ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

Evaluación de métodos de agrupamiento de perfiles de carga para gestión de demanda en clientes de media tensión

TRABAJO DE TITULACIÓN

Previo la obtención del título de:

MAGISTER EN ELECTRICIDAD, MENCIÓN EN SISTEMAS ELÉCTRICOS DE POTENCIA

Presentado por:

Sebastián Israel Segovia Ortega

GUAYAQUIL – ECUADOR

OCTUBRE 2021

DEDICATORIA

Este trabajo va dedicado a Dios, que siempre está presente en todos los días de mi vida, en mi caminar, bendiciéndome y dándome fuerzas para cumplir mis metas planteadas sin desfallecer. A mi Padre Ángel Segovia Ortega y mi Madre Narcisa Ortega Sacoto por ese apoyo incondicional, confianza y amor, por ser el motor de mis días.

AGRADECIMIENTOS

Agradezco a Dios por la vida, por permitirme llegar y cumplir con una de mis metas.

A mis padres por el amor, dedicación, confianza, paciencia, y consejos que los llevare siempre en mi corazón y sobre todo el aliento en cada paso, por ser los principales impulsadores de mis sueños y decirme que siempre hay solución.

A todos los docentes que formaron parte de esta maestría que entregan día a día su corazón para compartir sus conocimientos y experiencias para la formación de otros para que podamos desenvolvernos ante situaciones que se presenten en nuestras vidas, en especial a mi tutor que ha sido mi guía en este trabajo Ing. Holguer Noriega y coordinador de la materia Ing. Fernando Vaca.

A mis colegas y amigos que hemos estado en contacto desde el inicio desde esta etapa y han sido parte de mi vida profesional.

Gracias a ESPOL que ha sido parte de mi camino profesional, por haber permitido formarme en ella, por sembrar conocimientos de calidad.

Este trabajo es muy especial, ya que es el resultado de mucho esfuerzo y dedicación.

DECLARACIÓN EXPRESA

Los derechos de titularidad y exploración, me corresponde conforme al reglamento de propiedad intelectual de la institución; Sebastián Israel Segovia Ortega doy mi consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual.

Firmado electrónicamente por:
SEBASTIAN ISRAEL SEGOVIA ORTEGA

Sebastián Segovia Ortega

EVALUADORES

Fírmado electrónicamente por:
FERNANDO ARTURO VACA URBANO

Fernando Vaca Urbano

PROFESOR DE TRABAJO DE TITULACIÓN

Fírmado electrónicamente por:
HOLGUER HUMBERTO NORIEGA ZAMBRANO

Holger Noriega Zambrano

TUTOR DE TRABAJO DE TITULACIÓN

Fírmado electrónicamente por:
JOHNNY WLADIMIR RENGIFO SANTANA

Johnny Rengifo Santana

PROFESOR REVISOR

RESUMEN

El presente trabajo de titulación trata de la evaluación de diferentes métodos de agrupamiento (*clustering*) aplicados a una estructura de datos obtenida de medidores AMI instalados en clientes de media tensión de la concesionaria distribuidora CNEL E.P Unidad de Negocios Guayas-Los Ríos. Los datos analizados son medidos en intervalos de 30 días, es decir una medida representa un mes del año, a lo largo de dos años. El objetivo principal de este estudio es manejar los perfiles de carga de los clientes de media tensión mediante diferentes métodos de agrupamiento para la gestión de la demanda por parte de la empresa distribuidora.

La metodología aplicada en este trabajo para poder evaluar los métodos de agrupamiento incluye el manejo de datos en bruto de la empresa distribuidora así como el filtrado de estos, un análisis de sensibilidad de los parámetros iniciales relacionados directamente a cada algoritmo a implementar y la normalización de las curvas de consumo de los agrupamientos formados por cada técnica *clustering.*

Se evaluarán cuatro métodos de *clustering*: K-Means, Fuzzy C- Means, DBSCAN y AGNES. El método de validación del parámetro inicial k para la formación de grupos de los métodos que lo necesitan (K-Means y Fuzzy C- Means) se llevó a cabo mediante el método del codo; mientras que, la validación del parámetro ε (usado por DBSCAN) se llevará a cabo mediante una función de MATLAB®.

Al realizar el análisis de la estructura de datos de los clientes de media tensión mediante cada uno de los métodos se puede afirmar que el tiempo computacional para ejecutar los algoritmos se reduce cuando los datos de entrada se encuentran normalizados; así como que en la mayoría de los métodos *clustering*, a excepción de DBSCAN, hubo una tendencia a formar 3 o 4 cuatro agrupamientos sobre los que se puede realizar la gestión de demanda para los perfiles de consumo ingresados, teniendo como modelos más robustos los algoritmos AGNES y DBSCAN; mientras que, los más flexibles fueron Fuzzy C-Means y K-Means, los cuales fueron clasificados de esta forma en base a los resultados obtenidos.

Palabras Clave: métodos clustering, clientes de media tensión, gestión de demanda.

ABSTRACT

This study deals with the evaluation of different clustering methods applied to a data structure obtained from AMI metering devices installed in medium voltage customers from the DSO CNEL E.P Guayas-Los Ríos. The data analyzed are measured in 30-day intervals, i.e., one measurement represents one month of the year, through two years. The main objective of this work is to manage the load profiles of medium voltage customers using different grouping methods for demand management by the distribution company.

The methodology applied is the evaluation of the clustering methods including the management of raw data from the distribution company as well as the filtering of these data, a sensitivity analysis of the initial parameters directly related to each algorithm to be implemented and the normalization of the consumption curves of the groupings formed by each clustering technique.

Four clustering methods will be evaluated: K-Means, Fuzzy C- Means, DBSCAN and AGNES. The validation method of the initial parameter k for cluster formation of the methods that need it (K-Means and Fuzzy C- Means) will be carried out using the elbow method; whereas the validation of the parameter ε (used by DBSCAN) will be carried out using a MATLAB® function.

When analyzing the data structure of the medium voltage customers using each of the methods, it can be stated that the computational time to run the algorithms is reduced when the input data is normalized; and that in most of the clustering methods, except for DBSCAN, there was a tendency to form 3 or 4 clusters on which demand management can be performed for the consumption profiles entered, with the most robust models being the AGNES and DBSCAN algorithms; while the most flexible ones were Fuzzy C-Means and K-Means, which were classified in this way based on the results obtained.

Keywords: clustering methods, medium voltage customers, demand management.

ÍNDICE GENERAL

RESUMEN i

ABSTRACT ii

ÍNDICE GENERAL iii

ABREVIATURAS vi

ÍNDICE DE FIGURAS vii

ÍNDICE DE TABLAS x

CAPÍTULO 1 1

1 INTRODUCCION 1

1.1 DESCRIPCIÓN DEL PROBLEMA 1

1.2 JUSTIFICACIÓN DEL PROBLEMA 2

1.3 OBJETIVOS 3

1.3.1 OBJETIVO PRINCIPAL 3

1.3.2 OBJETIVOS SECUNDARIOS 3

CAPÍTULO 2 4

2 CONCEPTOS BÁSICOS DE AGRUPAMIENTO 4

2.1 DEFINICIÓN DE AGRUPAMIENTO 4

2.2 REQUERIMIENTOS PARA LOS MÉTODOS DE AGRUPAMIENTO 6

2.3 PROBLEMAS DE LOS MÉTODOS DE AGRUPAMIENTO 8

2.4 USOS DE AGRUPAMIENTO 9

2.5 MÉTODOS DE AGRUPAMIENTO 9

2.5.1 K-MEANS 10

2.5.2 K-MEDOIDS 12

2.5.3 K-NEAREST NEIGHBOR 14

2.5.4 HIERARCHICAL CLUSTERING 16

2.5.5 DENSITY BASED CLUSTERING (DBSCAN) 19

2.5.6 GAUSSIAN MIXTURE MODELS (GMMs) 21

2.5.7 FUZZY C-MEANS CLUSTERING (SOFT K-MEANS) 23

2.5.8 AGNES ALGORITHM (AGGLOMERATIVE NESTING) 25

CAPÍTULO 3 27

3 CONCEPTOS BÁSICOS DE MÉTODOS DE VALIDACIÓN PARA AGRUPAMIENTOS 27

3.1 COEFICIENTE DE SILUETA 27

3.2 COEFICIENTE DE SILUETA PROMEDIO 29

3.3 MÉTODO DEL DIAGRAMA DEL CODO 29

CAPÍTULO 4 32

4 METODOLOGÍA DE LOS AGRUPAMIENTOS DE PERFILES DE CARGA 32

4.1 PERSPECTIVA GENERAL DE LA METODOLOGÍA 33

4.2 ESCALAMIENTO DE LAS CURVAS DE SALIDA DE CADA GRUPO FORMADO 34

4.3 MANEJO DE LOS DATOS EN BRUTO DE LA DISTRIBUIDORA 35

4.4 ANÁLISIS DE SENSIBILIDAD BASADO EN LOS PARÁMETROS DE ENTRADA 36

4.5 VALIDACIÓN DE GRUPOS 37

CAPÍTULO 5 40

5 ANÁLISIS DE RESULTADOS 40

5.1 REPRESENTACIÓN GRÁFICA DE LOS CLÚSTER Y RANGOS 40

5.1.1 K-MEANS 40

5.1.2 FUZZY C-MEANS 43

5.1.3 AGNES ALGORITHM (AGGLOMERATIVE NESTING) 45

5.1.4 DBSCAN 47

5.2 CURVAS DE CONSUMO PROMEDIO 50

5.2.1 K-MEANS 50

5.2.2 FUZZY C-MEANS 54

5.2.3 AGNES ALGORITHM (AGGLOMERATIVE NESTING).....57

5.2.4 DBSCAN 60

5.3 DISCUSIÓN DE RESULTADOS.....61

CAPÍTULO 6 64

6 CONCLUSIONES Y RECOMENDACIONES.....64

6.1 CONCLUSIONES.....64

6.2 RECOMENDACIONES 65

ANEXOS 68

ANEXO A. CURVAS ESTÁNDAR BASADAS EN EL MÉTODO DE MEJOR RENDIMIENTO 68

REFERENCIAS.....69

ABREVIATURAS

CNEL EP:	Empresa Eléctrica Pública Estratégica, Corporación Nacional de Electricidad
GIS:	Sistema de información geográfica
k-NN:	k-Nearest Neighbor
SSE:	Error de suma cuadrada
GMM:	Gaussian Mixture Models
EM:	Expectation-Maximization (Esperanza de maximización)

ÍNDICE DE FIGURAS

Figura 2.1 Medidas de asociación para el análisis de clúster. 7
Figura 2.2 Clasificación y subclasificación de los métodos de análisis clúster [12]. 9
Figura 2.3 Aplicación gráfica del método K-means [15].....11
Figura 2.4 Aplicación del método K-medoids a un conjunto de datos [15]. 13
Figura 2.5 Aplicación del método k-NN en una estructura de datos con dos etiquetas [15]. 15
Figura 2.6 Aplicación del método jerárquico anglomerativo en una estructura de datos [15]. a) Muestra los clústeres iniciales, b), c) y d) muestran el proceso iterativo del algoritmo.....18
Figura 2.7 Aplicación del algoritmo DBSCAN con un parámetros de 3 puntos mínimos [28]. 20
Figura 2.8 Aplicación del método GMM en una estructura de datos con sus diferentes tipos de clústeres [15].....22
Figura 2.9 Clasificación rígida de una estructura de datos: los puntos azules o rojos solo pertenecen a un solo agrupamiento. [15].....24
Figura 2.10 Clasificación flexible de una estructura de datos: los puntos de dos colores tienen grados de pertenencia a dos agrupamientos distintos. [15].....24
Figura 2.11 Dendograma del algoritmo AGNES en MATLAB(R), ejemplo de gráfico con colores limitados por el usuario [44]. 26
Figura 3.1 Aplicación del método de la silueta a un grupo de conglomerados con k = 2 [45]. 29
Figura 3.2 Aplicación del método del codo a una estructura de datos [51].....30
Figura 4.1 Clasificación de los algoritmos a implementar para el agrupamiento de datos de los clientes de media tensión.....33
Figura 4.2 Épsilon óptimo para datos escalados 39
Figura 4.3 Épsilon óptimo para datos no escalados 39
Figura 5.1 Gráfica de KMeans de 4 clúster con datos escalados 41
Figura 5.2 Gráfica de K-Means de 4 clúster con datos no escalados.....42
Figura 5.3 Gráfica de Fuzzy C-Means de 4 clúster con datos escalados 43
Figura 5.4 Gráfica de Fuzzy C-Means de 4 clúster con datos no escalados 44
Figura 5.5 Gráfica de AGNES con datos escalados 45

Figura 5.6 Dendograma AGNES con datos escalados 45
Figura 5.7 Gráfica de AGNES con datos no escalados 46
Figura 5.8 Dendograma AGNES con datos no escalados 46
Figura 5.9 Gráfica de DBSCAN con datos escalados 48
Figura 5.10: Grafica de DBSCAN con datos no escalados 49
Figura 5.11 Curva de consumo promedio anual de clúster 1 creado por KMeans con datos escalados 50
Figura 5.12 Curva de consumo promedio anual de clúster 2 creado por KMeans con datos escalados 51
Figura 5.13 Curva de consumo promedio anual de clúster 3 creado por KMeans con datos escalados 51
Figura 5.14 Curva de consumo promedio anual de clúster 4 creado por k-means con datos escalados 52
Figura 5.15 Curva de consumo promedio anual de clúster 1 creado por KMeans con datos no escalados 52
Figura 5.16 Curva de consumo promedio anual de clúster 2 creado por KMeans con datos no escalados 53
Figura 5.17 Curva de consumo promedio anual de clúster 3 creado por KMeans con datos no escalados 53
Figura 5.18 Curva de consumo promedio anual de clúster 4 creado por KMeans con datos no escalados 53
Figura 5.19 Curva de consumo promedio anual de clúster 1 creado por Fuzzy C-Means con datos escalados 54
Figura 5.20 Curva de consumo promedio anual de clúster 2 creado por Fuzzy C-Means con datos escalados 54
Figura 5.21 Curva de consumo promedio anual de clúster 3 creado por Fuzzy C-Means con datos escalados 55
Figura 5.22 Curva de consumo promedio anual de clúster 4 creado por Fuzzy C-Means con datos escalados 55
Figura 5.23 Curva de consumo promedio anual de clúster 1 creado por Fuzzy C-Means con datos no escalados 56
Figura 5.24 Curva de consumo promedio anual de clúster 2 creado por Fuzzy C-Means con datos no escalados 56

Figura 5.25 Curva de consumo promedio anual de clúster 3 creado por Fuzzy C-Means con datos no escalados 56
Figura 5.26 Curva de consumo promedio anual de clúster 4 creado por Fuzzy C-Means con datos no escalados 57
Figura 5.27 Curva de consumo promedio anual de clúster 1 creado por AGNES con datos escalados 57
Figura 5.28 Curva de consumo promedio anual de clúster 2 creado por AGNES con datos escalados 58
Figura 5.29 Curva de consumo promedio anual de clúster 3 creado por AGNES con datos escalados 58
Figura 5.30 Curva de consumo promedio anual de clúster 1 creado por AGNES con datos no escalados 59
Figura 5.31 Curva de consumo promedio anual de clúster 2 creado por AGNES con datos no escalados 59
Figura 5.32 Curva de consumo promedio anual de clúster 3 creado por AGNES con datos no escalados 60
Figura 5.33 Curva de consumo promedio anual de clúster 1 creado por DBSCAN con datos escalados 60
Figura 5.34 Curva de consumo promedio anual de clúster 1 creado por DBSCAN con datos no escalados 61

ÍNDICE DE TABLAS

Tabla 4.1 Algoritmos de agrupamiento y parámetros de entrada requeridos. 36
Tabla 4.2 Valores obtenidos de método del codo. 37
Tabla 5.1 Rangos de clústeres creados por K-Means con datos escalados 41
Tabla 5.2 Rangos de clústeres creados por K-Means con datos no escalados 42
Tabla 5.3 Rangos de clústeres creados por Fuzzy C-Means con datos escalados 43
Tabla 5.4 Rangos de clústeres creados por Fuzzy C-Means con datos no escalados .44
Tabla 5.5 Rangos de clústeres creados por AGNES con datos escalados 46
Tabla 5.6 Rangos de clústeres creados por AGNES con datos no escalados 47
Tabla 5.7 Rangos de clústeres creados por AGNES con datos no escalados 48
Tabla 5.8 Rangos de clústeres creados por AGNES con datos no escalados 49
Tabla A.1 Curvas de demanda en p.u. del método de agrupamiento con mejor rendimiento: AGNES. 68

CAPÍTULO 1

1 INTRODUCCION

En la actualidad se conoce que el incremento de la demanda energética y los costos económicos se disparan teniendo mayor facturas a pagar debido a que no existe una planificación y estudios previos al comportamiento de la carga con el único fin que la demanda energética sea de forma segura y eficiente, por ello es importante conocer las principales características de los perfiles de carga que pueden ayudar a identificar soluciones óptimas para hacer más eficientes las redes de distribución, representado en una reducción de las planillas de electricidad tanto en el sector comercial como en el industrial.

Es por esta razón que este capítulo expone el problema actual que existe en las redes de distribución de media tensión, así como la propuesta de una solución óptima que permita realizar estudios con históricos para definir curvas de carga estándar que permitan la gestión de la demanda en los consumidores.

1.1 DESCRIPCIÓN DEL PROBLEMA

Conocer las principales características de los perfiles de carga pueden ayudar a identificar soluciones óptimas para hacer más eficientes las redes de distribución, representado en una reducción de las planillas de electricidad tanto en el sector comercial como en el residencial [1]. Para conocer el comportamiento eléctrico de los clientes, esta tesis analizará diferentes métodos de agrupamiento. El agrupamiento (*clustering* en inglés) es considerado uno de los más importantes problemas de aprendizaje no supervisado (unsupervised learning), el cual trata de encontrar una estructura de una colección de datos sin etiquetar [2].

En la literatura existen muchos casos de estudio usando métodos no supervisados. Autores en [3] segmentan hogares en grupos (clusters) basados en sus patrones de electricidad a través del día. Adicionalmente, los hogares son caracterizados basados en variaciones de demandas entre días, estaciones, meses. En [4] un procedimiento es analizado para determinar la forma de los

perfiles de carga por estaciones europeas. [5] presenta un algoritmo nuevo de agrupación basado en datos iniciales usando la distancia de Hausdorff. En [6] los autores analizan el consumo eléctrico de diferentes países europeos de diferentes características geográficas, resaltando los más importantes patrones.

Esta tesis evalua diferentes métodos de agrupamiento de carga a perfiles de carga de clientes de media tensión, información que será obtenida de CNEL E.P Unidad de Negocios Guayas- Los Ríos. Estos métodos son analizados y comparados entre sí. Basados en los diferentes grupos obtenidos a partir del método del codo (elbow method), se calcula un perfil de carga estándar como variables de salida del estudio.

1.2 JUSTIFICACIÓN DEL PROBLEMA

Para una buena planificación y operación en las distribuidoras de energía eléctrica, es importante conocer el comportamiento de la carga, un elemento difícil de estimar. Por tal razón se hace de vital importancia, utilizar métodos de Machine Learning como es el caso del agrupamiento de perfiles de carga [7].

Analizar las características comunes de grupos de perfiles de carga de clientes de media tensión mediante métodos de agrupación permite extraer la información que se relaciona con los usuarios, de tal modo que se pueden observar patrones de comportamiento en diferentes grupos de usuarios y de este modo utilizar dichas tendencias particulares de cada grupo para diferentes propósitos y beneficios de la empresa distribuidora, entre los cuales se pueden destacar la reducción de pérdidas no técnicas, predicción de carga a corto plazo o por intervalos de tiempo que pueden estar relacionados a días de la semana o estaciones del año, así como establecer normativas para fijación de precios dinámicos según el tiempo de uso de electricidad de los usuarios [7], [8].

1.3 OBJETIVOS

1.3.1 OBJETIVO PRINCIPAL

Evaluar métodos de agrupamiento (*Clustering* en inglés) para gestión de demanda de clientes en media tensión de una empresa distribuidora (en el caso de Ecuador, CNEL E.P.) mediante el manejo de sus datos de perfiles de carga.

1.3.2 OBJETIVOS SECUNDARIOS

- Revisar en la literatura de aplicativos de métodos de agrupamientos en manejo de datos de curvas de carga.

- Analizar el conjunto de perfiles de carga de clientes de media tensión y observar cómo se encuentran distribuidos por grupos.

- Definir las ventajas y desventajas de los diferentes métodos de agrupamientos de perfiles de carga de la distribuidora a analizar mediante una revisión bibliográfica.

- Analizar las fortalezas de cada método de agrupamiento por medio de los resultados obtenidos.

- Generar una curva estándar de carga por cada grupo, basado en el método de agrupamiento con mejor rendimiento para la gestión de la demanda de clientes en media tensión.

CAPÍTULO 2

2 CONCEPTOS BÁSICOS DE AGRUPAMIENTO

En el presente capítulo se detallan los conceptos básicos de la técnica de agrupamiento o *clúster*, así como los requerimientos y limitaciones que estas implican. Dado a que objetivo de esta tesis es evaluar métodos de agrupamiento se debe tener una clara idea de los fundamentos de esta técnica, sus limitantes, aplicaciones y métodos existentes mediante la revisión de la literatura para posteriormente en el Capítulo 4 definir una metodología y evaluar el desempeño de algunos de los métodos propuestos en esta sección.

