRBT3	542	0.0001%
TRBT2	456	0.0001%
AT5	326	0.0001%
BT5	85	0.0000%
TRAT2	78	0.0000%
TRAT3	30	0.0000%

En cuanto a la energía total facturada a los clientes regulados, durante el período de estudio, alrededor del 50.7% (198,068,287MWh) fue facturada a clientes regulados residenciales, siendo el 49.25% del total correspondientes a clientes residenciales con tarifa BT1a. En la Figura 2 se presenta la energía total facturada por tipo de cliente y por tipo de tarifa, de la cual se puede observar que, en cuanto a los clientes regulados no residenciales, aquellos con tarifas AT43 y BT43, fueron los de mayor energía eléctrica facturada.

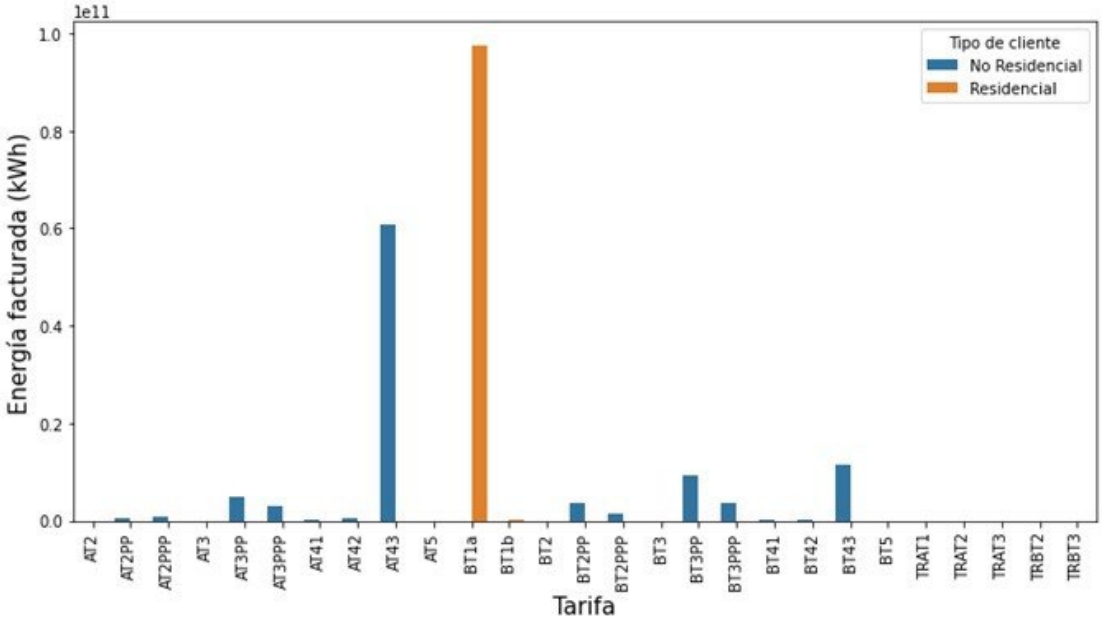

Figura 2. Energía total facturada por tipo de tarifa.
Fuente: Elaboración propia en base a datos de la CNE.

Por otra parte, la energía facturada total anual a clientes regulados para el año 2015, el primero del período de estudio, fue de 30,897,137MWh. Este valor aumentó 1.68% durante el año 2016, pero luego ha disminuido continuamente, 3.11% en el año 2017, 7.45% en el año 2018, 8.50% durante el año 2019, 0.01% durante el año 2020, y finalmente 0.69% durante el año 2021. En la Figura 3 se presenta la información completa, mostrando la energía facturada total anual por tipo de cliente regulado.

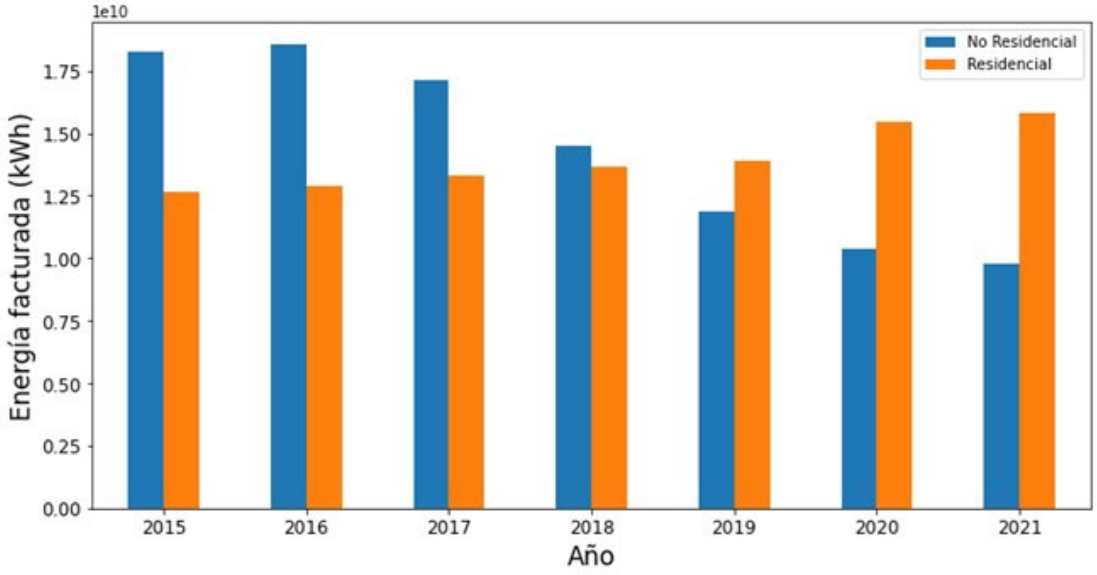

Figura 3. Energía total facturada anual por tipo de cliente.
Fuente: Elaboración propia en base a datos de la CNE.

