



DEPARTAMENTO DE COMPUTACIÓN E INFORMÁTICA

FACULTAD DE INGENIERÍA

UNIVERSIDAD DE PLAYA ANCHA

**PROYECTO DE TÍTULO PARA OPTAR AL TÍTULO DE
INGENIERO EN INFORMÁTICA Y AL GRADO ACADÉMICO DE
LICENCIADO EN CIENCIAS DE LA INGENIERÍA**

**CARACTERIZACIÓN DE LA GENERACIÓN REAL DE ENERGÍA
DEL SISTEMA ELÉCTRICO NACIONAL (SEN) MEDIANTE
AGRUPAMIENTO DE SERIES TEMPORALES: 2020-2023**

Alumno: José René Meza Villalón

Profesor Guía: Dr. Carlos Valle Vidal

Diciembre 2024

TABLA DE CONTENIDO

ÍNDICE DE TABLAS.....	iv
ÍNDICE DE FIGURAS.....	v
GLOSARIO.....	vi
Resumen	viii
Abstract.....	ix
Introducción.....	x
CAPITULO I: Antecedentes del estudio	1
1. Planteamiento del problema	2
1.1 Distribución de Tipos de Generadoras en las Regiones de Chile	5
1.2 Justificación	6
1.3 Preguntas de investigación	7
1.4 Objetivos del estudio	8
1.4.1 Objetivo general.....	8
1.4.2 Objetivos específicos.....	8
CAPITULO II: Marco Conceptual.....	9
2.1 Estado del arte	9
2.1.1 Agrupamiento en la Generación de Energía a Nivel Mundial.....	9
2.1.2 Estado Actual de las Energías Renovables en Chile	10
2.2 Técnicas de agrupamiento de series de tiempo	11
2.2.1 Algoritmos de aprendizaje automático en energía en Chile.....	12
CAPITULO III Propuesta de solución	14
3.1 Algoritmos y pseudocódigos propuestos.....	14
3.3.1 Algoritmo de clustering <i>K</i> -means.....	15
3.3.2 Método de agrupamiento jerárquico	16
3.2 Deep learning.....	19
3.2.1 Modelo de Autoencoder LSTM para clustering de Series Temporales.....	20
3.3 Herramientas y recursos	22
3.3.1 Descripción de la base de datos (Dataset).....	23
3.4 Análisis de la base de datos	24

3.5 Métricas de distancia, desempeño y visualización	26
3.5.1 Distancia Euclidiana	26
3.5.2 Dynamic Time Warping (DTW).....	26
3.5.3 Método de Ward.....	27
3.5.4 Método de Enlace Completo (Complete Linkage)	28
3.5.5 WSS para Selección del K	28
3.5.6 Predicción Strength.....	29
3.5.7 Índice de Silueta:.....	30
3.5.8 Visualización.....	32
3.6 Códigos.....	32
CAPÍTULO IV: Análisis de Clústeres e hiperparámetros.....	33
4.1 Análisis de hiperparámetros	33
4.1.1 Selección de hiperparámetros para K -means	34
4.1.2 Resultados de hiperparámetros para AHC.....	39
4.1.3 Resultados de hiperparámetros.....	41
4.2 Resultados del análisis de clústeres.....	45
4.2.1 Resultados de K -means	46
4.2.1.1 Parámetros y Ejecución DTW	46
4.2.1 Resultados del agrupamiento jerárquico aglomerativo (AHC)	48
4.2.3 Resultados del Autocodificador LSTM DTW	51
4.3 Análisis de Resultados.....	53
4.3.1 Resultados de K -means	53
4.3.2 Resultados del Agrupamiento Jerárquico Aglomerativo (AHC)	57
4.3.3 Resultados del Autoencoder LSTM	58
CAPITULO V: Conclusiones.....	61
5.1 Identificación de Patrones	61
5.2 Comparación de Métodos.....	61
5.3 Análisis de Tendencias y Ciclos	62
5.4 Recomendaciones para futuras investigaciones.....	63
5.5 Respuesta a las Preguntas de Investigación	64
BIBLIOGRAFIA.....	66

ÍNDICE DE TABLAS

Tabla 1: Tipo de generadora y cantidad existente a diciembre del año 2023	5
Tabla 2: Distribución de generadoras del SEN por tipo en regiones.....	6
Tabla 3: Componentes y descripción del algoritmo LSTM.....	21
Tabla 4: Cuadro comparativo métodos de agrupamiento	21
Tabla 5: Información de la base de datos	23
Tabla 6: Detalles de la base de datos.....	23
Tabla 7: Resultados de las métricas Euclidiana y DTW.....	36
Tabla 8: Comparativo de resultados de métricas con $K = 2, 6$ y 15	37
Tabla 9: Resultados de hiperparámetros para AHC	40
Tabla 10: Participación porcentual por tipo de generadora.....	46
Tabla 11: Generadoras por clúster en cantidad y porcentaje DTW	47
Tabla 12: Generadoras por clúster en cantidad y porcentaje AHC.....	49
Tabla 13: Generadoras por clúster en cantidad y porcentaje LSTM – DTW	51
Tabla 14: Distribución porcentual de generadoras por clúster con DTW	54
Tabla 15: Distribución en % de generadoras por clúster con Ward	58
Tabla 16: Distribución en % de generadoras por clúster métrica DTW.....	59

ÍNDICE DE FIGURAS

Figura 1: Porcentaje de energía renovable vs convencional 2020-2023.....	3
Figura 2: Generación de energía por tipo años 2020 a 2023.....	7
Figura 3: Algoritmo de clustering K-means	15
Figura 4: Métodos de agrupamiento	17
Figura 5: Algoritmo de agrupamiento jerárquico aglomerativo.....	18
Figura 6: Algoritmo Autoencoder LSTM.....	20
Figura 7: Resultados de hiperparámetros con WSS	34
Figura 8: Resultados de hiperparámetros en métricas Euclidiana y DTW	35
Figura 9: Resultados de Hiperparámetros en métricas Ward y Complete	39
Figura 10: Comportamiento del autoencoder LSTM vs Epoch	42
Figura 11: Fuerza de predicción vs clúster en autoencoder LSTM (Euclidiana).....	44
Figura 12: Fuerza de predicción vs clúster en autoencoder LSTM (DTW)	44
Figura 13: Generadoras por clúster con métrica DTW usando PCA.....	47
Figura 14: Generadoras por clúster con métrica DTW usando t-SNE	48
Figura 15: Dendrograma para agrupamiento jerárquico aglomerativo.....	49
Figura 16: Generadoras por clúster con método AHC usando PCA.....	50
Figura 17: Generadoras por clúster con método AHC usando t-SNE.....	51
Figura 18: Generadoras por clúster, método LSTM DTW usando PCA.....	52
Figura 19: Generadoras por clúster, método LSTM DTW usando t-SNE	53
Figura 20: Curvas representativas del Clúster 1 para tipo Térmica	55
Figura 21: Curvas representativas del Clúster 4 para tipo Eólica	55
Figura 22: Curvas representativas del Clúster 5 para tipo Hidráulica	56
Figura 23: Curvas representativas del Clúster 3 para tipo Solar.....	56
Figura 24: Curvas representativas del Clúster 2 para tipo Solar.....	57

GLOSARIO

- 1) **Agrupamiento (Clustering):** Método de agrupación de objetos similares en conjuntos (clústeres), en este caso, generadoras de energía, según sus características o comportamientos.
- 2) **Agrupamiento Jerárquico Aglomerativo (AHC):** Método de agrupamiento en el que los datos comienzan como clústeres individuales y luego se agrupan jerárquicamente en función de su similitud.
- 3) **Autoencoders:** Redes neuronales utilizadas para aprender una representación eficiente de los datos de entrada, que luego se utiliza para reducción de dimensionalidad y, en este caso, para el agrupamiento de series temporales.
- 4) **Dropout:** Técnica de regularización utilizada en redes neuronales para prevenir el sobreajuste (overfitting). Durante el entrenamiento, dropout aleatoriamente "apaga" un porcentaje de las neuronas en cada capa de la red, lo que obliga al modelo a aprender representaciones más robustas y generalizables al evitar depender demasiado de unas pocas neuronas.
- 5) **Dynamic Time Warping (DTW):** Métrica que mide la similitud entre dos secuencias temporales que pueden variar en velocidad o duración, especialmente útil en el análisis de series temporales. El término DTW se mantiene en inglés, ya que es el nombre técnico.
- 6) **Epoch:** En el contexto del entrenamiento de redes neuronales, una epoch es una iteración completa sobre todo el conjunto de datos de entrenamiento. Durante una epoch, el modelo ajusta sus pesos y parámetros para aprender los patrones presentes en los datos. Un modelo generalmente necesita múltiples epochs para alcanzar una convergencia óptima.
- 7) **Índice de Silueta (Silhouette Index):** Métrica que evalúa la calidad del agrupamiento, indicando cuán similar es un objeto con su propio clúster en comparación con otros clústeres, con valores que varían entre -1 (agrupamiento incorrecto) y 1 (agrupamiento correcto).

