https://doi.org/10.37815/rte.v34n4.963
Artículos originales

Uso de algoritmos de aprendizaje automático para analizar los datos de energía eléctrica facturada en la Región Metropolitana de Chile durante el período 2015-2021

Use of machine learning algorithms to analyze data on electricity billed in the Metropolitan Region of Chile during the period 2015-2021

César A. Yajure Ramírez[1] https://orcid.org/0000-0002-3813-7606

[1]*Universidad Central de Venezuela*, Caracas, Venezuela
cyajure@gmail.com

Enviado: 2022/08/18
Aceptado: 2022/12/01
Publicado: 2022/12/30

Resumen

En la presente investigación se hace el análisis de los datos de energía eléctrica facturada a los clientes regulados en la región metropolitana de Chile durante el período 2015-2021, con el fin de establecer las características de la estructura de los datos, la relación entre las variables, predecir las clases de los registros nuevos, e identificar los patrones subyacentes en los datos. Para ello se utilizó el análisis estadístico descriptivo y los algoritmos de aprendizaje automático K-Means y K-NN. Se pudo establecer que, para el período de estudio, el consumo de energía unitario promedio para clientes residenciales fue de 453 kWh, y de 10.315 kWh para clientes no residenciales. Asimismo, se estableció que hay dependencia entre el número de clientes y la energía eléctrica facturada, al igual que entre la comuna y la empresa distribuidora. Por otra parte, con el algoritmo K-Means se obtuvo un modelo que agrupó los datos de acuerdo con el tipo de cliente, así como también de acuerdo con el tipo de empresa de distribución eléctrica que abastece a los clientes regulados. Por medio del algoritmo K-NN se obtuvo un modelo para predecir el tipo de cliente de los registros nuevos, con una exactitud del 99,73%.

Palabras clave: Ciencia de datos, clientes regulados, tipos de clientes, K-NN, K-Means.

Abstract

This research presents the data analysis of electrical energy billed to regulated clients in the

Sumario: Introducción, Metodología, Resultados y Discusión y Conclusiones.

Como citar: Yajure, C. (2022). Uso de algoritmos de aprendizaje automático para analizar los datos de energía eléctrica facturada en la Región Metropolitana de Chile durante el período 2015-2021. *Revista Tecnológica - Espol, 34(4), 137-152.* http://www.rte.espol.edu.ec/index.php/tecnologica/article/view/963

metropolitan region of Chile during 2015-2021 to establish the characteristics of the data structure and the relationship between the variables. It also aims to predict the classes of new records, and to identify underlying patterns in the data. This study uses descriptive statistical analysis, and the K-Means and K-NN machine learning algorithms. For this study period, it was established that the average unit energy consumption for residential customers was 453 kWh, and 10,315 kWh for non-residential customers. Likewise, there is a dependency between the number of clients and the electricity billed, as well as between the commune and the distribution company. On the other hand, the K-Means algorithm suggests a model that groups the data according to the type of customer and the type of electricity distribution company that supplies regulated customers. The application of the K-NN algorithm resulted in a model to predict the type of client of the new records, with an accuracy of 99.73%.

Keywords: Data Science, regulated customers, types of clients, K-NN, K-Means.

Introducción

En Chile se tienen tres sistemas eléctricos independientes, el sistema eléctrico nacional 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 en la región del mismo nombre, y el Sistema de Magallanes que abarca la región de Magallanes y la Antártica Chilena.

Desde el punto de vista de los usuarios, la normativa chilena establece dos segmentos principales en el área de consumo de energía eléctrica: clientes regulados y clientes libres. De acuerdo con la Sociedad Alemana de Cooperación Internacional (2020, p. 30), "el segmento de clientes regulados lo conforman consumidores con una potencia conectada igual o inferior a 5 MW, pero aquellos con una potencia conectada entre 500 kW y 5 MW, y que están ubicados en el área de concesión de una empresa distribuidora, pueden optar a ser clientes libres". Por el contrario, el segmento de clientes libres está compuesto por consumidores cuya potencia conectada es superior a 5 MW, y que pueden pactar libremente los precios y condiciones con sus suministradores. Aquellos con potencia superior a 500 kW que opten a ser clientes libres, deben permanecer al menos 4 años en esta categoría. La principal razón que impulsa el traspaso de clientes regulados a clientes libres es el monto que se debe pagar por el consumo de electricidad, y según García Bernal (2019) desde el año 2018 el monto por el kWh de energía ha sido mayor en la tarifa de clientes regulados con respecto a la de los clientes libres, y se espera que se mantenga esa tendencia, por lo menos hasta el año 2028.

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 Argüello y García (2020, p. 1), "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 Argüello y García, 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, y define dos niveles de voltaje para los tipos de tarifas. Según (Azócar Rojas, 2018) 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 Mercados Energéticos Consultores (2014) 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.

Ahora bien, con el fin 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 desempeño 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, 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 región metropolitana 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. 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, y el algoritmo K-NN para predecir las categorías de nuevos datos.

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 Rajabi et al. (2020) 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. De igual forma, en (M. Shapi, Ramli, & Awalin, 2021) 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, (Yilmaz, Chambers, Li, & Patel, 2021) 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. Adicionalmente, (Valgaev, Kupzog, & Schme, 2017) 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.

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.

