Universidad de Costa Rica

Facultad de Ingeniería
Escuela de Ingeniería Eléctrica
IE0435 Inteligencia Artificial Aplicada a la Ingeniería Eléctrica
II ciclo 2023

Proyecto 1

Kevin Campos Castro, B91519
Profesor: Marvin Coto Jiménez

18 de setiembre

Índice

Índice de figuras

1. Introducción

El siguiente proyecto consiste en utilizar los datos del informe anual de calidad eléctrica del año 2021 del ARESEP y a partir de esta información, aplicar dos algoritmos de agrupamiento como los que se ha visto en clases, esto para poner en práctica lo aprendido durante el curso y conocer un poco acerca de las herramientas utilizadas para esta área.
Durante el proyecto se utilizaron 2 algoritmos, uno de ellos siendo el algoritmo de K-means y el otro algoritmo aplicado es el de mini batch k-means, para ambos casos se tuvo que tomar los datos del informe y convertirlo a un formato en el que se facilite trabajar, por lo general para el análisis de datos se prefieren archivos .csv, ya que los mismos se pueden utilizar con la librería de Pandas, una herramienta poderosa para el manejo de información. Una vez obtenido este archivo, los datos pueden ser interpretados por un lenguaje de programación que en este caso es Python.
Al final del proyecto se muestran los resultados obtenidos, se observan ciertos patrones basado en los parámetros utilizados por la ARESEP tales como DPI, número de abonados y FPI.

2. Elementos de análisis y discusión

1. ¿Por qué se requiere una autoridad como ARESEP para evaluar a las empresas distribuidoras de energía eléctrica?
   Es porque se requiere una entidad imparcial que vela por la distribución correcta de acuerdo con el estándar definido por la ley y además que, para una persona no le es fácil hacerlo porque no tiene las herramientas necesarios para llevar estudios que velen por la calidad de estos servicios proporcionado a la población. [1]

2. ¿Cuál es el propósito de utilizar varios métodos de agrupamiento para este problema de circuitos eléctricos de distribución?
   En general cada algoritmo ayuda a que se descubran patrones inherentes al conjunto de datos y esto se hace agrupando los datos que hay bajo las características que los hace semejante, cada uno de estos algoritmos va a generar resultados diferentes, por lo que hay que ver cuál es el que mejor se ajusta a lo que se estudia, que sería las fallas en los circuitos del servicio de distribución eléctrica de Costa Rica.

3. Mencione y describa un caso de métodos de agrupamiento que se encuentre en la literatura científica reciente relacionada con el sector eléctrico.
   La siguiente investigación se basa en el problema fundamental de que la fiabilidad de las empresas distribuidoras de energía eléctrica es sumamente importante y como mejorar estos, sin embargo, encontrar un modelo matemático que explique la demanda es muy difícil porque constantemente está cambiando, por lo que se utilizan métodos de agrupamiento basado en el consumo eléctrico de las personas. Con lo que se propuso en dicha investigación lograron mejorar la calidad de la distribución de la electricidad inclusive disminuyendo costos. Los parámetros utilizados para esta investigación fueron actividad económica, tamaño y consumo energético, basado en esto se agrupan los consumidores eléctricos a clases similares. Para el caso del método de agrupamiento, en la misma investigación se hace referencia a otro estudio en el que su utiliza el algoritmo que llaman "K-means híbrido". [2]

4. ¿Existe alguna anomalía en estos conjuntos de datos? ¿Como pueden detectarse anomalías con otros métodos además de agrupamiento?
   En términos de la electricidad hay muchas variantes externas que pueden afectar las mediciones de los datos, esto incluye error humanos, interferencias, ruido, etc. por lo que los datos no son completamente exactos, sin embargo es por esta razón que se toman diversas mediciones. Otro método para detectar anomalía que no sea agrupamiento puede ser "one class support vector machine". [3]

5. ¿Qué papel juega la normalización de los datos en este problema? ¿Existen otras formas de realizar normalización de datos? ¿Cambian los resultados si no se aplica la normalización?
   Es importante ya que, involucra cambiar la escala de cada clase de forma en que tengan una desviación estándar de 1, esto es importante porque a veces los datos con los que se trabaja tienen rangos de valores muy altos, en un algoritmo como el de KNN, se debe calcular la distancia en cada iteración y este cálculo es afectado por la escala de la distancia entre los datos, no solo eso, sino que a la hora de normalizarlo inclusive los cálculos de cada iteración son menores porque las distancias también lo son. Los resultados si cambian si no se aplica la normalización como se observa en [4].
   Existen muchas formas de normalizar, para mencionar algunos: Absolute Maximum, Min-Max Scaling, Normalization, Standardization, etc. [5]

6. ¿Qué significa y cuál es la fórmula utilizada para el cálculo de las variables FPI, DPI y cantidad de abonados?