2.1 DEFINICIÓN DE AGRUPAMIENTO

La definición del Método de Agrupamiento o *Clustering* actualmente no es fija debido a la interpretación de esta está envuelta en base a su aplicación, sin embargo, tal como indican [9] y [10] puede ser definido como una técnica cuya función es agrupar datos no etiquetados en un grupos que compartan la misma características (homogéneos) de sus elementos y sean diferentes de los otros grupos, permitiendo de esta forma descubrir la estructura de los datos para su respectiva clasificación.

El análisis *clúster* es una tarea descriptiva, es decir, permite la descripción datos existentes (histórico), pero no la predicción de estos; esto se debe a que la técnica se basa en criterios geométricos y estadística multivariante, de ahí que se utilice como una técnica para describir la estructura de los datos o agrupar datos, pero no para explicarlos ni mucho menos para su pronóstico [10], [11].

Lo que resulta atrayente del análisis *clúster* es que se debe tener poco o ningún conocimiento de la estructura de los datos que se desean agrupar, sin embargo, el analista debe de los datos debe estar lo suficientemente informado para poder definir cuáles son las características deseables e indeseables al momento de clasificar los grupos [12].

Debido a las múltiples aplicaciones que tiene esta técnica, las soluciones que brinda dependen mucho de la técnica elegida y del objetivo principal con el que se lo utilice, tal como índice [12] se pueden definir al menos cuatro:

- Desarrollar una topología de la estructura de datos.
- Investigar esquemas conceptuales para el agrupamiento de entidades.
- Generar hipótesis mediante la exploración de los datos.
- Contrastar hipótesis.

El uso típico que suele darse a los métodos de *clustering* es la creación de una estructura de datos o de clasificaciones, pero todos estos objetivos en conjunto pueden formar la base de estudio para un proyecto o investigación.

Los análisis de agrupamientos se dividen en dos grandes grupos, los cuales son: **jerárquicos y no jerárquicos**. El tipo de *análisis jerárquico* crea *clústeres* pequeños que son englobados por otros más grandes, estos últimos son considerados de mayor nivel en comparación a los primeros, formando así una estructura arborescente. Por otro lado, el tipo de *análisis no jerárquico* permite crear grupos claramente diferenciados, es decir, ningún agrupamiento (*clúster*) depende de otro, a estos grupos se los conoce como *clusters disjuntos*; mientras que, si algún elemento o varios elementos pertenecen a otro o varios grupos se conoce como *clusters solapados*, los cuales son empleados con rara frecuencia debido a su complicada interpretación [13].

Debido a lo expuesto en el párrafo anterior es importante tomar en cuenta tres pautas antes de iniciar un análisis con *clústeres* [13]:

- Selección de variables relevantes para identificar a los grupos que se desean clasificar.
- Elección de la medición de distancias entre puntos y proximidad entre los datos.
- Selección de un criterio para agrupar los datos en *clústeres.*

2.2 REQUERIMIENTOS PARA LOS MÉTODOS DE AGRUPAMIENTO

Los requerimientos para realizar un análisis por medio de la técnica *clúster* pueden resumirse en 3 puntos clave [12], estos son:

1. Elección de las variables de análisis.
2. Elección de la medida de asociación para los datos.
3. Elección de la técnica *clúster* a implementar en el estudio

La ***elección de las variables de análisis*** es el marco de referencia sobre el cual se va a construir los grupos, este requerimiento es clave porque define el rumbo del proyecto o de la investigación que se esté realizando, de modo que desde la elección de las variables se tiene una clasificación de datos, pues dependiendo de qué tan relevantes sean estas para el tipo de clasificación que se está buscando se podrá obtener al final del proceso un resultado significativo [12].

Dependiendo de la aplicación la variable puede ser de dos tipos: 1) Cualitativa (ordinal o nominal), o 2) Cuantitativa (discreta o continua) [13]. Esto se debe a diferentes situaciones: datos con unidades de medida diferente, datos con categorías predefinidas, datos con un valor de escala tipo intervalo, entre otros; de modo que, tipificar las variables antes de realizar un análisis es la solución general para evitar problemas al utilizar la técnica *clúster* al tratar como equivalentes a todas los parámetros de un conjunto de datos [12].

Es por esto por lo que al delimitar las variables a las más relevantes o las que posean mayor información útil para la investigación, de tal modo que se asegura que exista la mínima redundancia entre los datos seleccionados [9].

Por otro lado, la ***elección de la medida para la asociación de datos*** como define [9] es la cuantificación de qué tan similares o diferentes son los objetos analizados, es decir, permite medir la proximidad entre los mismos de manera cuantitativa. Usualmente las técnicas de creación de grupos involucran medidas del tipo coeficiente de correlación, las cuales tiene una interpretación sencilla o complicada en base al método que se utilice para realizar la misma [12].

Las medidas de asociación para los datos pueden clasificarse en dos: 1) Distancia, o 2) Similaridad. Cuando se selecciona como medida de asociación *distancia* se tiende a formar agrupamientos y en estos los individuos se caracterizan por tener una distancia pequeña entre ellos, esto se puede observar cuando se aplica la distancia euclidiana, por ejemplo; mientras que, cuando se selecciona *similaridad* como método de asociación los agrupamientos resultantes se caracterizan por tener un valor de similaridad alto entre los datos, por ejemplo esto se obtiene al aplicar un coeficiente de correlación [13]. La figura 2.1 muestra las diferentes medidas de asociación que pueden aplicarse a una estructura de datos.

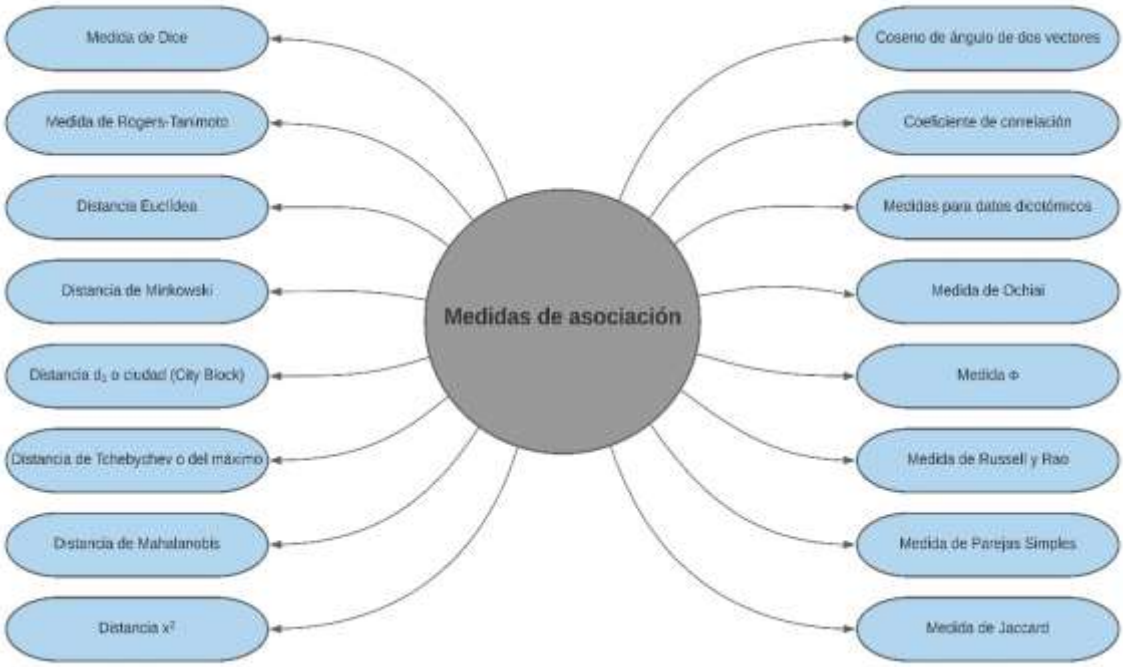

Figura 2.1 Medidas de asociación para el análisis de clúster.

Las figura 2.1 muestran las diferentes medidas de asociación existentes, las cuales se aplican según el tipo de análisis que va realizarse, la cual puede ser por tipo de variables o individuos, aunque en ambos casos se puede utilizar cualquier tipo de medidas, por lo que queda a elección del usuario cuál escoger [13].

Finalmente, ***la elección de la técnica clúster a implementar en el estudio*** permite al investigador tener su conclusiones en base a los datos a analizar con la

técnica usada. Si implementa un *clúster del tipo jerárquico* el investigador debe tener sus propias conclusiones en base a los resultados obtenidos, pues al implementar la técnica en el algoritmo este le mostrará cuántos *clústeres* puede contener la estructura de los datos ingresados; mientras que, si implementa un *clúster* del *tipo no jerárquico* al permitir ingresar de antemano el número de *clústeres*, tienen cierta flexibilidad en decidir el número de agrupaciones que se tendrán al final, por lo que es el investigador el que incide directamente en los resultados y puede tener conclusiones en base a las decisiones tomadas [12].

2.3 PROBLEMAS DE LOS MÉTODOS DE AGRUPAMIENTO

En base a los requerimientos se puede expresar que uno de los problemas en el método de agrupamiento es el número de variables que van a definir para poder implementar la técnica, ya que, si el investigador selecciona una cantidad amplia de estas puede ocasionar que la capacidad computacional de un ordenador no sea suficiente o que la estructura de los datos no quede bien definida debido a que las variables no son las adecuadas.

Otro de los problemas que relacionados a la selección de variables es que debido a la cantidad de datos que se ingresan, estas pueden tener diferentes dimensiones de medida, lo cual hace que se requiera de convertir las variables a binarias para calcular la similitud entre las mismas, causando que en este proceso se pueda perder información que resulte valiosa para los resultados de la investigación [12].

Por otro lado, debido a que no existe una pauta definida sobre qué técnica de agrupamiento aplicar, esto es jerárquico o no jerárquico, se deben probar varios métodos y técnicas implementadas en los algoritmos para que estos resultados se puedan contrastar, de tal modo que si resultan lo más parecido posibles demuestran la estructura natural de los datos y permite tener conclusiones válidas; mientras que, sino no son parecidos este indica que los datos no obedecen una estructura definida y por ende concluir que la información utilizada no es la adecuada para aplicar la técnica *clúster* [12].

2.4 USOS DE AGRUPAMIENTO

La técnica *clúster* tiene una gran variedad de campos de aplicación, entre ellos se encuentran los siguientes usos [14] :

- Reconocimiento de patrones
- Mapas temáticos (GIS)
- Segmentación de clientes (Marketing)
- Clasificación de documentos
- Análisis de web logs
- Minería de datos

2.5 MÉTODOS DE AGRUPAMIENTO

Los métodos de agrupamiento (*clustering*) son muy extensos, ya que existen múltiples técnicas para su implementación, debido a la aplicación práctica que tienen en diferentes ámbitos; sin embargo, en esta sección solo se mencionarán las que tienen desarrollos en herramientas computacionales como MATLAB®, R™, entre otros.

La clasificación parte de las técnicas jerárquicas o no jerárquicas mencionadas previamente en la primera sección del capítulo, pero estas pueden contener subclases tal como lo muestra la figura 2.2, en la cual se observa que las técnicas jerárquicas se dividen en dos subclases y la no jerárquicas en cuatro subclases.

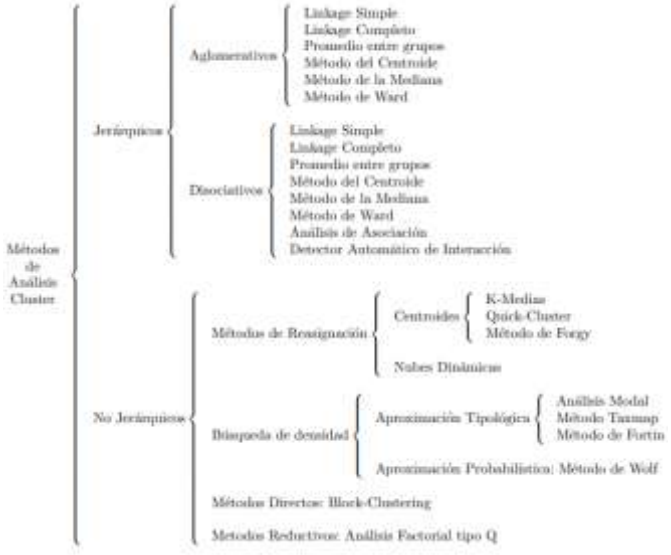

Figura 2.2 Clasificación y subclasificación de los métodos de análisis clúster [12].

Los algoritmos que implementan estas técnicas son variados, pero en el presente documento se presentará las técnicas de *clustering* más utilizadas, entre ellas destacan [15]:

- K-means
- K-medoids
- K-Nearest Neighbor
- Fuzzy C-means Clustering (Soft k-means)
- Density Base Clustering (DBSCAN)
- Gaussian Mixture Models (GMMs)
- Hierarchical cluestering

2.5.1 K-MEANS

Es un algoritmo encargado de agrupar los grupos de datos en un número k de *clústeres*, objetivos, variables, puntos, observaciones, entre otros; donde en cada grupo se busca minimizar la medida de disimilitud, siendo la distancia euclidiana medida típica utilizada para reducir esta disimilitud, buscando que los grupos tengan la menor varianza posible [16].

Se agrupan en función de una característica en común, siendo en este caso un centroide; un centroide es un valor aleatorio dado por el propio modelo, sobre el cual se le es asignado un nuevo objeto más cercano. El objetivo de esta metodología es reducir en lo más posible la varianza que se presentaría con la adición de nuevos puntos cercanos al centroide, a la varianza interna por cada *clúster* se la conoce también como inercia; como dato adicional, el centroide de un *clúster* es el promedio de los elementos de este [15].

El resultado de este método despende de la selección inicial del valor k, pues el usuario decidirá los k grupos/*clústeres* que debe asignar el algoritmo al conjunto de datos, de este modo, al asignar los k centroides se asociarán los datos más cercanos a estos. Una vez que se han fijado los centroides, al agregar nuevas observaciones estos se actualizan calculando el promedio del

grupo y midiendo las disimilitud (varianza) entre los datos causando el desplazamiento del centroide [16], [17].

Existen dos herramientas matemáticas usadas para determinar la varianza presente dentro de un *clúster*, las cuales son [15]:

- Suma de las distancias euclidianas al cuadrado entre cada objeto (x_i) y el centroide (u_k)

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2 \tag{2.1}$$

- Suma de las distancias euclidianas de todos los pares de observaciones asociadas a un *clúster*, de la cual se divide para la cantidad de dichas observaciones en el conjunto.

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - \mu_{i'j})^2 \tag{2.2}$$

En simples palabras podría decirse que, realiza un problema de optimización para buscar su solución debido a que busca minimizar la sumatoria de todas las posibles distancias euclidianas (suma de distancias cuadráticas) dadas entre un punto/objeto y su respectivo centroide asociado a cada *clúster*. Un ejemplo de agrupación del método K-means se muestra en la figura 2.3

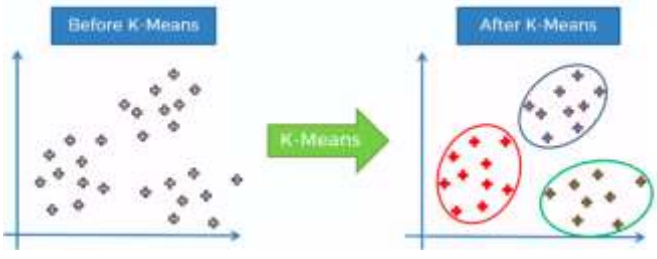

Figura 2.3 Aplicación gráfica del método K-means [15].

Una de las ventajas del método es que se considera este modelo como sencillo y rápido, además de presentar cierta robustez a dimensiones no tan elevadas. Dará un mejor resultado si el conjunto de datos de entrada permite

una densidad lo suficientemente espaciada entre los *clústeres*. La manera de definir qué tan cercano se encuentra una observación al centroide se lo realiza por medio de la distancia euclidiana, que de por sí es uno de los métodos más amigables para aplicar con respecto a las diferentes medida de asociación de datos que existen [15].

Por otro lado, la principal desventaja radica en el hecho de que depende totalmente de la cantidad óptima de k *clústeres*, no se puede realizar el modelo sin saber el coeficiente k, en otras palabras, dependerá de la selección inicial del usuario. Del mismo modo, otra de las limitantes al usar este tipo de algoritmos es que no funciona bien cuando la estructura de datos ingresada tiene una forma no esférica, causando problemas al momento de decidir los óptimos locales y globales para formar los agrupamientos. Adicional a esto, un problema que tiene una influencia directa en la presentación de los resultados es que K-Means es muy sensible al ruido (datos aberrantes) lo que le da un problema de sensibilidad [15], [17], [18].

2.5.2 K-MEDOIDS

La metodología de este algoritmo es idéntica a la de K-Means pues tiene como objetivo principal agrupar observaciones, puntos, objetos en k *clústeres* siendo denominados C, sin embargo, la diferencia radica en la forma que se calcula el centroide en la estructura de datos ingresadas; mientras que, K-Means está representado por un centroide, el cual es el promedio del *clúster*, el método K-Medoids representa cada *clúster* por una observación aleatoria existente [15], [18].

El desarrollo del modelo es el mismo que el ocurrido en K-Means, sólo que ahora se reducirá la distancia cuadrática en base al medoid por cada *clúster,* para que el mismo posea una distancia mínima entre todas las observaciones (si la distancia se puede minimizar, se actualiza el valor de medoids); en sí ambos métodos se pueden resumir de la siguiente manera: K-Means es a media (promedio); mientras que, K-Medoids es a mediana (frecuencia de ocurrencias) [15], [19].

En base a lo expresado, entonces puede definirse un medoid como el *objeto existente de un grupo* cuya varianza o disimilitud promedio con el grupo es mínima, siendo un objeto representativo de cada agrupamiento convirtiéndolo en el objeto con más similaridad de los datos de su *clúster*, pudiéndose tomar como referencia al igual que los centroides del K-Means [20].

Los objetos representativos iniciales (medoids iniciales) se seleccionan aleatoriamente del conjunto de datos ingresados, luego al igual que K-Means mediante un proceso iterativo se reemplazan los objetos no representativos por uno representativo, mejorando la calidad de la agrupación resultante [20].

La figura 2.4 muestra de manera gráfica la aplicación del método K-Medoids en un conjunto de datos, nótese que los *clúster* seleccionados tienen como centroide un dato existente (medoid).

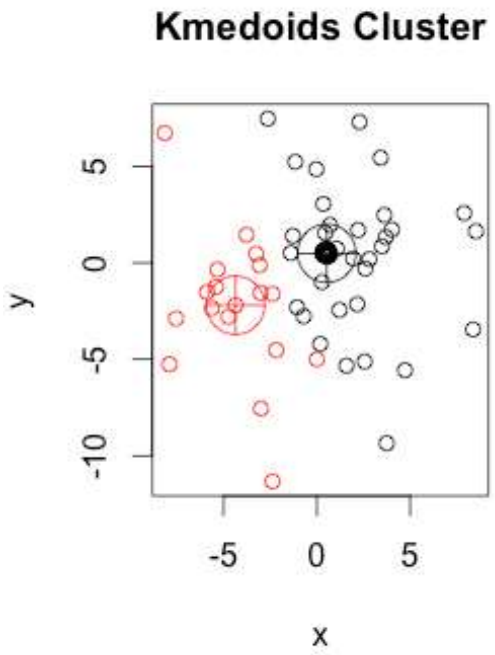

Figura 2.4 Aplicación del método K-medoids a un conjunto de datos [15].

Una de las principales ventajas de este método es que a diferencia del método K-Means es más robusto pues al seleccionar un objeto representativo como centro (medoid) no tiene el efecto de ruido, es decir, puede manejar los datos atípicos sin problemas, pero para poder realizar esto el investigador debe reemplazar el uso de la distancia Euclidiana por la de Manhattan, ya que, la

primera medida para la asociación de datos es afectada principalmente por el ruido [15], [20]. Otra de las ventajas es que el error cuadrático medio de los agrupamientos formados por un algoritmo de K-Medoids es mucho menor que el de los formados por el algoritmo K-Means, permitiendo así obtener *clústeres* óptimos [19].

Por otro lado, al igual que K-Means presenta un problema de sensibilidad con los datos iniciales, ya que, el investigador debe ingresar la cantidad k de *clúster* que desea realizar en la estructura de datos a analizar, lo cual en el mundo real es precisamente el dato que se desconoce. Del mismo modo, al adicionar la aleatoriedad de los medoids, esto causa que los agrupamientos resultantes sean diferentes en cada ejecución del algoritmo [18].

Otra de las limitantes y desventaja respecto a K-Means es que este algoritmo por lo general es adecuado solo para un conjunto pequeño de datos, ya que, cuando existe una gran estructura de datos representa un problema de carácter computacional [15], [20]. Aunque la solución a este problema es utilizar el algoritmo CLARA en vez de PAM [18].

2.5.3 K-NEAREST NEIGHBOR

El método de k-Nearest Neighbor (k-NN) permite la agrupación de datos mediante la diferencia de distancias cuadráticas al igual que el modelo k-means; sin embargo, la diferencia clave que existe es que k-NN clasifica los objetos de base a las etiquetas o límites (boundaries) más cercanas del objeto a procesar en su espacio de características; es decir, un objeto se clasifica por “el voto mayoritario de sus vecinos” de tal modo que este se asigna un agrupamiento por los k datos más cercanos [15], [21].