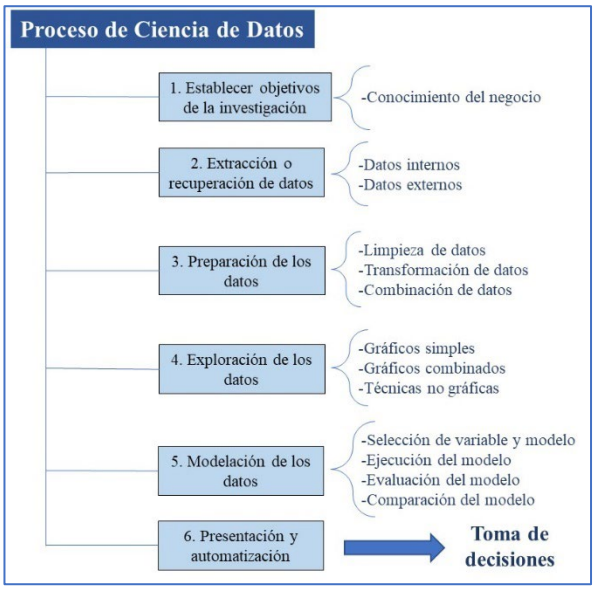
Metodología

Este trabajo tiene rasgos de una investigación descriptiva, asociados 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 Bernal (2010), en la investigación descriptiva se refieren las características del fenómeno objeto de estudio. Adicionalmente Bernal plantea que en la investigación de tipo explicativa se analizan causas y efectos de la relación entre variables existentes.

Por otra parte, para realizar el análisis de grandes cantidades de datos, con el fin de extraer de ellos la información pertinente para la toma de decisiones se utiliza lo que se conoce como la Ciencia de Datos. De acuerdo con Cielen y otros (2016), 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 usualmente se ejecuta 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 características. En ese sentido, las etapas que conforman un proceso de Ciencia de Datos se presentan en la Figura 1.

Figura 1

Etapas del proceso de la Ciencia de Datos



De la Figura 1 se puede observar que la primera etapa consiste en establecer los objetivos de la investigación, la cual requiere tener un conocimiento básico del negocio del que se generan los datos a analizar. En esta investigación se desarrollan las seis etapas del proceso y se aplican a los datos de energía eléctrica facturada mensual en la Región Metropolitana (RM) de Chile, durante el período 2015-2021. La etapa 2 se presenta en esta sección, y las restantes etapas se presentan en la siguiente sección.

Extracción y descripción del conjunto de datos

Los datos utilizados se extrajeron el 16/07/2022 de la plataforma online "Energía Abierta" de la Comisión Nacional de Energía de Chile (2022), 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.

El conjunto de datos tiene 338.652 filas y 11 columnas. Las columnas equivalen a las 11 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 comuna 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"), la energía eléctrica total en kWh facturada a los clientes regulados durante el período informado ("Energia_kwh"), el precio equivalente de la energía en pesos por kWh (PEE), el precio equivalente de la potencia en pesos por kW (PEP).

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.

Resultados y Discusión

Seguidamente, se aplican las etapas restantes del proceso de Ciencia de Datos, y se discuten los resultados obtenidos.

Preparación de los datos

La limpieza y preparación de los datos se hizo aplicando las técnicas mencionadas por (McKinney, 2018), utilizando el lenguaje de programación Python. Incluyó, entre otras técnicas, la verificación del formato adecuado de los datos, corrigiendo cuando era necesario, verificación de datos faltantes, y en caso de haberlos, aplicación de la técnica de imputación adecuada, verificación de datos duplicados, transformación de datos, y combinación de datos.

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, solo fue necesario ajustar el formato del número de clientes de decimal a entero.

Adicionalmente, se detectaron 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 correspondieron 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 fue de 4.468, representando sólo el 1,32% del total filas, por lo que fueron eliminadas del conjunto de datos, quedando 334.170 filas. Posteriormente, se filtraron los datos de manera tal de trabajar sólo con los de la Región Metropolitana, después de lo cual quedaron 50.960 filas. A continuación, se consideró que la energía eléctrica facturada se puede tomar como un proxy del consumo de energía, por lo que se combinaron las columnas "Número_Clientes" y "Energía_kwh" para calcular el consumo unitario en kWh y agregarlo como una columna adicional ("ConsUnit_kwh"). Finalmente, se agregó una columna con la empresa distribuidora de electricidad correspondiente a cada grupo de clientes abastecidos.

Análisis exploratorio de los datos

Consistió en el desarrollo de un análisis descriptivo de los datos, utilizando tanto herramientas visuales como analíticas, con el fin de obtener un mayor entendimiento de éstos, y de la interacción entre las variables. El conjunto de datos disponibles en este punto está compuesto por 50.960 filas y 13 columnas, correspondientes a los datos de la energía eléctrica facturada en la RM.

En primer lugar, se comprueba que en los datos hay tarifas para clientes residenciales y para clientes no residenciales. Para clientes residenciales se tiene únicamente la tarifa BT1a. Para clientes no residenciales se tienen las tarifas: BT2, BT3, BT4.1, BT4.2, BT4.3, AT2, AT3, AT4.1, AT4.2, AT4.3.

Del conjunto de datos analizados se puede establecer que, durante el período de estudio, se abastecieron mensualmente, en promedio, 2.741.233 clientes regulados, equivalentes a un promedio anual de 29.654.800 clientes. Del total del período, el 97,22% correspondióa clientes regulados residenciales con el tipo de tarifa BT1a, mientras que sólo el 1,04% correspondió a clientes regulados con tarifa BT2. En la Tabla 1 se muestran los datos para todo el período de estudio.

