Elaboración de modelo de clusterización de clientes asociados al servicio de energía en Antioquia

Seminario de Analítica y Ciencia de Datos

# Resumen Descriptivo del Proyecto

Diseñar un modelo de clusterización de clientes de Empresas Públicas de Medellín (EPM) del servicio de energía en el departamento de Antioquia, tomando como fuente de información archivos suministrados desde la empresa de servicios públicos con los cuales se aplicará un modelo de analítica de datos que permitan generar los insumos necesarios para el establecimiento de estrategias comerciales enfocando esfuerzos en el mejoramiento de la experiencia del cliente.

Marco Teórico:

1. Revisión de Literatura.

El análisis del problema de la presente monografía destaca la necesidad de generar soluciones que aporten a mejorar la experiencia del cliente en EPM. En la actualidad, las experiencias y percepciones sobre una marca han adquirido una relevancia superior a los beneficios inherentes de los productos o servicios, debido a las nuevas expectativas del consumidor [1] y en la que no se contempla los cambios que surgen a partir del aumento del acceso a la información y las comunicaciones [2]. Teniendo en cuenta lo anterior, se hace necesario que una empresa como EPM enfoque sus esfuerzos para mejorar la experiencia del cliente. En este sentido, la analítica de datos emerge como una herramienta clave para dicho propósito. Se pueden generar modelos de predicción con base en datos de experiencia del cliente mediante depuración de información aplicando los métodos correspondientes llegando a una efectividad del 84% [3]. Ahora bien, en el contexto de EPM, el consumo es insumo fundamental que indica cómo se podrían comportar los clientes, en [4] se aplican métodos analíticos con una previa preparación de los datos en la que se concluye que los modelos de Random Forest y Decision Tree Classifier debido a que tienen la mejor capacidad para equilibrar la precisión en la identificación de fraudes y se establece relación significativa entre los patrones de consumo de los clientes y los fraudes.

De acuerdo con lo planteado anteriormente, se propone como solución el desarrollo de una clusterización de clientes basado principalmente en los consumos de energía eléctrica. Existen diversos estudios en los que se realizan clusterizaciones de clientes con base en los consumos de energía eléctrica.

[5] Emplea el uso de algoritmos para clasificar y analizar el comportamiento de consumo de energía eléctrica, para la agrupación de los datos se utilizó metodología de K-means y para determinar el número de agrupaciones se utilizó el método del codo, dentro de dicho estudio, se encontró la heterogeneidad entre los segmentos de consumidores lo que permite el desarrollo de intervenciones específicas con el fin de promover la eficiencia energética. Similar al estudio anterior, [6] establece un análisis de comportamiento de consumos de energía eléctrica por medio de la metodología de K-means, en el que se calcularon dos grupos definidos donde se plantea que dicha información sirve para optimizar políticas públicas y privadas sobre la distribución de energía eléctrica. La información de los dos estudios anteriores se obtiene con base en medidores inteligentes de energía.

[7] Realiza una clusterización de clientes con base en sus hábitos de demanda de energía realizando un comparativo de la segmentación de clientes calculada frente a una segmentación basada en la curva típica característica de consumo concluyendo que es más significativa la segmentación calculada en el estudio porque permite dirigirse a grupos de clientes más pequeños.

[8] Propone realizar un modelo de segmentación de clientes aplicando un caso de uso en Sudáfrica con base en medidas de evaluación internas y externas, la medida externa se obtuvo por medio de entrevistas no estructuradas realizadas con expertos y la medida interna mediante matriz de puntuación de agrupaciones, con la concatenación de dichas variables se pudo determinar la mejor agrupación que se ajustaba a los requisitos de la aplicación. Se debe tener en cuenta que en dicho trabajo se considera que Sudáfrica presenta volatilidad económica, diversidad de ingresos y diversidad geográfica.

En línea con el estudio anterior y aterrizando dichos conceptos a Colombia, se encuentran diferentes estudios en ese sentido, [9] establece una caracterización de clientes con base en su comportamiento de consumo que presentan medición inteligente con el apoyo de la empresa CELSIA, inicialmente realizaron una vigilancia tecnológica en la que se observó un aumento en el interés en el comportamiento del consumo de los usuarios para la incorporación adecuada de programas de respuesta de la demanda, se utilizó una metodología de clusterización de K-means para identificar segmentos de clientes de acuerdo con la respuesta de la demanda de energía.

Manteniendo el mismo enfoque anterior, [10] propone una caracterización de clientes como base de implantación de programas de respuesta de la demanda de la energía con el fin de que coincida con la energía disponible. Para realizar dicha clusterización utilizó dos metodologías las cuales corresponden a combinación del método Ward, el cual es de aglomeración y el método de K-means, que no es jerárquico.