Este es un algoritmo de aprendizaje supervisado más simple de todos los algoritmos de aprendizaje automático (Machine Learning Algorithms) y es debido a esta característica que como dato de entrada es necesaria una etiqueta correcta que es previamente conocida por el investigador, esta categorización o etiqueta es lo que debe contener cada columna de la estructura de datos a analizar para poder realizar los agrupamientos [21].

El valor de k en este algoritmo ya no se refiere al número de agrupamientos que se van a realizar, más bien ahora este parámetro se refiere al número de "vecinos" u observaciones más cercanas tomando en cuenta la medida de distancia óptima del *clúster*, esto se consigue a través de la distancia euclidiana o mediante algún otro método de distancia, formando así los límites o superficie sobre el que se etiquetará un patrón; por lo tanto, es de suponer que los patrones de la estructura de datos que se encuentran cercanos entre sí tendrán la misma etiqueta, teniendo en cuenta un número mínimo de objetos para poder crear un *clúster*, es decir, bajo esta técnica son los objetos vecinos los que deciden si un objeto nuevo pertenece o no a dicha etiqueta/clase [15], [21].

La figura 2.5 muestra la aplicación del método k-NN en un conjunto de datos mostrando cómo varía la superficie de una etiqueta/clase al aumentar el valor del parámetro k, que como se explicó previamente es el número de vecinos mínimos para crear una agrupación.

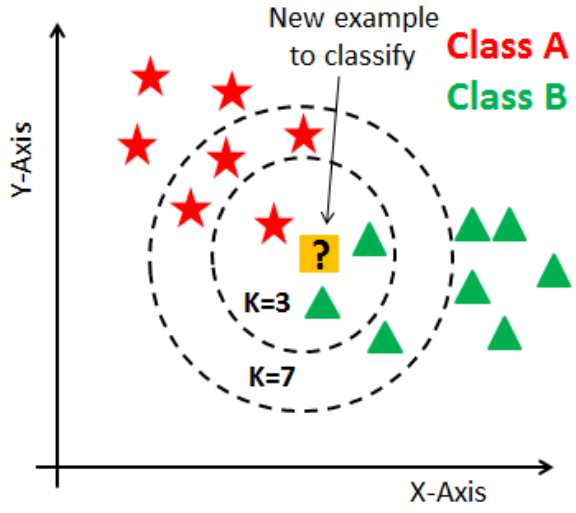

Figura 2.5 Aplicación del método k-NN en una estructura de datos con dos etiquetas [15].

La principal ventaja de este algoritmo es un algoritmo robusto que permite manejar específicamente problemas de clasificación, ya que, su ingreso de la estructura de datos deben ser datos claramente etiquetados. Adicional a esto, gracias a su flexibilidad y sencilles en comparación a otros algoritmos de este tipo permite puede interpretarse como un clasificador empírico de Bayes, permitiendo así obtener también estimaciones de los datos o estimación de

probabilidades, teniendo incluso una tasa de error de clasificación menos del doble que la de Bayes [15], [22], [23].

Del mismo modo, debido a la teoría simple e intuitiva del algoritmo este puede ser aplicado en muchos campos logrando los objetivos esperados, actualmente ha sido implementado en situaciones relacionadas al flujo del tráfico, en la predicción de estudios biológicos relacionados al cactus, en la ingeniería de sistemas biológicos para detectar enfermedades en las hojas de arroz, entre otros [24].

Por otro lado, como desventaja cuenta con una limitante a grupos de datos pequeños, esto impide su aplicación a análisis de tiempo real de una estructura de datos u otras aplicaciones donde se necesite agrupar una gran cantidad de objetos; es decir, a medida que aumenta la dimensión de los datos ingresados se pueden producir problemas de carácter computacional [15]. Un ejemplo de esto lo menciona [22] en donde se evalúa el desempeño de este algoritmo en la detección de perturbaciones en la calidad de energía en tiempo real, encontrándose con deficiencias de funcionamiento cuando la estructura de los datos es muy grande y mostrando un aumento en la precisión del agrupamiento a medida que se reducen las dimensiones de los mismos.

Otro de los parámetros que afectan a la eficiencia del método es el valor de k, el cual debe determinarse de manera empírica para poder obtener resultados significativos [24].

2.5.4 HIERARCHICAL CLUSTERING

Este tipo de algoritmo tal como indica su nombre, agrupamiento jerárquico, clasifica los objetos de tal modo que un *clúster* está anidado con el siguiente de la secuencia, poniéndolos desde un nivel mayor a un nivel menor, es decir, un *clúster* con *subclústeres*. A diferencia de otras técnicas este no necesita el parámetro inicial k que define el número de agrupamientos que se desea obtener en la estructura de datos [15], [25].

La clasificación de los métodos jerárquicos se puede realizar mediante dos grandes grupos: 1) Algoritmos anglomerativos, y 2) Algoritmos divisivos. En los *algoritmos anglomerativos* cada objeto de la estructura de datos inicialmente representa un *clúster*, luego de manera iterativa los *clústeres* procederán a unirse hasta formar un árbol completo; mientras que, en los *algoritmos divisivos* la estructura de datos se la considera como un solo *clúster* de manera inicial y de manera iterativa este se va separando son *subclústeres* [25].

Pese a la existencia de estos dos grandes grupos de los algoritmos de jerarquización, en términos de calidad de agrupación los anglomerativos resultan más eficientes que sus homólogos divisivos [25]. Debido a esto, es importante entender la secuencia de funcionamiento de esta clasificación del algoritmo, la cual se describe a continuación tal y como detalla [26]:

1. Inicialmente considera cada objeto de la estructura de datos como un *clúster* o grupo.
2. Calcula las distancias entre dos agrupamientos y combina los dos conglomerados más cercanos posibles, es decir, los que tienen más similitud entre sí.
3. Repite de manera iterativa el paso 2 hasta que el número de *clústeres* obtenidos sea igual al 10% del total de agrupamientos iniciales, es decir, hasta que el número de *clúster* sea igual al 10% de los objetos contenidos en el grupo de datos.

Los criterios de medidas de asociación de datos utilizados en este algoritmo son la distancia mínima, la distancia máxima, la distancia media y la distancia promedio [25].

Una manera gráfica de modelar los resultados de este algoritmo es mediante el dendograma, el cual en la parte inferior del mismo cuenta con los *clústeres* iniciales del paso 1 (*llamados hojas)* y a medida que el algoritmo repite su proceso iterativo para verificar la similitud de los objetos se van creando nuevos *clústeres* (*llamados ramas)*, subido así hasta niveles superiores finalizando con una estructura arbórea [15], [25].

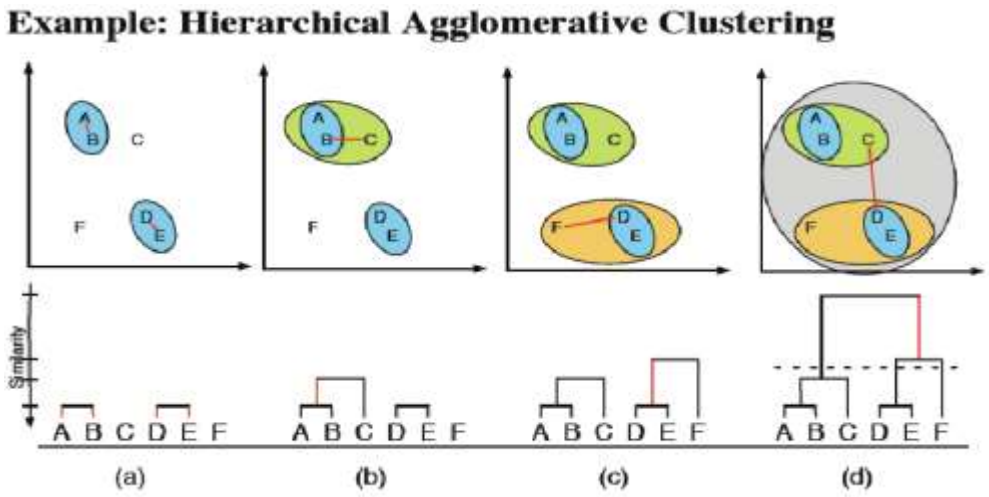

Figura 2.6 Aplicación del método jerárquico anglomerativo en una estructura de datos [15]. a) Muestra los *clústeres* iniciales, b), c) y d) muestran el proceso iterativo del algoritmo.

Una de las principales ventajas de los algoritmos jerárquicos anglomerativos es que permiten obtener de manera visual un resultado de fácil interpretación para el investigador, pues debido a su alta precisión se puede obtener agrupamientos de estructuras con cualquier tipo de forma, es por esto que es usualmente usado cuando se desea definir la cantidad de *clústeres* deseados en un conjunto de datos [15], [26].

Debido a su sencilla aplicación, al igual que k-means, y su alta precisión ha tenido algunas aplicaciones prácticas, tal como en la búsqueda de imágenes web facilitándole así la búsqueda de los usuarios, también ha sido aplicado para controlar la trayectoria de vehículos de evacuación después de un desastre, así como en redes de logística en las que se tiene que entregar/recoger artículos en casos de desastre, análisis de datos de energía fotovoltaica, entre otros [27].

Por otro lado, actualmente existen varias desventajas al utilizar este tipo de técnica *clúster*. La principal desventaja es que estos no vuelven a revisar los agrupamientos creados con el objetivo de mejorar la presentación final de la estructura de los datos, es decir, los *clústeres* y *subclústeres*. Otra de las desventajas es que la complejidad del tiempo computacional es elevada y a medida que la dimensión de la estructura de datos crece, se complica manejar la información para crear los agrupamientos. Del mismo modo, respecto al ingreso de datos, no es recomendable si la estructura contiene datos nulos/vacíos o la

presencia de muchos datos atípicos, pues el algoritmo no se comporta correctamente [15], [25], [26].

2.5.5 DENSITY BASED CLUSTERING (DBSCAN)

El modelo DBSCAN es un algoritmo de agrupación de densidad relativamente reciente, presentado en el año de 1996. Esta técnica de agrupamiento es capaz de reconocer la forma de la estructura de los datos aunque esta sea arbitraria y dentro de las mismas contengan una alta presencia de datos atípicos, a diferencia de k-means y sus asociados este algoritmo puede identificar los *clústeres* de manera análoga a que un investigador seleccione de manera manual los agrupamientos de una estructura de datos [15], [28].

La implementación del algoritmo DBSCAN tiene como punto de partida la inicialización de los parámetros Epsilon (ε) y los puntos mínimos (MinPts) para formar los agrupamientos, donde el primero se define como el radio de la superficie que albergará las observaciones vecinas y el segundo es el número mínimo de objetos presentes dentro de la superficie creada para crear un agrupamiento. Estos dos parámetros son esenciales, pues determinan el carácter del agrupamiento de la estructura de los datos a analizar durante la ejecución del algoritmo [15], [28].

Una forma sencilla de describir el proceso del algoritmo es la que plantea [29] y se detalla a continuación:

1. Selecciona de manera arbitraria un objeto central sin una categoría/*clúster* como semilla para la búsqueda de los agrupamientos.
2. Encuentra todos los conjuntos de muestras cuyo objeto central, semilla, puede alcanzar la densidad de datos seleccionada (puntos mínimos), es decir, forma un *clúster*.
3. Repite el paso 1 para encontrar otro objeto que no ha sido procesado y posteriormente el paso 2 de manera iterativa para obtener el resto de los *clústeres* hasta que todos los objetos centrales (semillas) que el algoritmo designe tengan la densidad de datos alcanzada.

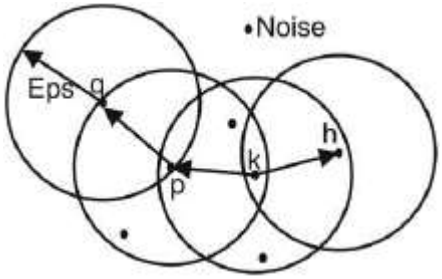

Figura 2.7 Aplicación del algoritmo DBSCAN con un parámetros de 3 puntos mínimos [28].

La figura 2.7 muestra de manera gráfica el funcionamiento del algoritmo DBSCAN con un parámetro de 3 puntos mínimos. Inicialmente, comienza con la sección de un punto aleatorio k. Luego, este verifica si existen objetos alrededor de la superficie de radio ε, en caso de que existan estos y cumplan la densidad alcanzable de datos (puntos mínimos) se forma un *clúster*; por otro lado, si k es un punto fronterizo y no hay se cumple el requisito de puntos mínimos se pasa al siguiente punto de la base de datos para tomarlo como semilla e iterar el algoritmo nuevamente. Si el objeto no es una semilla y no se puede alcanzar por ninguna de las antes seleccionadas, este datos es un ruido y por lo tanto el algoritmo lo ignora [28].

Una de las principales ventajas de usar esta técnica *clúster* es que se pueden realizar *clústeres* sin importar la estructura que tenga la base de datos y si existen altas condiciones de ruido (datos atípicos), ya que, a diferencia de k-means y sus algoritmos modificados, este no cuenta con la limitante de operación cuando la base de datos tiene forma circular o convexa, es decir, formaciones cuasi-circulares. Adicional a esto, al igual que los algoritmos jerárquicos no necesita el parámetros inicial del número de *clúster* para realizar sus iteraciones [15], [28] .

Algunas de sus aplicaciones han sido para la clasificación de señales de radar para el reconocimiento de ruidos, para el agrupamiento de errores de manipulación de transporte ferroviario, para el control de tráfico en las costas debido a la densidad de barcos, entre otros [30], [31], [32].

Por otro lado, existen dos principales dificultades al implementar el método. La primera es que la elección de los parámetros externos, épsilon y puntos

mínimos, debe ser ingresadas por el usuario sin un conocimiento previo; dado a que estos afectan de manera directa los resultados de los *clúster*, este inconveniente puede producir diferentes agrupamientos en base a los parámetros ingresados que no necesariamente son los más óptimos, lo cual resulta en una rigurosa experimentación del usuario para determinar los valores apropiados en base a la estructura de datos que desea analizar, de modo que estos parámetros sean establecidos artificialmente en base a la experiencia, lo cual reduce su confiabilidad. La segunda desventaja es que su tiempo de funcionamiento computacional aumenta a medida que aumenta la dimensión de la base de datos ingresada debido a su método para calcular la similaridad de los objetos de un agrupamiento [29], [33].

2.5.6 GAUSSIAN MIXTURE MODELS (GMMs)

Este Modelo de Mezcla Gaussiana (GMM) se puede definir como una función de densidad de probabilidad paramétrica con la que el algoritmo estima las diferentes agrupaciones a partir de los datos de entrada, esto lo logra a través de una combinación probabilística de diferentes/múltiples distribuciones normales, de modo que lo que se realiza para obtener un *clúster* es la superposición de las curvas normalizadas, basándose en la probabilidad de que uno de los objetos de la estructura de datos pertenezca a un agrupamiento, dejando así de lado los centroides formados por la distancia media de los objetos del *clúster* que se utilizan en métodos como k-means [15], [34].

Previo a definir los factores importantes que permiten funcionar esta técnica se debe tener en cuenta que este modelo trabaja con modas para datos desagrupados, ya que, el objetivo es formar grupos a partir de la estructura de datos; estas para este tipo de datos se clasifican en: unimodal, bimodal, multimodal o puede carecer de moda. El principal interés de este algoritmo son las modas unimodales y multimodales, las primeras se obtienen a partir de un solo dato que se repite varias veces en una estructura de datos; mientras que, las últimas se obtienen cuando más de dos datos se repiten con frecuencia [35].

Un factor importante que se debe tener en cuenta en esta técnica es que la definición de los centroides se lleva a cabo de representaciones unimodales

sobrepuestas sobre un modelo mixto que agrupa múltiples distribuciones normales, esto es lo que define los nuevos “centroides”. El encargado de verificar que esta representación multimodal probabilística encaje con la estructura de datos ingresada, es el que se conoce como *Expetation-Maximitation* (EM), que es prácticamente un algoritmo que maximiza la verosimilitud de los parámetros del GMM, o sea un parámetro de ajuste [15], [34].

Existen 4 tipos de *clúster* asociados a este modelo como muestra la figura 2.8, los cuales se describen a continuación [15]:

1. Tied: Todos los *clústeres* poseen una misma matriz de covarianza (a diferencia de k-means, la representación multimodal aumenta la dimensionalidad del problema).

2. Diagonal: Las dimensiones de cada *clúster* pueden variar, dependiendo de la dimensión en que se encuentren.

3. Spherical: Misma dimensionalidad de la matriz de covarianza, indiferente de la dimensión en que se encuentren.

4. Full: Independencia de dimensión y orientación, modelado como una elipse.

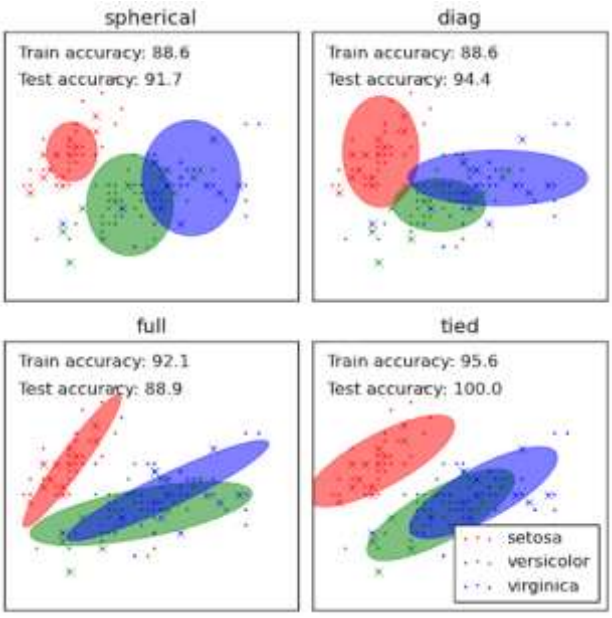

Figura 2.8 Aplicación del método GMM en una estructura de datos con sus diferentes tipos de clústeres [15].

La principal ventaja de esta técnica es que, al aplicar una distribución normalizada, se asegura que las observaciones asociadas a un *clúster* no se repitan o compartan características de otros "centroides". Del mismo modo, al depender de funciones de probabilidad el agrupamiento de datos se pude llevar a cabo sin importar la estructura de datos [15].

Actualmente tiene varias aplicaciones prácticas debido a la forma en la que agrupa los datos, entre ellas destacan el uso en vigilancia de tráfico, en la codificación o restauración de imágenes, en clasificación de clips de audio, así como el administración de conjuntos desequilibrados, entre otras [34], [36]–[38].

Por otro lado, la principal desventaja es que a diferencia del modelo de k-means y k-medoids, al basarse en métodos probabilísticos, la complejidad del modelo aumenta considerablemente y eso sin considerar la afinación de parámetros que también se asocia a metodologías descriptivas (EM), lo cual también representa una dificultad para el investigador al implementar esta técnica *clúster* [15].

2.5.7 FUZZY C-MEANS CLUSTERING (SOFT K-MEANS)

El algoritmo Fuzzy C-Means toma como base al algoritmo de la técnica K-Means, en la cual se asociada el parámetro k como el número de *clúster* a realizar en la estructura de datos, así como el hecho de que los centroides que se forman son el promedio de las distancias de cada *clúster* encontrado; la única diferencia con respecto al método K-Means es la manera de agrupar los datos [15].

La diferencia en la asignación del *clúster* para un objeto radica en que el método, ya no es rígido como tal, es decir, el objeto de análisis ya no tiene una valoración binaria (0 o 1) para pertenecer a determinado agrupamiento, sino que ahora existe la posibilidad de que un objeto tenga un grado de pertenencia a cada *clúster*, causando que este método sea flexible al permitir que un dato pueda pertenecer a uno o más *clústeres* con diferentes grados de pertenencia. Cabe recalcar que la suma de los grados de pertenencia debe sumar 1, pues el

de manera práctica lo que calcula el algoritmo es la probabilidad de pertenencia, la cual va entre 0 y 1, de un objeto con respecto a los diferentes *clústeres* [39], [40].

Este nuevo enfoque permite que el centroide sea actualizado constantemente en base a los datos que contienen varios grados de pertenencia y un mejor agrupamiento de los datos. De modo que, la "suavidad" o "flexibilidad" que le otorga este nuevo algoritmo a los datos es tener diversos grados de membresía para los k *clústeres* un lugar de una sola asignación como lo hace el método k-means ordinario, esto es posible gracias a la definición de un nuevo parámetro β, el cual es el parámetro de rigidez que afecta a la probabilidad de pertenencia de un dato a un agrupamiento [40], [41].

La figura 2.10 muestra de manera gráfica la aplicación del método Soft k-means en la cual se pueden observar una comparación del método k-means (figura 2.11) con este, resaltando de manera clara los objetos de la estructura de datos en la técnica soft k-means que pertenecen a más de un agrupamiento con sus respectivos grados de pertenencia.

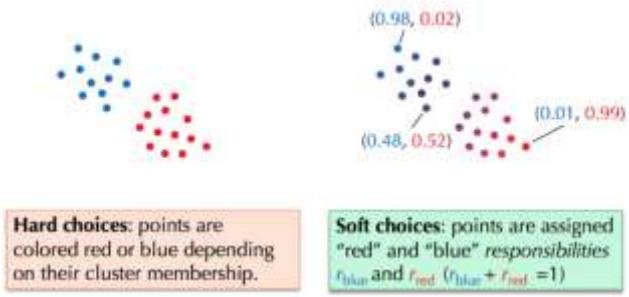

Figura 2.9 Clasificación rígida de una estructura de datos: los puntos azules o rojos solo pertenecen a un solo agrupamiento. [15].