Tabla 1

Cantidad de clientes abastecidos por tipo de tarifa

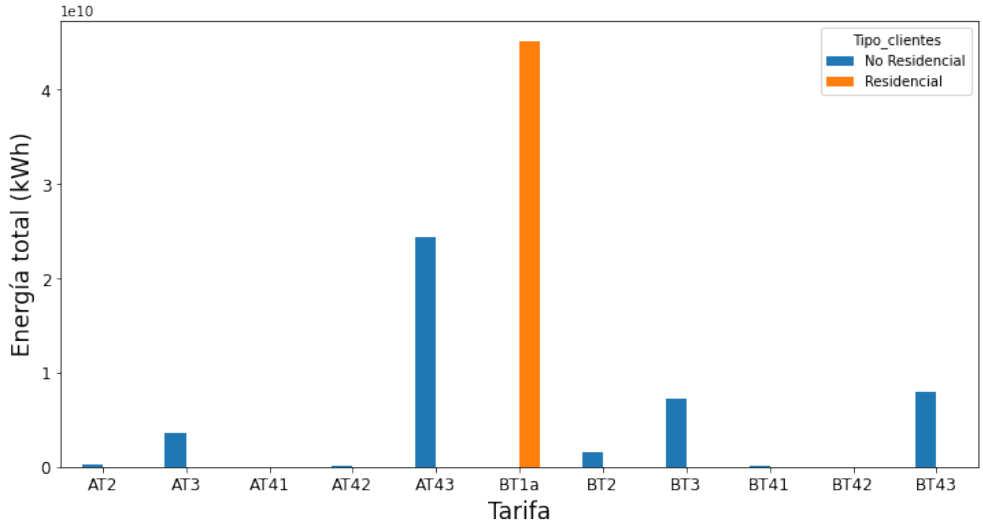
Tarifa	Numero_Clientes	%
BT1a	201.809.155	97,218
BT3	2.166.553	1,0437
BT43	1.474.378	0,7103
BT2	1.137.571	0,5480
AT43	550.715	0,2653
AT3	240.765	0,1160
AT2	175.825	0,0847
BT41	14.605	0,0070
AT42	9.146	0,0044
BT42	2.750	0,0013
AT41	2.135	0,0010

Por otra parte, durante el período de estudio se facturó un total de 90.230.754 MWh, siendo aproximadamente 50% a clientes regulados no residenciales, y el otro 50% a clientes residenciales. El consumo unitario promedio para clientes residenciales fue de aproximadamente 453 kWh, mientras que para los clientes no residenciales fue 10.315 kWh. Sin embargo, es importante mencionar que entre los años 2015 y 2019 el consumo unitario

residencial fue en promedio de 224 kWh, pero en el año 2020 en el que ocurrieron las cuarentenas por la pandemia de la Covid-19, subió a 1.759,28 kWh, cayendo nuevamente en el año 2021 hasta 258,92 kWh. El aumento del consumo durante el año 2020 coincide con lo presentado por (Moreno, y otros, 2020), quienes indican que, durante el mes de junio del 2020 el consumo residencial aumentó 17% con respecto al mismo mes del año 2019. 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

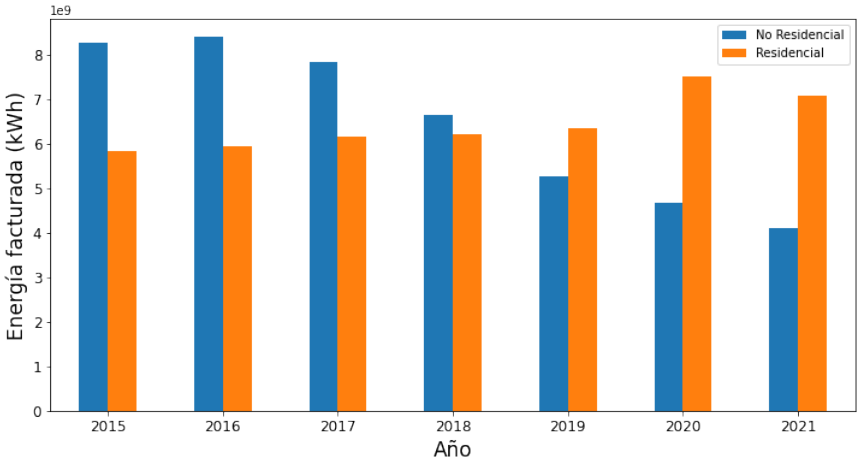
Energía facturada por tarifa y tipo de cliente



Por otra parte, la energía facturada total anual a clientes regulados para el año 2015, el primero del período de estudio, fue de 14.325.878,7 MWh. Este valor aumentó 1,48% durante el año 2016, pero luego ha disminuido continuamente, 2,37% en el año 2017, 8,14% en el año 2018, y 9,69% durante el año 2019. Luego aumentó 4,87% durante el año 2020 impulsada por el sector residencial, pero finalmente cayó 8,09% durante el año 2021. En total, durante el período de estudio, cayó casi 21%. En la Figura 3 se presenta la información completa, mostrando la energía facturada total anual por tipo de cliente regulado.