La FPI o frecuencia promedio de las interrupciones por abonado se refiere a la cantidad promedio de veces que un abonado o usuario percibe una interrupción en el suministro eléctrico. Matemáticamente se expresa como:

$$\textbf{FPI} = \frac{\text{Cantidad de interrupciones}}{\text{Abonados}}$$

La DPI o tiempo promedio de interrupción por abonado contabiliza el tiempo promedio en que el servicio eléctrico no le fue suministrado.

$$\textbf{DPI} = \frac{\text{Duración de la interrupción}}{\text{Tiempo total suministrado}}$$

De acuerdo al informe de la ARESEP [1], ambos datos se calculan tomando en consideración todas las interrupciones de duración superior a los 5 minutos y que se suscitan a lo largo de las redes de distribución, hasta los transformadores que sirven la energía en baja tensión y también considerar que las expresiones matemáticas son por usuario.
Finalmente la cantidad de abonados son las personas que reciben el servicio, que en este caso es de electricidad.

3. Procedimiento

Este proyecto se inició tomando los datos proporcionados por el estudio de calidad de energía eléctrica de la ARESEP, puesto que los mismo se encontraban en un PDF hubo que cambiarlos a formato csv para poder trabajarlo en Python, por lo que se utilizó el programa de Word para convertir el PDF a un archivo tipo docx y después revisar que los datos se transfirieron de forma correcta en Excel.

Uno vez obtenido los datos, se procedió a determinar cuales algoritmos utilizar para poder realizar el agrupamiento, en este caso se utilizó el de k-means y mini batch k-means para observar si los datos introducidos por medio del informe de la ARESEP con respecto a la calidad de energía eléctrica tienen algún comportamiento de interés.

Antes de iniciar con el algoritmo, puesto que este es uno que trabaja con distancias, es importante normalizar los datos en especial este caso, ya que hay una variación grande de magnitud entre los abonados y otros datos, para ello se utiliza las herramientas que ofrece la librería de sklearn, en este caso se importó "StandarScaler" para poder normalizar la información.

Seguidamente, ahora que se posee los datos de la forma ideal, se continua con el algoritmo, se sabe que k-means y mini batch k-means trabaja con clusters, por lo que lo primero que se hacer es aplicar el método codo para determinar la cantidad de clusters (k). Para ello se obtienen los siguientes resultados 1 y 2 para los datos del anexo A y las figuras 5 y 6 para los del anexo B, sin embargo, al final se determinó el valor de Silhouette, mientras más cercano sea este valor a 1, mejor la consistencia de datos, por lo que se realizaron diversas pruebas para determinar el número de clusters a utilizar, los resultados de la tabla 1, basado en esta tabla se aplica 2 clusters con el algoritmo de K-means para ambos datos y para mini batch k-means se aplican 3 para el anexo A y 2 para el anexo B.

Finalmente, volviendo a utilizar la librería de sklearn, se procede a utilizar el algoritmo de Kmeans para ver como se agrupa la información, los resultados para el anexo A son 3 y 4 y para los del anexo B se tienen las figuras 7 y 8, para ambos casos se tienen los resultados de k-means y mini batch k-means respectivamente.