- 8) **K-means:** Un algoritmo de agrupamiento particional que organiza los datos en un número predefinido de clústeres basados en la proximidad a los centroides. El nombre *K-means* se mantiene en inglés, ya que es el término estándar.
- 9) **LSTM (Long Short-Term Memory):** Un tipo de red neuronal recurrente (RNN) utilizada para predecir secuencias, eficaz en el análisis de datos temporales o series de tiempo. El término LSTM se mantiene en inglés, ya que es el nombre específico del modelo.
- 10) **Redes Neuronales Recurrentes (RNN - Recurrent Neural Networks):** Tipo de red neuronal diseñada para trabajar con datos secuenciales o de series temporales. A diferencia de las redes neuronales tradicionales, las RNN tienen conexiones que permiten mantener información sobre entradas previas, lo que les permite capturar patrones de temporalidad y dependencias a largo plazo en los datos. Las RNN son útiles en tareas como el procesamiento de lenguaje natural, predicción de series temporales y análisis de secuencias.
- 11) **WSS (Suma de Errores Cuadráticos dentro del Clúster):** Métrica que evalúa la compactación de los clústeres al medir la distancia de los puntos al centroide dentro de cada clúster.

Resumen

La creciente transición hacia fuentes de energía renovables plantea desafíos significativos en la planificación y operación de redes eléctricas, como el Sistema Eléctrico Nacional (SEN) de Chile, especialmente en el contexto de su compromiso hacia una matriz 100% renovable para 2030. Si bien estudios previos han aplicado técnicas de agrupamiento en datos de series temporales, la mayoría se ha enfocado en un número limitado de métodos y la falta de información al manejar datos complejos. Este trabajo aplica y compara tres enfoques diferenciados: el algoritmo K -means, el agrupamiento jerárquico aglomerativo (AHC) y un modelo basado en autoencoders LSTM, utilizando dos métricas para cada método.

Los datos, proporcionados por el Coordinador Eléctrico Nacional, incluyen registros horarios de generación de energía entre 2020 y 2023 de 623 generadoras seleccionadas. Los resultados experimentales destacan que K -means con métrica DTW formó clústeres estables. AHC fue capaz de encontrar subgrupos dentro de los grupos, permitiendo una separación más detallada de los datos. Por su parte, los autoencoders LSTM fueron menos efectivos en términos generales, aunque lograron detectar patrones específicos en las generadoras térmicas.

Abstract

The growing transition towards renewable energy sources poses significant challenges in the planning and operation of electrical grids, such as Chile's National Electric System (SEN), especially in the context of its commitment to achieving a 100% renewable energy matrix by 2030. While previous studies have applied clustering techniques to time series data, most have focused on a limited number of methods and lack detailed insights when handling complex data. This work applies and compares three distinct approaches: the K -means algorithm, agglomerative hierarchical clustering (AHC), and a model based on LSTM autoencoders, using two metrics for each method.

The data, provided by the National Electric Coordinator, includes hourly energy generation records from 2020 to 2023 for 623 selected generators. The experimental results highlight that K -means with DTW metrics formed stable clusters. AHC was able to identify subgroups within the clusters, enabling a more detailed separation of the data. In contrast, LSTM autoencoders were generally less effective, though they succeeded in detecting specific patterns in thermal generators.

Introducción

La transición hacia un sistema energético sostenible representa un desafío global crítico para mitigar el cambio climático y reducir la dependencia de combustibles fósiles. En Chile, este compromiso se traduce en la ambiciosa meta de alcanzar una matriz energética 100% renovable para 2030, liderada por el Sistema Eléctrico Nacional (SEN). Esta transición no solo implica una integración significativa de fuentes renovables, sino también la necesidad de superar los desafíos asociados a su variabilidad e intermitencia.

El presente trabajo aborda la caracterización de los patrones de generación de energía en el SEN mediante técnicas de agrupamiento de series temporales, para identificar patrones en los datos históricos de generación y para facilitar una comprensión más profunda de la dinámica operativa del sistema. El estudio evalúa y compara tres métodos principales: el algoritmo *K*-means, agrupamiento jerárquico aglomerativo (AHC) y autoencoders LSTM, utilizando dos métricas para cada método.

Con un enfoque en los datos proporcionados por el Coordinador Eléctrico Nacional, que abarcan el período 2020-2023, de este se seleccionaron 623 generadoras. Este análisis busca responder preguntas sobre el uso de estas técnicas para caracterizar la generación de energía en el SEN. Los resultados no tienen implicaciones para la planificación energética nacional, sino para generar conocimiento sobre el comportamiento y la eficacia de estas herramientas en el análisis de patrones en los datos de generación eléctrica.

Este documento se organiza en cinco capítulos: antecedentes del estudio, marco conceptual, propuesta de solución, análisis de resultados, y conclusiones. A través de ellos, se detallan los objetivos, metodología, análisis y hallazgos que subrayan el uso de las técnicas de agrupamiento como herramientas que sirven para la optimización del sistema eléctrico.

CAPITULO I: Antecedentes del estudio

La transición hacia fuentes de energía renovables es un componente fundamental en los esfuerzos globales por reducir las emisiones de gases de efecto invernadero y combatir el cambio climático. En Chile, esta transición ha sido especialmente relevante, con el compromiso del país de alcanzar una matriz energética 100% renovable para el año 2030. Este proceso ha implicado una transformación en el Sistema Eléctrico Nacional (SEN), que ha integrado de manera significativa energías renovables como la solar y eólica, junto a fuentes tradicionales como la térmica e hidráulica. Sin embargo, la variabilidad y complejidad asociadas a la generación renovable imponen desafíos para la planificación y estabilidad de la red.

La motivación detrás de este proyecto radica en la necesidad de optimizar el funcionamiento del SEN ante la creciente integración de energías renovables, que, aunque beneficiosas, son intermitentes y dependen de factores climáticos. Una caracterización efectiva de los patrones de generación de energía permitiría una gestión más eficiente de la red, favoreciendo su estabilidad y facilitando la toma de decisiones estratégicas.

Este trabajo tiene como objetivo explorar y analizar patrones en la generación de energía en el SEN durante el período 2020-2023, mediante técnicas avanzadas de agrupamiento de series temporales. Las preguntas de investigación incluyen la eficacia de técnicas como K -means, agrupamiento jerárquico y Deep learning con redes LSTM para identificar patrones de generación, y la comparación de estas en términos de eficiencia y calidad en la segmentación de datos.

Para llevar a cabo este análisis, se utilizará una metodología que incluye el preprocesamiento de datos históricos de generación, proporcionados por el Coordinador Eléctrico Nacional (CEN), seguido de la aplicación de distintos métodos de agrupamiento. Se evaluarán los resultados en función de métricas de desempeño, como la distancia Euclidiana y Dynamic Time Warping (DTW),

permitiendo identificar patrones relevantes en las series temporales y facilitando el análisis comparativo entre los métodos aplicados.

1. Planteamiento del problema

El crecimiento de la complejidad y la variabilidad en el Sistema Eléctrico Nacional (SEN) de Chile, impulsado por la integración de diferentes tipos de generadoras de energía renovable, presenta desafíos significativos para la operación y optimización de la red. La transición energética es un proceso crucial en la lucha contra el cambio climático y la reducción de la dependencia de combustibles fósiles. En este contexto, el Sistema Eléctrico Nacional (SEN) de Chile ha logrado avances significativos. En la Cuenta Pública 2023 del Coordinador Eléctrico Nacional (CEN, 2023), se destacó que el SEN alcanzó un 64% de generación renovable en el año, con un máximo horario del 94%, y un 71% si se consideran únicamente las fuentes renovables variables, como las plantas solares fotovoltaicas y eólicas. Estas cifras reflejan los esfuerzos continuos para incorporar fuentes de energías renovables en la matriz de generación eléctrica. Además, las condiciones hidrológicas favorables durante el año permitieron una reducción en los niveles de generación térmica convencional, contribuyendo a la seguridad del servicio del SEN, por lo que el éxito de estas iniciativas es atribuible a las acciones adoptadas por toda la industria, que está integrando nuevas fuentes de generación renovable. La meta es operar un sistema 100% renovable hacia el año 2030, según la Hoja de Ruta para una Transición Energética Acelerada presentada en 2022 (CEN, 2022). Este objetivo requiere un cambio de paradigma en el desarrollo y expansión de la red eléctrica, integrando nuevos recursos y tecnologías para asegurar una operación confiable y segura.

La Figura 1 muestra la evolución de la generación de energía por tipo Renovable o Convencional para el periodo 2020 a 2023.