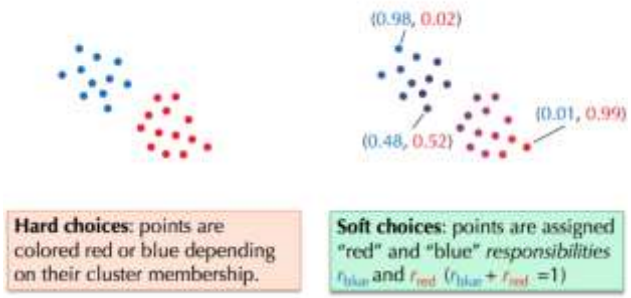

Figura 2.10 Clasificación flexible de una estructura de datos: los puntos de dos colores tienen grados de pertenencia a dos agrupamientos distintos. [15].

La principal ventaja de este método es que al implementar un parámetro de flexibilidad se permite tener una partición de datos más clara, ya que, al permitir que un objeto pueda pertenecer a varios *clústeres* la precisión al momento de clasificar los datos aumenta en comparación a la técnica K-Means [39] .

Por otro lado, su principal desventaja sigue siendo la definición de los parámetros iniciales, que en este caso es el número de *clústeres* k, adicionando el parámetro β. Ambos parámetros tienen una repercusión directa en los resultados por lo que el investigador necesita varias experimentaciones para poder llegar a un agrupamiento óptimo [15], [40].

2.5.8 AGNES ALGORITHM (AGGLOMERATIVE NESTING)

El algoritmo AGNES pertenece a los algoritmos jerárquicos expuestos en la sección 2.5.4, el cual tiene un uso de anidación aglomerativa, por lo que es de esperarse que su resultado se presente a través de un dendograma con un conjunto k de bloques, los cuales van desde 1 hasta n. La ventaja de este algoritmo es que es fácil de implementar, tal como es el caso de K-Means o K-Medoids [42].

El hecho de que sea una anidación anglomerativa, causa que inicialmente cada objeto de la estructura de datos de entrada sea asignado a un *clúster*, condición a partir de la cual comienza a crear los grupos en base a la similitud de cada uno de los objetos hasta llegar a un solo grupo principal [43].

Aunque es cierto que al ser un algoritmo del tipo jerárquico y no necesita un parámetro de entrada, cabe recalcar que en cada lenguaje de programación se pueden utilizar líneas de código para poder identificar con colores los principales grupos que se pueden observar en un Dendograma, en el caso de MATLAB® esto es posible, tal como se muestra en el ejemplo de la figura 2.11 donde se pueden identificar 3 principales grupos para la estructura de datos ingresada [44].

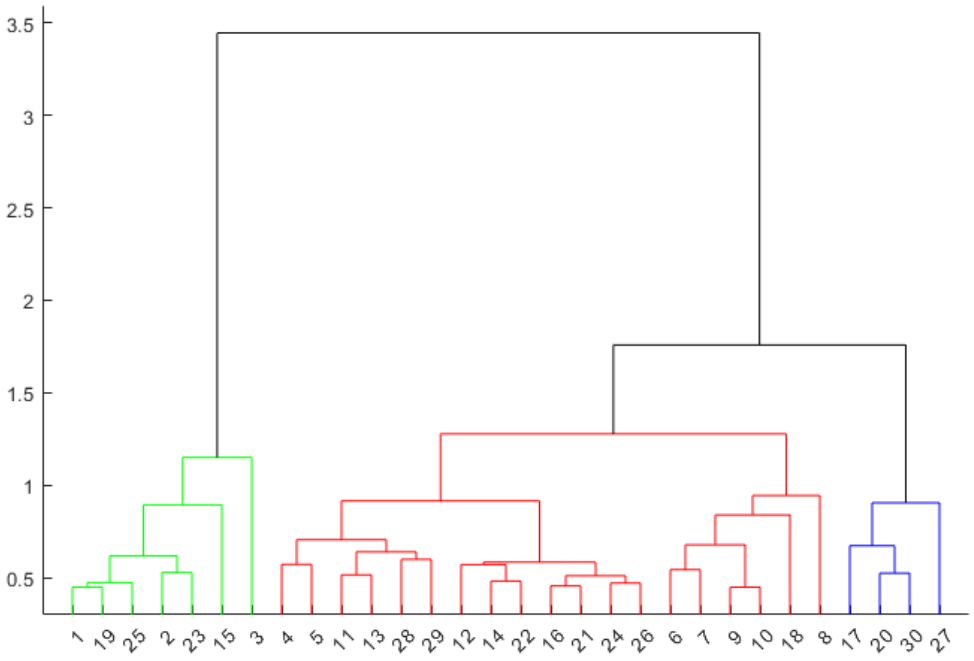

Figura 2.11 Dendograma del algoritmo AGNES en MATLAB(R), ejemplo de gráfico con colores limitados por el usuario [44].

CAPÍTULO 3

3 CONCEPTOS BÁSICOS DE MÉTODOS DE VALIDACIÓN PARA AGRUPAMIENTOS

En el presente capítulo se discute la importancia de evaluar los resultados de los algoritmos de técnicas *clúster* analizados en el Capítulo 2, ya que, una vez realizado los agrupamientos se deben cuantificar la calidad de estos. La importancia de estos métodos radica en que como algunos algoritmos necesitan de determinados parámetros iniciales para poder realizar los *clústeres*, independientemente de que sean jerárquicos o no, y dado a que estos en muchas ocasiones se establecen por la experiencia del usuario al experimentar con los algoritmos, resulta en que los agrupamientos obtenidos no siempre serán los óptimos debido a la baja confiabilidad que reporta este tipo de selección.

En la presente sección se analiza tres métricas para la validación de los agrupamientos obtenidos por las técnicas *clúster*, las cuales son las siguientes: 1) Coeficiente de silueta, 2) Coeficiente de silueta promedio, y 3) Método del diagrama de codo.

3.1 COEFICIENTE DE SILUETA

El coeficiente de silueta es una métrica para evaluar la calidad de los agrupamientos obtenidos el cual tiene como objetivo principal identificar el número óptimo de *clústeres* que deben crear los algoritmos [45].

Este parámetro como se ha analizado previamente en secciones anteriores es el problema de muchos de los algoritmos existentes, pues sea que necesite el valor k como un parámetro inicial (k-means, k-medoids) o lo determine automáticamente (DBSCAN, AGNES, GNN) la técnica implementada, el resultado final debido a la repercusión directa de otros valores iniciales no siempre es el ideal. De ahí que, sea necesario este tipo de indicadores para calcular el número óptimo de *clústeres* en base a la estructura de datos a analizar [45].

La evaluación de la calidad de los agrupamientos se realiza utilizando la distancia interna e interdistancia que existen entre los objetos que son parte de una estructura de datos, entre más alto es el valor de este indicador mayor es la probabilidad de que ese sea el número ideal de *clústeres* que se deben ingresar cómo parámetro inicial en alguna técnica de agrupamiento que lo requiera o esperar en los resultados presentados por los algoritmos que lo calculan automáticamente [45], [46].

El coeficiente de silueta para un objeto i se define tal como se muestra en la ecuación 3.1, donde a es el promedio de disimilitud o distancias del objeto i con respecto a los demás objetos de la estructura de datos que pertenecen al *clúster* i; mientras que, b es la distancia mínima del objeto i con respecto a su *clúster* más cercano que es diferente al *clúster* que pertenece la observación [46], [47].

$$s(i) = \frac{b - a}{máx(a, b)} \tag{3.1}$$

Los valores que se pueden obtener de este índice para los objetos evaluados están comprendidos entre -1 y 1, de lo cual se puede inferir lo siguiente respecto a la asignación de los datos [47]:

- Si $s(i) \approx 1$, la observación i ha sido bien asignada al *clúster* en el que se encuentra.
- Si $s(i) \approx 0$, la observación i se encuentra ubicada entre dos agrupamientos y no solo entre uno.
- Si $s(i) \approx -1$, la observación i no ha sido bien asignada al *clúster* en el que se encuentra.

La representación gráfica de este método se puede observar en la Figura 3.1, en la cual se puede observar este índice cómo evalúa cada observación de una estructura de datos en base a su *clúster*. Para este caso en específico el ejemplo ha sido realizado con un k = 2, en donde se puede apreciar que uno de los objetos del *clúster* rojo tiene un valor menor a cero, lo cual puede indicar que el coeficiente inicial para los agrupamientos no se encuentra bien definido [45].

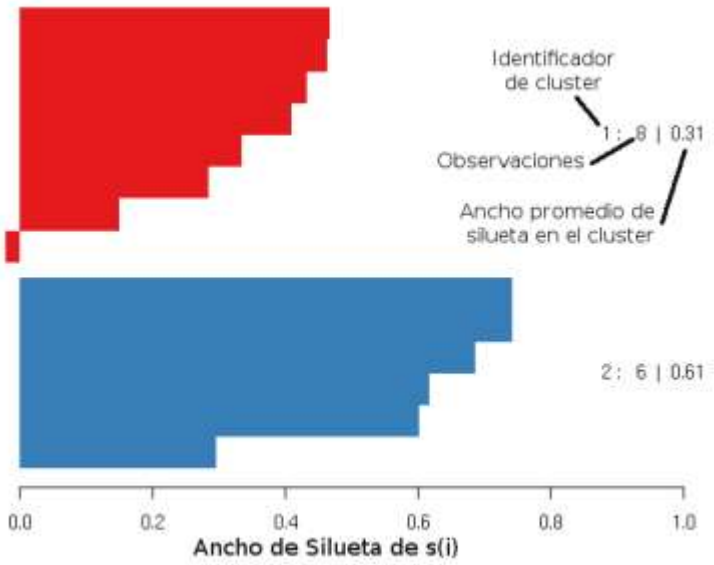

Figura 3.1 Aplicación del método de la silueta a un grupo de conglomerados con k = 2 [45].

3.2 COEFICIENTE DE SILUETA PROMEDIO

Tal como su nombre lo indica es un promedio de todas las expresiones individuales $s(i)$, el cual se define como el coeficiente de silueta de un objeto que pertenece a un *clúster* tal como se mostró en la sección 3.1. Al realizar esto se puede inferir si el conjunto de datos de un *clúster* está bien definido, dado a que este índice también se encuentra entre un valor de -1 y 1, indicando un valor en el extremo de -1 que el modelo obtenido por un algoritmo es pobre; mientras que, un valor de 1 un valor excelentes, es decir, el número de *clústeres* es el adecuado [48].

3.3 MÉTODO DEL DIAGRAMA DEL CODO

Es un algoritmo que permite la medición de la media de los grupos de tal modo que se permite determinar el número de grupos n eficientes para poder realizar la técnica de agrupamiento, esto lo logra mediante un proceso iterativo del algoritmo k-means [49].

La inercia de los objetos es el valor de error de suma cuadrada (SSE), la cual a su vez representa la suma de la distancia euclidiana media de cada punto de un k

clúster con respecto a su centroide, esto se repite iterativamente partiendo desde un k = 2 hasta llegar a un valor de k_n, en donde el valor de k variando en pasos de 1 en uno [50]. La ecuación que representa la inercia de los objetos del conjunto de datos está dada por la ecuación 3.2 [51].

$$Inercia = \sum_{i=0}^{k} ||x_i - \mu||^2 \tag{3.2}$$

Una vez que se aplica el proceso iterativo de método k-means este se puede representar de manera gráfica, representando el número de *clústeres* que deben implementarse como dato inicial del algoritmo de *clúster* a usar, en caso de necesitarlo. La figura 3.2 muestra un ejemplo del diagrama del codo, en donde se puede observar que tiene como eje x el número de *clúster* y como eje y la inercia de las distancias de los agrupamientos de cada *clúster*; el punto donde se produce un cambio brusco en la inercia de los objetos a medida que aumenta el parámetro k se conoce como "codo" y es justo para este número de agrupamientos que el modelo obtenido en los resultados es el más óptimo [51].

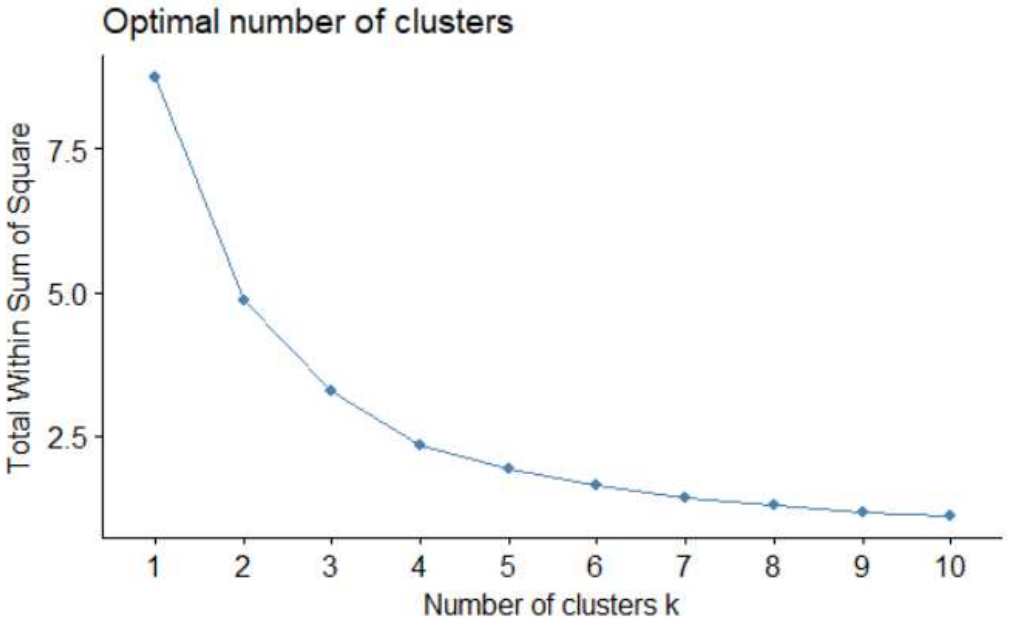

Figura 3.2 Aplicación del método del codo a una estructura de datos [51].

Como se puede observar en la figura 3.2 este método muestra una menor inercia o SSE al aumentar el número de *clústeres*, pues al aumentar este

parámetro la muestra será más refinada; por otro lado, cuando el valor de este es pequeño se obtiene un número elevado de SSE, de lo cual se puede inferir que los resultados obtenidos son pobres. En términos prácticos esto es lo que permite determinar el valor del codo, pues al ir aumentando k el SSE sufrirá un cambio brusco, ya que este irá disminuyendo, sin embargo, llegará a un valor en el cual los cambios sean tan pequeños que gráficamente el SSE se ha estabilizado, el resultado de esta gráfica en conjunto es lo que permite determinar el valor óptimo para k [52].

CAPÍTULO 4

4 METODOLOGÍA DE LOS AGRUPAMIENTOS DE PERFILES DE CARGA

El presenta capítulo detalla la metodología empleada de este trabajo de titulación para desarrollar y lograr los objetivos planteados. Dado a que el objetivo principal de este proyecto es la gestión de la demanda mediante la aplicación de técnicas de agrupamiento, se ha definido trabajar con cuatro algoritmos *clustering*, uno del tipo jerárquico, dos del tipo no jerárquico y uno basado en la densidad espacial, de modo que se pueda observar y analizar las bondades de cada tipo de métodos para la generación de las curvas estándar de demanda de los clientes de media tensión.

A lo largo de esta sección se aplica lo citado en la literatura de los Capítulos 2 y 3. En la sección 4.1 se detalla de manera general los algoritmos a utilizar, así como sus requerimientos para su aplicación y en la sección 4.2 la forma en la que se deben tratar los datos de los clientes de media tensión de para su correcta implementación. Por otro lado, las secciones 4.3 y 4.4 explica de manera general los parámetros relacionados a los algoritmos y cuáles deben modificarse para poder obtener un resultado óptimo. Finalmente, la sección 4.5 muestra los criterios utilizados para poder brindar una curva estándar por cada agrupación generada por los algoritmos *clúster*.

4.1 PERSPECTIVA GENERAL DE LA METODOLOGÍA

En esta metodología se presenta los métodos de agrupamiento implementados para la gestión de demanda de los clientes de media tensión, los cuales son los presentados en la figura 4.1.

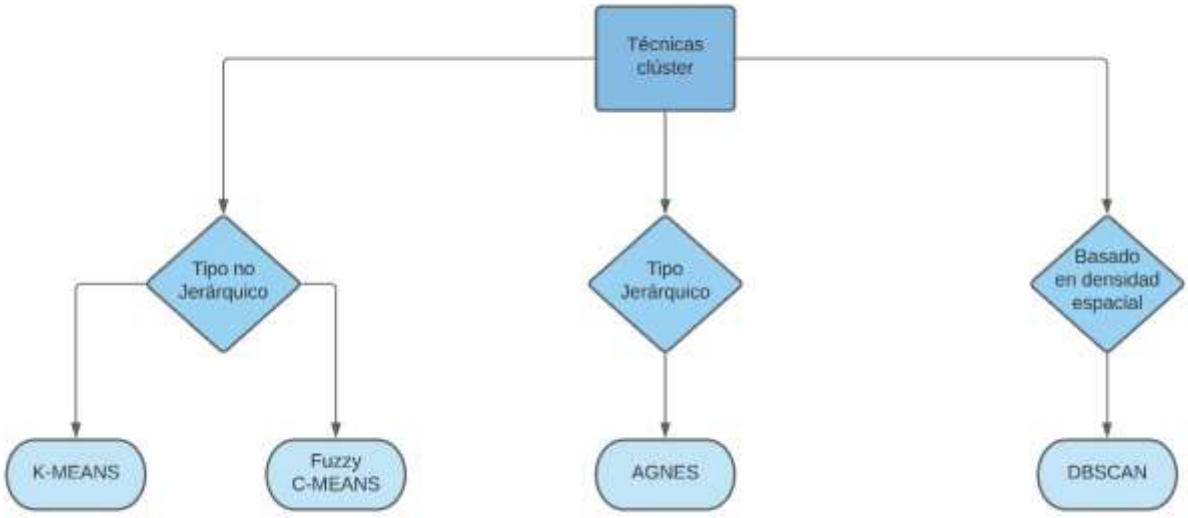

Figura 4.1 Clasificación de los algoritmos a implementar para el agrupamiento de datos de los clientes de media tensión.

Estos algoritmos han sido seleccionados debido a su aplicabilidad práctica en diferentes ámbitos de las ciencias, en especial con el objeto de comprobar la estructura de la base de datos de los clientes de media tensión se han implementado tres tipos de técnicas *clúster*: jerárquicas, no jerárquicas y basadas en densidad espacial.

Los requerimientos para ejecutar el agrupamiento de principalmente dos: 1) La base de datos de clientes de media tensión, y 2) La ejecución del algoritmo en un lenguaje de programación. En el caso del primer requerimiento, la base de datos es obtenida a través de un convenio de la ESPOL con la empresa pública distribuidora de el país, CNEL EP Unidad de Negocio Guayas- Los Ríos; mientras que, la herramienta computacional sobre la que se realizará la programación de los algoritmos es MATLAB®.

4.2 ESCALAMIENTO DE LAS CURVAS DE SALIDA DE CADA GRUPO FORMADO

Para la normalización de datos, se usó una técnica llamada normalización basada en la unidad o escalamiento de datos, que consiste en representar los datos en una escala entre valores de cero y uno, para una mejor comprensión del comportamiento de estos. Su aplicación se llevó acabo de la siguiendo los siguientes pasos para los grupos formados por los algoritmos, recordando que las filas representan a los medidores y las columnas al consumo registrado en un mes:

- Primero se determina el consumo promedio por mes de cada una de las observaciones pertenecientes a un mismo grupo, es decir se obtiene un promedio por columna.
- Se extrae el valor promedio máximo (X_{max}) y mínimo (X_{min}) del paso anterior.
- Para cada uno de los valores del primer paso (X), se aplica la siguiente formula:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

4.3 MANEJO DE LOS DATOS EN BRUTO DE LA DISTRIBUIDORA

Los datos entregados por la empresa distribuidora contienen información de ubicación, tarifa, registro mensual desde septiembre de 2019 hasta agosto de 2021 de consumo de energía, demanda mensual facturable y energía reactiva mensual de 4692 medidores de clientes de media tensión, entre otros datos. Como información útil se consideró solo a las columnas que tienen el registro de consumo de energía. La muestra inicialmente se componía de 4692 filas que representan a los medidores por 24 columnas que representan los meses de los 2 años de registro, la cual fue filtrada bajo los siguientes criterios para obtener un mejor desempeño de los modelos de agrupamiento aplicados:

- Se eliminó a usuarios (filas) que les faltara un registro de consumo, es decir que posean celdas vacías.
- Se eliminó a usuarios (filas) que tuvieran más de 12 registros con 0.
- Se eliminó a usuarios (filas) que presentaran errores considerables en su registro.

Luego de la etapa de filtrado, la muestra quedo reducida 4131 datos de medidores de usuarios, que representa el 83.25% de la muestra original, es decir, se conserva la suficiente información para que los modelos de agrupamiento puedan operar sin los inconvenientes provocados por las celdas vacías o también llamados datos *NaN* (*Not a Number*), o que la calidad de los resultados se vea comprometida por poseer valores atípicos o aberrantes.