Para el análisis de consumos es relevante tener en cuenta información temporal, por lo que [11] realiza una clusterización de series de tiempo, después de la aplicación de técnicas para la preparación de datos, se aplican técnicas de aprendizaje no supervisado cuyo objetivo sea el clustering de datos, como lo son los algoritmos DTW (Dynamic Time Warping) y SDTW (Soft Dynamic Time Warping) teniendo mejor desempeño el algoritmo SDTW.

En Colombia, existen diferentes empresas multiservicios públicos domiciliarios, es decir, ofrecen la comercialización de diferentes servicios tales como energía, gas, acueducto, alcantarillado, entre otras. Una de las empresas más robustas por los servicios que ofrece y por la cantidad de clientes es EPM, en la que según el informe de sostenibilidad del año 2023 se tienen 2,8 millones de puntos de prestación del servicio de energía [12]. Teniendo en cuenta la condición multiservicio de EPM, se tomarán como insumos adicionales para la presente monografía, la información de consumo de los servicios de acueducto, alcantarillado y gas para su uso como variables adicionales de los clientes del servicio de energía. Los autores [5, 6, 7, 9, 10] utilizan la metodología de análisis de consumo de energía de medidores inteligentes teniendo en común que el objetivo es determinar segmentos que permitan conocer el comportamiento de consumo para proponer mejoras en la demanda de energía, pero no se tienen en cuenta el consumo de otros servicios para dicho análisis, el cual puede servir de insumo para encontrar una correlación y poder encontrar caminos en la propuesta de la solución del problema de negocio.

La medición inteligente o avanzada permite generar un análisis más detallado de los clientes de acuerdo con [5, 6, 7, 9], la cual utiliza como fuente de información el consumo obtenido de medidores inteligentes a nivel horario, la realidad actual es que en Colombia gran parte de los clientes aún no cuentan con dicha medición. Sin embargo, de acuerdo con la resolución 40072 del año 2018 del ministerio de minas y energía de Colombia establece que para el año 2030, el mínimo porcentaje de usuarios conectados en un mercado de comercialización, con infraestructura de medición avanzada (AMI) (por sus iniciales en inglés) debe ser del 75% [13] por lo que dichos estudios permiten tener un avance en el tema, pero con el fin de cubrir las necesidades de negocio actuales se realizará el análisis en el presente trabajo con base en consumos mensuales.

Todos los estudios anteriormente mencionados toman como base principal el consumo de energía a excepción del autor [8], el cual establece un punto interesante, correspondiente a tener información de expertos para considerar dentro de las variables de análisis. Lo anterior, en concordancia a lo que se va a realizar en la presente monografía asociado al análisis de consumo de energía teniendo en cuenta variables de consumo de otros servicios (acueducto, alcantarillado y gas) y adicionalmente variables de interacciones con los clientes.

1. Modelos y Métodos ya existentes.

Según [14], el proceso de segmentación de patrones de consumo energético puede dividirse en cinco etapas: recolección de datos, reducción dimensional y selección o definición de características, clustering, evaluación del clustering y generación de clases de clientes. Cada una de estas etapas puede adaptarse según los objetivos y requerimientos del proyecto. En este caso, nos enfocaremos en analizar algunos de los métodos empleados en las etapas de reducción dimensional y/o selección de características y clustering.

El método de Principal Components Analysis (PCA), es una técnica que permite transformar los datos a un sistema de coordenadas tal que los primeros componentes principales retienen la mayor parte de la variación presente en todas las variables originales [15]. Aunque es posible ver el uso de este método aplicado en distintos estudios dentro del contexto de consumo energético [16–18], no se resalta una razón específica para su selección. Es probable que su uso se deba a su popularidad en diferentes disciplinas y su capacidad para manejar conjuntos de variables con alta multicolinealidad [19], la cual es una propiedad común en datos sobre consumo energético [16]. Esta técnica podría ser una primera aproximación debido a su amplia documentación y facilidad de implementación, pero su utilidad deberá ser evaluada de acuerdo a los patrones de consumo obtenidos y métricas de evaluación de los clusters generados.