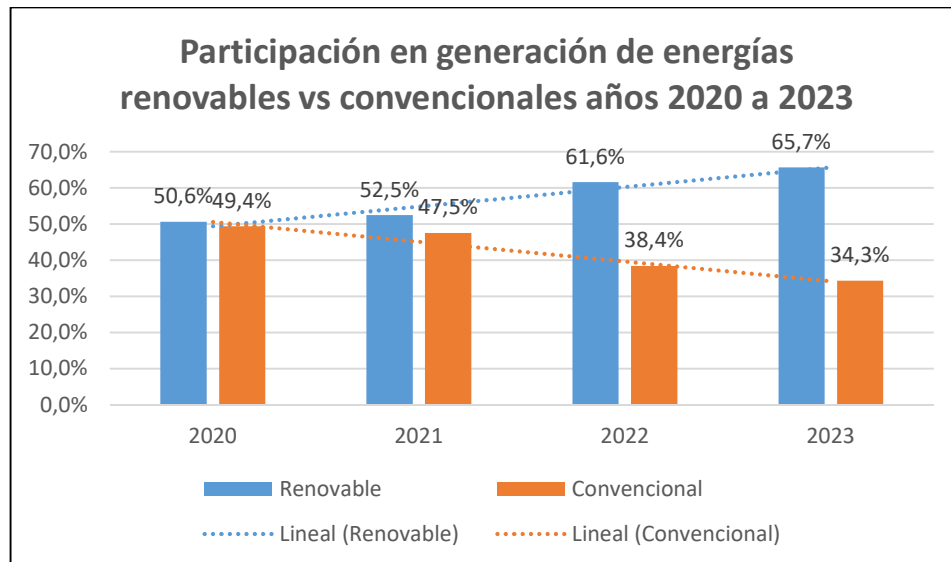


Figura 1: Porcentaje de energía renovable vs convencional 2020-2023

Fuente: elaboración propia

Chile, tras su compromiso en el Acuerdo de París conocido como COP 21, tiene la obligación de reducir las emisiones de gases de efecto invernadero (GEI). Las autoridades buscan alcanzar este objetivo principalmente mediante la disminución del uso de combustibles fósiles, como carbón, petróleo y gas, tanto en la industria como en la vida cotidiana. La electricidad, siendo un componente esencial, constituye un sector clave en la reducción de GEI. Se ha observado un incremento en las inversiones en energías renovables no convencionales (ERNC) con el fin de disminuir la producción de GEI. No obstante, la intermitencia y la variabilidad de las ERNC, junto con las tecnologías necesarias para su utilización eficiente, presentan ciertos desafíos. A pesar de estos retos, se han desarrollado tecnologías que permiten que la generación de electricidad a partir de ERNC sea competitiva con las energías tradicionales, especialmente los combustibles fósiles, y que también puedan sustituir a la energía hídrica, la cual ha disminuido drásticamente debido al cambio climático. Por lo que Chile enfrenta el desafío de reducir sus emisiones de GEI, principalmente mediante la disminución del uso de combustibles fósiles y el fomento de las ERNC. Aunque

existen obstáculos, el desarrollo tecnológico permite que las ERNC sean una alternativa viable y competitiva (Torres y García, 2021).

El Coordinador Eléctrico Nacional ha implementado adecuaciones organizacionales, como la creación de una Subgerencia de Estudios Eléctricos, una Unidad de Innovación y una Unidad de Regulación. Estas unidades trabajan en colaboración con universidades nacionales e internacionales para impulsar la innovación en el sector. En el año 2023, el SEN alcanzó 34.321 MW de capacidad instalada, con un 64% proveniente de fuentes renovables. La demanda máxima fue de 11.549 MW y la producción anual de energía fue de 83.637 GWh, un crecimiento del 0,8% respecto al año anterior. La red de transmisión nacional se extendió a 37.353 kilómetros. Las mejoras hidrológicas y la incorporación de nuevas unidades de Energía Renovable Variable redujeron significativamente las congestiones en el sistema de transmisión, especialmente en la zona norte durante el horario solar.

A pesar de los avances logrados, el SEN enfrenta un problema crítico: la creciente complejidad y variabilidad de la red eléctrica debido a la alta integración de fuentes de energía renovable. Esto plantea desafíos significativos para la planificación, operación y estabilidad del sistema. La variabilidad en la generación de energía renovable requiere sistemas de gestión avanzados y tecnologías innovadoras para asegurar un suministro eléctrico continuo y confiable. Además, la dependencia de condiciones climáticas variables añade una capa adicional de incertidumbre que debe ser gestionada eficazmente para evitar interrupciones y asegurar la resiliencia del sistema eléctrico nacional.

El Sistema Eléctrico Nacional (SEN) de Chile es un sistema eléctrico complejo compuesto por diversas plantas de generación de energía con diferentes tecnologías, capacidades y patrones de generación. El SEN fue establecido en 2017 y conecta el país desde Arica hasta Chiloé. Está formado por los antiguos








sistemas Interconectado Central (SIC) e Interconectado del Norte Grande (SING), además del Sistema de Aysén (SEA) y el Sistema de Magallanes (SEM).

Para comprender mejor el funcionamiento del SEN y optimizar su operación, es necesario caracterizar eficientemente estas plantas. Este informe presenta un estado del arte sobre las técnicas de agrupamiento de series de tiempo y sus aplicaciones en la caracterización de plantas de generación eléctrica en el SEN de Chile.

1.1 Distribución de Tipos de Generadoras en las Regiones de Chile

La Tabla 1 nos proporciona un panorama general de los distintos tipos de plantas generadoras que conforman el SEN, indicando la cantidad total de cada una. Esto nos permite identificar las tecnologías predominantes en la generación de electricidad a nivel nacional.

Tabla 1: Tipo de generadora y cantidad existente a diciembre del año 2023

ITEM	TIPO Y CANTIDAD
	Térmica (537): Generación de energía a partir de fuentes térmicas (combustibles fósiles).
	Hidráulica (193): Generación de energía a partir de fuentes hidroeléctricas.
	Solar (598): Generación de energía a partir de la radiación solar.
	Eólica (73): Generación de energía a partir del viento.
	Bess (12): Sistema de almacenamiento de energía en baterías.
	Geotérmica (5): Generación de energía mediante fuentes geotérmicas.
	Termosolar (1): Generación de energía térmica a partir de la radiación solar.

Fuente: con datos del CEN

La Tabla 2 complementa la información anterior, presentando una distribución geográfica de estas plantas por región. Esta tabla nos revela las regiones con mayor concentración de cada tipo de generadora, lo que a su vez refleja las

particularidades geográficas y climáticas de cada zona, así como las políticas energéticas implementadas.

Tabla 2: Distribución de generadoras del SEN por tipo en regiones

NUM	Nombre Región	Total							
XV	Arica y Parinacota	14	6	2	4	0	2	0	0
I	Tarapacá	52	30	5	17	0	0	0	0
II	Antofagasta	261	174	0	63	11	7	5	1
III	Atacama	89	30	1	45	10	3	0	0
IV	Coquimbo	98	16	3	66	13	0	0	0
V	Valparaíso	140	53	8	79	0	0	0	0
RM	Metropolitana de Santiago	187	52	23	112	0	0	0	0
VI	Libertador General O'Higgins	134	22	17	88	7	0	0	0
VII	Maule	127	24	29	73	1	0	0	0
XVI	Ñuble	49	12	2	35	0	0	0	0
VIII	Biobío	133	62	38	15	18	0	0	0
IX	La Araucanía	34	10	16	1	7	0	0	0
XIV	Los Ríos	38	13	25	0	0	0	0	0
X	Los Lagos	63	33	24	0	6	0	0	0
	TOTALES	1419	537	193	598	73	12	5	1

Fuente: con datos del CEN

1.2 Justificación

El análisis de los patrones de generación de energía eléctrica en Chile, podría ser fundamental en la planificación energética y consiguiente mejora de la gestión de la red eléctrica. En este contexto, se plantea la necesidad de caracterizar de manera efectiva la generación de energía a través de técnicas de agrupamiento con el fin de identificar patrones y agrupaciones relevantes. Esta caracterización permitiría no solo comprender la distribución de las fuentes de energía, sino también optimizar la operación del sistema y avanzar hacia una red más sostenible y eficiente (Berzal, F. 2018, p 43).

La Figura 2 da una idea de la generación total de energía en el periodo de años 2020 a 2023 por tipo de planta, datos que se buscara caracterizar.