De la Figura 3 se puede observar que entre el año 2015 y el año 2018, los clientes no residenciales tuvieron una mayor facturación de energía eléctrica. A partir del año 2019 la situación cambió, siendo los clientes residenciales los que tuvieron una mayor facturación. De hecho, desde el año 2017, la energía facturada a los clientes no residenciales ha disminuido constantemente, mientras que la facturación de energía a los clientes residenciales ha aumentado. En cuanto al número de clientes no residenciales, para el año 2021 hay casi 11% de clientes menos de lo que había en el año 2019. Estos resultados coinciden con lo presentado por [19], quien en su investigación plantea que, durante el año 2017, hasta 1100 clientes con potencia instalada entre 500kW y 5000 kW emigraron desde el segmento de clientes regulados al segmento de clientes libres.

En cuanto a la energía facturada mensual, en promedio se muestra un mayor consumo de energía durante los meses de junio, julio y agosto, siendo la mayor facturación en el mes de julio. Los meses de menor facturación promedio corresponden a los meses del verano, es decir, entre noviembre y marzo, siendo febrero el mes de menor facturación promedio de energía eléctrica, durante el período de estudio. La información completa se presenta en la Figura 4.

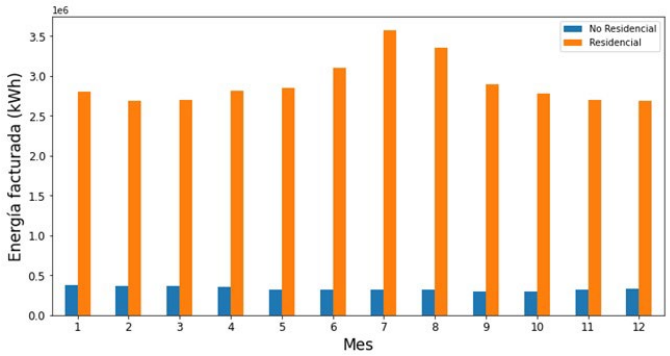

Figura 4. Energía promedio mensual facturada por tipo de cliente.
Fuente: Elaboración propia en base a datos de la CNE.

De la Figura 4 también se puede observar que la energía facturada promedio mensual para los clientes no residenciales se mantiene aproximadamente constante, y la variación mensual de la energía promedio la establecen los clientes residenciales, esto coincide con lo mostrado en [20] en cuanto al consumo promedio de energía eléctrica de los clientes residenciales. Es importante indicar que para los clientes no residenciales se nota una pequeña reducción durante los meses con alto número de feriados, por ejemplo, el mes de septiembre.

De igual manera, se tiene que la Región Metropolitana (RM) lideró las regiones en cuanto a la energía facturada total con un 45.56% del total, seguida por la Región de Valparaíso (V) con un 9.94%, y la Región de Coquimbo (IV) con un 9.18%. La información completa se presenta en la Tabla 2, de la cual se observa que la Región de Aysén (XI) tiene la menor energía facturada durante el período de estudio, con 967,923 MWh.

Tabla 2. Energía facturada por región

Región	Energía Facturada (MWh)	%
RM	90,230,754	45.56
V	19,680,697	9.94
IV	18,184,633	9.18
VIII	14,347,307	7.24
VII	11,488,534	5.80
X	10,113,303	5.11
IX	8,174,859	4.13
II	6,000,846	3.03
XVI	4,266,850	2.15
XIV	4,031,180	2.04
III	3,484,006	1.76
I	3,034,719	1.53
XII	2,102,951	1.06
XV	1,959,725	0.99
XI	967,923	0.49

Al discriminar la energía facturada por el tipo de cliente, se obtiene la Figura 5, de la cual se puede ver que en la RM los clientes residenciales y no residenciales tienen aproximadamente, la misma cantidad de energía facturada, siendo la única región en donde presenta esa paridad.

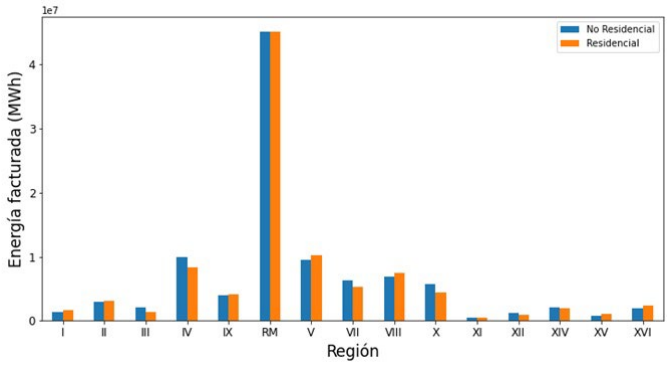

Figura 5. Energía total facturada por región y tipo de cliente.
Fuente: Elaboración propia en base a datos de la CNE.

Sin embargo, lo anterior no se refleja en la cantidad de usuarios por región, en los que los clientes residenciales superan con creces a los clientes no residenciales. Por ejemplo, en la RM el número de clientes residenciales es de 201,809,163, mientras que hay 5,775,938 clientes no residenciales. Este resultado tiene sentido puesto que la carga eléctrica instalada en un local comercial o industrial, por lo general, será mayor a la carga eléctrica instalada en una residencia.

3.3 Aplicación de algoritmos de aprendizaje automático

Se aplicó el algoritmo de agrupamiento K-Means para generar un modelo que permite detectar patrones dentro del conjunto de datos. Adicionalmente, se aplicó el algoritmo de predicción K-NN, para generar un modelo que permite predecir la clase de los datos nuevos que se incorporen al conjunto. Por último, se aplicó el análisis de componentes principales para determinar las variables del conjunto de datos original, que mejor explican la varianza de dicho conjunto.