4. Resultados

4.1. Resultados Anexo A

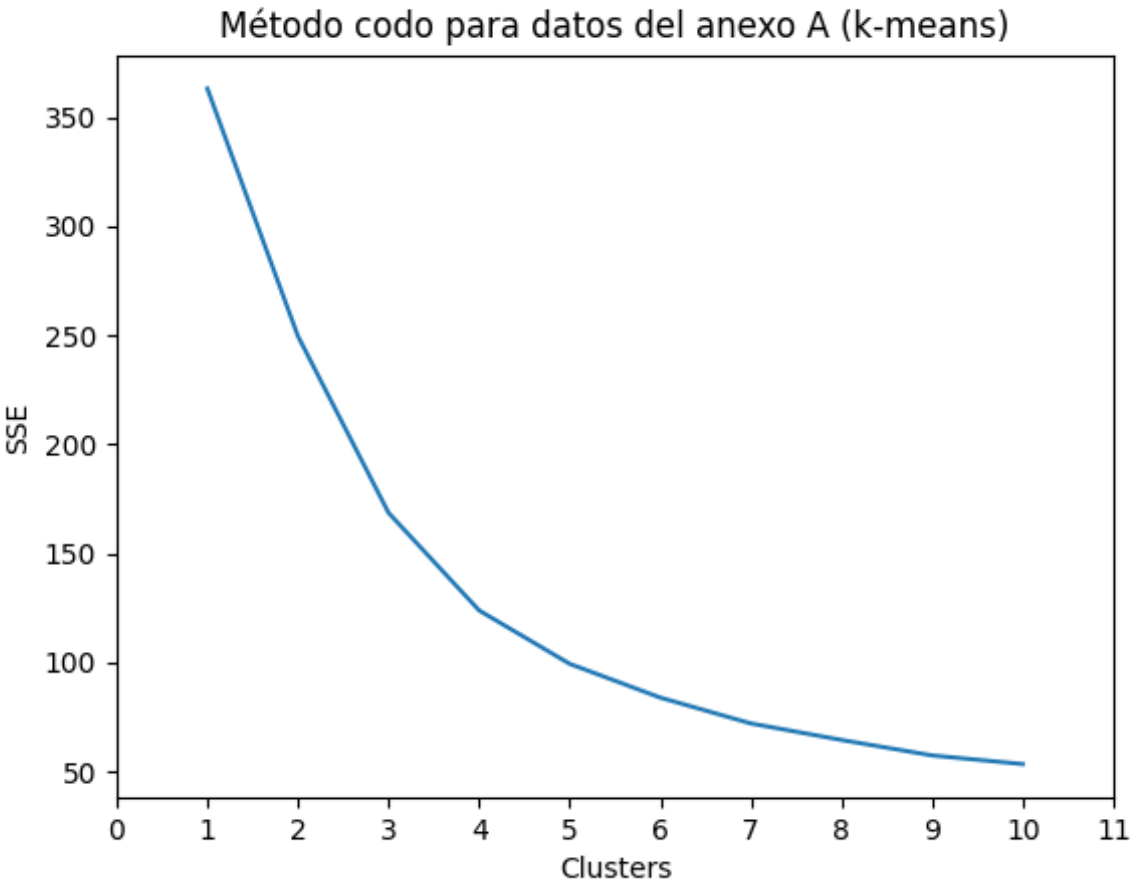

Figura 1: Resultado de método codo con los datos del Anexo A (k-means).