Figura 3

Energía facturada por año y tipo de cliente

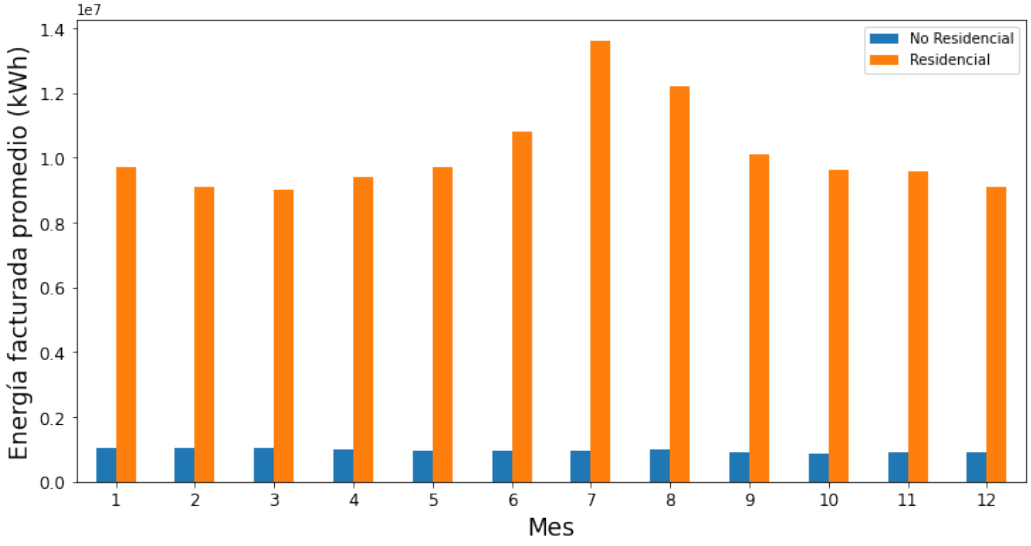


De la Figura 3 se puede observar que entre el año 2015 y el año 2018, los clientes no residenciales tuvieron una mayor facturación de energía eléctrica. A partir del año 2019 la situación cambió, siendo los clientes residenciales los que tuvieron una mayor facturación. De hecho, desde el año 2017, la energía facturada a los clientes no residenciales ha disminuido constantemente, mientras que la facturación de energía a los clientes residenciales ha aumentado desde el año 2015. En cuanto al número de clientes no residenciales, para el año 2021 hay 9,7% menos de lo que había en el año 2019. Estos resultados coinciden con lo presentado por Salazar Córdova (2018), quien en su investigación plantea que, durante el año 2017 hasta 1100 clientes con potencia instalada entre 500 kW y 5000 kW emigraron desde el segmento de clientes regulados al segmento de clientes libres.

Respecto 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, específicamente los meses de diciembre, febrero y marzo, siendo marzo 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 facturada promedio por mes y tipo de cliente



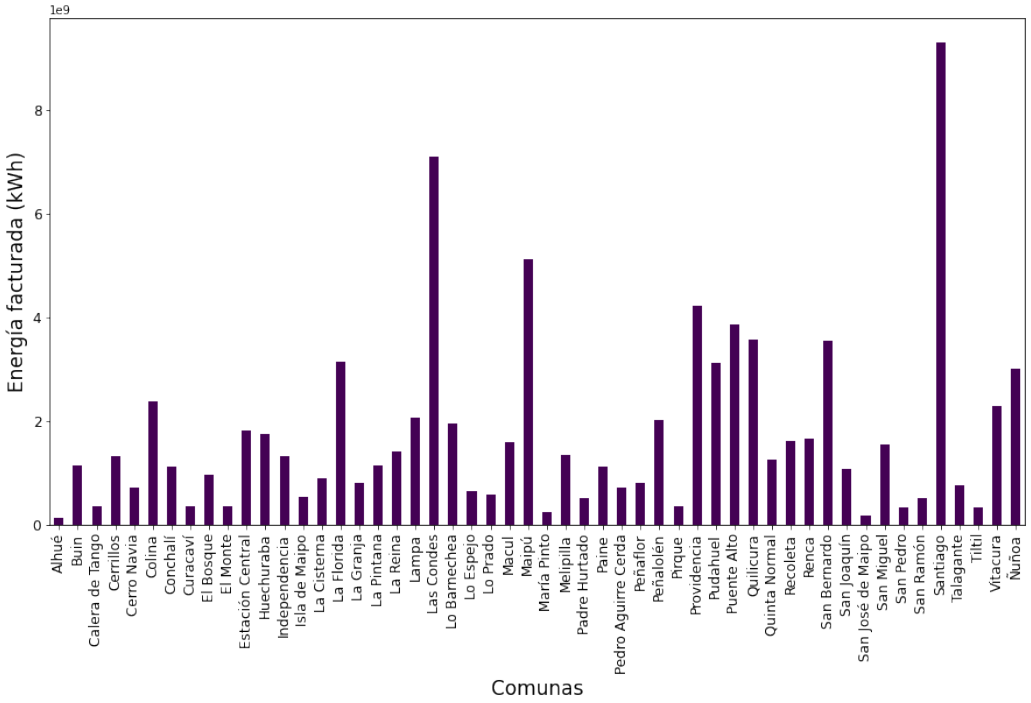
De la Figura 4 también se puede observar que la energía facturada promedio mensual para los clientes no residenciales se mantiene aproximadamente constante, y la variación mensual de la energía promedio la establecen los clientes residenciales; esto coincide con lo mostrado en (Mellado Leal, 2021) en 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.

Referente a las comunas, se tiene que, durante el período de estudio, Santiago es la comuna en la que ha habido mayor energía eléctrica facturada con un total de 9.317.092,7 MWh y un promedio para el año 2021 de 21.590 clientes, seguida de Las Condes con 7.113.678,2 MWh y un promedio para el año 2021 de 13.811,7 clientes. La comuna de Alhué es la que ha tenido la menor cantidad de energía facturada con 131.683,4 MWh para un

promedio de 218,7 clientes para el año 2021. En la Figura 5 se presenta la información completa sobre la energía eléctrica facturada por comuna durante el período de estudio.