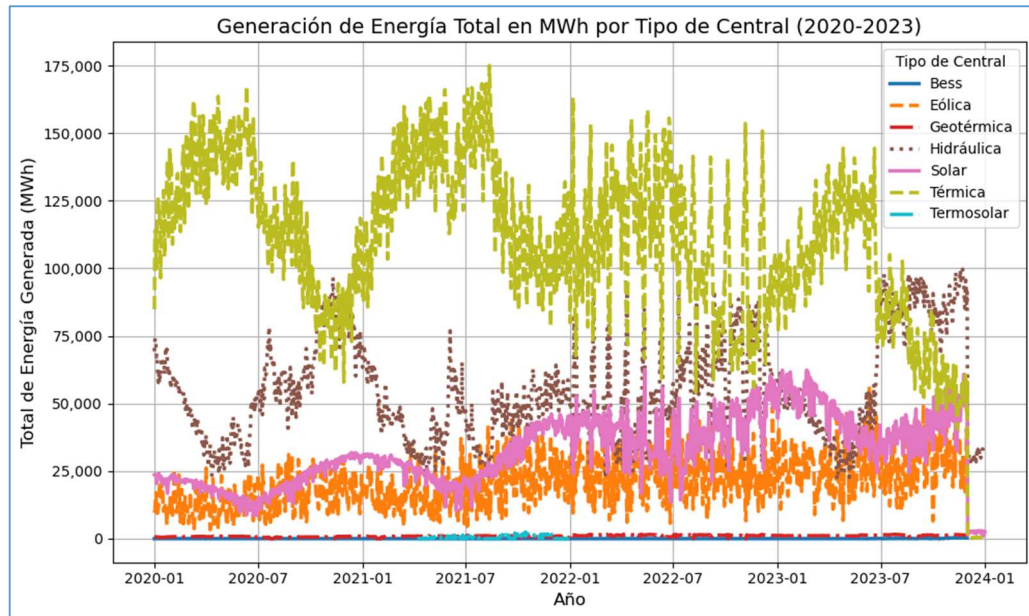


Figura 2: Generación de energía por tipo años 2020 a 2023

Fuente: elaboración propia

1.3 Preguntas de investigación

1. ¿Cómo se pueden aplicar técnicas de agrupamiento de series temporales para caracterizar de manera efectiva los patrones de generación de energía real del SEN?
2. *K*-means, Agrupamiento Jerárquico o Deep learning, ¿Cuál de estas tres técnicas es más adecuada para identificar patrones y agrupaciones relevantes en la generación de energía en el SEN de Chile?
3. ¿Cómo se comparan las herramientas en términos de eficiencia computacional y calidad de los agrupamientos generados? ¿Qué prioriza cada uno a la hora de agrupar?

1.4 Objetivos del estudio

1.4.1 Objetivo general

Explorar y comprender los patrones de generación de energía eléctrica en Chile durante el período 2020-2023, para evaluar su comportamiento y desempeño, utilizando técnicas de agrupamiento de series temporales y evaluando su desempeño con métricas de calidad, utilizando datos históricos proporcionados por el Coordinador Eléctrico Nacional (CEN).

1.4.2 Objetivos específicos

- 1) Identificar patrones en los datos de generación de energía del Sistema Eléctrico Nacional (SEN), mediante técnicas de agrupamiento como *K-means*, agrupamiento jerárquico y Deep learning de series temporales, para clasificar y caracterizar fuentes renovables y convencionales.
- 2) Comparar y evaluar los resultados obtenidos con los diferentes métodos aplicados, considerando sus similitudes y/o diferencias, utilizando la "distancia euclidiana" o "Dynamic Time Warping", para determinar el método más efectivo en la identificación de patrones.
- 3) Analizar las tendencias y ciclos identificados en los datos de generación de energía del SEN, utilizando técnicas avanzadas de análisis de series temporales como el coeficiente de Silhouette para comprender mejor los patrones subyacentes y su comportamiento a lo largo del tiempo.

CAPITULO II: Marco Conceptual

En este capítulo se desarrollan los conceptos y antecedentes teóricos necesarios para entender el contexto de este estudio. Se profundiza en las técnicas de agrupamiento de series temporales aplicadas al análisis de la generación de energía eléctrica, y se revisan las investigaciones previas que han utilizado estas metodologías, tanto a nivel mundial como en Chile.

2.1 Estado del arte

2.1.1 Agrupamiento en la Generación de Energía a Nivel Mundial

El uso de técnicas de clustering ha sido ampliamente explorado en el ámbito de la generación de energía, demostrando su capacidad para resolver problemas asociados con la integración de fuentes renovables y la gestión eficiente de sistemas eléctricos. Estas técnicas permiten identificar patrones en grandes volúmenes de datos, optimizando tanto la planificación como la operación de redes energéticas.

Un ejemplo destacado es el estudio realizado en Colombia, donde se emplearon métodos de clustering, como *K*-means y Fuzzy *C*-means, para analizar el potencial fotovoltaico en ciudades como Bogotá, Manizales, Pasto y Cúcuta. Este enfoque permitió identificar patrones en datos meteorológicos, como radiación solar y temperatura, que son fundamentales para optimizar la planificación y diseño de sistemas fotovoltaicos. Además, el uso de MATLAB® facilitó la implementación de estos algoritmos, destacando su aplicabilidad en regiones con alta variabilidad climática (Ramírez-Murillo, Torres-Pinzón, & Forero-García, 2019).

De manera similar, en Europa, el clustering ha sido utilizado en sistemas de control predictivo para redes eléctricas de gran escala. Estas aplicaciones incluyen la agrupación de generadores distribuidos y sistemas de almacenamiento para equilibrar variaciones inesperadas de carga en tiempo real.

Este enfoque ha mejorado significativamente la estabilidad y eficiencia de las redes eléctricas, mostrando cómo el clustering puede ser integrado con éxito en la operación de sistemas energéticos complejos (Xu et al., 2020). Por otra parte, en el diseño de sistemas energéticos de bajo carbono, se han implementado técnicas de clustering para seleccionar períodos representativos en análisis temporales. Esto ha permitido reducir la complejidad computacional y mejorar la precisión en la optimización de estos sistemas. Este enfoque es particularmente útil en proyectos de transición energética que requieren modelar grandes volúmenes de datos históricos y evaluar escenarios de planificación sostenible (Zhou et al., 2022).

Estas investigaciones demuestran cómo las técnicas de clustering han contribuido al análisis y la optimización de sistemas energéticos a nivel mundial, destacando su aplicabilidad en la caracterización de patrones temporales y en la toma de decisiones estratégicas para integrar fuentes renovables.

2.1.2 Estado Actual de las Energías Renovables en Chile

En Chile, las técnicas de clustering han comenzado a integrarse en el análisis energético como herramientas para enfrentar los desafíos asociados con la variabilidad de las Energías Renovables No Convencionales (ERNC) y la optimización de la operación del sistema eléctrico nacional. Estas metodologías permiten identificar patrones y segmentar datos energéticos complejos, contribuyendo a la transición hacia una matriz energética más sostenible.

Un ejemplo significativo en el contexto chileno es el uso de K -means y Análisis de Componentes Principales (PCA) para analizar datos de consumo eléctrico en distintas regiones del país. Este enfoque permitió segmentar consumidores según patrones de comportamiento energético entre 2015 y 2021, optimizando la planificación de recursos y ofreciendo una visión más detallada del uso energético. La aplicación del algoritmo K -means para la clasificación de clientes

residenciales utilizando datos de medidores inteligentes ha demostrado ser efectiva, identificando grupos de consumidores con patrones de consumo similares. Este análisis permite una mejor gestión y planificación de los recursos energéticos, favoreciendo tanto a los consumidores como a las distribuidoras de energía (Marrero, Carrizo, García-Santander, & Ulloa-Vásquez, 2021).

En el ámbito de la generación de energía, el clustering ha sido utilizado para analizar patrones en la producción de energía solar y eólica, especialmente en zonas como el norte del país, donde la radiación solar y las corrientes de viento ofrecen un alto potencial para proyectos renovables. Mediante el uso de técnicas de agrupamiento, se han identificado períodos de alta y baja generación, mejorando así la integración de estas fuentes en el Sistema Eléctrico Nacional (SEN) y reduciendo la incertidumbre asociada a su variabilidad.

Además, el clustering ha demostrado ser una herramienta valiosa en estudios preliminares para proyectos de hidrógeno verde, al permitir agrupar datos relacionados con la generación renovable y las condiciones óptimas para la producción y almacenamiento de hidrógeno. Esto refuerza la posición de Chile como líder regional en esta tecnología emergente, alineando sus esfuerzos con las tendencias globales hacia la sostenibilidad energética.

Estos casos muestran cómo las técnicas de clustering no solo permiten abordar desafíos específicos del sector energético chileno, sino también alinearse con las metas nacionales de descarbonización y transición energética, demostrando su relevancia como herramienta estratégica para el análisis y la planificación energética en el país.

2.2 Técnicas de agrupamiento de series de tiempo

A partir del análisis del estado del arte, tanto a nivel mundial como nacional, se evidencia que las técnicas de clustering son herramientas clave para caracterizar

patrones energéticos. En este capítulo, se describen las herramientas y métodos utilizados en el presente proyecto, adaptados a las particularidades del sistema energético chileno.

Los métodos de agrupamiento, conocidos también como agrupamiento, son utilizados en informática para agrupar objetos en conjuntos similares, buscando que los objetos dentro de un mismo grupo sean más similares entre sí que con los de otros grupos. Esta técnica es fundamental para encontrar patrones en conjuntos de datos complejos.