Un método para reducción dimensional y selección de características que vale la pena resaltar es el Sparse Denoising Autoencoder (SDAE), implementado en [20], el cual permite la extracción de características y capturar relaciones no lineales entre las variables, gracias a que aprende a codificar los datos en una representación comprimida y luego los reconstruye, minimizando la pérdida de información entre la entrada y la salida [21]. Esta técnica podría ser especialmente útil en el caso propuesto, ya que se pretenden utilizar datos no solo de los consumos energéticos sino también de los consumos de otros servicios (acueducto, alcantarillado y gas), interacciones, ubicación, estrato y demás características del cliente que no necesariamente son un indicador directo de su consumo. Por lo tanto, una técnica de este tipo podría ser prometedora para encontrar relaciones complejas que no son fácilmente identificables entre las diferentes características que describen cada registro.

En la etapa de clustering, hay mayor variedad en las alternativas de los modelos utilizados para el análisis de patrones de consumo energético [14]. Entre las técnicas empleadas se encuentran: Spectral Clustering (SC), Fuzzy CMeans (FCM), K-means Clustering (KMC), Agglomerative Clustering (AC), entre otros. El SC es utilizado para identificar patrones de consumo en conjuntos de datos con altas resoluciones temporales [20], debido a que trata las series de tiempo como grafos, segmentándolas de acuerdo a los distintos patrones que estas siguen. Para nuestro caso de estudio, en donde se tienen consumos mensuales, es posible que esta alternativa no se desempeñe de la mejor manera ya que la resolución temporal de los datos no es tan fina y los patrones pueden ser más simples en comparación a los de una resolución temporal en horas. Los modelos de FCM han sido también utilizados para la identificación de patrones de consumo energético, estos modelos traen consigo el beneficio de poder asignar un grado de pertenencia a todos los segmentos a la vez, pero esto viene con el costo de identificar un parámetro de difusividad apropiado para el caso de estudio, de igual forma estos modelos son de búsqueda local por lo que suelen tener problemas de convergencia y pueden terminar atascados en mínimos locales [22]. En nuestro caso no se espera asignar un grado de pertenencia para todos los segmentos encontrados sino más bien determinar únicamente un segmento para cada usuario. El uso de KMC se puede observar en diferentes estudios [23 – 25], donde sugieren que su uso se debe a la simplicidad y eficiencia, escalabilidad, clara asignación de segmentos y adaptabilidad a diferentes tipos de datos [26]. Para nuestro caso, en donde se tienen no solo datos temporales, sino también características externas al consumo energético de los usuarios, podría verse este modelo como una buena alternativa por su adaptabilidad y uso en distintos contextos y tipos de datos.

Cabe resaltar que si bien existen estas y muchas otras técnicas para la identificación de patrones de consumo similares, comúnmente se nota que técnicas diferentes se ajustan mejor a diferentes conjuntos de datos [18], esta decisión se toma a partir de la selección de un Cluster Validity Index (CVI). El uso de uno u otro índice de validación puede resultar recomendando diferentes modelos dependiendo de la métrica sobre el clúster que el índice seleccionado priorice [27], para la resolución de estos posibles conflictos de resultados entre distintos índices, es importante tener en cuenta heurísticas de los datos para tomar esta decisión, por ejemplo, si se desea tener un número fijo de grupos o que cada grupo cumpla ciertas características de tamaño o densidad [18].

De los trabajos revisados en este proyecto, se observa un factor común: los datos utilizados para generar las segmentaciones se limitan al consumo energético de cada usuario y al análisis de series de tiempo derivadas de estos datos. Nuestra propuesta ofrece un valor adicional, ya que además de los consumos energéticos mensuales de cada cliente, contamos con información de consumos de otros servicios (acueducto, alcantarillado y gas), ubicación, estrato socioeconómico e interacciones con EPM de cada cliente. Aunque estas características pueden no parecer directamente relacionadas con el consumo energético, podrían contener relaciones no lineales subyacentes que se podrían aprovechar mediante técnicas avanzadas de extracción de características como las que hemos descrito.

1. Diferenciación del Proyecto

La solución que se pretende desarrollar en la monografía es innovadora debido a la integración de variables adicionales al consumo energético y la ampliación de la perspectiva del mismo. De acuerdo a la revisión de literatura, las clusterizaciones que se realizan en dichos estudios se enfocan principalmente en el consumo de energía eléctrica, con pocas excepciones que consideran variables adicionales.Nuestra propuesta, rompe con este enfoque tradicional al incorporar una visión más holística del comportamiento de consumo energético, debido a que se visualiza el comportamiento del cliente al analizar el consumo de otros servicios públicos, como acueducto, alcantarillado y gas. Además, consideramos variables propias de los clientes como el municipio donde se encuentra ubicado, el estrato socioeconómico y las interacciones de los clientes con EPM.