Ya con los datos filtrados y para poder comparar el desempeño de cada algoritmo consigo mismo, se procedió a crear dos muestras que se llamaron “Datos No Escalados” y “Datos Escalados”, cada una con el registro de consumo de los 4131 medidores, pero con una escala diferente. La primera muestra son los datos sin modificar obtenidos después del filtrado, mientras que la segunda muestra se la creo aplicando la técnica que se explica en la sección 4.5 y cuya escala de valores esta entre cero y uno. Este paso se realizó para poder observar si existe diferencia en la creación de los grupos hechos por un mismo algoritmo y ver con que tipo de escala los algoritmos trabajan mejor.

4.4 ANÁLISIS DE SENSIBILIDAD BASADO EN LOS PARÁMETROS DE ENTRADA

Los algoritmos descritos en la figura 4.1 requieren de determinados parámetros iniciales para poder realizar la clasificación de los datos, los cuales se describen en la tabla 4.1.

Tabla 4.1 Algoritmos de agrupamiento y parámetros de entrada requeridos.

Algoritmo de agrupamiento	Parámetro de entrada	Descripción
$\boldsymbol{k-Means}$	k	Número de centroides para realizar las agrupaciones.
$\boldsymbol{Fuzzy\ C-Means}$	k	
$\boldsymbol{AGNES}$	-	No aplica.
$\boldsymbol{DBSCAN}$	ε	Radio de superficies que alberga las observaciones de cada clúster.
	$MinPtos$	Mínimo número de puntos presentes dentro de la superficie creada para el agrupamiento.

Los parámetros indicados en la tabla son los que deben modificarse en cada método para poder obtener un resultado significativo. Cabe recalcar que la modificación del parámetro inicial k para los algoritmos K-Means y Fuzzy C-Means tienen una repercusión directa en disminuir el error cuadrático de la suma de los k grupos, la cual es inversamente proporcional a la variación de este, es decir, mientras mayor sea el número de grupo, menor será el SSE; sin embargo, el aumento arbitrario de este parámetro reduce la confiabilidad del algoritmo, de ahí que se necesite el método del codo para poder tener un agrupamiento de datos óptimo.

Por otro lado, el algoritmo AGNES no necesita parámetros de entrada al ser del tipo jerárquico anglomerativo; mientras que, en el caso de DBSCAN el parámetro que tiene definirse en base a la experiencia del usuario es ε, pero una vez más esto causa que el modelo no sea confiable, por lo que la selección de un valor de referencia para poder validar los agrupamientos obtenidos se realiza mediante

una función de MATLAB® que permite obtener el épsilon óptimo, la cual es expuesta en la sección 4.4.

4.5 VALIDACIÓN DE GRUPOS

Los métodos de K-Means y Fuzzy C-Means son los únicos métodos que por la forma en la que funcionan es necesario especificar cuantos *clúster* o grupos se quiere crear. Debido a que este parámetro es especificado por el usuario de cierta forma se puede forzar a crear *clúster* donde no los hay. Debido a lo subjetivo que puede llegar a ser la elección de este parámetro, se utilizó el método del codo o *elbow method*, explicado en capítulos anteriores. La aplicación del método en Matlab y su consecuente resultado se apreciar en la siguiente tabla:

Tabla 4.2 Valores obtenidos de método del codo.

# Clúster	Valores de Criterio
1	NaN
2	24771
3	30351
4	30402
5	30268
6	29587
7	27505
8	25619
9	23597
10	22353

Los valores mostrados en la tabla 4.2, se obtuvieron usando la función *OptimalK* donde se le indico que pruebe hasta con diez grupos máximo, en ella se aprecia que el mayor valor de criterio se produce cuando se crean cuatro

clústeres, por lo que este valor se usara como parámetro de entrada en la implementación de los modelos anteriormente mencionados.

El algoritmo de AGNES no requiere que se especifique ningún parámetro de entrada en especial, más allá del método para determinar la distancia entre *clústeres,* que en este caso se usó “Ward” que la distancia al cuadrado inferior que a su vez es un algoritmo de mínima varianza.

Para el método DBSCAN se requiere especificar dos parámetros de entrada, la cantidad de puntos mínimos para crear un *clúster* y épsilon que es la distancia radial máxima que debe existir entre puntos de un agrupamiento. La cantidad de puntos mínimo se fijó en 4 para evitar crear muchos *clústeres* pequeños; mientras que, para determinar épsilon se usó una función *custerDBSCAN.estimateEpsilon* incorporada en MATLAB®. Esta función busca en un rango de valores el valor de épsilon óptimo para los datos de entrada y entrega como único resultado una gráfica donde se aprecia este valor épsilon como una línea roja y curvas de color azul que representan al rango de valores con los que probó para llegar a ese resultado, tal como se ve en las figuras 4.3 y 4.4.

Debido a que se usaron dos muestras con diferente escala, el valor de épsilon óptimo difiere entre ellas; para la muestra con datos escalados la épsilon óptima estimada es 0.159 y para los datos no escalados es 1436.4. En las figuras 4.3 y 4.4 podemos apreciar que se realizaron pruebas con 6 valores, y el valor épsilon óptimo estimado recae en el eje de las abscisas en el punto de inflexión aproximado donde las curvas de color azul presentan un cambio en su comportamiento.

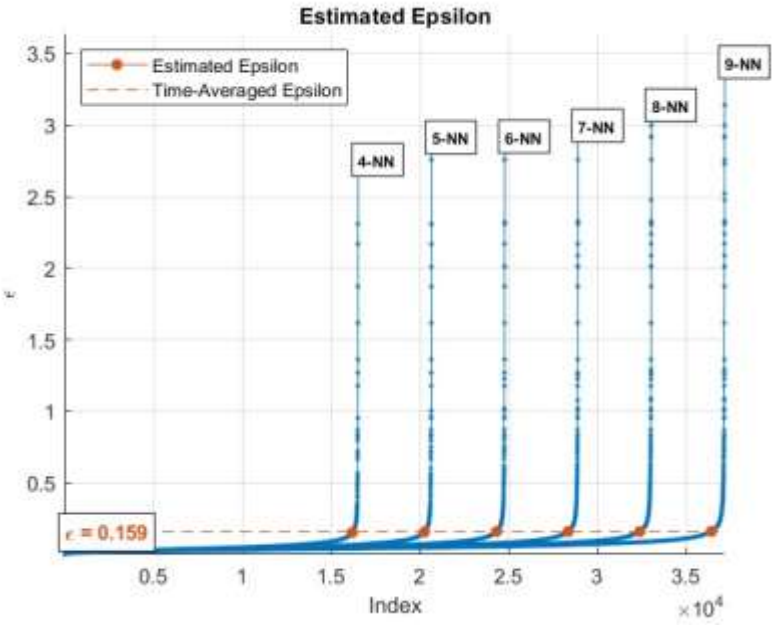

Figura 4.2 Épsilon óptimo para datos escalados

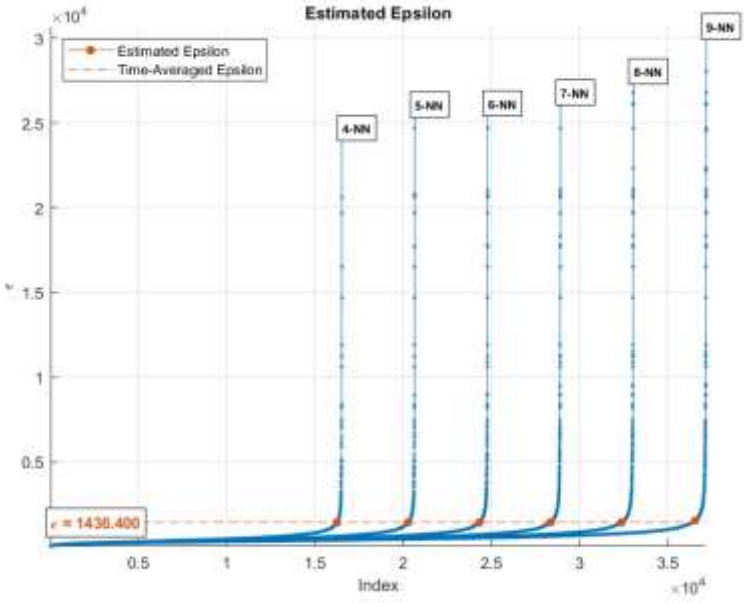

Figura 4.3 Épsilon óptimo para datos no escalados

CAPÍTULO 5

5 ANÁLISIS DE RESULTADOS

En la presente sección se detallan los resultados de los modelos de *clustering* que se plantearon en el capítulo anterior, empezando por la representación gráfica de los *clústeres* creados por cada modelo, hasta las curvas de consumo anual promedio de los grupos formados.

5.1 REPRESENTACIÓN GRÁFICA DE LOS CLÚSTER Y RANGOS

Para la representación gráfica del *clúster* en un plano XY o en 2D, se usó como eje de las abscisas a la media de consumo de cada observación, mientras que para el eje de las ordenadas se usó a la desviación estándar de consumo de la misma observación, teniendo un total de 4131 puntos a graficar que representan al total de observación introducidas a los modelos. Una vez ejecutado los algoritmos, se crea una gráfica que representa a los *clúster*es creados y una matriz cuyos datos fueron tabulados para una mejor comprensión que contiene información de cuantas observaciones fueron asignadas por grupo y el rango de valores entre los que se encuentra cada grupo.

5.1.1 K-MEANS

Para el algoritmo K-MEANS se realizaron pruebas con 4 *clúster*, siendo este el valor óptimo de grupos a crear según el método del codo (Tabla 4.2), y se aplicó este algoritmo a las dos muestras que tenemos, una con datos escalados entre cero y uno; y la otra con datos no escalados. En estas graficas los “*” de colores representan a las observaciones y las “x” son los centroides del *clúster*. A continuación, se muestran las gráficas y tablas correspondientes a la creación de 4 *clúster*.

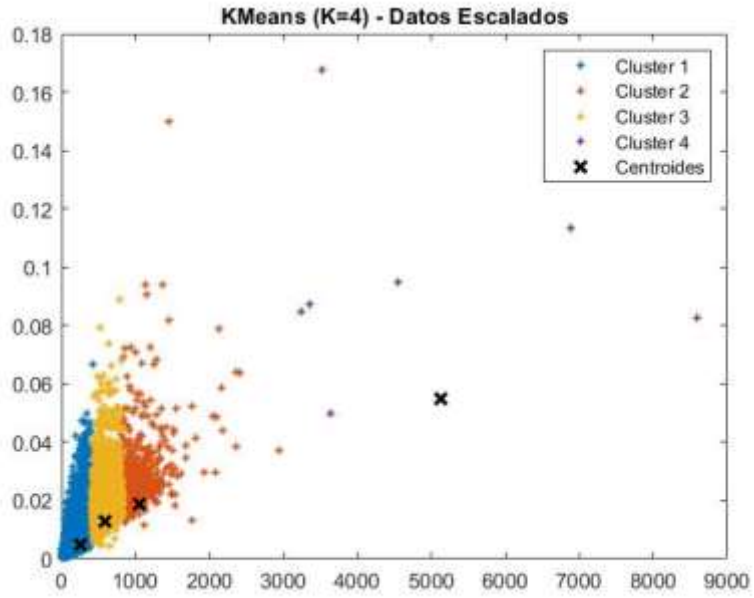

Figura 5.1 Gráfica de KMeans de 4 *clúster* con datos escalados

Tabla 5.1 *Rangos de clústeres creados por K-Means con datos escalados*

# Clúster	Cantidad de observaciones	Valor mínimo	Valor máximo
1	1915	0.79	438.75
2	516	516	2948.3
3	1692	412.25	836.5
4	8	3244	8598.6

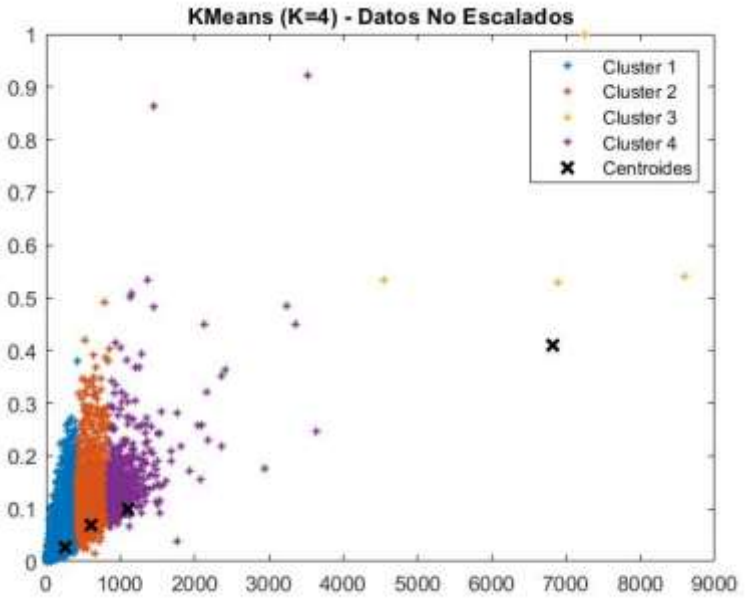

Figura 5.2 Gráfica de K-Means de 4 *clúster* con datos no escalados

Tabla 5.2 *Rangos de clústeres creados por K-Means con datos no escalados*

# Clúster	Cantidad de observaciones	Valor mínimo	Valor máximo
1	1963	0.79	453.46
2	1709	420.96	862.54
3	4	4552.6	8598.6
4	455	846.46	3636.7

Cabe resaltar que el método K-Means y posteriores métodos, el número de *clúster* es asignado de manera aleatoria cada vez que se ejecuta el código, por lo que no es factible hacer una comparación con base a este valor. En las tablas 5.1 y 5.2 que contienen la información de los *clúster*es creados con las dos muestras, apreciamos que algunas observaciones son reasignadas a otros grupos cuando se usan los datos no escalados, lo cual modifica los rangos de consumo de los grupos.

5.1.2 FUZZY C-MEANS

Para el algoritmo Fuzzy C-Means se especificó la creación de 4 *clúster* al igual que el modelo anterior, tanto para la muestra con datos escalados como para los no escalados. En estas graficas los puntos de colores representan a las observaciones y las “x” son los centros del *clúster*.

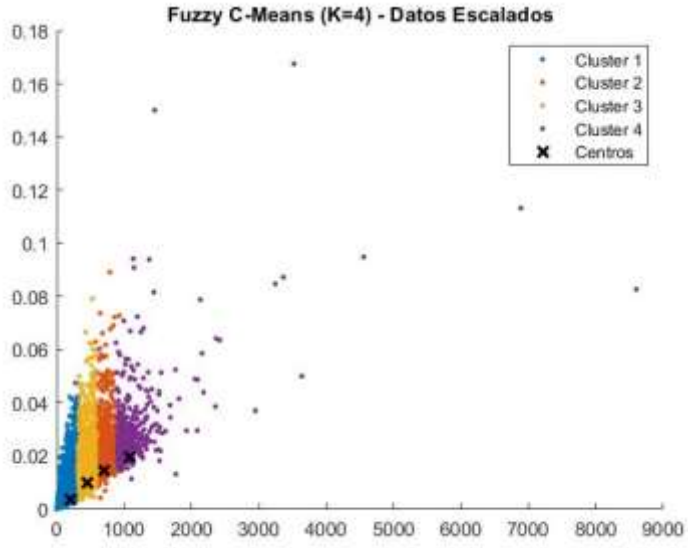

Figura 5.3 Gráfica de Fuzzy C-Means de 4 *clúster* con datos escalados

Tabla 5.3 *Rangos de clústeres creados por Fuzzy C-Means con datos escalados*

# Clúster	Cantidad de observaciones	Valor mínimo	Valor máximo
1	1423	311.58	602.08
2	375	890.54	8598.6
3	992	559.25	904.75
4	1341	0.791	349.54

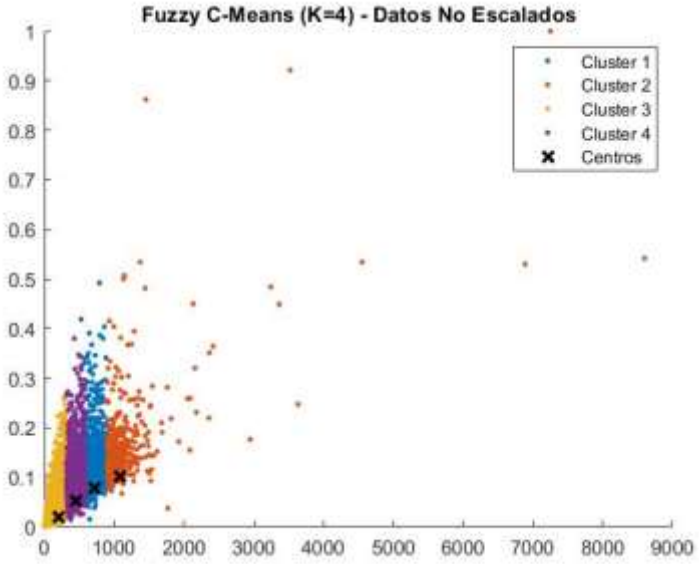

Figura 5.4 Gráfica de Fuzzy C-Means de 4 *clúster* con datos no escalados

Tabla 5.4 *Rangos de clústeres creados por Fuzzy C-Means con datos no escalados*

# Clúster	Cantidad de observaciones	Valor mínimo	Valor máximo
1	992	559.25	904.75
2	375	890.54	8598.6
3	1431	0.791	349.54
4	1423	311.58	602.08

Similar al método K-Means, cuando se usan datos no escalados cambia la cantidad de observaciones de cada grupo, y por consecuente el rango de cada *clúster* se ve modificado, ya que los *clústeres* que agregan más observaciones expanden su rango, mientras los que disminuyen observaciones su rango se reduce.

5.1.3 AGNES ALGORITHM (AGGLOMERATIVE NESTING)

Para el algoritmo de AGNES no se requiere especificar el número de *clúster* a crear debido a la forma en la que funciona. En su lugar se obtienen dos gráficas, la primera representa a las observaciones agrupadas en *clústeres* de diferentes colores; mientras que, la segunda gráfica es un dendograma que muestra de manera visual como se realizan la agrupación. A continuación, se muestran los resultados para las dos muestras que se tiene:

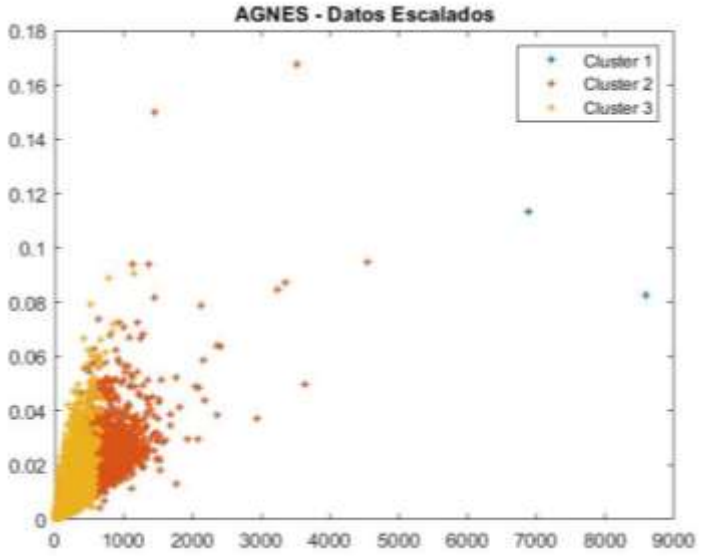

Figura 5.5 Gráfica de AGNES con datos escalados

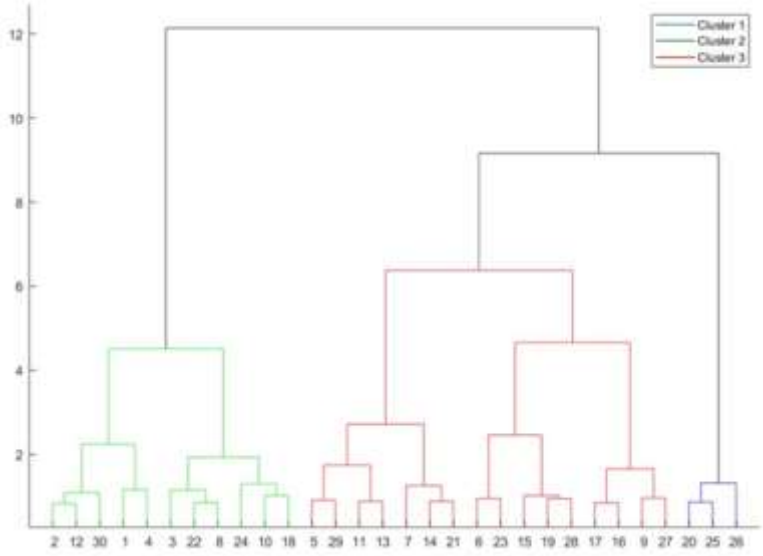

Figura 5.6 Dendograma AGNES con datos escalados

Tabla 5.5 *Rangos de clústeres creados por AGNES con datos escalados*

# *Clúster*	Cantidad de observaciones	Valor mínimo	Valor máximo
1	3	6887.3	8598.6
2	1578	386.08	4552.6
3	2550	0.79	1148.5

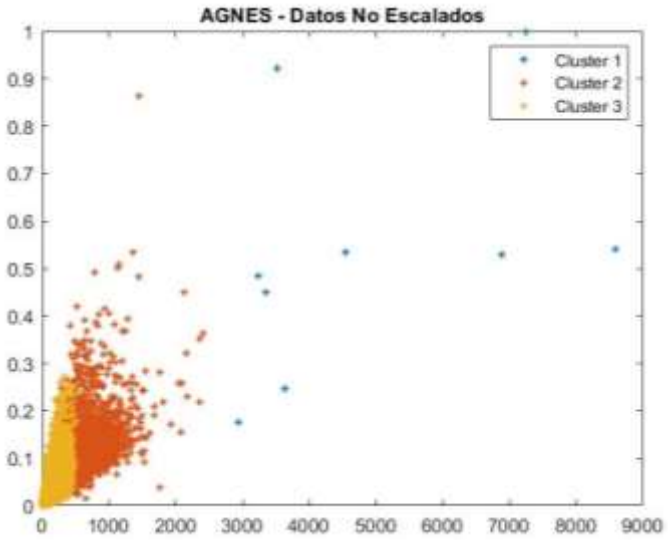

Figura 5.7 Gráfica de AGNES con datos no escalados

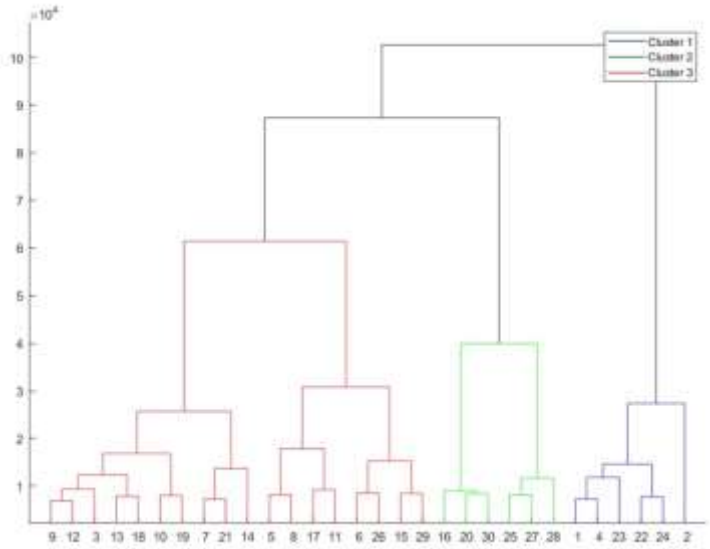

Figura 5.8 Dendograma AGNES con datos no escalados

Tabla 5.6 *Rangos de clústeres creados por AGNES con datos no escalados*

# Clúster	Cantidad de observaciones	Valor mínimo	Valor máximo
1	9	2948.3	8598.6
2	1578	259.21	8598.6
3	2550	0.79	538.63

Comparando los *clúster*es creados por el método AGNES que se ven las tablas 5.5 y 5.6 al igual que en los métodos anteriores, podemos apreciar que, al usar los datos no escalados, los rangos de los grupos cambian ya que algunas observaciones son reasignadas a otros *clústeres*, por ejemplo el *clúster* 1 pasa de tener tres observaciones a tener nueve, y su rango se extiende, pasando de ser [6887.3 – 8598.6] a [2948.3 – 8598.6].