3.3.1 Aplicación de algoritmo K-Means

El algoritmo de agrupamiento o clustering K-Means, es un algoritmo de aprendizaje no supervisado que busca principalmente definir grupos de tal forma que cada dato dentro de un grupo tenga una variación mínima respecto a los otros integrantes del grupo. De acuerdo con [21], el agrupamiento por K-Means consiste en agrupar juntos objetos que sean similares entre sí. Puede haber más de un grupo, siempre y cuando los objetos de un mismo grupo o clúster sean similares entre sí, y los objetos de grupos diferentes tengan características diferentes entre sí.

En la presente investigación se utiliza K-Means para detectar patrones en los datos, tal como lo hace [22] pero ellos utilizan datos de consumo diario de energía. Ahora, previo a la aplicación del algoritmo, se hace un análisis de correlación entre las variables numéricas para reducir la dimensionalidad del conjunto datos. Como no se tiene un conocimiento previo de la posible normalidad de los datos, se procede a realizar el análisis de correlación considerando tres métodos: Pearson, Spearman y Kendall. Según lo planteado por [23], el coeficiente de Pearson funciona bien para datos cuantitativos y distribuidos normalmente, pero cuando no se cumple la condición de normalidad se deben utilizar alternativas no paramétricas como el estadístico Rho de Spearman o el estadístico Tau de Kendall.

Luego de realizar el análisis, se encontró que hay una alta correlación entre las variables: “Energia_kwh”, “E1_kwh”, y “Numero_Clientes”. Este resultado se obtiene para cada uno de los tres métodos aplicados, y era de esperarse puesto que la energía facturada se mueve en la misma dirección que se mueve el número de clientes que consumen dicha energía. Adicionalmente, la energía facturada base es la componente principal de la energía total facturada. Los resultados completos se presentan en la Tabla 3.

Tabla 3. Coeficientes de correlación con respecto a la energía total

VARIABLES	Pearson	Spearman	Kendall
Energia_kwh	1.00	1.00	1.00
E1_kwh	1.00	1.00	1.00
Numero_Clientes	0.78	0.87	0.69
E2_kwh	0.45	0.34	0.28
Year	0.03	0.03	0.02
Mes	0.01	0.01	0.00

Posteriormente, se desarrolla un análisis de dependencia de las variables categóricas, puesto que se presume que hay dependencia entre los tipos de clientes (Residencial o No Residencial), y las tarifas. Adicionalmente, se presume que hay dependencia entre las regiones y las comunas. Para llevar a cabo el análisis, se crean tablas de contingencia entre cada par de variables, y a cada una de esas tablas se les aplica la Prueba de Independencia de Chi-Cuadrado para variables

categóricas. Se concluye, con un nivel de significancia del 5%, que las variables "Tipo_clientes" y "Tarifa" son dependientes, así como también las variables "Region" y "Comuna".

Por consiguiente, para la aplicación del algoritmo K-Means se descartan las variables "E1_kwh" y "Número_Clientes", debido al análisis de correlación. Además, se descartan las variables: "Tarifa" y "Comuna", debido al análisis de dependencia de las variables categóricas.

El algoritmo tiene como hiperparámetro el número de clústers K, cuyo valor debe fijarlo el usuario. Sin embargo, tal como lo indican en [24], se puede utilizar una metodología para obtener el valor óptimo de K. Ésta se conoce como el "método del codo", para lo cual debe definirse una métrica de optimización. Según lo indicado por [25], la inercia es una métrica muy popular, que se utiliza para obtener el valor óptimo de K, y no es más que el cuadrado de la distancia euclidiana entre cada punto del clúster y su centroide. En [26] se utiliza el método del codo con la inercia como métrica para seleccionar el K óptimo, pero también utilizan la técnica de la métrica Silhouette. En esta investigación, luego de aplicar el método del codo utilizando la inercia como métrica, se obtiene que el valor óptimo de K es 10, la ilustración del método se presenta en la Figura 6.