Al tener en cuenta todas las variables anteriormente mencionadas, junto con un análisis ajustado a la realidad actual de la mayoría de los clientes en el cual la información de consumo disponible es de carácter mensual a diferencia de los estudios previamente descritos en los que la información de consumo es a nivel horario, permite tener una solución de clusterización para la actualidad con el mayor nivel de profundidad posible a la información disponible con el fin de generar la clusterización de los clientes en donde se tenga una mejor comprensión de las necesidades de los mismos.

1. Referencias.

[1] D. Le, N. Scott, and G. Lohmann, “Applying experiential marketing in selling

tourism dreams,” Journal of Travel & Tourism Marketing, vol. 36, no. 2, pp. 220–235, Feb. 2019, doi: 10.1080/10548408.2018.1526158.

[2] L. F. Sabogal Russi and S. P. Rojas-Berrio, “Evolución del marketing experiencial:

una aproximación teórica a su definición - Revisión Sistemática de Literatura,” Entramado, vol. 16, no. 1, pp. 94–107, Dec. 2019, doi: 10.18041/1900-3803/entramado.1.6140.

[3] A. Felipe, A. Álvarez, and S. Betancur Téllez, “Santander Customer Satisfaction:

un caso práctico de analítica predictiva,” 2021. [Online]. Available: https://hdl.handle.net/10495/21866

[4] S. Maria Alvarez Castillo Luis Miguel Román Pereira, A. Walter Mauricio Villa

Acevedo, and D.-I. Álvaro Jaramillo Duque, “Detección de instalaciones fraudulentas utilizando datos históricos de consumo energéticos,” 2024. [Online]. Available: https://hdl.handle.net/10495/40450

[5] Y. Guan, H. Zhang, Y. Jin, and S. Zhou, “Classification and Analysis of Users’

Electricity Consumption Behavior Using Cluster Analysis Algorithm,” J. Electrical Systems, vol. 20, no. 9, pp. 1396–1402, 2024.

[6] L. Marrero, D. Carrizo, L. García-Santander, and F. Ulloa-Vásquez, “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, Dec. 2021, doi: 10.4067/S0718-33052021000400778.

[7] A. Camero, G. Luque, Y. Bravo, and E. Alba, “Customer Segmentation Based on

the Electricity Demand Signature: The Andalusian Case,” Energies, vol. 11, no. 7, p. 1788, Jul. 2018, doi: 10.3390/en11071788.

[8] W. Toussaint and D. Moodley, “Clustering Residential Electricity Consumption

Data to Create Archetypes that Capture Household Behaviour in South Africa,” South African Computer Journal, vol. 32, no. 2, Dec. 2020, doi: 10.18489/sacj.v32i2.845.

[9] D. Echeverry, S. Ospina, J. Zapata, and K. Menco, “Customers characterization:

A pilot study for the incorporation of demand response programs,” in 2019 FISE-IEEE/CIGRE Conference - Living the energy Transition (FISE/CIGRE), IEEE, Dec. 2019, pp. 1–6. doi: 10.1109/FISECIGRE48012.2019.8985014.

[10] M. A. Arroyo Bernal, “Análisis de la información y caracterización del consumo

de energía eléctrica de los estratos 1 y 2 caso EMCALI, como base para la implantación de programas de respuesta de la demanda,” 2016. [Online]. Available: http://hdl.handle.net/1992/14889

[11] D. Z. Chaves, “Clustering de series temporales pertenecientes al consumo de

productos para la agrupación por patrones,” 2023. [Online]. Available: https://hdl.handle.net/10495/35519

[12] Empresas Públicas de Medellín - EPM, “Informe Sostenibilidad EPM 2023.”

[Online]. Available: https://sostenibilidadgrupoepm.com.co/wp-content/uploads/2024/08/5\_INFORME-SOSTENIBILIDAD-ESP\_2023\_unif.pdf

[13] Ministerio Minas y Energía Colombia, “Resolución 40072 de 2018 .” [Online].

Available: <https://gestornormativo.creg.gov.co/gestor/entorno/docs/resolucion_minminas_40072_2018.htm>

[14] A. Rajabi, M. Eskandari, M. J. Ghadi, L. Li, J. Zhang, and P. Siano, “A comparative study of clustering techniques for electrical load pattern segmentation,” *Renewable and Sustainable Energy Reviews*, vol. 120, p. 109628, Mar. 2020, doi: 10.1016/j.rser.2019.109628.

[15] I. T. Jolliffe, “Principal Component Analysis and Factor Analysis,” 1986, pp. 115–128. doi: 10.1007/978-1-4757-1904-8\_7.