Figura 5

Energía facturada por comuna



De la Figura 5 se observa que la comuna de Maipú es la tercera con mayor energía eléctrica facturada con 5.136.314,6 MWh y con un promedio para el año 2021 de 13.373,8 clientes, el cual es mayor al de Las Condes que sin embargo tiene un 27,8% más de energía facturada. En ese sentido, el consumo unitario de energía promedio mensual en la comuna de Las Condes es de 262,27 kWh para clientes residenciales y de 20.495,51 kWh para clientes no residenciales, mientras que en la comuna de Maipú es de 207,02 kWh y 18.168,24 kWh, respectivamente. Para el caso de la comuna de Santiago se tienen 185,41 kWh y 14.566,21 kWh, y en la comuna de Alhué se tienen 199,17 kWh y 13.789,7 kWh, respectivamente.

Aplicación de algoritmos de aprendizaje automático

A continuación, se presenta la aplicación de los algoritmos de aprendizaje automático al conjunto de datos, y los resultados correspondientes. En ese sentido, 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 registros nuevos que se incorporen al conjunto.

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 dentro de los datos, de tal forma que cada dato dentro de un grupo tenga una variación mínima respecto a los otros integrantes del grupo. De acuerdo con Igual y Seguí (2017), 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 hacen en Pizarro Herrera (2017), con la salvedad que 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 Amat Rodrigo (2022), 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 (mayor a 0,65 en magnitud) 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 valores de coeficiente de correlación con respecto a la energía total facturada ("Energía_kwh") se presentan en la Tabla 2.

Tabla 2

Coeficientes de correlación con la variable Energía_kwh

Variable	Pearson	Spearman	Kendall
Energia_kwh	1,0000	1,0000	1,0000
E1_kwh	0,9995	1,0000	0,9996
Numero_Clientes	0,7714	0,8474	0,6668
E2_kwh	0,4781	0,4223	0,3413
ConsUnit_kwh	0,1662	0,2970	0,2439
Year	0,0299	0,0231	0,0165
PEE	0,0196	0,0231	0,0168
PEP	0,0111	0,0012	0,0009
Mes	0,0052	0,0088	0,0061

Posteriormente, se desarrolla un análisis de dependencia de las variables categóricas, puesto que se presume que hay dependencia entre los tipos de clientes y las tarifas, así como entre las distribuidoras 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 "Distribuidora" y "Comuna".

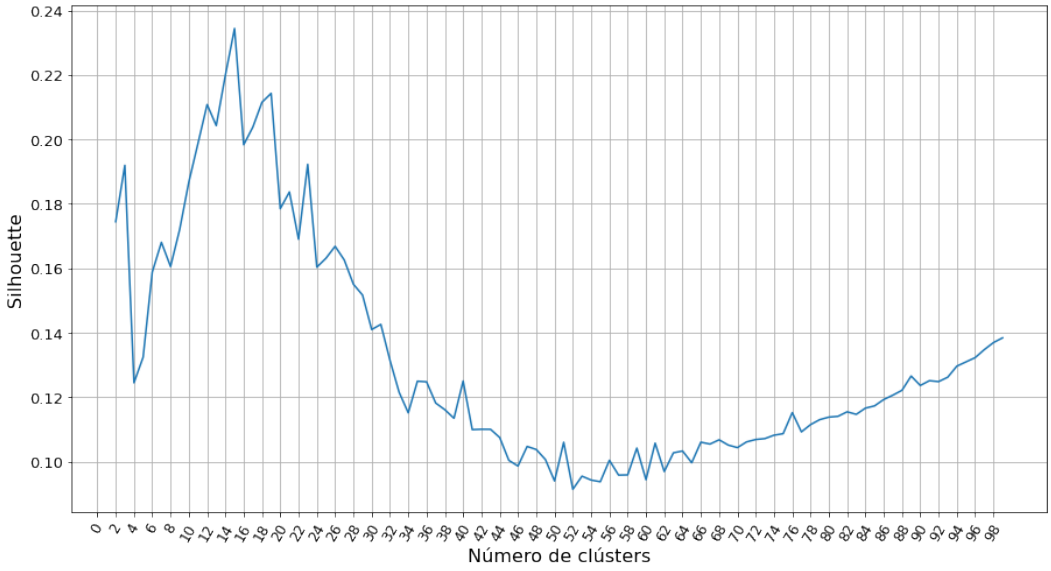
Por consiguiente, para la aplicación del algoritmo K-Means se descartan las variables "E1_kwh" y "Numero_Clientes", debido al análisis de correlación. Adicionalmente, se descartan las variables: "Tarifa" y "Comuna", debido al análisis de dependencia de las variables categóricas.

El algoritmo K-Means tiene como hiperparámetro el número de clústers K, cuyo valor debe ser definido por el usuario. Sin embargo, tal como lo indican en Umargono et al. (2020), 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 Russano y Ferreira (2020), 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 su investigación, Kong et al. (2021) utilizan el método del codo con la inercia como métrica para seleccionar el K óptimo, pero también utilizan la técnica de maximizar el valor 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 16. De igual manera, se utilizó la métrica Silhouette para obtener el K óptimo, resultando el valor de 15. La ilustración del método utilizando la métrica Silhouette se presenta en la Figura 6.