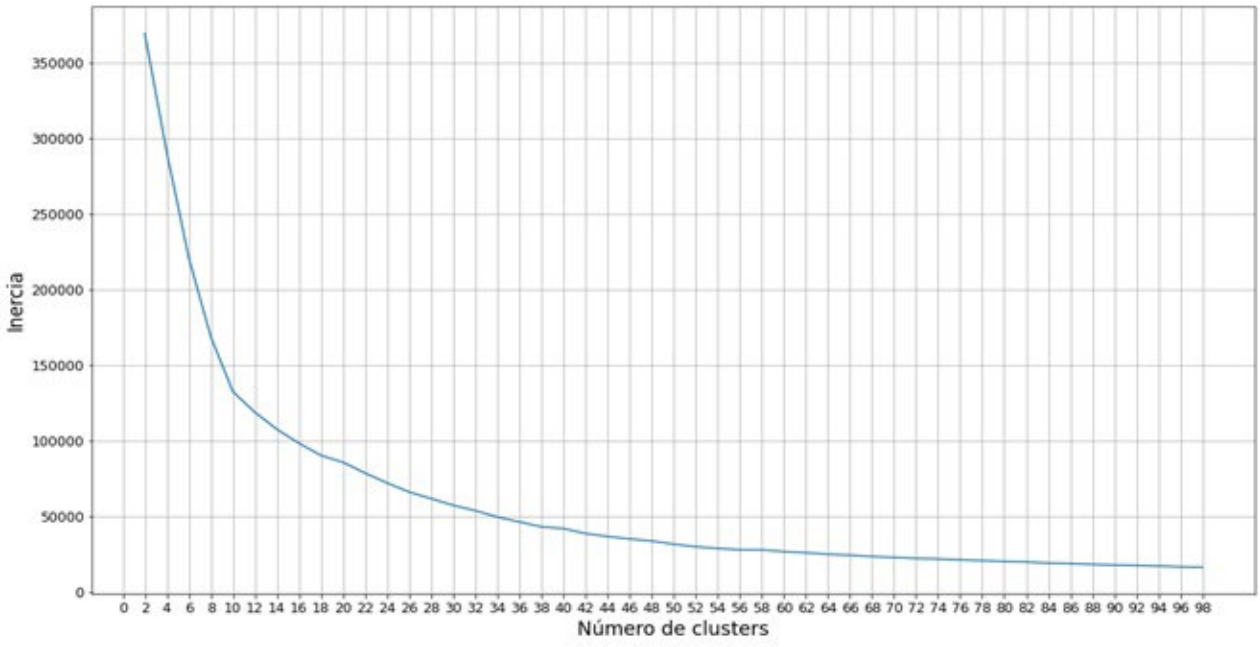

Figura 6. Ilustración del método del codo.
Fuente: Elaboración propia en base a datos de la CNE.

De la Figura 6 se puede observar que se escogió un mínimo local de la inercia y no el mínimo global , debido a que el valor mínimo global ocurre para un valor demasiado alto para el número de clústers. Con ese valor de K=10, se aplica el algoritmo K-Means para detectar patrones en los datos.

En la Figura 7, se presentan los clústers obtenidos y su relación con las regiones. Se puede observar que, en ocho de los diez clústers se tienen datos de una sola Región. Por lo que se puede decir que, los datos se pueden agrupar de acuerdo con la Región geográfica en donde se ubiquen los clientes. Por ejemplo, en el clúster 0 sólo hay clientes de la Región del Biobío (VIII), en el clúster 1 sólo hay clientes de la Región Metropolitana, y así con otras regiones.

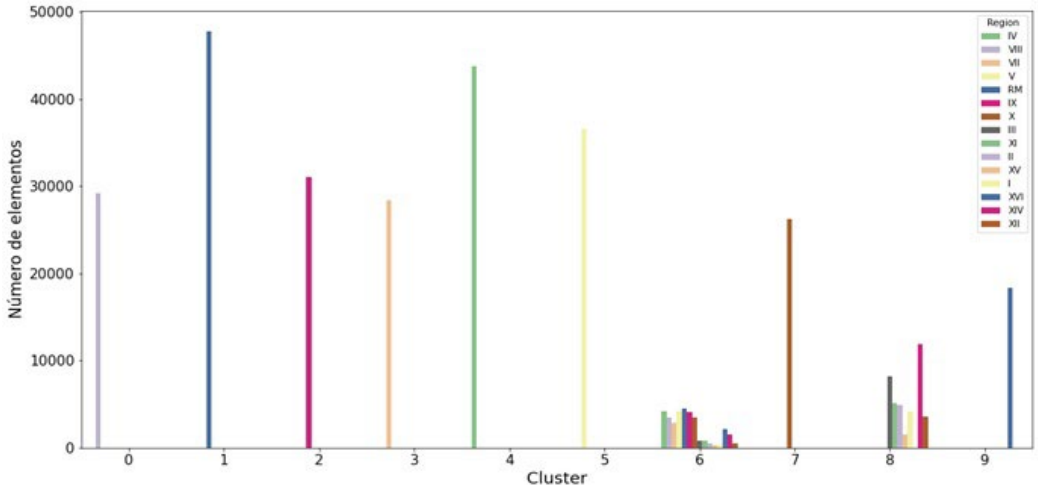

Figura 7. Clústers y su relación con las regiones.
Fuente: Elaboración propia en base a datos de la CNE.

Sin embargo, también se observa que en los clústers 6 y 8 hay clientes de varias regiones, lo cual pudiera indicar que además de las regiones, hay otros posibles patrones en los datos. Profundizando un poco más, se detecta que en el clúster 8 sólo hay clientes no residenciales de las regiones que no están solas en un clúster, mientras que en el clúster 6 sólo se encuentran clientes residenciales. Las regiones que no están solas en un clúster son exclusivamente de los extremos norte y sur del país, lo cual ocurre porque para los extremos del país hay muchos menos habitantes que en las regiones centrales, y por lo tanto menos clientes, y al tratar el algoritmo de que las cantidades por clúster sean similares, completa estos dos clústers con muestras de varias regiones.

Seguidamente, se presentan los clústers obtenidos y su relación con el tipo de clientes. La información se presenta en la Figura 8, de la cual se puede observar que todos los clientes residenciales se encuentran ubicados en el clúster 6, y todos los clientes no residenciales están repartidos en los restantes nueve clústers, por lo que los datos se agrupan de acuerdo con el tipo de cliente.