5.1.4 DBSCAN

Finalmente, de manera similar al algoritmo AGNES, DBSCAN no se requiere especificar el número de *clúster* a crear debido a la forma en la que funciona. De este método solo se obtiene la gráfica que representa a los *clústeres* creados y su correspondiente tabla con la información que compone cada agrupación. a diferencia de los métodos anteriores, el *clúster* "-1" corresponde a datos clasificados como ruido por el algoritmo, por lo que no se lo tendrá en cuenta para posteriores análisis. Para los datos normalizados, los resultados son los siguientes:

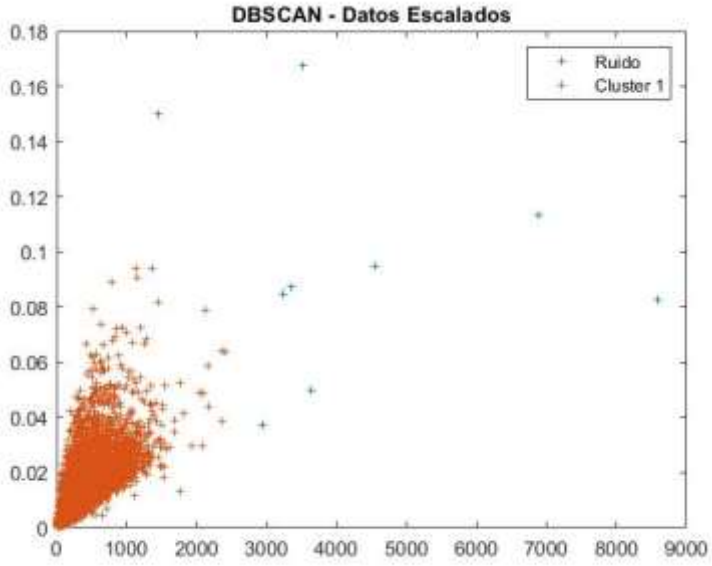

Figura 5.9 Gráfica de DBSCAN con datos escalados

Tabla 5.7 *Rangos de clústeres creados por AGNES con datos no escalados*

# *Clúster*	Cantidad de observaciones	Valor mínimo	Valor máximo
-1	10	1453.8	8598.6
1	4121	0.79	2419.9

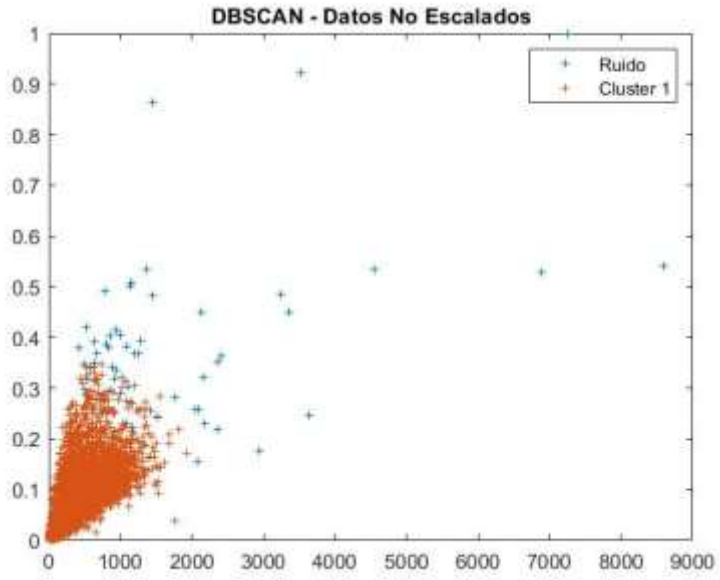

Figura 5.10: Grafica de DBSCAN con datos no escalados

***Tabla* 5.8** *Rangos de clústeres creados por AGNES con datos no escalados*

# Clúster	Cantidad de observaciones	Valor mínimo	Valor máximo
-1	58	429.63	8598.6
1	4073	0.79	1927.6

Como se puede apreciar tanto en las figuras 5.16 y 5.17, existen datos agrupados como ruido o cuyo número de *clúster* es “-1” según las tablas 5.7 y 5.8, estas observaciones son clasificadas así por el algoritmo ya sea porque en su radio de longitud épsilon no se encuentran más observaciones dentro, es decir están aislados; o porque a pesar de existir observaciones dentro de su radio de longitud épsilon, no son el mínimo de cuatro observaciones que se requieren para crear un *clúster*. Además, podemos apreciar que solo crea un grupo, debido a que las observaciones están tan cerca unas de las otras y concentradas en una sola región del plano XY, que las interpreta como un único grupo.

Comparando las tablas 5.7 y 5.8, que se obtiene de los datos escalados y los datos no escalados respectivamente, podemos apreciar que existe en la diferencia en el grupo creado, ya que con los datos no escalados 48 observaciones pasan a ser consideradas ruido, por lo que el rango del *clúster* 1 se reduce.

5.2 CURVAS DE CONSUMO PROMEDIO

A partir de los *clústeres* creados por los 4 modelos implementados, se puede trazar una curva de consumo anual promedio de todas las observaciones contenidas de en un mismo grupo por modelo cuando al modelo se le entregan datos normalizados y datos reales.

5.2.1 K-MEANS

A continuación, se muestras las curvas de consumo anual promedio correspondientes a los grupos creados modelo K-Means cuando se le entregan datos escalados.

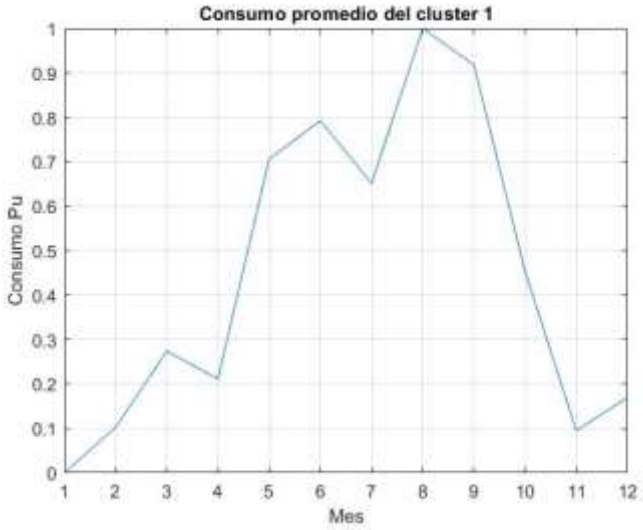

Figura 5.11 Curva de consumo promedio anual de *clúster* 1 creado por KMeans con datos escalados

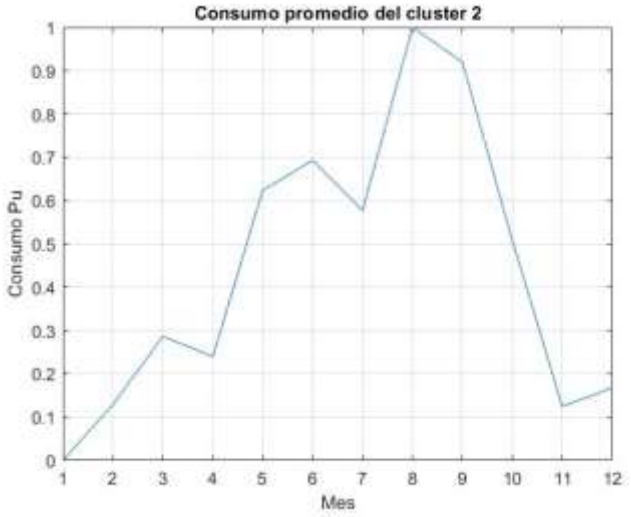

Figura 5.12 Curva de consumo promedio anual de *clúster* 2 creado por KMeans con datos escalados

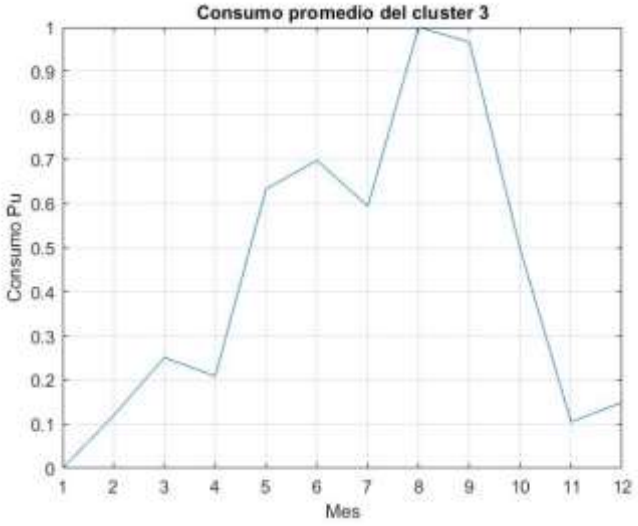

Figura 5.13 Curva de consumo promedio anual de *clúster* 3 creado por KMeans con datos escalados

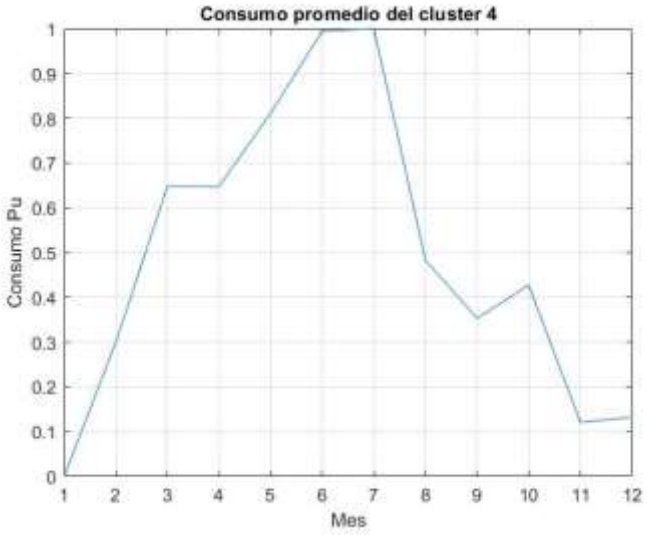

Figura 5.14 Curva de consumo promedio anual de *clúster* 4 creado por k-means con datos escalados

A continuación, se muestras las curvas correspondientes al modelo cuando se le entregan datos no escalados.

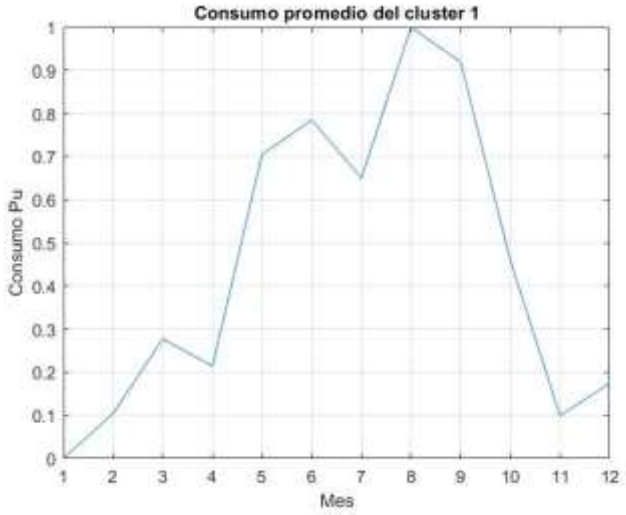

Figura 5.15 Curva de consumo promedio anual de *clúster* 1 creado por KMeans con datos no escalados

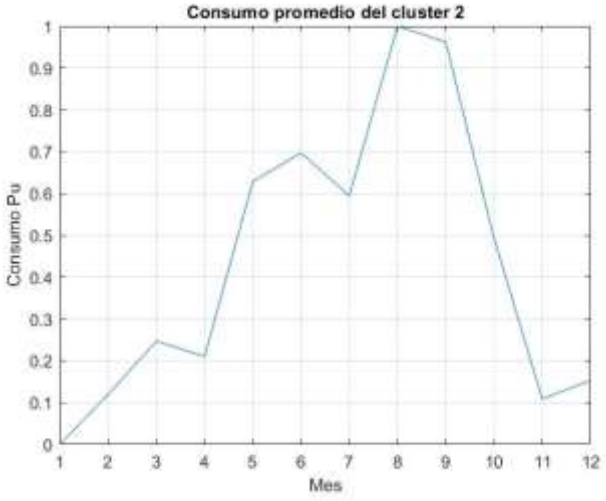

Figura 5.16 Curva de consumo promedio anual de clúster 2 creado por KMeans con datos no escalados

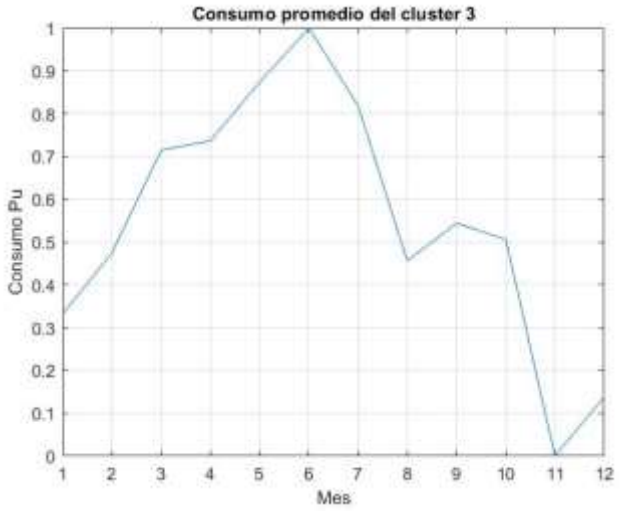

Figura 5.17 Curva de consumo promedio anual de clúster 3 creado por KMeans con datos no escalados

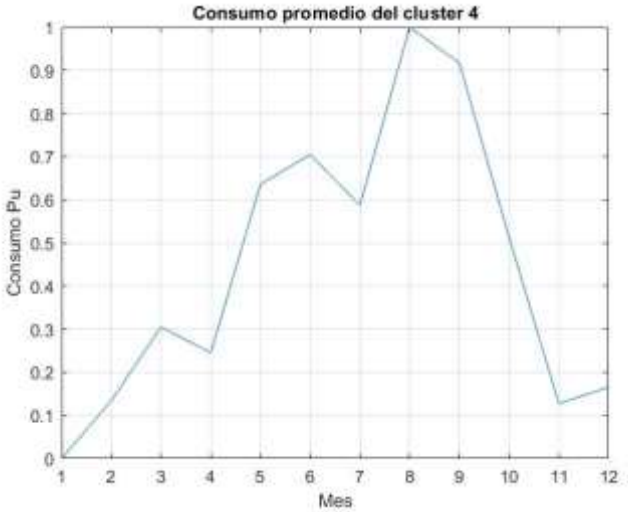

Figura 5.18 Curva de consumo promedio anual de clúster 4 creado por KMeans con datos no escalados

5.2.2 FUZZY C-MEANS

De manera similar al caso anterior, se muestran las curvas de consumo anual promedio por *clúster* creados por el modelo Fuzzy C-Means cuando se entregan datos escalados.

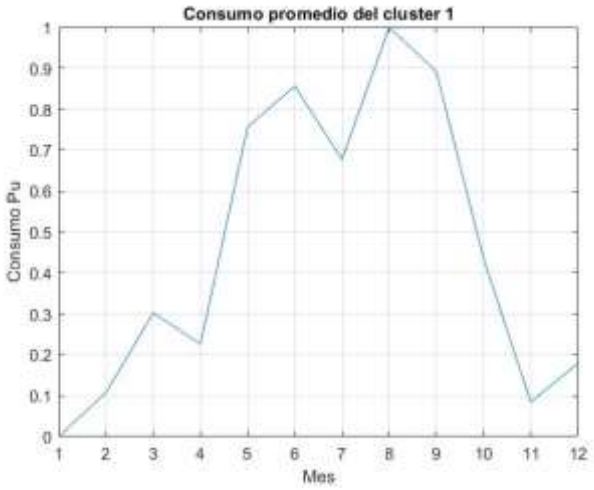

Figura 5.19 Curva de consumo promedio anual de *clúster* 1 creado por Fuzzy C-Means con datos escalados

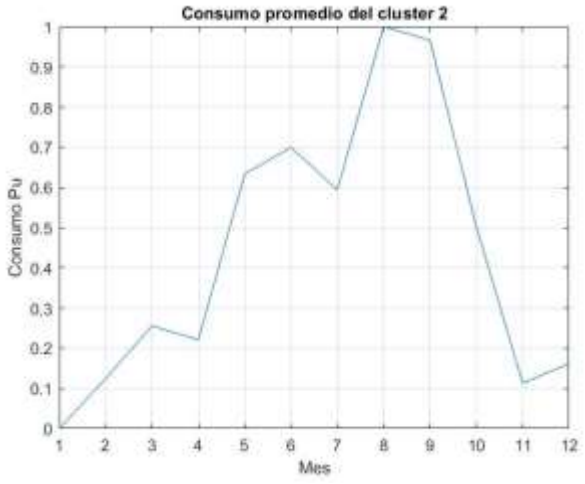

Figura 5.20 Curva de consumo promedio anual de *clúster* 2 creado por Fuzzy C-Means con datos escalados

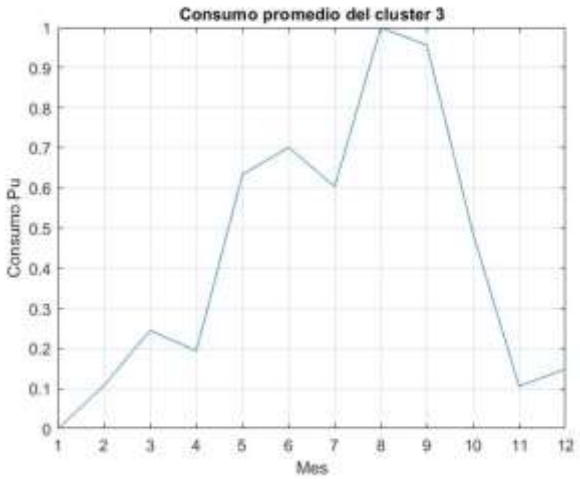

Figura 5.21 Curva de consumo promedio anual de *clúster* 3 creado por Fuzzy C-Means con datos escalados

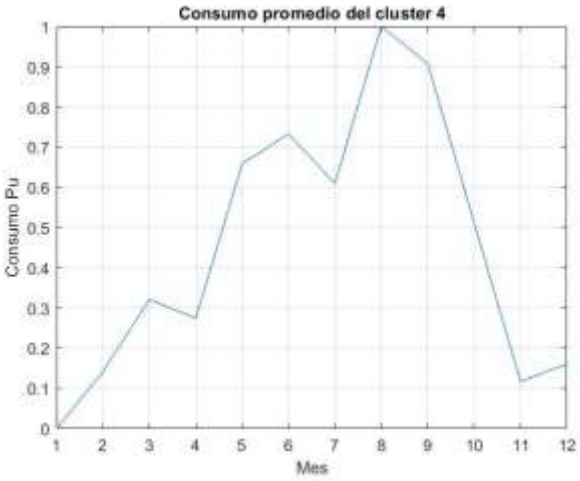

Figura 5.22 Curva de consumo promedio anual de *clúster* 4 creado por Fuzzy C-Means con datos escalados

A continuación, se muestran las curvas de consumo anual promedio por *clúster* creados por el modelo Fuzzy C-Means cuando se entregan datos no escalados.