La elección del algoritmo de agrupamiento, el conjunto de datos y la medida de similitud son factores clave que determinan los resultados del agrupamiento. La medida de similitud se define generalmente como una distancia entre los objetos, que puede variar según la interpretación semántica de los datos y el problema específico. La forma ideal de agrupar datos no tiene una respuesta absoluta, ya que el proceso de agrupamiento es subjetivo. Sin embargo, se considera que un buen agrupamiento es aquel donde los objetos dentro de un mismo clúster son muy similares entre sí y diferentes de los objetos de otros clústeres. Esto se logra minimizando la distancia intraclúster y maximizando la distancia interclúster.

Los algoritmos de agrupamiento varían en su capacidad para manejar diferentes tipos de atributos (numéricos, categóricos), su eficiencia computacional, tolerancia al ruido y capacidad de identificar clústeres con formas arbitrarias. Algunos de los métodos más comunes incluyen los métodos particionales como *K*-means, jerárquicos, basados en densidad y agrupamiento en subespacios.

2.2.1 Algoritmos de aprendizaje automático en energía en Chile

El uso de algoritmos de Deep learning ha demostrado ser una herramienta eficaz en el análisis del sector energético en Chile. Un ejemplo destacado es el estudio de Yajure-Ramírez, C. A. (2022), que aplicó técnicas como *K*-means y Análisis de Componentes Principales (PCA) para segmentar datos de energía eléctrica

facturada entre 2015 y 2021. Este enfoque permitió identificar patrones relevantes y variables significativas en los datos, optimizando el análisis y la interpretación de la información disponible.

El estudio destacó que *K*-means fue efectivo para agrupar consumidores eléctricos en diferentes segmentos según su comportamiento, mientras que PCA se utilizó para reducir la dimensionalidad de los datos, identificando las variables que más influían en el consumo energético. Estas metodologías han sentado las bases para trabajos actuales en el análisis de series temporales y la caracterización de patrones energéticos en Chile.

Estos avances refuerzan la relevancia de los algoritmos de agrupamiento en el contexto energético nacional, especialmente en proyectos que buscan identificar patrones complejos en datos históricos. El éxito de estos enfoques respalda su inclusión en el presente proyecto, aplicándolos a un dominio temporal más reciente y con datos más detallados.

CAPITULO III Propuesta de solución

La creciente integración de fuentes de energía renovables en el Sistema Eléctrico Nacional (SEN) de Chile presenta tanto oportunidades como desafíos para su gestión y operación eficiente. Dado que las energías renovables, como la solar y eólica, son intermitentes y dependen de condiciones climáticas variables, es crucial contar con herramientas analíticas que permitan identificar patrones en los datos de generación de energía para optimizar la estabilidad y el funcionamiento de la red. En este contexto, este capítulo propone el uso de técnicas avanzadas de agrupamiento de series temporales, específicamente los algoritmos *K*-means, agrupamiento jerárquico aglomerativo (AHC) y autoencoders LSTM, con el objetivo de caracterizar los patrones de generación de energía del SEN durante el período 2020-2023.

Estas técnicas permiten clasificar las generadoras en función de sus patrones de comportamiento a lo largo del tiempo, facilitando la identificación de grupos con características similares. El uso de métricas de distancia, como la Euclidiana y el Dynamic Time Warping (DTW), permite una comparación más precisa entre las series temporales, considerando las variaciones en la generación de energía de cada tipo de generadora. En este capítulo se detallan los algoritmos propuestos, el preprocesamiento de los datos y las herramientas utilizadas para llevar a cabo el análisis. Además, se presentan las métricas de desempeño que se emplean para evaluar la calidad de los clústeres obtenidos, proporcionando así una base para comparar los métodos de agrupamiento seleccionados.

3.1 Algoritmos y pseudocódigos propuestos

Se presentan a continuación los algoritmos de los tres métodos que se proponen finalmente para este proyecto, a saber: agrupamiento *K*-means, agrupamiento jerárquico y Deep learning.

3.3.1 Algoritmo de clustering *K*-means

Los métodos particionales, por ejemplo, como *K*-means que corresponde a un método particional que agrupa series de tiempo en un número predefinido de clústeres, asignando centroides que minimizan las distancias a los puntos del clúster (Ikotun et al., 2023), requieren especificar el número de clústeres (*K*) y asignan centroides a cada clúster para minimizar la distancia de los puntos al centroide correspondiente. En la Figura 3 se presenta el agrupamiento *K*-means.

<p>Algorithm 1 Algoritmo de Clustering <i>K</i>-means</p> <hr/> <p>Require: <i>K</i> (número de clústeres), dataset $X = \{x_1, x_2, \dots, x_n\}$</p> <p>1: Inicializar <i>K</i> clústeres con sus centroides $\mu_1, \mu_2, \dots, \mu_K$ de forma aleatoria</p> <p>2: while no converge do</p> <p>3: for cada $x_i \in \text{dataset}$ do</p> <p>4: Asignar x_i al clúster C_k más cercano, es decir, $C_k := \arg \min_k \ x_i - \mu_k\ ^2$</p> <p>5: end for</p> <p>6: for cada clúster <i>K</i> do</p> <p>7: Recalcular el centroide $\mu_k := \frac{1}{ C_k } \sum_{x_i \in C_k} x_i$</p> <p>8: end for</p> <p>9: end while</p>
--

Figura 3: Algoritmo de clustering *K*-means

Fuente: Gironés Roig, *et al.* (2017, p 117)

De la Figura 3, el proceso del algoritmo *K*-means se desarrolla de la siguiente manera: en el Paso 1, durante la inicialización, se seleccionan aleatoriamente los *K* centroides dentro del rango de los datos. Luego, en el Paso 4, cada punto del dataset se asigna al clúster cuyo centroide esté más cercano, utilizando la distancia euclidiana como métrica. A continuación, en el Paso 7, correspondiente a la etapa de actualización de centroides, se recalculan los centroides de cada clúster como la media de todos los puntos asignados a ellos. Este proceso de asignación y actualización se repite en el bucle while hasta que se cumpla el

criterio de convergencia (Paso 9), el cual evalúa si los centroides han cambiado significativamente respecto a la iteración anterior; si el cambio es menor a una tolerancia predefinida, el algoritmo termina, de lo contrario, se repite el ciclo.

3.3.2 Método de agrupamiento jerárquico

Los métodos jerárquicos construyen una jerarquía de clústers donde cada clúster puede contener subclústeres, permitiendo una exploración detallada de la estructura de los datos (Murtagh & Contreras, 2017). Estas construcciones, denominados dendrogramas muestran la similitud entre objetos y pueden ser aglomerativos (de abajo hacia arriba) o divisivos (de arriba hacia abajo). En los métodos basados en densidad identifican clústeres como regiones densas de puntos separadas por regiones menos densas, lo que los hace eficientes para agrupamientos no globulares y robustos frente a ruido y outliers. Los métodos de agrupamiento en subespacios abordan la alta dimensionalidad de los datos al buscar patrones en subconjuntos de atributos. Estos métodos, como CLIQUE (Clustering In QUEst), exploran distintas combinaciones de atributos para encontrar clústeres en espacios de datos complejos (Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. 1998). Además, las reglas de asociación, como las utilizadas en la minería de datos para identificar patrones en grandes bases de datos, permiten descubrir correlaciones entre conjuntos de elementos, como el clásico ejemplo de la cerveza y los pañales en supermercados, lo cual puede influir en las estrategias de marketing y disposición de productos.

Estos métodos son clave en minería de datos y análisis de patrones, cada uno con sus ventajas y limitaciones dependiendo del tipo de datos y los objetivos del análisis. Aunque los métodos de agrupamiento y las reglas de asociación proporcionan herramientas poderosas para explorar datos sin etiquetar, es esencial adaptar la metodología al contexto específico. Además, hay que tener en cuenta las limitaciones propias de cada técnica para obtener resultados que sean significativos y aplicables en la práctica.

Los métodos de agrupamiento, mostrados en la Figura 4, se clasifican en dos categorías: jerárquicos y no jerárquicos. Su principal diferencia reside en la flexibilidad para definir el número y tamaño de los grupos. Los métodos jerárquicos permiten explorar distintos niveles de agrupamiento mediante un dendrograma, donde la altura del corte determina la cantidad y tamaño de los grupos. Por otro lado, los métodos no jerárquicos requieren definir el número y tamaño de los grupos antes del análisis.