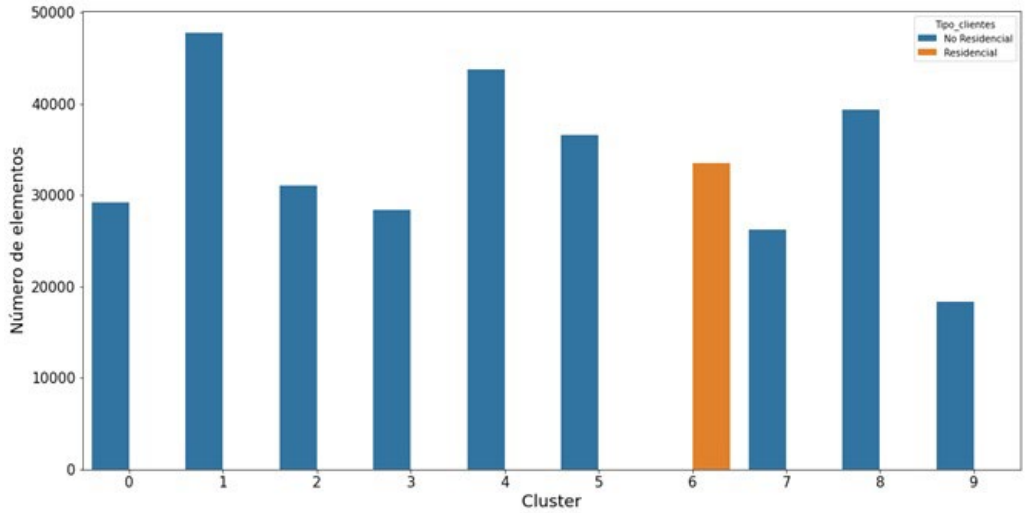

Figura 8. Clústers y su relación con el tipo de clientes.
Fuente: Elaboración propia en base a datos de la CNE.

De la misma forma, se presentan los clústers obtenidos y su relación con el tipo de tarifa de los clientes. Los resultados se presentan en la Figura 9, de la cual se puede observar que el clúster 4 es el que tiene menos variedad en los tipos de tarifa, e indagando con mayor profundidad, se detecta que en dicho clúster hay sólo tarifas asociadas a clientes residenciales, lo que refuerza el hecho que los datos se agrupan de acuerdo con el tipo de clientes.

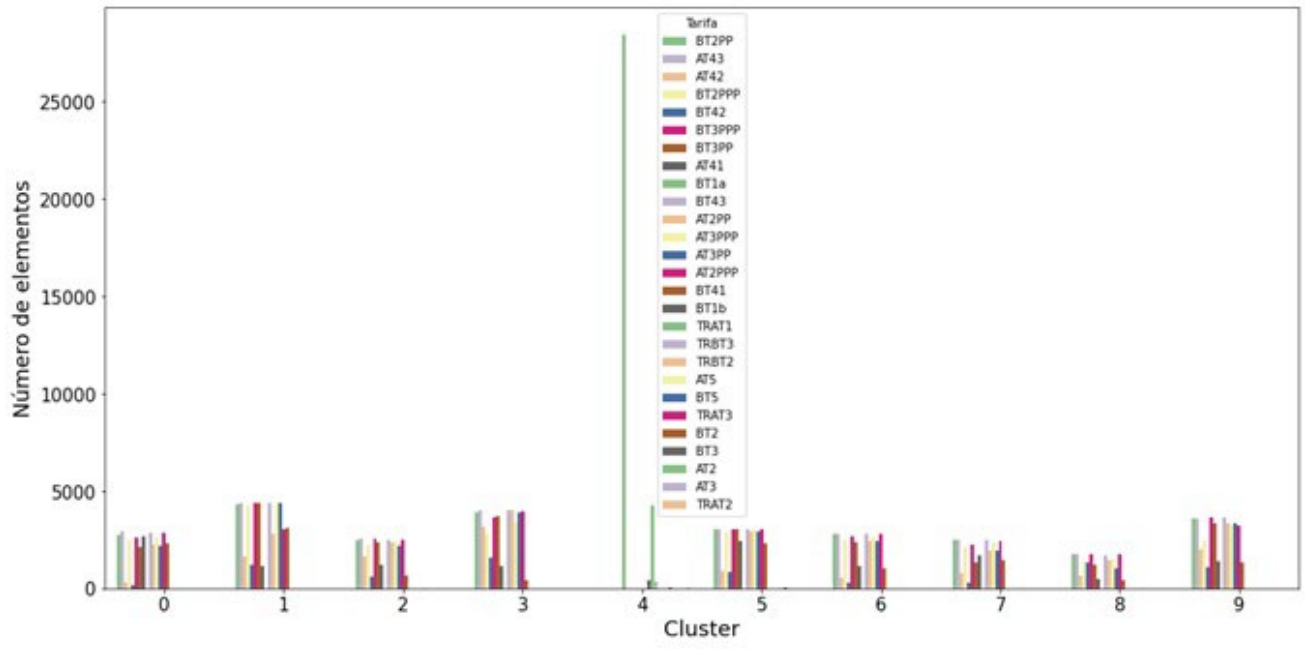

Figura 9. Clústers y su relación con los tipos de tarifas.
Fuente: Elaboración propia en base a datos de la CNE.

En este punto, es importante recordar, que los elementos de los clústers están compuestos por cada una de las filas del conjunto de datos, y que cada fila está asociada a un lote de usuarios y no a un usuario en particular.

3.3.2 Aplicación del algoritmo K-NN

El algoritmo de los K vecinos más cercanos, K-NN, es un algoritmo de aprendizaje supervisado para clasificación, mediante el cual se busca predecir la clase o categoría de un conjunto de datos, a partir de un conjunto de variables predictoras. De acuerdo con [27], K-NN es uno de los algoritmos más simples dentro de los algoritmos de aprendizaje supervisado para clasificación. Funciona comparando la distancia entre cada instancia de referencia y las otras muestras del set de entrenamiento, seleccionando los K vecinos más cercanos a ellas. En [28] se plantea que es un algoritmo que no genera una función discriminativa para clasificar los puntos de datos nuevos.