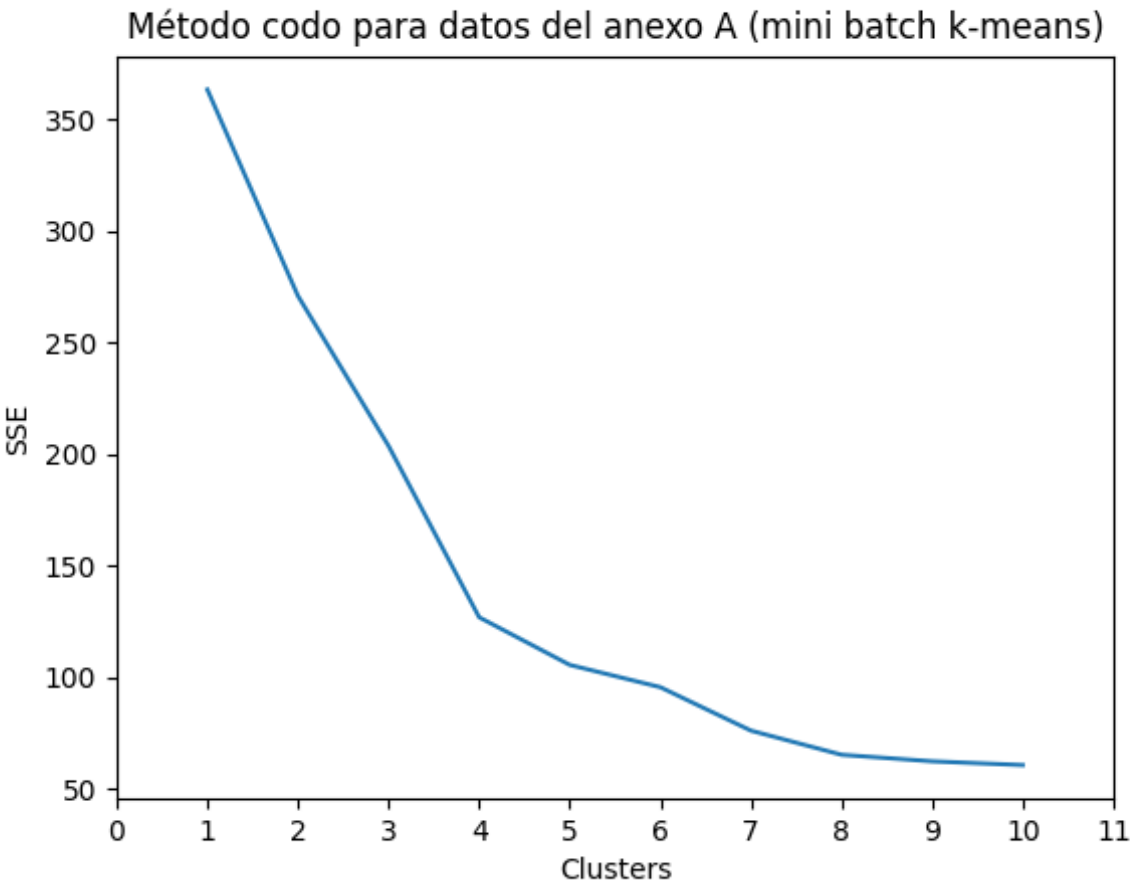

Figura 2: Resultado de método codo con los datos del Anexo A (mini batch k-means).