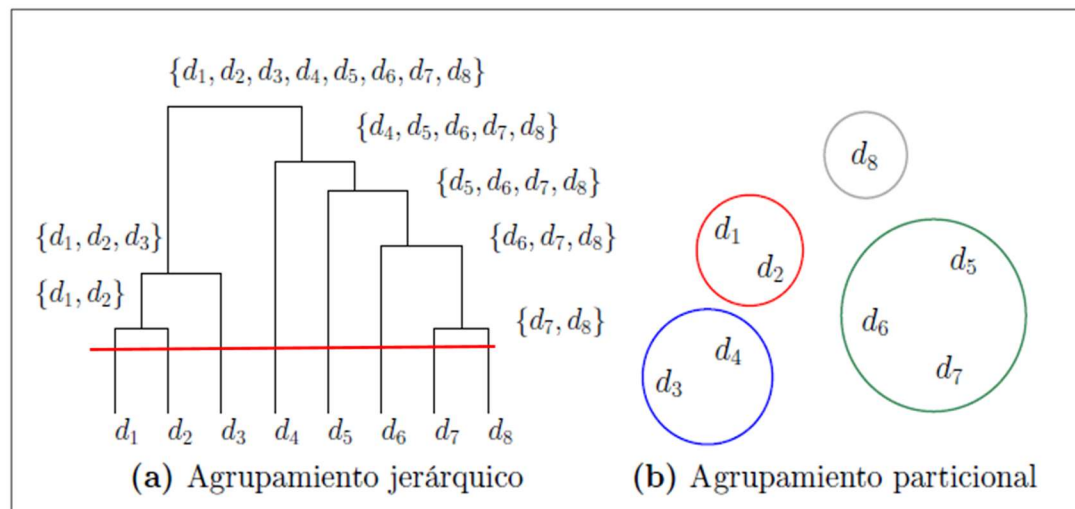


Figura 4: Métodos de agrupamiento

Fuente: Gironés Roig. *et al* (2017, p 40).

En la Figura 4 observamos los siguientes aspectos característicos del agrupamiento jerárquico (dendrograma).

Eje X: Representa los diferentes elementos o categorías analizados, los cuales se agrupan en función de sus características o propiedades.

Eje Y: Indica la distancia de fusión entre los grupos en el proceso de agrupamiento jerárquico. Cuanto mayor es la altura de un enlace, mayor es la disimilitud entre los grupos combinados.

Línea de Corte (línea roja): Marca un umbral que define el número de grupos o clústeres, permitiendo clasificar los elementos según su similitud.

Se ha optado por utilizar el método de agrupamiento jerárquico aglomerativo (AHC) debido a su capacidad para identificar estructuras jerárquicas en los datos, lo que resulta especialmente útil en el análisis de series temporales complejas como las asociadas al Sistema Eléctrico Nacional (SEN). A continuación, se presenta el pseudocódigo correspondiente, que describe el proceso iterativo de fusión de clústeres en función de una métrica de similitud predefinida. En la Figura 5 se ilustra el algoritmo de agrupamiento jerárquico aglomerativo, detallando los pasos clave de este enfoque.

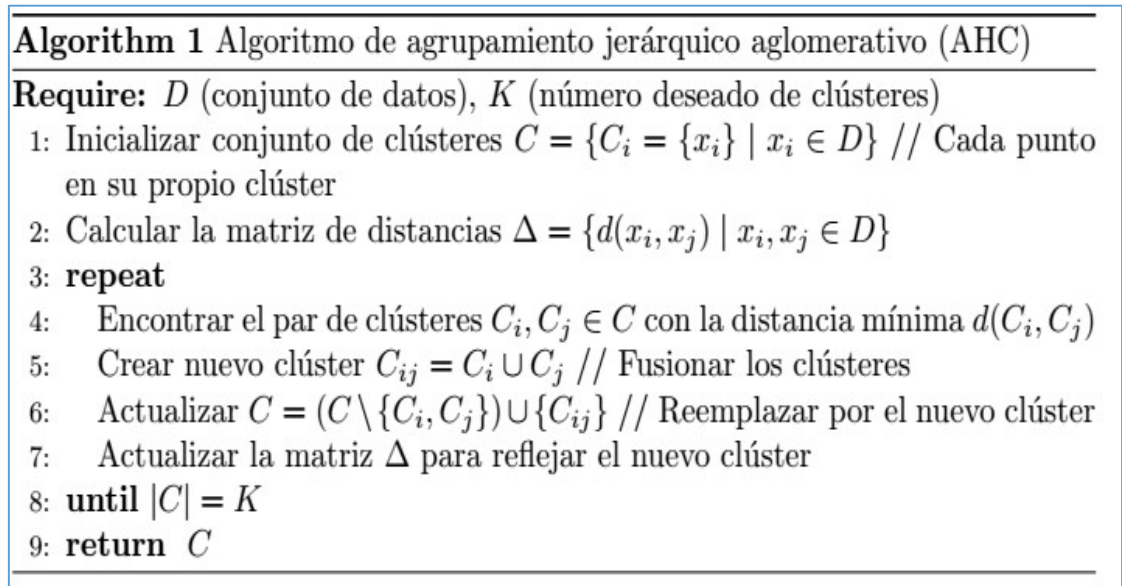


Figura 5: Algoritmo de agrupamiento jerárquico aglomerativo

Fuente: Zaki, M. J., & Meira, W., Jr. (2014).

En la Figura 5, se presentan los pasos con su correspondiente descripción del algoritmo de agrupamiento jerárquico aglomerativo (AHC). El desarrollo del algoritmo sigue los siguientes pasos: en el Paso 1, durante la etapa de inicialización, cada dato se considera como un clúster independiente. En el Paso 2, se calcula la matriz de distancias Δ , que contiene las distancias entre todos los puntos o clústeres actuales. En el Paso 3, se identifica el par de clústeres C_i y C_j

con la menor distancia $d(C_i, C_j)$. Luego, en los Pasos 4-6, estos clústeres se fusionan en un nuevo clúster C_{ij} , actualizando tanto el conjunto de clústeres como la matriz de distancias Δ . Este proceso se repite iterativamente (repeat) hasta que el número de clústeres $|C|$ es igual a K , el número deseado de clústeres (Paso 7). Finalmente, en el Paso 9, se retorna el conjunto final de clústeres C . Este enfoque crea una jerarquía de clústeres que organiza los datos en función de sus similitudes, ofreciendo una representación jerárquica.

3.2 Deep learning

El Deep learning, también conocido como aprendizaje profundo, representa una avanzada técnica dentro del campo del aprendizaje automático que ha revolucionado la capacidad de los sistemas informáticos para aprender a partir de datos no estructurados o complejos. Se basa en redes neuronales artificiales compuestas por múltiples capas de unidades de procesamiento, inspiradas en el funcionamiento del cerebro humano. Desde un punto de vista conceptual, las redes neuronales en el Deep learning operan mediante la representación de conceptos a través de patrones de actividad distribuida en la red. A diferencia de los modelos simbólicos que explicitan el conocimiento mediante reglas lógicas, las redes neuronales aprenden de manera implícita ajustando los pesos de las conexiones entre las neuronas durante el entrenamiento. Este enfoque sub-simbólico permite a las redes neuronales generalizar automáticamente ante nuevos datos, una capacidad crucial para resolver problemas complejos de aprendizaje (Heaton, J. 2018).

El Deep learning ha logrado avances significativos gracias al refinamiento de los algoritmos de entrenamiento y al aumento de la disponibilidad de grandes conjuntos de datos. A partir de los años 2000, se produjo un salto cualitativo con el desarrollo de técnicas más complejas que permiten construir modelos con múltiples capas de neuronas. Este avance ha facilitado la aplicación efectiva del Deep learning en diversas áreas como el reconocimiento de voz, la visión por

computadora y el procesamiento de datos no estructurados como imágenes y señales (Chollet, F. 2021).

Podemos inferir que la integración de técnicas de agrupamiento de series temporales con Deep learning, ofrecen una metodología robusta y avanzada para explorar, modelar y analizar datos temporales complejos. Mientras que los métodos de agrupamiento tradicionales permiten segmentar inicialmente los datos, mientras que el Deep learning proporciona una capacidad adicional para aprender representaciones profundas de los datos, mejorando así la precisión y la capacidad predictiva en problemas complejos de series temporales.

Se hará uso del método Autoencoder Long Short-Term Memory (LSTM) que son modelos de autoencoder para series temporales

3.2.1 Modelo de Autoencoder LSTM para clustering de Series Temporales

En la Figura 6, se presenta el Algoritmo 1 del Autoencoder Long Short-Term Memory (LSTM) para agrupamientos con series de tiempo.

<p>Algorithm 1 Autoencoder LSTM para series de tiempo con clustering</p> <p>Require: <i>input_data</i> (serie de tiempo de entrada)</p> <p>Ensure: Serie reconstruida y clústeres asignados</p> <ol style="list-style-type: none">1: Preprocesamiento:2: Cargar y preparar <i>input_data</i> (normalización y transformación a tensores)3: Definir el Autoencoder LSTM:4: Crear el Encoder para obtener una representación comprimida de las series de tiempo5: Crear el Decoder para reconstruir las series de tiempo a partir de la representación comprimida6: Entrenamiento:7: Entrenar el autoencoder para minimizar la pérdida de reconstrucción8: Obtención de Representaciones Latentes:9: Pasar los datos a través del encoder y almacenar las representaciones latentes10: Clustering:11: Aplicar KMeans en las representaciones latentes12: return Serie reconstruida y etiquetas de clúster

Figura 6: Algoritmo Autoencoder LSTM

Fuente: recuperado de Tavakoli et al. (2020).