Para este algoritmo, se trabaja con la variable "Tipo_clientes" como variable objetivo, es decir, el modelo obtenido debe predecir si la instancia que se pruebe pertenece a clientes regulados residenciales o no residenciales. Para generar el modelo, el número de vecinos K debe establecerse inicialmente, pero se puede obtener el valor de K más adecuado optimizando alguna métrica de desempeño. En esta investigación se utilizó la métrica exactitud (accuracy), la cual de acuerdo con [29] "es la métrica que tenemos para evaluar que tan bien nuestra conjetura o predicción coincide con la realidad".

Los resultados obtenidos para obtener el K óptimo se presentan en la Figura 10, de la cual se puede observar que el valor óptimo de K es 6, pues es el valor para el cual se alcanza el valor máximo posible de la exactitud, el cual es 98.4%.

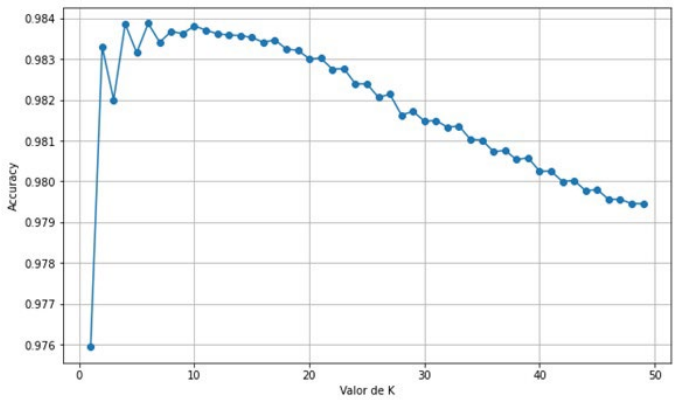

Figura 10. Exactitud vs. Número de vecinos K.
Fuente: Elaboración propia en base a datos de la CNE.

Una vez obtenido el K óptimo, se aplica el algoritmo K-NN para generar el modelo de predicción del tipo de clientes de cada una de las instancias. El conjunto de datos se divide en: el set de entrenamiento, correspondiente al 80% de los datos, y el set de prueba correspondiente al 20% restante de los datos. Con el set de entrenamiento se genera el modelo, y con el set de prueba se evalúa el modelo. Como resultado de la evaluación, se obtiene la matriz de confusión, que es una matriz cuadrada, en la que las celdas tienen la siguiente información: los verdaderos negativos y los verdaderos positivos en la diagonal principal, y los falsos negativos y los falsos positivos en las otras celdas. Para nuestro caso se obtuvo una matriz de 2x2 ya que se tienen sólo dos categorías para clasificar los datos. La matriz se presenta en la Tabla 4.

Tabla 4. Matriz de confusión

Clasificación real	**No Residencial**	60,102	48
	Residencial	1,030	5,654
		No Residencial	**Residencial**
		Predicción	

De la Tabla 4 se puede decir que el conjunto de prueba estuvo compuesto por 66,834 filas del conjunto de datos original, el cuál es el 20% de los datos originales. Adicionalmente, 60,102 filas eran de clientes no residenciales y el modelo los clasificó de esa forma, pero 48 no residenciales fueron clasificados como residenciales. Por otra parte, 5,654 filas eran de clientes residenciales y el modelo los clasificó de esa manera, pero 1,030 residenciales fueron clasificados como no residenciales.

Para efectos de comparación, se generó un nuevo modelo utilizando como objetivo la variable "Tarifa". Es decir, el modelo debe predecir a cuál tarifa pertenece la instancia que se pruebe a través del modelo. Sin embargo, dado que la exactitud máxima obtenida fue del 30%, este modelo se descartó. De igual manera, se generó otro modelo utilizando como variable objetivo "Region". Por tal razón, el modelo debe predecir a cuál región del país pertenece la instancia que se pruebe a través del modelo. Sin embargo, la exactitud máxima que se pudo obtener fue de alrededor de 19%, por consiguiente, este modelo también se descartó. Esto era de esperarse puesto que los datos están claramente distribuidos de acuerdo con el tipo de cliente, pero no es así para las clases de tarifa en la que predomina una sobre las restantes dieciocho, ni para las clases de región en la que predomina una región sobre las quince restantes.

3.3.3 Análisis de componentes principales

El análisis de componentes principales (PCA) es una técnica que nos permite reducir la dimensionalidad de un caso de estudio, en términos de reducir la cantidad de variables que conforman un conjunto de datos, dejando sólo las variables que tienen una mayor incidencia en la explicación de la varianza del sistema. PCA realmente transforma los datos originales en un nuevo conjunto de características 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. Según lo indicado en [30], es una técnica utilizada para reducción de dimensionalidad,

lo cual se lleva a cabo por medio de transformaciones lineales. Esta reducción, lleva a una mejor interpretabilidad de los datos, al encontrar estructuras significativas en los mismos. De igual manera, es muy utilizada para el preprocesado de predictores en el ajuste de modelos de aprendizaje supervisado. De acuerdo con [31], PCA se utiliza en un gran número de aplicaciones, previo a la aplicación de K-Means, con el fin de reducir la cantidad de columnas del conjunto de datos.

Para generar el modelo PCA no se indica el número de componentes principales deseadas, de modo que por defecto será igual al menor valor entre el número de filas del conjunto de datos original menos uno y el número de columnas del mencionado conjunto. En este caso, se obtienen seis componentes principales, las cuales se muestran en la Tabla 5. Cada celda de la tabla tiene un número entre -1 y 1, el cual representa el peso que tiene la variable original son la respectiva componente principal. Mientras más cerca este ese número del valor absoluto de uno, mayor peso tendrá la variable correspondiente. Por ejemplo, el valor 0.09 que está en la primera fila de la componente "0", significa que el año tiene un peso muy poco significativo para esa componente.