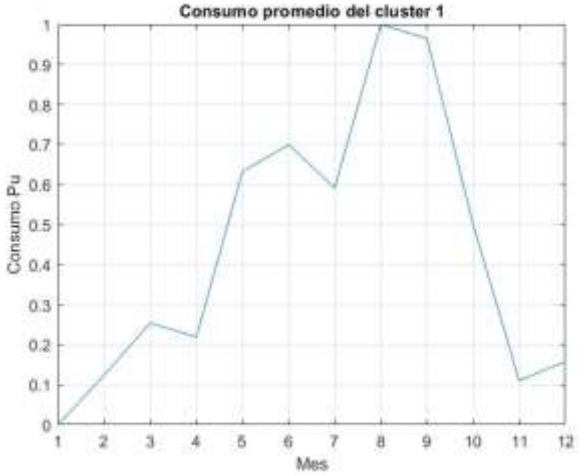

Figura 5.23 Curva de consumo promedio anual de clúster 1 creado por Fuzzy C-Means con datos no escalados

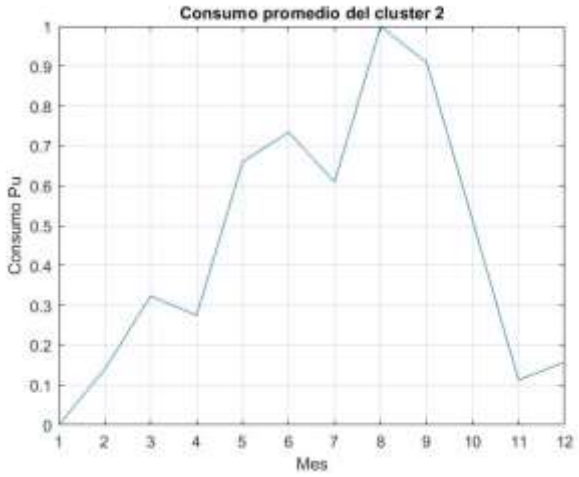

Figura 5.24 Curva de consumo promedio anual de clúster 2 creado por Fuzzy C-Means con datos no escalados

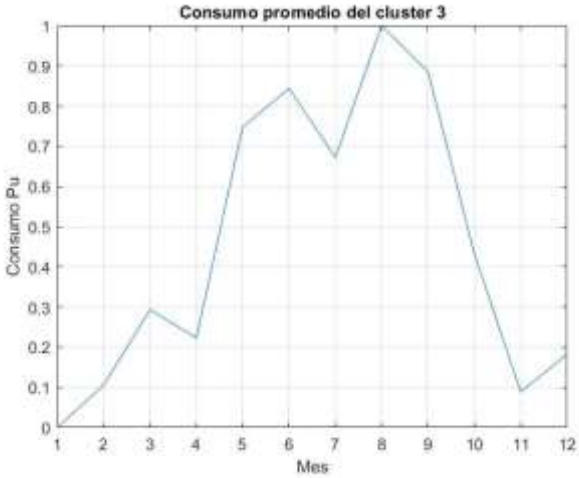

Figura 5.25 Curva de consumo promedio anual de clúster 3 creado por Fuzzy C-Means con datos no escalados

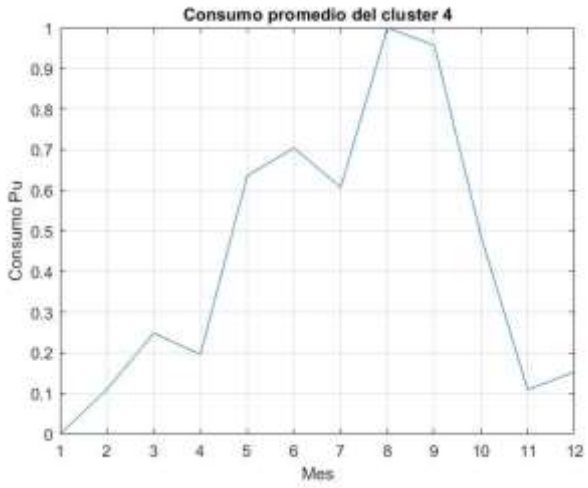

Figura 5.26 Curva de consumo promedio anual de clúster 4 creado por Fuzzy C-Means con datos no escalados

5.2.3 AGNES ALGORITHM (AGGLOMERATIVE NESTING)

A partir de la agrupación hecha por el algoritmo de AGNES, se procede a crea una curva de consumo anual promedio de cada *clúster* creado cuando se le introducen datos escalados al modelo.

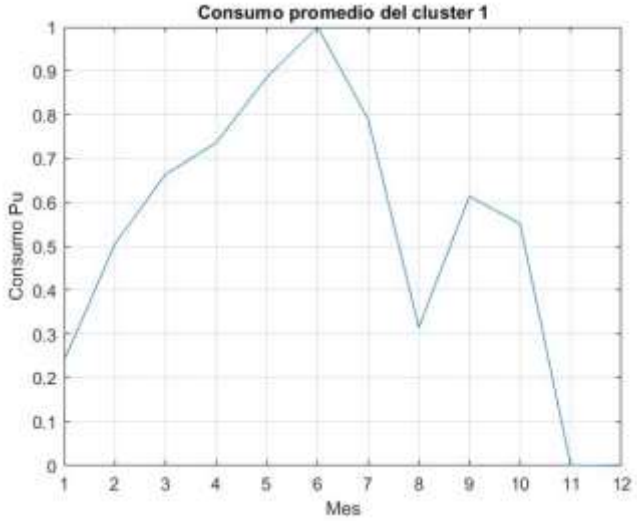

Figura 5.27 Curva de consumo promedio anual de clúster 1 creado por AGNES con datos escalados

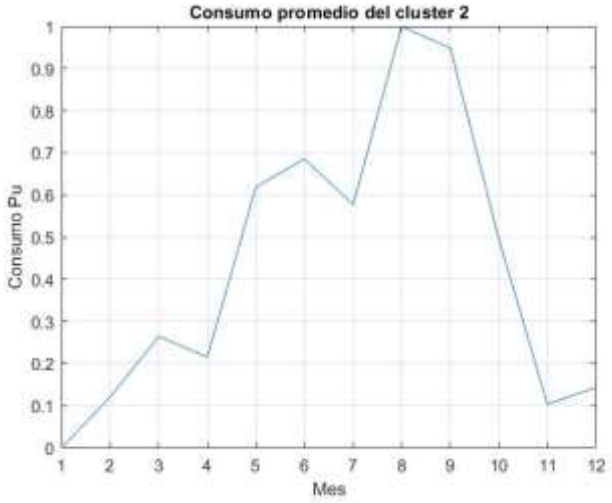

Figura 5.28 Curva de consumo promedio anual de clúster 2 creado por AGNES con datos escalados

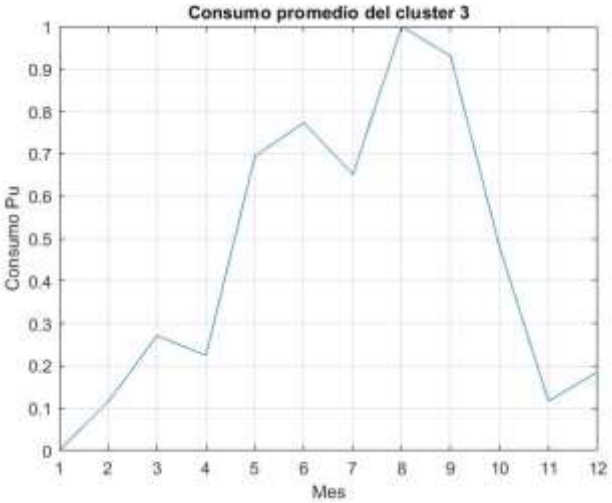

Figura 5.29 Curva de consumo promedio anual de clúster 3 creado por AGNES con datos escalados

De forma similar, se muestras las curvas de consumo anual promedio por *clúster*, pero cuando se le entregan datos no escalados al modelo.

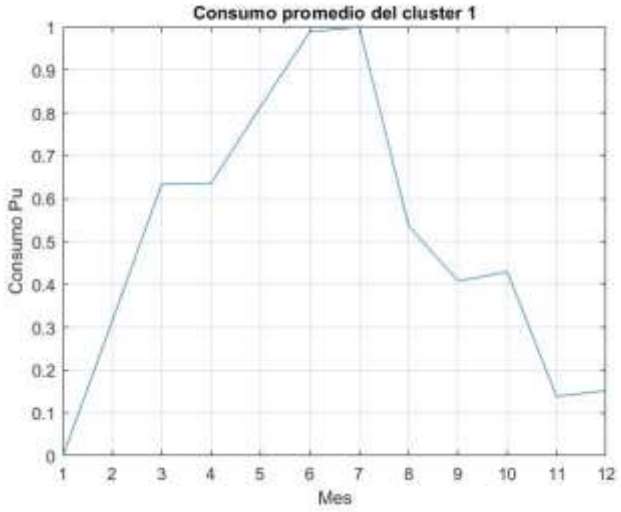

Figura 5.30 Curva de consumo promedio anual de clúster 1 creado por AGNES con datos no escalados

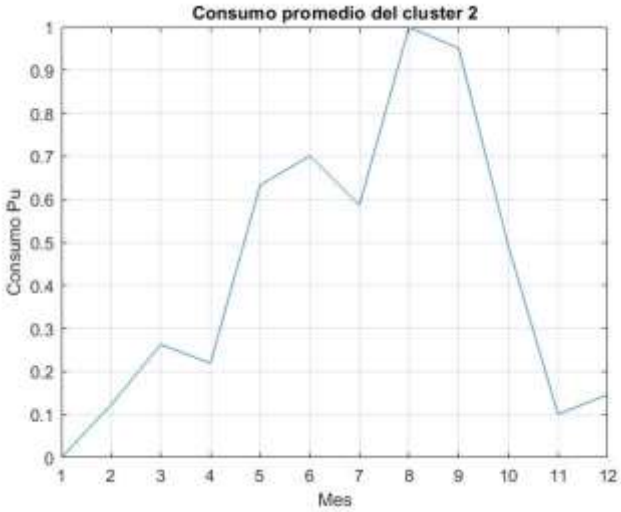

Figura 5.31 Curva de consumo promedio anual de clúster 2 creado por AGNES con datos no escalados

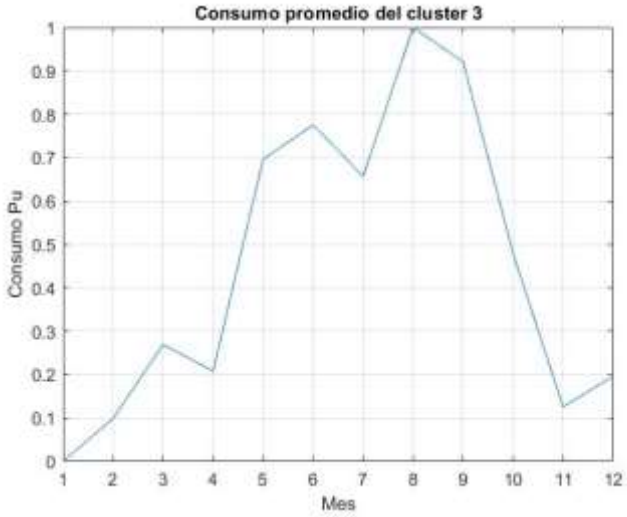

Figura 5.32 Curva de consumo promedio anual de clúster 3 creado por AGNES con datos no escalados

5.2.4 DBSCAN

En el caso de este algoritmo solo es de interés hacer la curva promedio del *clúster* que no es clasificados como ruido, que para el caso de cuando se introducen datos escalados es la siguiente gráfica.

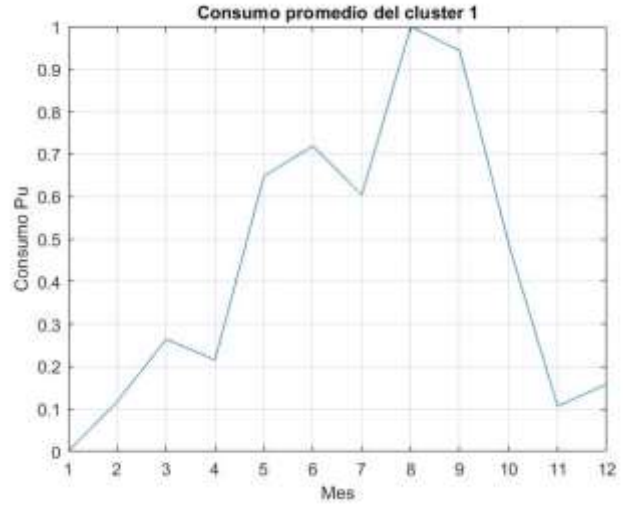

Figura 5.33 Curva de consumo promedio anual de clúster 1 creado por DBSCAN con datos escalados

Para el caso de cuando se introducen datos no escalados se tiene la siguiente gráfica

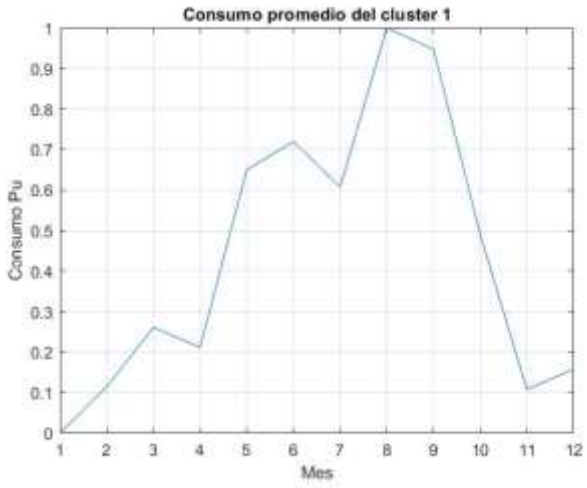

Figura 5.34 Curva de consumo promedio anual de clúster 1 creado por DBSCAN con datos no escalados

5.3 DISCUSIÓN DE RESULTADOS

Una de las comparaciones que resalta es la cantidad de *clúster* que crea cada modelo, en el caso de K-Means y Fuzzy C-Means se especifica cuantos *clústeres* se desea crear que se especificó en cuatro por el método del codo, mientras que el algoritmo AGNES y DBSCAN crearon tres y uno respectivamente por su cuenta.

Entre los resultados de un mismo modelo ejecutado con datos escalados y con datos no escalados, como se explicó por cada método explicó la sección 5.2, existen diferencias debido a que se redistribuyen algunas observaciones entre los *clústeres* lo cual provoca una variación en los limites superior e inferior de los grupos, pero esta diferencia tiene un impacto significativo en la forma de onda del consumo promedio anual. Otra diferencia entre ellos es el tiempo de ejecución, siendo que para datos escalados los resultados se obtenían más rápido que con los datos no escalados.

Entre los modelos K-Means y Fuzzy C-Means, a pesar de crear la misma cantidad de *clústeres*, la distribución de las observaciones dentro de los grupos es diferente y en consecuencia el rango en el que están las mediciones difiere también. Sin embargo, la forma de onda de tres de los cuatro *clústeres* es similar, tanto si la comparación se hace cuando se usan datos escalados, como cuando se usan datos no escalado.

A pesar de que numero de *clúster* creados es distinto, las curvas de consumo promedio creadas con los algoritmos K-Means y el algoritmo AGNES presentan cierta similitud en la forma de onda tanto si la compara cuando se usa la muestra con datos escalados como con los no escalados. Tomando como base los resultados evaluados con datos escalados, las Figuras 5.11, 5.12 y 5.13 provenientes de K-Means presentan similitud en forma y tendencia con las figuras 5.28 y 5.29 de AGNES, mientras que las figura 5.14 y 5.27 presentan menor similitud en la forma, pero una tendencia parecida. Estas diferencia se debe a la cantidad de observaciones cada clúster, en la segunda comparación (Figura 5.14 y 5.27) los clústeres están formados por 8 y 3 observaciones respectivamente, mientras que en la primera comparación, cada clúster está conformado por más de 1000 observaciones.

El algoritmo AGNES es el más robusto de los modelos implementados, ya que sus resultados en la creación de grupos no varían en comparación a los otros modelos que cada vez que se vuelven a ejecutar, tienden a reasignar algunas observaciones a otros grupos, ya que no depende de una componente aleatoria como si K-Means y Fuzzy C-mean.

A diferencia de los otros métodos, DBSCAN solo es capaz de encontrar un *clúster*, esto se debe a que la distribución de las observaciones está concentrada en una sola región, por lo que el algoritmo no es capaz de encontrar algún otro *clúster* cuya densidad de puntos cumpla con los parámetros de entrada establecidos y finalmente, clasifica a las observaciones restantes como ruido.

Entre los cuatro modelos evaluados, la mejor opción a implementar es AGNES, teniendo en cuenta que se requiere un análisis previo con del dendograma para poder determinar cuales los *clústeres* más grandes formados.

A partir de los resultados del modelo AGNES con datos escalados y teniendo en cuenta que la información es tomada a partir del mes de septiembre, se puede obtener la siguiente información:

- Dos de los tres *clústeres* creados presentan un incremento progresivo de consumo desde el mes de septiembre hasta mayo, teniendo una reducción

de consumo en los meses de enero y marzo. A partir de septiembre presentan una reducción en su consumo hasta noviembre. (figuras 5.28 y 5.29)

- El *clúster* restante presenta incremento desde septiembre hasta marzo, De ahí el reduce su consumo hasta mayo, presenta un incremento nuevamente en junio y decae lo que queda del año. (figura 5.27).
- Debido a que las curvas representan el promedio de mediciones de consumo de dos años atípicos por la pandemia, el comportamiento de las curvas creadas presenta picos y es diferentes a las de años anteriores.
- La tendencia creciente entre los meses de septiembre y abril de las tres curvas puede deberse a factores como:
    - El incremento de las temperaturas por el cambio de estación, que teniendo en cuenta que la zona de donde se tomó los datos pertenece a un sector con un ingreso per cápita medio-alto, se podría relacionar a un aumento en el uso de Aires Acondicionados.
    - Mayor cantidad de personas en casa debido a las medidas de confinamiento tomadas por la pandemia del COVID 19, conlleva a un aumento en el consumo general de un hogar por el uso de más electrodomésticos y dispositivos electrónicos.
- Por otro lado, la reducción en el consumo a partir de los meses de mayo y marzo puede deberse al cambio a la estación fría, que reduciría el uso de aires acondicionados, y a la relajación de las medidas de confinamiento por la pandemia, que permitirá que las personas estar por más tiempo fuera de sus hogares.

De los métodos aplicados, cada uno con sus ventajas y desventajas, se pueden hacer agrupaciones de usuarios con base no solo a su consumo mensual, sino a otras variables numéricas, como si poseen o no deudas, el valor de estas, tarifa, etc. Lo que puede permitir una mejor proyección de la demanda basado en grupos de usuarios, picos máximos y mínimos de consumo, fijación de tarifas, entre otros en cuestión de minutos y solo con poseer la data necesaria, algo que con métodos convencionales, podría ser un trabajo de muchas horas y personas, sujeto a la subjetividad de quien o quienes realicen el trabajo y a errores humanos.

CAPÍTULO 6

6 CONCLUSIONES Y RECOMENDACIONES

6.1 CONCLUSIONES

De los métodos de agrupación implementados, el algoritmo AGNES y DBSCAN son modelos más robustos ya que sus resultados en la conformación de grupos, en los rangos de estos y las forma de la curva estándar no varían cuando se vuelve a ejecutar el programa, mientras que Fuzzy C-Means y K-Means tienden a variar un poco en la conformación de los *clústeres* y por ende cambian ligeramente los rangos de los mismo, pero sin afectar visiblemente a la curva estándar a partir de los *clústeres*.

El análisis de los grupos de media tensión mediante los diferentes algoritmos de agrupamiento demostraron que el número de *clústeres* óptimos para los datos brindados por CNEL EP puede ser 3 o 4, esto se considera evaluando los algoritmos K-Means, Fuzzy C-Means y AGNES. Este resultado es útil para la planeación y operación del sistema de distribución, pues al clasificar los 4.131 clientes validados en 3 o 4 grandes usuarios, se pueden definir tarifas para la venta de energía mediante la curva estándar de cada uno, establecer políticas de uso de la energía, así como predecir el consumo a corto, mediano y largo plazo, resultando en una información útil para la distribuidora, ya que, les puede permitir establecer un precio por temporada en base a los datos que se proporcionaron (mensuales).