Figura 6

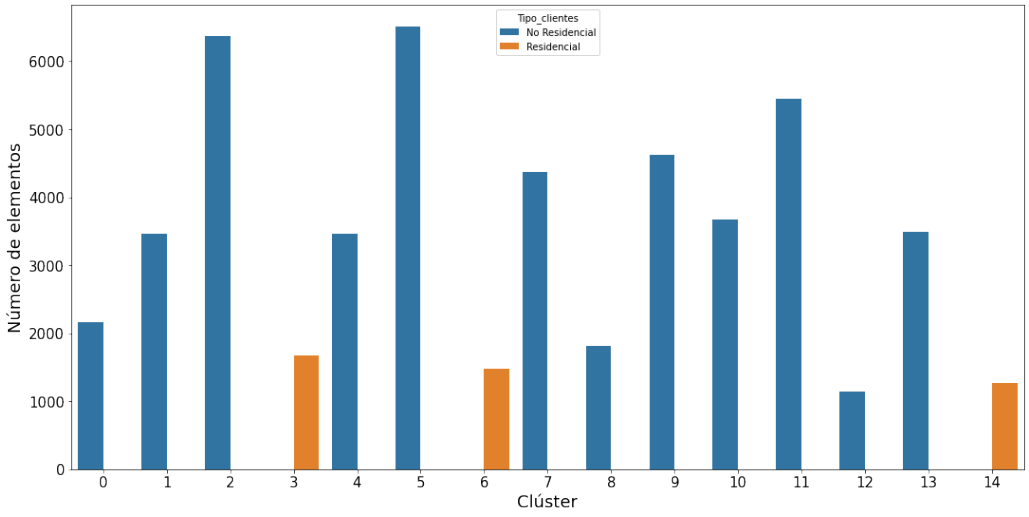
Curva para obtener el número óptimo de clústers K



De la Figura 6 se puede observar que el valor máximo de la métrica se alcanza cuando el número de clústers es igual a 15. Con ese valor de K=15, 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 los tipos de clientes. Se puede observar que los clientes residenciales se agrupan únicamente en los clústers 3, 6 y 14. Los clientes no residenciales se agrupan en el resto de los clústers.

Figura 7

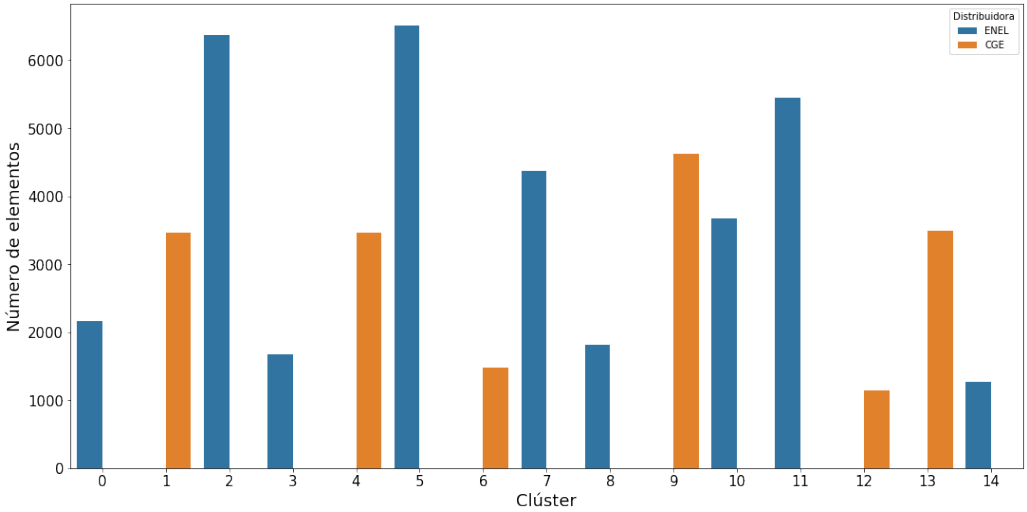
Clústers vs. Tipo de clientes



Igualmente, la forma en que se relacionan las empresas distribuidoras con los clústers se muestran en la Figura 8. Se puede ver en la figura que las empresas distribuidoras se agrupan en clústers diferentes. Los registros asociados a la empresa CGE se agrupan en los clústers 1, 4, 6, 9, 12, y 13. Los registros correspondientes a la empresa ENEL se agrupan en los restantes clústers.

Figura 8

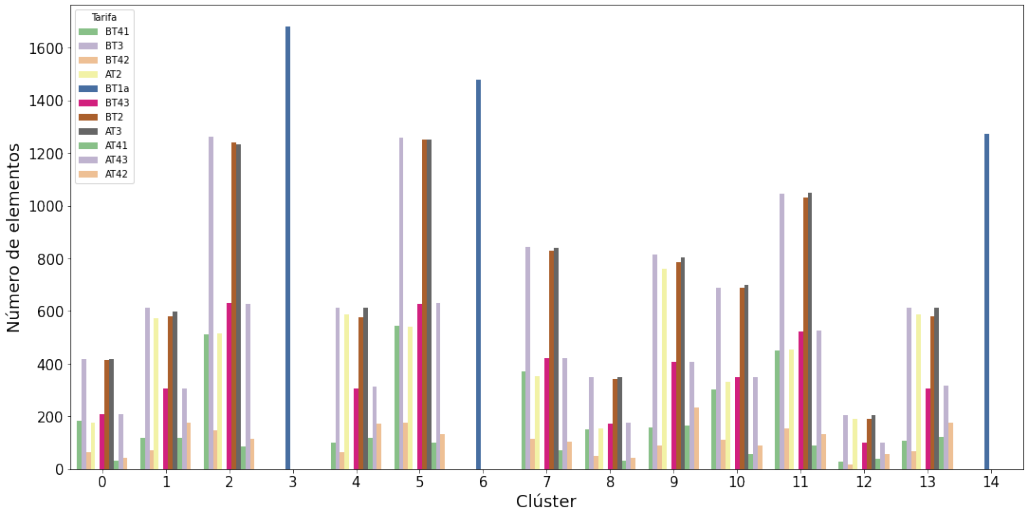
Clústers vs. Distribuidora



Dado que las tarifas se relacionan con los tipos de clientes, en los clústers 3, 6 y 14 hay sólo registros de clientes residenciales con la tarifa BT1a, tal como se observa en la Figura 9.