Tabla 5. Matriz de componentes principales

Variable vs Componente	0	1	2	3	4	5
Year	0.09	-1.00	-0.03	0.00	0.00	0.00
Mes	0.00	0.03	-1.00	0.00	0.00	0.00
E2_kwh	0.01	0.00	0.00	0.37	0.93	0.00
Energia_kwh	0.02	0.00	0.00	0.93	-0.37	0.00
Tipo_clientes_No Residencial	-0.70	-0.07	-0.01	0.01	0.00	0.71
Tipo_clientes_Residencial	0.70	0.07	0.01	-0.01	0.00	0.71

Por otra parte, en la Figura 11 se presenta la proporción de varianza que explica cada componente principal, de la cual se observa que las tres primeras componentes principales ya explican el 99% de la varianza del sistema. La componente principal "0" explica el 46% de la varianza, y las variables que tienen mayor peso en esa componente es el tipo de clientes. Para la componente principal "1" la variable con mayor peso es el año (Year), y para la componente principal "2" la variable más significativa es el mes.

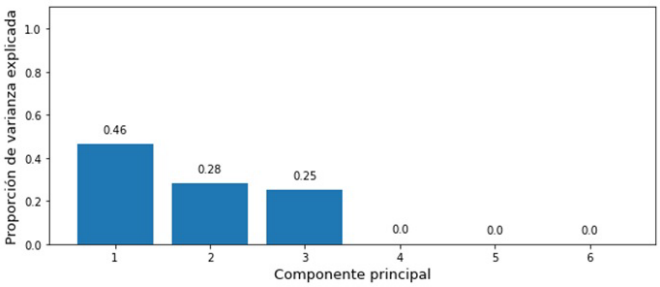

Figura 11. Proporción de varianza explicada por cada componente principal.
Fuente: Elaboración propia en base a datos de la CNE

Este resultado está acorde con lo obtenido con las otras técnicas, es decir, el tipo de cliente es la variable de mayor impacto en el conjunto de datos, explicando la mayor parte de la variabilidad presente. Seguidamente, tenemos el año como la otra variable importante, y tiene sentido puesto que, como se vio previamente, es dinámico año a año, con una tendencia ascendente. Por último, se tiene a la variable mes, que también explica parte de la variabilidad de los datos, al estar presente la estacionalidad mensual del consumo de acuerdo con la estación del año.

4. Conclusiones

Más del 96% de los clientes presentes en el conjunto de datos son de tipo residencial, quienes consumieron un poco más del 50% de la energía total facturada durante el período de estudio, y que por lo tanto tuvieron un consumo unitario más bajo, en comparación con los clientes no residenciales.

Los datos de energía facturada promedio mensual presentan una estacionalidad, con una mayor facturación entre los meses de abril y octubre, resaltando el mes de julio como el de mayor facturación de energía eléctrica a los clientes regulados. La estacionalidad es significativamente establecida por los clientes residenciales.

Desde un punto de vista geográfico, en la Región Metropolitana se facturó el 45,56% de toda la energía facturada durante el período de estudio, y los tipos de clientes se dividen el total de facturación en

aproximadamente partes iguales. No obstante, en esta región el número de clientes no residenciales representa sólo el 2,9% del total de clientes residenciales.

En la aplicación del algoritmo K-Means se obtuvo que el valor óptimo del número de clústers es diez, y que estos se agrupan principalmente de acuerdo con el tipo de cliente.

Luego de aplicar el algoritmo K-NN, se logró obtener un modelo para predecir el tipo de cliente de los datos, con una exactitud del 98%.

El análisis de componentes principales nos indica que las variables del conjunto de datos que mejor explican su varianza son el año, el mes y el tipo de cliente. De las tres variables mencionadas, el tipo de clientes es la más significativa de todas.

Se recomienda continuar la investigación generando modelos de regresión lineal múltiple para predecir el consumo de energía eléctrica por tipo de cliente en el corto plazo, y comparar sus resultados con un modelo de serie de tiempo.

AGRADECIMIENTOS

A la Corporación de Fomento de la Producción de Chile y a la iniciativa público-privada Talento Digital, por permitir actualizar mis competencias en el último año, pues éstas contribuyeron al desarrollo de esta investigación.

CONFLICTO DE INTERESES

El autor declara no tener algún conflicto de interés.

CONTRIBUCIÓN Y APROBACIÓN DE LOS AUTORES

El autor contribuyó con el 100% de la investigación realizada.

El autor afirma que leyó y aprobó la versión final de este artículo.

REFERENCIAS

[1] Sociedad Alemana de Cooperación Internacional, "Las energías no renovables en el mercado eléctrico chileno," Deutsche Gesellschaft für Internationale Zusammbenardeit (GIZ) GmbH, Santiago de Chile, 2020.

[2] S. Arguello V. y N. García B., "Componentes y determinación de la tarifa eléctrica para los clientes regulados ," Biblioteca del Congreso Nacional de Chile, Santiago de Chile, 2020.

[3] M. A. Azócar, "Estudio y análisis del Nuevo Decreto Tarifario 11 T. Aplicable a los suministros sujetos a precios," Tesis de Pregrado, Pontificia Universidad Católica de Valparaíso, Valparaíso, 2018.

[4] Mercados Energéticos Consultores, "Análisis de consumo eléctrico en el corto, mediano y largo plazo," Mercados Energéticos Consultores, Santiago de Chile, 2014.

[5] A. Rajabi, M. Eskandari, M. J. Ghadi, L. Li, J. Zhang y P. Siano, "A comparative study of clustering techniques for electrical load pattern segmentation," Renewable and Sustainable Energy Reviews, vol. 120, 2020. https://doi.org/10.1016/j.rser.2019.109628.