Respecto a los tres grandes grupos construíos a partir del algoritmo AGNES y considerando que los datos se tomaron a partir del mes de septiembre, se tienen las siguientes características: Dos grupos tienen su pico máximo de consumo en el mes 8 (mayo), teniendo esta característica como similitud al igual que todo el perfil de consumo, con la excepción del mes 6 (marzo) en donde un grupo de consumidores tiene un pico de consumo del 0.7 p.u. y otro el de 0.8. p.u., sin embargo, a pesar de que el perfil de consumo es parecido, no pueden llegar a ser considerados o confundidos como un solo grupo y deben ser considerados por separado para cualquier tipo de proyección o estudio que se haga, ya que la escala real de consumo de cada grupo es totalmente diferente, así como la

cantidad de usuarios. El grupo restante de análisis tiene un perfil de consumo distinto a los anteriores, su pico máximo de consumo en el mes 6 (febrero), seguido otro incremento producido durante el mes 9 (junio), estas diferencias se deben a que es el grupo con mayor escala de consumo y menor cantidad de clientes, pudiéndose tratar de algún tipo de cliente especial cuya actividad se concentra más en un determinado periodo del año, en este caso entre los meses de septiembre y febrero, a diferencia de los otros grupos que su consumo se distribuye durante todo el año. Estas características de los grupos energéticos formados permiten establecer pliegos tarifarios por temporadas de manera anual, más no para definir si los grupos de comportamiento pertenecen a un cliente industrial o residencial, dado a que estas características de curvas típicas suelen estar medidas alrededor de un día y no durante los meses del año.

Se aprecia una mejora en el rendimiento de los modelos cuando estos trabajan con datos cuyos valores se encuentran entre cero y uno, ya que el número de iteraciones requeridas para que converjan es menor que cuando se usan Los datos originales, lo que se puede llegar a traducir en un menor tiempo de ejecución si la cantidad de observaciones se llegase a incrementar.

De la revisión de literatura se pudo seleccionar diferentes tipos de algoritmos para el agrupamiento de datos, así como una metodología que permita justificar tanto la selección de los valores iniciales en cada algoritmo, así como la validación de los agrupamientos obtenidos por los diferentes métodos.

6.2 RECOMENDACIONES

Realizar un tratamiento de datos previo a la ejecución de los modelos, ya que la muestra entregada a los modelos debe además de ser representativa debe estar completa, es decir que no existan registros vacíos, ya que estos ocasionan errores de ejecución en los modelos si ha estos no se les indica como tratar con estos datos faltantes.

En caso de utilizar datos que contengan más variables o columnas, se recomienda implementar algún tipo de metodología para la reducción de

dimensionalidad antes de ingresar estos datos, ya que algunos de los métodos de *clustering* o agrupamiento sufren de algo llamado la maldición de la dimensionalidad, lo cual limitaría su aplicación por su elevado costo computacional, se puede hacer un Principal Component Analysis.

Este estudio sirve de base para realizar un análisis más complejo como lo es la predicción de carga, ya que, al observar los diferentes grupos de comportamiento energético las características de cada uno de estos se pueden calcular a partir de su histórico o de su curva estándar para los intervalos de tiempo que se han trabajado, en este caso, dicha predicción de carga se puede realizar de forma anual por meses, pero con los datos adecuados esta podría calcularse en intervalos más pequeños de hasta 15 minutos, los cuales pueden ser almacenados en un servidor para grabar la información y evitar su perdida.

Otra aplicabilidad práctica que puede tener este estudio es la selección de posibles clientes elegibles para poder aplicar el programa de Respuesta a la Demanda, de tal modo que se puede crear una mayor estabilidad en la red de la concesionaria distribuidora al tener posibles clientes comerciales e industriales a los que se les puede aumentar o disminuir el consumo en base a las necesidades de la red.

ANEXOS

ANEXO A. CURVAS ESTÁNDAR BASADAS EN EL MÉTODO DE MEJOR RENDIMIENTO

Tabla A.1 Curvas de demanda en p.u. del método de agrupamiento con mejor rendimiento: AGNES.

Mes	Grupo 1 [p.u.]	Grupo 2 [p.u.]	Grupo 3 [p.u.]
1	0	0	0
2	0.31	0.12	0.09
3	0.63	0.26	0.27
4	0.64	0.21	0.21
5	0.81	0.63	0.70
6	0.98	0.70	0.78
7	1	0.59	0.66
8	0.53	1	1
9	0.41	0.95	0.92
10	0.42	0.49	0.47
11	0.13	0.10	0.13
12	0.15	0.15	0.19

REFERENCIAS

[1] A. Bosisio *et al.*, “A GIS-based approach for high-level distribution networks expansion planning in normal and contingency operation considering reliability”, *Science Direct*, vol. 190, núm. 106684, ene. 2021. Consultado: ago. 23, 2021. [En línea]. Disponible en: https://doi.org/10.1016/j.epsr.2020.106684

[2] M. Khanum, T. Mahboob, W. Imtiaz, H. Abdul Ghafoor, y R. Sehar, “A Survey on Unsupervised Machine Learning Algorithms for Automation, Classification and Maintenance”, *International Journal of Computer Applications*, vol. 119, núm. 13, jun. 2015. Consultado: ago. 23, 2021. [En línea]. Disponible en: https://doi.org/10.5120/21131-4058

[3] F. McLoughlin, A. Duffy, y M. Conlon, “A clustering approach to domestic electricity load profile characterisation using smart metering data”, *Science Direct*, vol. 141, pp. 190–199, mar. 2015.

[4] C. Bucher y G. Andersson, “Generation of Domestic Load Profiles - An Adaptive Top-Down Approach”, en *Proceedings of PMAPS*, Istanbul, Turkey, jun. 2012, vol. 12. Consultado: ago. 23, 2021. [En línea]. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.723.2362&rep=rep1&type=pdf

[5] I. Benítez, J.-L. Díez, A. Quijano, y I. Delgado, “Dynamic clustering of residential electricity consumption time series data based on Hausdorff distance”, *Science Direct*, vol. 140, pp. 517–526, nov. 2016.

[6] A. Kumar Tanwar, E. Crisostomi, P. Ferraro, M. Raugi, M. Tucci, y G. Guinta, “Clustering analysis of the electrical load in european countries”, en *International Joint Conference on Neural Networks (IJCNN)*, Killarney, Ireland, jul. 2015, pp. 1–8. doi: 10.1109/IJCNN.2015.7280329.

[7] R. D. Trejos Ramírez, “Metodología para la detección de pérdidas no técnicas en sistemas de distribución utilizando métodos de mineria de datos”, Programa de Maestría en Ingeniería Eléctrica, Universidad Tecnológica de Pereira, Pereira, Colombia, 2014. Consultado: ago. 09, 2021. [En línea]. Disponible en: https://core.ac.uk/download/pdf/71398073.pdf

[8] G. Shamim y M. Rihan, “Novel Technique for Feature Computation and Clustering of Smart Meter Data”, en *International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, Aligarh, India, nov. 2019, pp. 1–5. doi: 10.1109/UPCON47278.2019.8980078.

[9] M. S. Amoroso Ordoñez y H. S. Ávila Flores, “Aplicación de las técnicas de agrupamiento para la distribución cuasi-óptima de una red híbrida WDM-TDM/PON en cascada multinivel que da soporte a una smart grod o smart city”, Trabajo de titulación de pregrado, Universidad Politécnica Salesiana Sede Cuenca, Cuenca, Ecuador, 2015. Consultado: ago. 23, 2021. [En línea]. Disponible en: https://dspace.ups.edu.ec/bitstream/123456789/7736/1/UPS-CT004587.pdf

[10] L. F. Camacho Ortega, “Estimación de la demanda Eléctrica de la cocina de inducción por análisis clúster”, Trabajo de titulación de pregrado, Universidad Nacional de Loja, Loja, Ecuador, 2017. Consultado: ago. 23, 2021. [En línea].

Disponible en: https://dspace.unl.edu.ec/jspui/bitstream/123456789/18662/1/Camacho%20Ortega%2C%20Leoncio%20Francisco.pdf

[11] B. J. Mayorga Márquez, “Pronóstico espacial de demanda eléctrica mediante la técnica de agrupamiento (clustering) de curvas históricas - aplicación a la Empresa Eléctrica Ambato Regional Centro Norte S.A.”, Trabajo de titulación de pregrado, Escuela Politécnica Nacional, Quito, Ecuador, 2018. Consultado: ago. 23, 2021. [En línea]. Disponible en: https://bibdigital.epn.edu.ec/bitstream/15000/19297/1/CD-8665.pdf

[12] M. Á. Gallardo Vigil, “Introducción al análisis clúster. Consideraciones generales.” Universidad de Granada, 2009. Consultado: ago. 23, 2021. [En línea]. Disponible en: https://www.ugr.es/~gallardo/pdf/cluster-1.pdf

[13] S. De la Fuente Fernández, “Análisis conglomerados”. Universidad Autónoma de Madrid, 2011. Consultado: ago. 23, 2021. [En línea]. Disponible en: https://www.estadistica.net/Master-Econometria/Analisis_Cluster.pdf

[14] M. del P. Landa Baella y D. C. Villagómez Véliz, “Estadística computacional | Desarrollo de análisis clúster en R”. 2015.

[15] J. Amat Rodrigo, “Clustering y heatmaps: aprendizaje no supervisado”. sep. 2017. Consultado: ago. 23, 2021. [En línea]. Disponible en: https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps

[16] S. A. Azad, A. B. M. Shawakat Ali, y P. Wolfs, “Identification of typical load profiles using K-means clustering algorithm”, en *Asia-Pacific World Congress on Computer Science and Engineering*, Nadi, Fiji, nov. 2014, pp. 1–6. doi: 10.1109/APWCCSE.2014.7053855.

[17] N. Sapkota, A. Alsadoon, P. W. C. Prasad, A. Elchouemi, y A. Kumar Singh, “Data Summarization Using Clustering and Classification: Spectral CLustering Combined with k-Means Using NFPH.”, presentado en International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, feb. 2019. doi: 10.1109/COMITCon.2019.8862218.

[18] A. Sabzi, Y. Farjami, y M. ZiHayat, “An improved fuzzy k-medoids clustering algorithm with optimized number of clusters”, en *International Conference on Hybrid Intelligent Systems (HIS)*, Melacca, Malaysia, dic. 2011, vol. 11th, pp. 206–210. doi: 10.1109/HIS.2011.6122106.

[19] N. Arbin, N. Suhailayani Suhaimi, N. Zafirah Mokhtar, y Z. Othman, “Comparative Analysis between K-Means and K-Medoids for Statistical Clustering”, en *International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)*, Kota Kinabalu, Malaysia, dic. 2015, vol. 3erd, pp. 117–121. doi: 10.1109/AIMS.2015.82.

[20] D. Cao y B. Yang, “An improved k-medoids clustering algorithm”, en *International Conference on Computer and Automation Engineering (ICCAE)*, Singapore, feb. 2010, vol. 2nd, pp. 132–135. doi: 10.1109/ICCAE.2010.5452085.

[21] H.-L. Shieh, C.-C. Kuo, y F.-H. Chen, “Two-phases clustering algorithm based on subtractive clustering and k-nearest neighbors”, en *International Conference on*

Machine Learning and Cybernetics, Tianjin, China, jul. 2013, pp. 1802–1806. doi: 10.1109/ICMLC.2013.6890889.

[22] D. Pan, Z. Zhao, L. Zhang, y C. Tang, “Recursive clustering K-nearest neighbors algorithm and the application in the classification of power quality disturbances”, en *IEEE Conference on Energy Internet and Energy System Integration (EI2)*, Beijing, China, nov. 2017, pp. 1–5. doi: 10.1109/EI2.2017.8245652.

[23] H.-L. Shieh, “Semi-supervised Clustering Based on K-Nearest Neighbors”, en *International Conference on Digital Manufacturing & Automation*, Guilin, China, ago. 2012, vol. 3th, pp. 759–762. doi: 10.1109/ICDMA.2012.179.

[24] C. Jie Ma y Z. Sheng Ding, “Improvement of k-nearest neighbor algorithm based on double filtering”, en *International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, Harbin, China, dic. 2020, vol. 5th, pp. 1567–1570. doi: 10.1109/ICMCCE51767.2020.00343.

[25] Z. Nazari, D. Kang, M. Reza Asharif, Y. Sung, y S. Ogawa, “A new hierarchical clustering algorithm”, en *International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS)*, Okinawa, Japan, nov. 2015, pp. 148–152. doi: 10.1109/ICIIBMS.2015.7439517.

[26] Y. Rong y Y. Liu, “Staged text clustering algorithm based on K-means and hierarchical agglomeration clustering”, en *IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, Dalian, China, jun. 2020, pp. 124–127. doi: 10.1109/ICAICA50127.2020.9182394.

[27] Y. B. Park y S. Park, “Photovoltaic power data analysis using hierarchical clustering”, en *International Conference on Information Networking (ICOIN)*, Chiang Mai, Thailand, ene. 2018, pp. 727–731. doi: 10.1109/ICOIN.2018.8343214.

[28] S. Babichev, V. Lytvynenko, y V. Osypenko, “Implementation of the objective clustering inductive technology based on DBSCAN clustering algorithm”, en *International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT)*, Lviv, Ukraine, sep. 2017, vol. 12th, pp. 479–484. doi: 10.1109/STC-CSIT.2017.8098832.

[29] L. Ma, “An improved and heuristic-based iterative DBSCAN clustering algorithm”, en *Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Chongqing, China, mar. 2021, vol. 5th, pp. 2709–2714. doi: 10.1109/IAEAC50856.2021.9390918.

[30] X. Chen, D. Liu, X. Wang, Y. Chen, y S. Cheng, “Improved DBSCAN Radar Signal Sorting Algorithm Based on Rough Set”, en *International Conference on Big Data and Informatization Education (ICBDIE)*, Hangzhou, China, abr. 2021, vol. 2nd, pp. 398–401. doi: 10.1109/ICBDIE52740.2021.00096.

[31] J. Wang, H. Shu, y G. Yan, “Clustering analysis of manipulate errors for railway transport based on DBSCAN”, en *International Conference on Computer Science and Service System (CSSS)*, Nanjing, China, jun. 2011, pp. 3080–3082. doi: 10.1109/CSSS.2011.5972071.

[32] Z. Zhang, G. Ni, y Y. Xu, “Comparison of Trajectory Clustering Methods based on K-means and DBSCAN”, en *International Conference on Information*

Technology,Big Data and Artificial Intelligence (ICIBA), Chongqing, China, nov. 2020, pp. 557–561. doi: 10.1109/ICIBA50161.2020.9277214.

[33] Q. Zhu, X. Tang, y Z. Liu, “Revised DBSCAN Clustering Algorithm Based on Dual Grid”, en *Chinese Control And Decision Conference (CCDC)*, Hefei, China, ago. 2020, pp. 3461–3466. doi: 10.1109/CCDC49329.2020.9163926.

[34] B. Pal y M. Kumar Paul, “Gaussian mixture based semi supervised boosting for imbalanced data classification”, en *International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE)*, Rajshahi, Bangladesh, dic. 2016, vol. 2nd, pp. 1–4. doi: 10.1109/ICECTE.2016.7879620.

[35] S. Ayala Hernández, “Moda para datos no agrupados y para datos agrupados”, México, jun. 26, 2020. Consultado: ago. 15, 2021. [En línea]. Disponible en: https://www.uaeh.edu.mx/division_academica/educacion-media/repositorio/2010/6-semestre/estadistica/moda-para-datos-agrupados-y-desagrupados.pdf

[36] F. Najar, S. Bourouis, N. Bouguila, y S. Belghith, “A Comparison Between Different Gaussian-Based Mixture Models”, en *International Conference on Computer Systems and Applications (AICCSA)*, Hammamet, Tunisia, nov. 2017, vol. 14th, pp. 704–708. doi: 10.1109/AICCSA.2017.108.

[37] S. Chandrakala y C. Chandra Sekhar, “Model Based Clustering of Audio Clips Using Gaussian Mixture Models”, en *International Conference on Advances in Pattern Recognition*, Kolkata, India, feb. 2009, vol. 7th, pp. 47–50. doi: 10.1109/ICAPR.2009.92.

[38] G. Qiu, Y. Wei, H. Yang, Y. Wang, y P. Luo, “Image Restoration Based on Improved Patch Clustering in Gaussian Mixture Models”, en *Chinese Automation Congress (CAC)*, Hangzhou, China, nov. 2019, pp. 4502–4505. doi: 10.1109/CAC48633.2019.8997202.

[39] X. Bai, S. Luo, y Y. Zhao, “Entropy based soft K-means clustering”, en *IEEE International Conference on Granular Computing*, Hangzhou, China, ago. 2008, pp. 107–110. doi: 10.1109/GRC.2008.4664627.

[40] B. Zhu, E. Bedeer, H. H. Nguyen, R. Barton, y J. Henry, “Improved Soft-k-Means Clustering Algorithm for Balancing Energy Consumption in Wireless Sensor Networks”, *IEEE Internet of Things Journal*, vol. 8, núm. 6, pp. 4868–4881, mar. 15, 2021.

[41] R. Biddle, S. Liu, y G. Xu, “Semi-Supervised Soft K-Means Clustering of Life Insurance Questionnaire Responses”, en *International Conference on Behavioral, Economic, and Socio-Cultural Computing (BESC)*, Kaohsiung, Taiwan, nov. 2018, vol. 5th, pp. 30–31. doi: 10.1109/BESC.2018.8697227.

[42] A. V. Dian Sano, T. Daud Imanuel, M. Intanadias Calista, H. Nindito, y A. Raharto Condrobimo, “The Application of AGNES Algorithm to Optimize Knowledge Base for Tourism Chatbot”, en *International Conference on Information Management and Technology (ICIMTech)*, Jakarta, Indonesia, sep. 2018, pp. 65–68. doi: 10.1109/ICIMTech.2018.8528174.

[43] J. A. Gómez Sánchez, “Análisis comparativo de diferentes métodos de agrupación para el tratamiento de datos de expresión genética.”, Trabajo de titulación para optar por el Título de Máster en Bioinformática y Bioestadística,

Universidad Oberta de Catalunya, Barcelona, España, 2018. Consultado: sep. 15, 2021. [En línea]. Disponible en: http://openaccess.uoc.edu/webapps/o2/bitstream/10609/82006/7/jgomezsanchezTFM0618memoria.pdf

[44] MathWorks(R), "Linkage", *Linkage*, 2021. https://la.mathworks.com/help/stats/linkage.html (consultado sep. 15, 2021).

[45] S. Banchero, "Calidad del agrupamiento: Coeficiente de Silueta". Universidad Nacional de Lujan, oct. 2015. Consultado: sep. 15, 2021. [En línea]. Disponible en: http://www.labredes.unlu.edu.ar/sites/www.labredes.unlu.edu.ar/files/site/data/bdm/coeficiente-silueta.pdf

[46] E. Fadila Sirat, B. Darma Setiawan, y F. Ramdani, "Comparative Analysis of K-Means and Isodata Algorithms for Clustering of Fire Point Data in Sumatra Region", en *International Symposium on Geoinformatics (ISyG)*, Malang, Indonesia, vol. 4th, pp. 1–6. doi: 10.1109/ISYG.2018.8611879.

[47] D. F. Vallejo Huanga, "Clustering de documentos con restricciones de tamaño", Trabaja de titulación de postgrado, Universidad Politécnica de Valencia, Valencia, España, 2016. Consultado: sep. 15, 2021. [En línea]. Disponible en: http://mugi.webs.upv.es/wp-content/uploads/2016/11/TFM-Diego-Vallejo-MUGI.pdf

[48] J. Rivera Fonseca, "Agrupamiento de enzimas similares de la familia GH-70 utilizando descriptores libres de alineamient", Trabajo de diplomado, universidad Central "Marta Abreu" de Las Villas, Santa Clara, Cuba, 2019. Consultado: sep. 15, 2021. [En línea]. Disponible en: https://dspace.uclv.edu.cu/bitstream/handle/123456789/11417/Tesis%20de%20Pregrado%20Jerry%20Rivera%20Fonseca.pdf?sequence=1&isAllowed=n

[49] E. D. Matrínez Ruíz, "Metodología para la selección mínima de coordinaciones necesarias para proteger una microrred", Trabajo de titulación de pregrado, Universidad Tecnológica de Pereira, Pereira, Colombia, 2019. Consultado: sep. 15, 2021. [En línea]. Disponible en: http://repositorio.utp.edu.co/dspace/bitstream/handle/11059/10818/T621.319%20M386.pdf?sequence=1&isAllowed=y

[50] D. Marutho, S. Hendra Handaka, E. Wijaya, y Muljono, "The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News", en *International Seminar on Application for Technology of Information and Communication*, Semarang, Indonesia, sep. 2018, pp. 533–538. doi: 10.1109/ISEMANTIC.2018.8549751.

[51] A. San Suárez, "Evaluación de una nueva metodología para la estimación del transvase de votos entre elecciones", Trabajo de titulación de pregrado, Universidad Politécnica de Madrid, Madrid, España, 2018. Consultado: sep. 15, 2021. [En línea]. Disponible en: https://oa.upm.es/49662/1/TFG_ALFONSO_SANZ_SUAREZ.pdf

[52] Y. Tao, J. Deng, y C. Xingshen, "Drug Audit Based on Bisecting K-Means Clustering Algorithm", en *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, Guilin, China, oct. 2019, pp. 265–270. doi: 10.1109/CyberC.2019.00052.