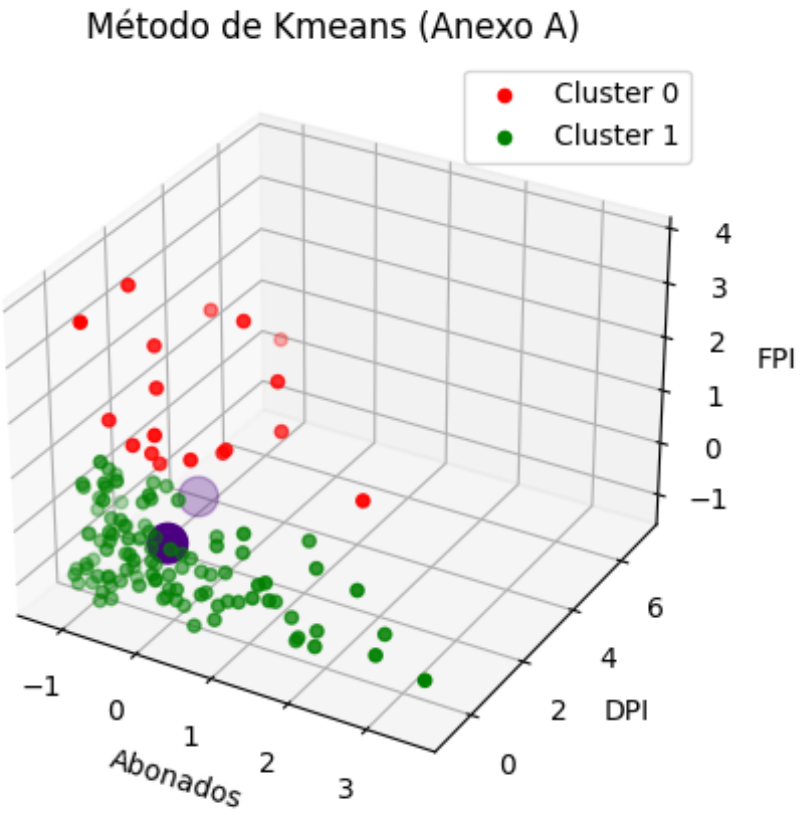

Figura 3: Resultado de aplicar el algoritmo K-means con los datos del Anexo A

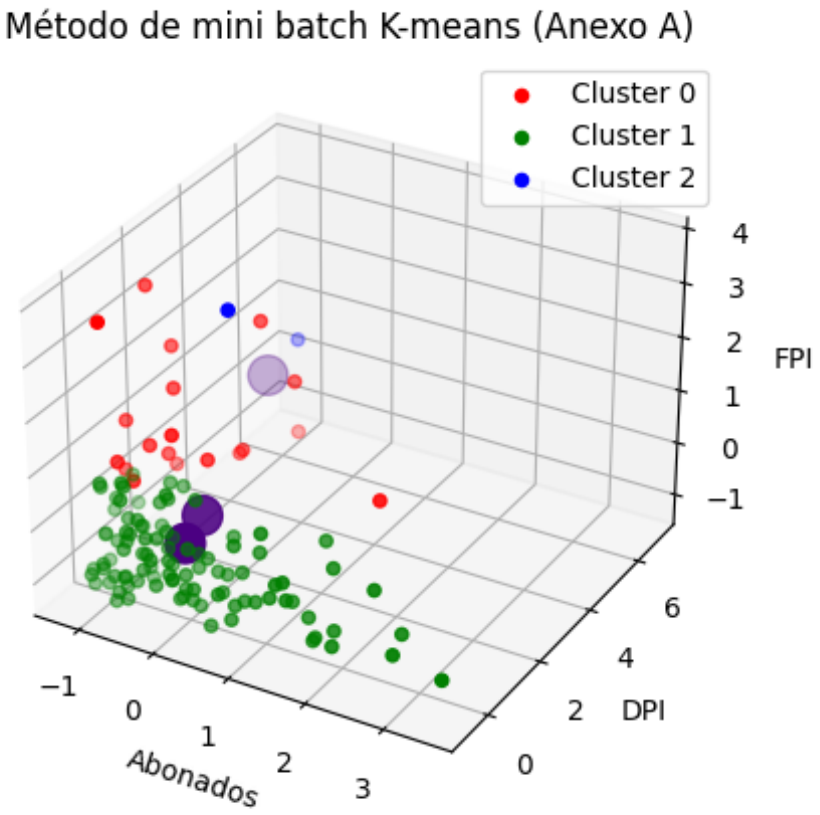

Figura 4: Resultado de aplicar el algoritmo mini batch K-means con los datos del Anexo A