[6] M. Lester, D. Carrizo, F. Ulloa-Vásquez y L. García-Santander, "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, vol. 29, nº 4, pp. 778-787, 2021. http://dx.doi.org/10.4067/S0718-33052021000400778.

[7] T. Parhizkar, E. Rafieipour y A. Parhizkar, "Evaluation and improvement of energy consumption prediction models using principal component analysis based feature reduction," Journal of Cleaner Production, vol. 279, 2021. https://doi.org/10.1016/j.jclepro.2020.123866

[8] M. K. M. Shapi, N. A. Ramli y L. J. Awalin, "Energy consumption prediction by using machine learning for smart building: Case study in Malaysia ," Developments in the Built Environment, vol. 5, 2021. https://doi.org/10.1016/j.dibe.2020.100037.

[9] S. Yilmaz, J. Chambers, X. Li y M. K. Patel, "A comparative analysis of patterns of electricity use and flexibility potential of domestic and non-domestic building archetypes through data mining techniques," Journal of Physics: Conference Series, vol. 2042, 2021. DOI:10.1088/1742-6596/2042/1/012021.

[10] E. Ruiz, R. Pacheco-Torres y J. Casillas, "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), pp. 1-7, 2017. doi: 10.1109/ETFA.2017.8247667.

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

[12] O. F. Núñez-Barrionuevo, E. A. Llanes-Cedeño, J. Martinez-Gomez, J. I. Guachimboza-Davalos y J. Lopez-Villada, "Clustering Analysis of Electricity Consumption of Municipalities in the Province of Pichincha-Ecuador Using the K-Means Algorithm," Proceedings of ICCIS 2020

Springer, vol. 1273, pp. 187-195, 2020. https://doi.org/10.1007/978-3-030-59194-6_16.

[13] O. Valgaev, F. Kupzog y H. Schmeck, "Building power demand forecasting using K-nearest neighbours model – practical application in Smart City Demo Aspern project," CIRED - Open Access Proceedings Journal, vol. 2017, nº 1, p. 1601 – 1604, 2017. DOI:10.1049/oap-cired.2017.0419.

[14] S. Pazi, C. M. Clohessy y G. D. Sharp, "A framework to select a classification algorithm in electricity fraud detection," South African Journal of Science, vol. 116, nº 9-10, pp. 1-7, 2020. http://dx.doi.org/10.17159/sajs.2020/8189.

[15] D. Cielen, A. D. B. Meysman y M. Ali, Introducing Data Science, Shelter Island, NY: Manning Publications Co., 2016.

[16] C. A. Bernal, Metodología de la investigación, Bogotá: Pearson Educación, 2010.

[17] Comisión Nacional de Energía de Chile, "Energía Abierta," 10 March 2022. [En línea]. Available: http://energiaabierta.cl/categorias-estadistica/electricidad/?sf_paged=2. [Último acceso: 16 July 2022].

[18] W. McKinney, Python for Data Analysis, Sebastopol, CA: O'Reilly Media, Inc., 2018.

[19] M. A. Salazar Córdova, "Impactos de la emigración de clientes regulados al mercado libre. Catastro, evolución y efectos en los clientes y en las empresas proveedoras (generación y distribución)," Tesis de Maestría, Universidad Técnica Federico Santa María, Santiago de Chile, 2018.

[20] B. M. Mellado Leal, "Aplicaciones de Data Science para la mejora de la medición y cobro de la distribución de la energía eléctrica en contextos de pandemia mundial," Tesis de Pregrado, Universidad de Chile, Santiago de Chile, 2018.

[21] L. Igual y S. Seguí, Introduction to Data Science - A Python Approach to Concepts, Techniques and Applications, Switzerland: Springer International Publishing, 2017.

[22] G. N. Pizarro Herrera, "Reconocimiento de patrones y pronóstico de consumo eléctrico," Tesis de Pregrado, Pontificia Universidad Católica de Valparaiso, Valparaíso, 2017.

[23] J. Amat Rodrigo, "Ciencia de Datos, Estadística, Machine Learning y Programación," Joaquin Amat Rodrigo, [En línea]. Available: https://www.cienciadedatos.net/documentos/pystats05-correlacion-lineal-python.html. [Último acceso: 16 Julio 2022].

[24] E. Umargono, J. E. Suseno y V. Gunanwan S.K, "K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula," Advances in Social Science, Education and Humanities Research, vol. 474, 2019. DOI:10.2991/assehr.k.201010.019.

[25] E. Russano y E. Ferreira Avelino, Fundamentals Of Machine Learning Using Python, Cánada: Arcler Press, 2020.

[26] W. Kong, Y. Wang, H. Dai, L. Zhao y C. Wang, "Analysis of energy consumption structure based on K-means clustering algorithm," de E3S Web of Conferences 267, 01054 (2021), Beijing, 2021. https://doi.org/10.1051/e3sconf/202126701054

[27] W.-M. Lee, Python Machine Learning, Indianapolis: John Wiley & Sons, Inc., 2019.

[28] S. Raschka y V. Mirjalili, Python Machine Learning - Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow, Birmingham: Packt Publishing Ltd., 2017.

[29] M. E. Fenner, Machine Learning with Python for Everyone, Boston: Pearson Education, Inc., 2020.

[30] S. Shalev-Shwartz y S. Ben-David, Understanding Machine Learning - From Theory to Algorithms, Cambridge: Cambridge University Press, 2014.

[31] R. D. Dana, D. Soilihudin, R. H. Silalahi, D. Kurnia y U. Hayati, "Competency test clustering through the application of Principal Component Analysis (PCA) and the K-Means algorithm," de IOP Conf. Series: Materials Science and Engineering 1088 (2021) 012038, Cirebon, 2021. doi:10.1088/1757-899X/1088/1/012038.