Figura 9

Clústers vs. Tipo de tarifa



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.

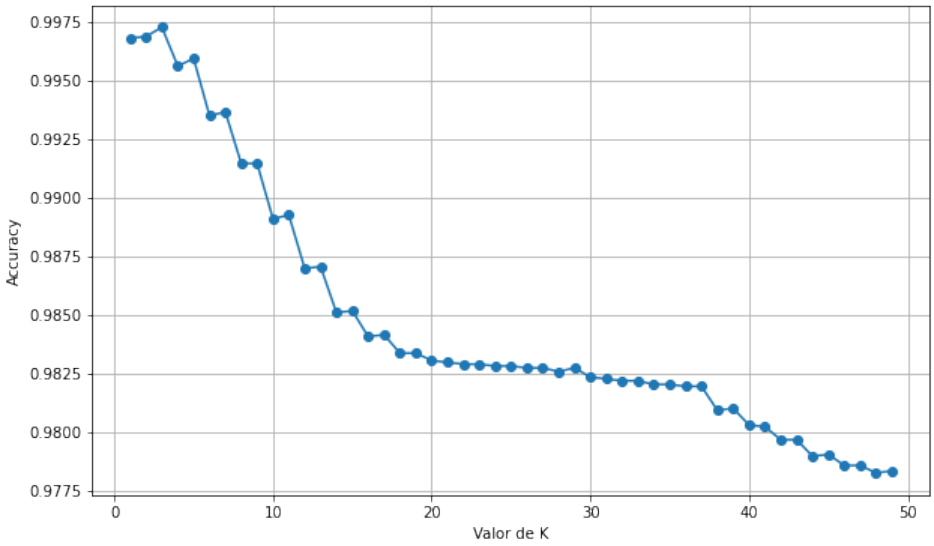
Aplicación del algoritmo K-NN

El algoritmo de los K vecinos más cercanos K-NN, es un algoritmo de aprendizaje automático supervisado para clasificación, mediante el cual se busca predecir la clase o categoría de un conjunto de datos, a partir de un grupo de variables predictoras. De acuerdo con Lee (2019), K-NN es uno de los algoritmos más simples dentro de los algoritmos de aprendizaje automático 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 su investigación, Raschka y Mirjalili (2017) plantean 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, inicialmente se debe establecer el número de vecinos K, 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 Fenner (2020, p. 163) "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 3, pues es el valor para el cual se alcanza el valor máximo posible de la exactitud, cuyo valor es 99,73%.

Figura 10

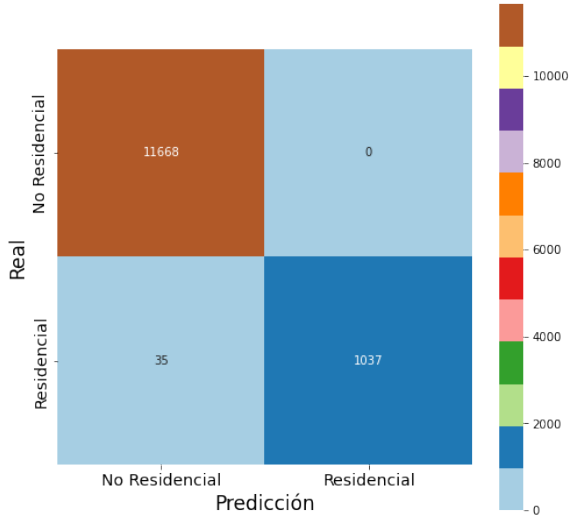
Curva para obtener el número óptimo de vecinos K



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 75% de los datos, y el set de prueba correspondiente al 25% 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 Figura 11.

Figura 11

Matriz de confusión



De la Figura 11 se puede decir que el conjunto de prueba estuvo compuesto por 12.740 filas del conjunto de datos, el cuál es el 25% de los datos originales. Adicionalmente, 11.668 filas eran de clientes no residenciales y el modelo clasificó a todas estas filas de manera correcta. Por otra parte, 1.037 filas eran de clientes residenciales y el modelo los clasificó de esa manera, pero 35 filas de clientes residenciales fueron clasificadas como no residenciales. Este modelo podría ser muy útil, por ejemplo, para determinar el tipo de cliente, y por lo tanto el tipo de tarifa a aplicar, cuando por alguna razón no se tiene esa información.

Conclusiones

La energía eléctrica facturada por los clientes regulados, durante el período de estudio, se dividió en partes iguales entre los clientes residenciales y los clientes no residenciales. Los datos presentan una estacionalidad mensual, presentándose una mayor facturación durante los meses de la estación de invierno, en comparación con los otros meses del año. La estacionalidad fue impuesta por la facturación de los clientes residenciales.

El consumo unitario promedio de los clientes residenciales fue de 453 kWh durante el período de estudio, mientras que los clientes no residenciales tuvieron un consumo unitario promedio de 10.315 kWh. Los clientes residenciales tuvieron un consumo unitario promedio anual alrededor de 225 kWh durante la mayor parte del período de estudio, pero durante el año 2020 se disparó a 1.759,28 kWh, cayendo nuevamente para el año 2021.