4.2. Resultados Anexo B

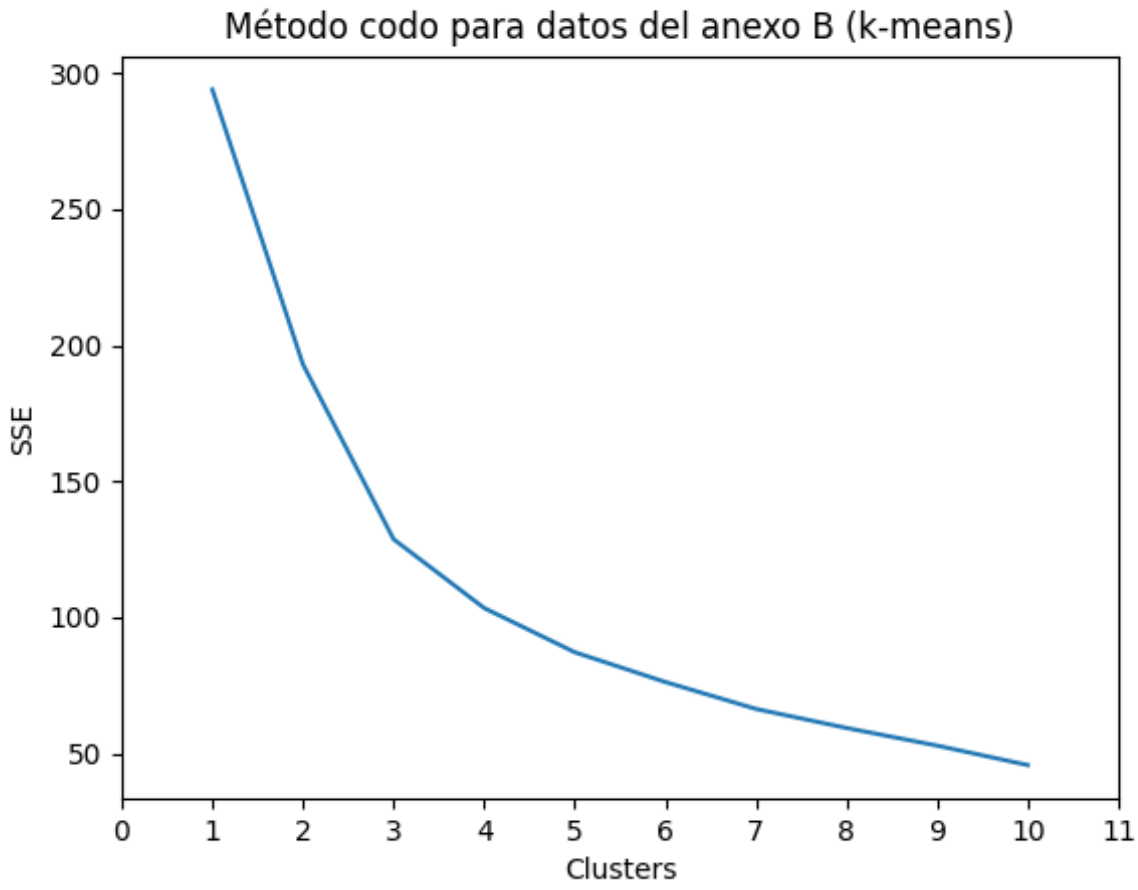

Figura 5: Resultado de método codo con los datos del Anexo B (k-means).

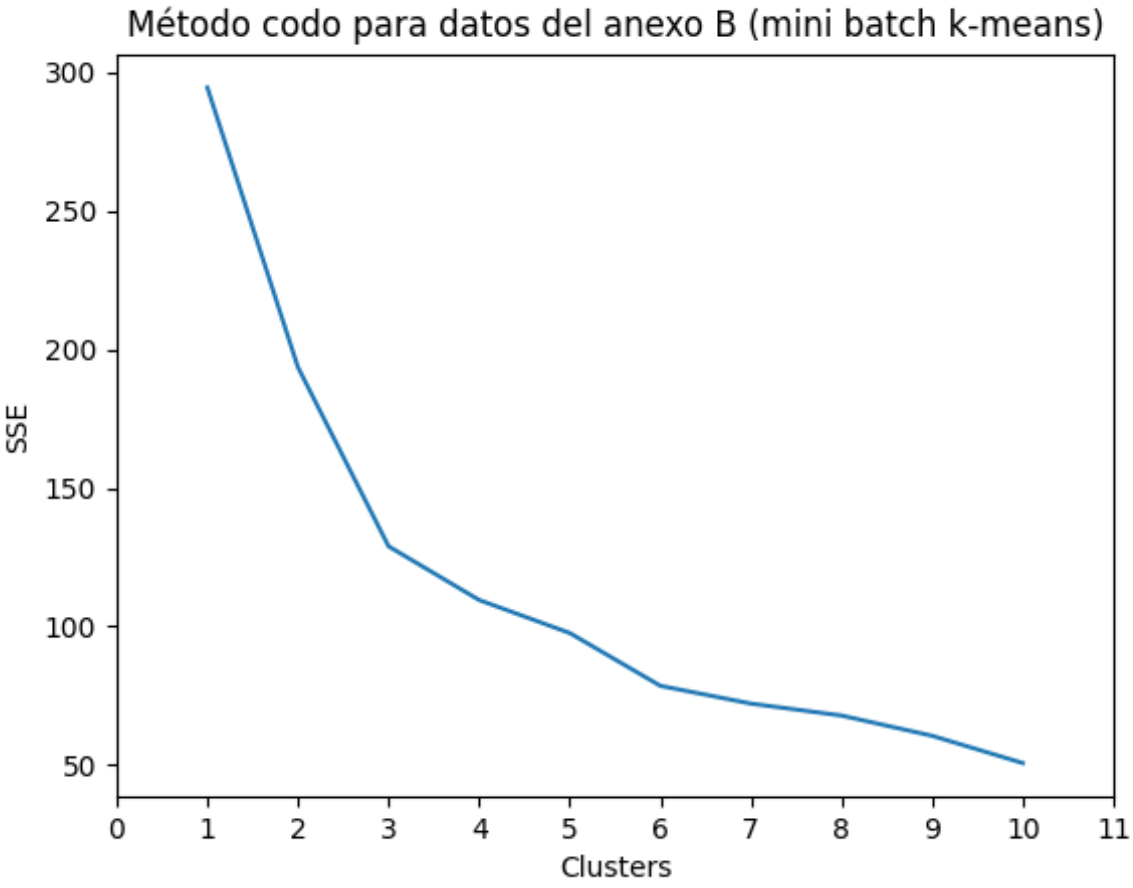

Figura 6: Resultado de método codo con los datos del Anexo B (mini batch k-means).