[16] R. Kaur and D. Gabrijelčič, “Behavior segmentation of electricity consumption patterns: A cluster analytical approach,” *Knowledge-Based Systems*, vol. 251, p. 109236, Sep. 2022, doi: 10.1016/j.knosys.2022.109236.

[17] A. M. S. Ferreira, C. A. M. T. Cavalcante, C. H. O. Fontes, and J. E. S. Marambio, “A new method for pattern recognition in load profiles to support decision-making in the management of the electric sector,” *International Journal of Electrical Power & Energy Systems*, vol. 53, pp. 824–831, Dec. 2013, doi: 10.1016/j.ijepes.2013.06.001.

[18] M. Jain, T. AlSkaif, and S. Dev, “Validating Clustering Frameworks for Electric Load Demand Profiles,” *IEEE Transactions on Industrial Informatics*, vol. 17, no. 12, pp. 8057–8065, Dec. 2021, doi: 10.1109/TII.2021.3061470.

[19] H. Abdi and L. J. Williams, “Principal component analysis,” *WIREs Computational Statistics*, vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/wics.101.

[20] Y. Huang, Z. Yao, and Q. Xu, “Classification model of electricity consumption behavior based on sparse denoising autoencoder feature dimensionality reduction and spectral clustering,” *International Journal of Electrical Power & Energy Systems*, vol. 158, p. 109960, Jul. 2024, doi: 10.1016/j.ijepes.2024.109960.

[21] A. Ng, “CS294A Lecture notes: Sparse Autoencoder,” Stanford. Accessed: Oct. 09, 2024. [Online]. Available: <https://web.stanford.edu/class/cs294a/sparseAutoencoder.pdf>

[22] K. Zhou, S. Yang, and Z. Shao, “Household monthly electricity consumption

pattern mining: A fuzzy clustering-based model and a case study,” *Journal of Cleaner*

*Production*, vol. 141, pp. 900–908, Jan. 2017, doi: 10.1016/j.jclepro.2016.09.165.

[23] J. L. Viegas, S. M. Vieira, R. Melício, V. M. F. Mendes, and J. M. C. Sousa, “Classification of new electricity customers based on surveys and smart metering data,” *Energy*, vol. 107, pp. 804–817, Jul. 2016, doi: 10.1016/j.energy.2016.04.065.

[24] Y.-I. Kim, J.-M. Ko, and S.-H. Choi, “Methods for generating TLPs (typical load profiles) for smart grid-based energy programs,” in *2011 IEEE Symposium on Computational Intelligence Applications In Smart Grid (CIASG)*, IEEE, Apr. 2011, pp. 1–6. doi: 10.1109/CIASG.2011.5953331.

[25] T.-S. Xu, H.-D. Chiang, G.-Y. Liu, and C.-W. Tan, “Hierarchical K-means Method for Clustering Large-Scale Advanced Metering Infrastructure Data,” *IEEE Transactions on Power Delivery*, vol. 32, no. 2, pp. 609–616, Apr. 2017, doi: 10.1109/TPWRD.2015.2479941.

[26] M. E. Celebi, H. A. Kingravi, and P. A. Vela, “A comparative study of efficient initialization methods for the k-means clustering algorithm,” *Expert Systems with Applications*, vol. 40, no. 1, pp. 200–210, Jan. 2013, doi: 10.1016/j.eswa.2012.07.021.

[27] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On Clustering Validation Techniques,” *Journal of Intelligent Information Systems*, vol. 17, no. 2/3, pp. 107–145, 2001, doi: 10.1023/A:1012801612483.

# Rúbricas

R01: Completitud y consistencia: el estudiante realiza una entrega del momento evaluativo con todos los elementos desarrollados, manteniendo una relación coherente entre los elementos solicitados. El estudiante usa una consistencia argumentativa que incluye el uso de referencias cuando hay mérito de autoría entre otros.

R02: Línea Base: el estudiante es capaz de establecer una clara línea de trabajos previos y teorías basadas alrededor de su proyecto de clase. Esto permite evidenciar que la solución propuesta en el trabajo precio es viable y tiene una base sólida teórica detrás de su posible solución.

R02: Innovación: dentro del ámbito de la analítica y la ciencia de datos el estudiante muestra, basado en la literatura existente, una clara diferenciación de su proyecto frente las soluciones tecnologico-metodológicas encontradas. Se tiene en cuenta la inclusión de la base de datos de CESET para la rúbrica.

R04: Referencias: hay referencias válidas según la descripción del capítulo y están adecuadamente vinculadas al estilo narrativo que permite una clara expresión de los resultados de la investigación de la línea base y el marco tecnológico-metodológico.