La energía eléctrica facturada por los clientes regulados residenciales aumentó continuamente desde el año 2015 hasta el año 2020, cuando alcanzó su valor máximo durante el período de estudio, impulsado por las cuarentenas impuestas ese año debido a la pandemia de la Covid-19. Por el contrario, la energía facturada a los clientes regulados no residenciales ha disminuido constantemente desde el año 2017 al año 2021, coincidiendo con las estadísticas de traspaso de clientes regulados a libres, permitido por la normativa vigente a los clientes no residenciales.

Para la aplicación del algoritmo K-Means, se obtuvo el número óptimo de clústers igual a 15, maximizando la métrica Silhouette. El modelo obtenido a través del algoritmo agrupó perfectamente los datos de acuerdo con el tipo de cliente de cada uno de los registros. De igual

manera, todos los datos dentro de un clúster particular pertenecen a una sola empresa de distribución de electricidad.

Para el desarrollo del modelo de predicción del tipo de clientes, los datos se dividieron en dos partes, 75% para entrenar el modelo y 25% para evaluar el modelo. Al aplicar el algoritmo K-NN se obtuvo un modelo que permite predecir, con una exactitud del 99,73%, el tipo de cliente regulado para los registros nuevos que requieran ser evaluados. Se obtuvo el valor óptimo del hiperparámetro K igual a 3, al maximizar la métrica exactitud.

Referencias

Amat Rodrigo, J. (s.f.). *Ciencia de Datos, Estadística, Machine Learning y Programación*. (Joaquin Amat Rodrigo) Recuperado el 16 de Julio de 2022, de https://www.cienciadedatos.net/documentos/pystats05-correlacion-lineal-python.html

Argüello Verbanaz, S., & García Bernal, N. (2020). *Componentes y determinación de la tarifa eléctrica para los clientes regulados.* Santiago de Chile: Biblioteca del Congreso Nacional de Chile.

Azócar Rojas, M. A. (2018). *Estudio y análisis del Nuevo Decreto Tarifario 11 T. Aplicable a los suministros sujetos a precios.* Valparaíso: Tesis de Pregrado, Pontificia Universidad Católica de Valparaíso.

Bernal, C. A. (2010). *Metodología de la Investigación - administración, economía, humanidades y ciencias sociales.* Bogotá: Pearson Educación.

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

Comisión Nacional de Energía. (16 de Julio de 2022). *Estadísticas.* Obtenido de Energía Abierta: http://energiaabierta.cl/categorias-estadistica/electricidad/

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

García Bernal, N. (2019). *Traspaso de clientes regulados a libres.* Valaparaíso: Biblioteca del Congreso Nacional de Chile.

Igual, L., & Seguí, S. (2017). *Introduction to Data Science - A Python Approach to Concepts, Techniques and Applications.* Switzerland: Springer International Publishing.

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

Lee, W.-M. (2019). *Python Machine Learning.* Indianapolis: John Wiley & Sons, Inc.

M. Shapi, M. K., Ramli, N. A., & Awalin, L. J. (2021). Energy consumption prediction by using machine learning for smart building: Case study in Malaysia. *Developments in the Built Environment*. https://doi.org/10.1016/j.dibe.2020.100037.

McKinney, W. (2018). *Python for Data Analysis.* Sebastopol, CA: O'Reilly Media, Inc.

Mellado Leal, B. M. (2021). *Aplicaciones de Data Science para la mejora de la medición y cobro de la distribución de la energía eléctrica en contextos de pandemia mundial.* Santiago de Chile: Tesis de Pregrado, Universidad de Chile.

Mercados Energéticos Consultores. (2014). *Análisis de consumo eléctrico en el corto, mediano y largo plazo.* Santiago de Chile: Mercados Energéticos Consultores.

Moreno, R., Sánchez, M., Suazo, C., Negrete, M., Olivares, D., Alvarado, D., . . . Basso, L. (2020). Impactos del COVID-19 en el Consumo Eléctrico Chileno. *Revista Ingeniería de Sistemas*.

Pizarro Herrera, G. N. (2017). *Reconocimiento de patrones y pronóstico de consumo eléctrico.* Valparaíso: Tesis de Pregrado, Pontificia Universidad Católica de Valparaiso.

Rajabi, A., Eskandari, M., Jabbari Ghadi, M., Li, L., & Zhang, J. (2020). A comparative study of clustering techniques for electrical load pattern segmentation. *Renewable and Sustainable Energy Reviews*. https://doi.org/10.1016/j.rser.2019.109628.

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning - Machine Learning and Deep Learning with Python, Scikit-Learn, and TensorFlow.* Birmingham: Packt Publishing Ltd.

Russano, E., & Ferreira Avelino, E. (2020). *Fundamentals of Machine Learning Using Python.* Oakville, Canadá: Arcler Press.

Salazar Córdova, M. A. (2018). *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).* Santiago de Chile: Tesis de Maestría, Universidad Técnica Federico Santa María.

Sociedad Alemana de Cooperación Internacional. (2020). *Las Energías No Renovables en el Mercado Eléctrico Chileno.* Santiago de Chile: Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) GmbH.

Umargono, E., Suseno, J. E., & S.K, V. G. (2020). 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, volume 474*. DOI:10.2991/assehr.k.201010.019.

Valgaev, O., Kupzog, F., & Schme, H. (2017). Building power demand forecasting using K-nearest neighbours model – practical application in Smart City Demo Aspern project. *CIRED, Open Access Proc. Journal* (págs. 1601–1604). IET. DOI:10.1049/oap-cired.2017.0419.

Yilmaz, S., Chambers, J., Li, X., & Patel, M. K. (2021). 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*. DOI:10.1088/1742-6596/2042/1/012021.