En la Tabla 3, se presentan los componentes con su correspondiente descripción del algoritmo de Deep learning con Autoencoder LSTM.

Tabla 3: Componentes y descripción del algoritmo LSTM

Componente	Paso	Descripción
Inicialización de Datos	1	Cargar y preparar input_data, incluyendo normalización y transformación a tensores.
Definición del Autoencoder	3-5	Crear un encoder para obtener una representación comprimida de la serie de tiempo y un decoder para reconstruirla.
Entrenamiento del Modelo	6-7	Entrenar el autoencoder para minimizar la pérdida de reconstrucción entre la entrada y la salida.
Extracción de Representaciones Latentes	8-9	Pasar los datos a través del encoder y almacenar las representaciones latentes.
Clustering	10	Aplicar K -means sobre las representaciones latentes obtenidas.
Salida	12	Retornar la serie reconstruida y las etiquetas de clúster asignadas.

Fuente: elaboración propia

En la Tabla 4 se presenta un cuadro comparativo con las principales características de los tres métodos propuestos.

Tabla 4: Cuadro comparativo métodos de agrupamiento

Característica	K -means	Jerárquico	Deep autoencoding
Descripción	<p>Inicialización: Se seleccionan K centroides iniciales y se asignan series de tiempo al clúster cuyo centroide es más cercano.</p> <p>Asignación: Cada serie de tiempo se asigna al clúster según la distancia euclidiana.</p> <p>* Actualización: Se recalculan los centroides como el promedio de las series de tiempo asignadas.</p> <p>Iteración: Se repiten asignación y actualización hasta convergencia o máximo de iteraciones.</p>	<p>Método Aglomerativo: Inicia con cada serie de tiempo como un clúster separado y fusiona los más similares.</p> <p>Método Divisivo: Inicia con todos en un clúster y divide. Utiliza distintas métricas de distancia (single-link, complete-link, average-link).</p>	<p>Redes Neuronales Recurrentes (RNN): Manejan datos secuenciales, capturan dependencias temporales.</p> <p>Redes LSTM: Aprenden dependencias a largo plazo, mitigando problemas de gradientes desvanecientes.</p> <p>Autoencoders: Aprenden representaciones eficientes.</p>

Fortalezas	Simplicidad: Fácil de entender e implementar. Rapidez: Converge rápidamente. Escalabilidad: Funciona bien con grandes volúmenes de datos.	Flexibilidad: No requiere K predefinido. Visualización: Produce dendrogramas. Variedad de formas de clústeres: Encuentra formas y tamaños variados de clústeres.	Capacidad para Modelar Patrones Complejos: Captura relaciones no lineales. * Adaptabilidad: Ajustable a diversas tareas. Aprendizaje de Características: Aprende representaciones de datos no supervisadas.
Debilidades	Necesidad de K predefinido: Requiere especificar el número de clústeres. Sensibilidad a la inicialización: Resultados dependen de la elección inicial de centroides. Assume clústeres esféricos: No adecuado para datos con formas irregulares.	Computacionalmente Intensivo: Menos eficiente con grandes conjuntos de datos. Sensibilidad a ruido y outliers: Influencia negativa en la jerarquía.	Requiere grandes cantidades de datos: Necesita volúmenes significativos para entrenamiento. Computacionalmente Intensivo: Demanda poder de cómputo y tiempo para entrenar. Complejidad: Más difícil de implementar y entender.

Fuente: elaboración propia basado Berzal, (2018).

3.3 Herramientas y recursos

Existen diversas herramientas y recursos disponibles para implementar técnicas de series de tiempo, como:

Panda - Python: Pandas es una biblioteca del lenguaje de programación Python, dedicada exclusivamente a la Ciencia de Datos. Proporciona estructuras de datos y herramientas de análisis de alto rendimiento y fáciles de usar. Es esencial para el manejo y análisis de datos en Python McKinney, W. (2022).

Jupyter Notebook: es una herramienta interactiva ampliamente utilizada para el análisis y visualización de datos. Permite combinar código, texto, ecuaciones y visualizaciones en un único documento, convirtiéndolo en un recurso importante para explorar y comprender información compleja McKinney, W. (2022).

3.3.1 Descripción de la base de datos (Dataset)

En la Tabla 5 se muestran los datos esenciales de la base de datos obtenida desde el CEN. Los datos consisten en filas de generadoras con su producción en MWh durante un día.

Tabla 5: Información de la base de datos

Título:	Generación Eléctrica por Hora en Centrales Generadoras entre los años 2020 y 2023
Resumen:	Este dataset contiene información detallada sobre la generación de energía eléctrica por hora en diversas centrales generadoras. Los datos incluyen el tipo de tecnología utilizada, la región geográfica, el tipo de combustible empleado y la generación de energía en MWh para cada hora del día.
Fuente de los Datos:	Los datos fueron obtenidos del Coordinador Eléctrico Nacional de Chile y están disponibles en su sitio web oficial.
Fecha de Recopilación:	La información fue recopilada el 12 de abril de 2024. Los datos cubren el periodo comprendido entre los años 2020 y 2023.
Alcance y Cobertura:	Geográfico: Chile Temporal: Datos horarios con cobertura a nivel diario
Formato de los Datos:	El dataset está en formato CSV (Comma-Separated Values).
Estructura del Dataset:	Número de registros: 1.373.907 Número de columnas: 33

Fuente: elaboración propia

A continuación, la Tabla 6 muestra el detalle de la base de datos en cuanto a sus columnas y las correspondientes descripciones:

Tabla 6: Detalles de la base de datos

Campo	Tipo de Dato	Diccionario de Datos
NOMBRE CENTRAL	Texto	Nombre de la central generadora.
LLAVE NOMBRE	Texto	Identificador único de la central generadora.
TIPO	Texto	Tipo de tecnología de la central generadora (e.g., hidroeléctrica, térmica).

SUBTIPO	Texto	Tipo de combustible utilizado por la central generadora (si aplica).
REGIÓN	Texto	Región donde se ubica la central generadora.
ERNC/Convencional	Texto	Indica si la central es de Energías Renovables No Convencionales (ERNC) o convencional.
Factor ERNC	Float	Factor de conversión ERNC valores de 1 y 0.
FECHA	Fecha	Fecha de la generación (YYYY-MM-DD).
HORA 1 a HORA 24	Float	Generación de energía eléctrica en MWh para las horas de 1 a 24 del día.
HORA 25	Float	Valores raros o errores en la recopilación de datos.
TOTAL	Float	Total de energía generada en el día.

Fuente: elaboración propia

3.4 Análisis de la base de datos

Método de recopilación

Los datos fueron recopilados y reportados por el Coordinador Eléctrico Nacional a través de sistemas de monitoreo y reportes de las centrales generadoras.

Preprocesamiento y limpieza de los datos

La base de datos original contiene 1.373.907 registros distribuidos en 33 columnas, que incluyen tanto registros horarios como acumulados diarios de generación para cada día del período de estudio (2020-2023). Para este análisis, se optó por utilizar el total diario de generación (columna TOTAL) en lugar de los datos horarios (columnas HORA1 a HORA25). Esta decisión permite simplificar el análisis al enfocarse en patrones globales diarios, los cuales son suficientes para los objetivos planteados, como la caracterización de fuentes de generación y la agrupación de generadoras con comportamientos similares. Aunque esta elección implica la pérdida de granularidad, particularmente relevante en fuentes más variables como la solar, se consideró adecuada debido al enfoque del estudio y a la necesidad de reducir la complejidad computacional.

Dado que algunos datos presentaban valores faltantes o anómalos, se realizó un proceso de limpieza de datos para garantizar la calidad y utilidad de la información.

El preprocesamiento realizado fue el siguiente:

- Los valores NaN en la columna TOTAL fueron reemplazados por la sumatoria de los valores de las columnas HORA1 a HORA25.
- En la columna REGIÓN, se eliminaron las filas sin valor.
- En la columna TIPO, se eliminaron las filas sin valor.

Tras agrupar los datos utilizando la columna LLAVE NOMBRE, se obtuvo un total de 1.419 generadoras. No obstante, algunas de estas generadoras no operaron durante el período de interés (2020-2023), por lo que se decidió trabajar únicamente con aquellas que tuvieron producción durante dicho período. Esto resultó en un total de 623 generadoras, componiéndose cada una de un registro de generadora con su serie temporal diaria de producción dentro del período de tiempo del estudio.

Todos los cambios fueron guardados en un archivo CSV, que se utilizó para el análisis, ya que este formato facilita la manipulación y análisis de datos tabulares.

Licencia y Restricciones de Uso

Los datos están disponibles públicamente a través del Coordinador Eléctrico Nacional de Chile. Es recomendable consultar las políticas de uso del sitio web para asegurarse del cumplimiento de cualquier restricción o requisito de atribución.