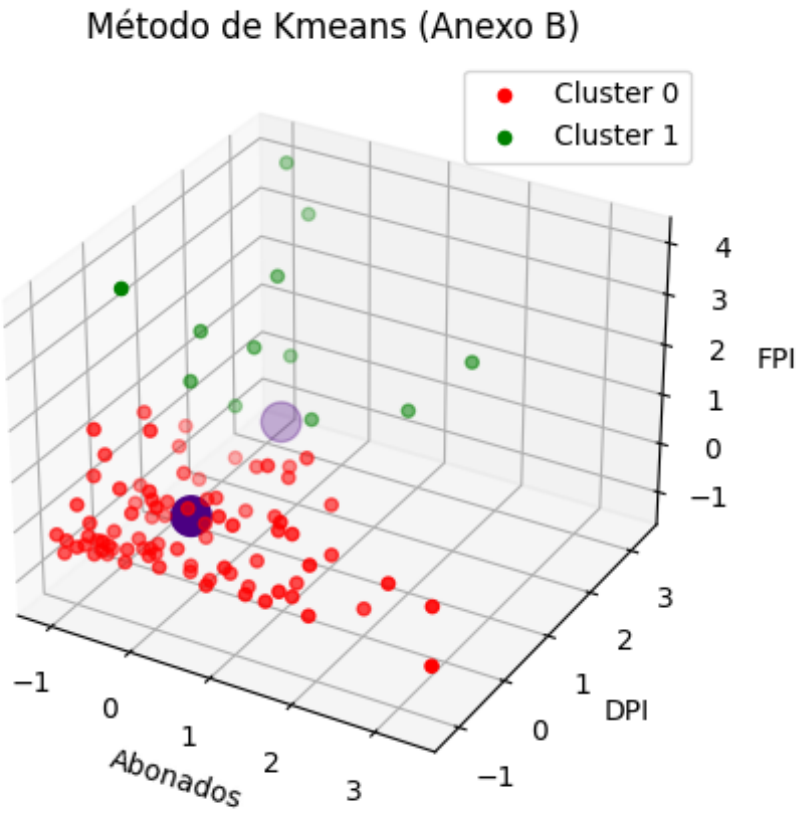

Figura 7: Resultado de aplicar el algoritmo K-means con los datos del Anexo B

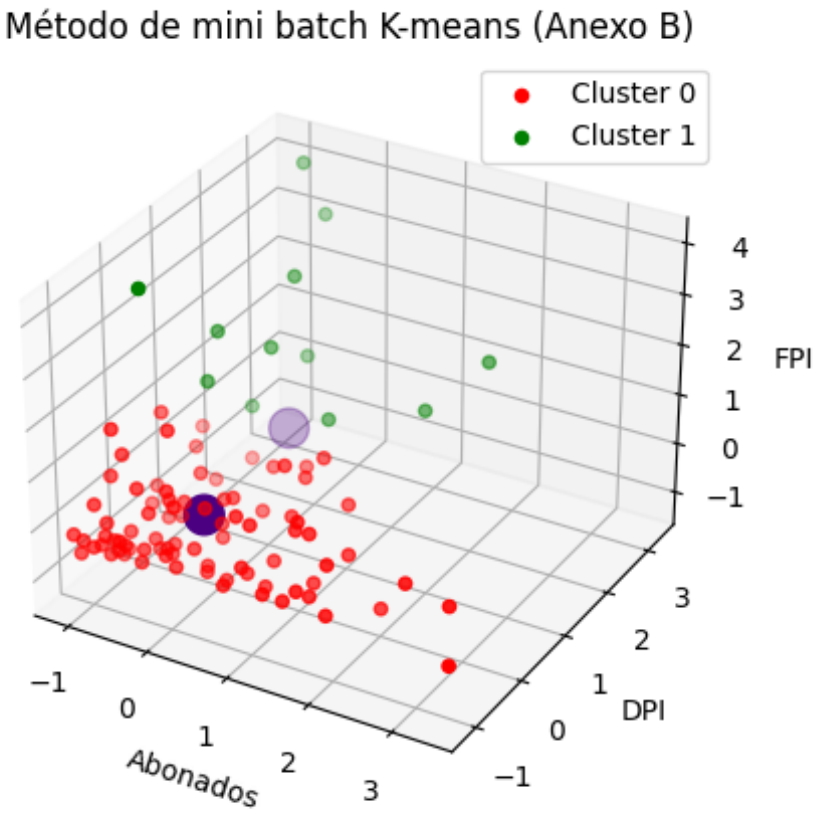

Figura 8: Resultado de aplicar el algoritmo mini batch K-means con los datos del Anexo B

K-means		
Clusters	Anexo A	Anexo B
2	0.469842927	0.487419336
3	0.399664994	0.369291121
4	0.391005108	0.375359225
5	0.355943634	0.31601784
Mini batch K-means		
2	0.263591114	0.487419336
3	0.414058683	0.369291121
4	0.377840644	0.291812644
5	0.371315706	0.2524759

Tabla 1: Resultados del valor Silhouetta

En base a los resultados obtenidos, lo más observable a simple vista en la mayoría de casos (con excepción al de la figura 4) de los datos usados, parece que se pueden clasificar en 2 grupos (o clusters en este caso), 1 de los grupos de datos se puede clasificar en grupos con una baja cantidad de abonados, DPI y FPI y los otros sería los que tienen una mayor cantidad de FPI y parece incluir, sin embargo, mantiene una cantidad pequeña de abonados con un leve incremento en DPI. Otra cosa por considerar es que no se sabe si se está trabajando con anomalías, por lo que no se sabe que tan preciso sean los datos, habría que implementar otro

algoritmo para ello. En base a los valores silhouetta obtenidos, es probable que no se utilizaron los algoritmos correctos para este problema porque no llegaron ni a 0.5, siendo 1 el mejor valor posible. Otra consideración es que en el anexo A, el circuito con el mayor valor de DPI, es Santarita islachira, esto tiene sentido porque es un lugar remoto, en este caso un dato, por lo que podría haber una relación entre la ubicación y DPI o FPI, sin embargo como no se utilizó el circuito como dato, es difícil afirmarlo.

5. Conclusiones

La normalización de datos en áreas como machine learning es importante, especialmente para algunos algoritmos como K-means, además que puesto que se trabaja con distancia hace que se tengan que hacer menos cálculos y que los resultados generados podrían cambiar si este proceso no se realiza.
Hay una gran cantidad de herramientas para aplicar algoritmos de inteligencia artificial, en este proyecto se utilizó la librería de sklearn, el mismo incluía los dos algoritmos en este trabajo, sin embargo, hay otros.
En este problema se utilizaron estos métodos con la intención de estudiar la calidad eléctrica de Costa Rica, sin embargo, en base a lo encontrado, no se puede confirmar que hay un relación concreta entre los datos, pero si se encuentran patrones, como por ejemplo, los circuitos con pocos abonados tienden a tener valores bajos de FPI y DPI y a medida que incrementa el DPI también tiende a aumentar el FPI también.
El agrupamiento es importante ya que, permite encontrar una relación entre datos que tal vez parezcan no tener relación alguna y encontrar patrones que podrían ayudar a solucionar un problema.

Referencias

[1] Misión, visión y valores. Disponible en `https://aresep.go.cr/mision-vision-valores/`, 2023.

[2] Thiago Gomes, André Borniatti, Vinícius Garcia, Laura Santos, Nelson Knak Neto, and Rui Garcia. Clustering electrical customers with source power and aggregation constraints: A reliability-based approach in power distribution systems. *Energies*, 16:2485, 03 2023.

[3] Y. Wang, J. Wong, and A. Miner. Anomaly intrusion detection using one class svm. In *Proceedings from the Fifth Annual IEEE SMC Information Assurance Workshop, 2004.*, pages 358–364, 2004.

[4] Tyler Lanigan, Sebastian Raschka, and Arturo Amor. Importance of feature scaling. Disponible en `https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html`, 2023.

[5] Feature scaling techniques in python – a complete guide. Disponible en `https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html`, 2022.