Contacto

Para más información, se puede contactar al Coordinador Eléctrico Nacional de Chile a través de su sitio web. <https://www.coordinador.cl>

3.5 Métricas de distancia, desempeño y visualización

Las métricas de distancia, desempeño y visualización son herramientas esenciales para evaluar la calidad y la efectividad de los algoritmos de clustering. Estas métricas permiten interpretar los resultados obtenidos, medir la proximidad entre elementos, y comprender la estructura subyacente en los datos. Entre las métricas más comunes se encuentran las medidas de distancia, las evaluaciones de cohesión y separación, y los métodos de representación gráfica que facilitan el análisis visual de los clústeres generados.

3.5.1 Distancia Euclidiana

La distancia Euclidiana es una de las métricas más simples y comunes para medir la proximidad entre dos puntos en un espacio geométrico (Liberti, L., & Lavor, C. (2017)). En el contexto de algoritmos como *K*-Medias, esta métrica se utiliza para calcular la distancia entre puntos y los centroides de los clústeres.

La fórmula para calcular la distancia Euclidiana en un espacio *n*-dimensional es la indicada en la fórmula (1):

$$d(p, q) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}. \quad (1)$$

Donde *p* y *q* son dos puntos en el espacio, y q_i y p_i son las coordenadas de los puntos en la dimensión *i*. La distancia Euclidiana es una medida de la "línea recta" entre dos puntos.

3.5.2 Dynamic Time Warping (DTW)

El Dynamic Time Warping (DTW) es una métrica de similitud ampliamente utilizada en series temporales que permite medir la distancia entre dos secuencias, incluso cuando estas tienen diferentes longitudes, variaciones en su escala temporal o desfases. Esta flexibilidad lo hace especialmente útil en el

análisis de datos energéticos, donde los patrones temporales pueden variar debido a factores como las condiciones climáticas o las diferencias en la operación de las centrales generadoras (Müller, 2007).

Definición de DTW:

Dadas dos series temporales $X = (x_1, x_2, \dots, x_n)$ y $Y = (y_1, y_2, \dots, y_m)$, DTW encuentra una alineación no lineal entre ambas que minimiza la distancia acumulativa, lo que se muestra en la fórmula (2):

$$DTW(X, Y) = \min_P \sum_{(i,j) \in P} d(x_i, y_j). \quad (2)$$

Donde:

P es una ruta válida que respeta las restricciones de monotonía (no retrocede en el tiempo) y continuidad (pasos consecutivos) (Müller, 2007).

$d(x_i, y_j)$ es la distancia local entre los puntos x_i y y_j , comúnmente la distancia Euclidiana.

La matriz de costos acumulativos $D(i, j)$ se calcula recursivamente (Berndt & Clifford, 1994), lo que se muestra en la fórmula (3):

$$D(i, j) = d(x_i, y_j) + \min \begin{cases} D(i-1, j) \\ D(i, j-1) \\ D(i-1, j-1) \end{cases} \quad (3)$$

La solución es la ruta P que minimiza $D(n, m)$, siendo n y m las longitudes de X e Y , respectivamente.

3.5.3 Método de Ward

El método de Ward minimiza la varianza total dentro de los clústeres. La idea principal es que, en cada fusión, se seleccionan los clústeres cuya combinación resulta en el menor aumento en la suma de cuadrados de las diferencias dentro

de los clústeres. Este enfoque busca generar clústeres lo más homogéneos posible (Ward, J. H. 1963).

La fórmula utilizada por Ward para calcular la distancia entre dos clústeres C_i y C_j se muestra en la fórmula (4):

$$d_{Ward}(C_i, C_j) = \frac{|C_i||C_j|}{|C_i|+|C_j|} \times \|\mu_i - \mu_j\|^2, \quad (4)$$

donde:

$|C_i|$ es el tamaño del clúster C_i ,

μ_i y μ_j son las medias de los clústeres C_i y C_j ,

$\|\mu_i - \mu_j\|^2$ es la distancia Euclidiana entre las medias de los clústeres.

El objetivo de este método es reducir la varianza interna dentro de cada clúster, favoreciendo agrupaciones compactas y homogéneas.

3.5.4 Método de Enlace Completo (Complete Linkage)

El método de enlace completo, o Complete Linkage, calcula la distancia entre dos clústeres como la máxima distancia entre cualquier par de puntos, uno de cada clúster. En otras palabras, considera la mayor distancia entre todos los puntos de los clústeres que se están comparando (Defays, D. 1977).

La fórmula para calcular la distancia completa entre dos clústeres C_i y C_j es la que se muestra en la fórmula (5):

$$d_{complete}(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y), \quad (5)$$

donde $d(x, y)$ es la distancia entre un punto x del clúster C_i y un punto y del clúster C_j .

3.5.5 WSS para Selección del K

El método del Within-Cluster Sum of Squares (WSS) es una técnica comúnmente utilizada para seleccionar el número óptimo de clústeres K en algoritmos de clustering como K -means. El WSS mide la compactación interna de los clústeres,

calculando la suma de las distancias cuadradas entre los puntos y sus respectivos centroides dentro de cada clúster. A medida que K aumenta, la WSS disminuye, ya que los clústeres tienden a ser más pequeños y más ajustados a los datos. Sin embargo, existe un punto en el que agregar más clústeres no reduce significativamente la WSS, lo que sugiere un valor óptimo de K . Este punto es conocido como la “rodilla” en la curva de WSS versus K , y su identificación puede realizarse visualmente mediante el método del codo (Elbow Method) (Kodinariya & Makwana, 2013).

El uso del WSS permite a los investigadores seleccionar un K que equilibre la complejidad del modelo y su capacidad para capturar patrones significativos en los datos sin sobreajustar el modelo. Este enfoque es particularmente efectivo en escenarios donde los datos son homogéneos o esféricos, ya que las distancias Euclidianas tienden a funcionar bien en estos casos (Tibshirani, Walther, & Hastie, 2001).

3.5.6 Predicción Strength

La Predicción Strength es una métrica estadística diseñada para evaluar la estabilidad y la validez de los resultados generados por algoritmos de clustering. Introducida por Tibshirani y Walther (2005), esta métrica tiene como objetivo medir la capacidad de un modelo de clustering para reproducir patrones consistentes cuando se aplican perturbaciones en los datos o cuando se entrena y valida en subconjuntos independientes del conjunto de datos.

Definición Matemática

La métrica de Predicción Strength (PS_k) para un número dado de clústeres k , lo que se muestra en la fórmula (6), se define como:

$$PS_k = \min_{1 \leq j \leq k} \text{mean}_{x_i \in C_j, x_j \in C_j} [I(C_i = C_j \mid \text{training}) \cdot I(C_i = C_j \mid \text{validation})]. \quad (6)$$

Metodología de cálculo

1. Dividir los datos en dos subconjuntos, uno de entrenamiento (D_{train}) y otro de validación (D_{val}).
2. Entrenar el modelo de clustering en D_{train} para obtener las asignaciones de clústeres.
3. Validar el modelo prediciendo las asignaciones de clústeres en D_{val} .
4. Comparar las asignaciones obtenidas en D_{train} y D_{val} para calcular PS_k .

Interpretación y aplicaciones

La Predicción Strength es particularmente útil en:

- Selección del número óptimo de clústeres (k): Observando cómo evoluciona PS_k a medida que aumenta k , se puede identificar un punto de saturación donde PS_k se estabiliza, indicando el valor óptimo de k (Tibshirani & Walther, 2005).
- Evitar el sobreajuste: En combinación con técnicas como PCA o autoencoders, asegura que los patrones detectados sean representativos y no artefactos del conjunto de datos original (Van der Maaten et al., 2009).
- Comparación de algoritmos de clustering: Permite evaluar cuál algoritmo ofrece mejor capacidad de generalización.

3.5.7 Índice de Silueta:

Este índice mide cuán similares son los objetos dentro de un mismo clúster en comparación con objetos de otros clústers. Sus valores van de -1 a 1, donde valores cercanos a 1 indican que los clústers están bien definidos y los puntos están más cerca de los puntos de su propio clúster que de otros clústers (Tan, P.-N., Steinbach, M., & Kumar, V., 2014, p. 535).

El coeficiente de Silueta es una métrica utilizada para evaluar la calidad del agrupamiento en algoritmos de clustering, ayudando a identificar el número óptimo de agrupamientos. Propuesto por Rousseeuw en 1987, este coeficiente es particularmente útil en algoritmos de aprendizaje no supervisado, donde la

cantidad de grupos puede ser un parámetro de entrada o determinado automáticamente por el algoritmo. En casos como el algoritmo *K-Mean*, el número óptimo de clúster debe ser determinado externamente, y el coeficiente de Silueta sirve como un indicador del número ideal de clústeres, siendo un valor más alto indicativo de un agrupamiento más adecuado (Rousseeuw, 1987).

La fórmula del coeficiente de Silueta para una observación i , lo que se muestra en la fórmula (7), denotado como $s(i)$, es:

$$s(i) = \frac{b-a}{\max\{a(i), b(i)\}}. \quad (7)$$

Donde:

- a : es el promedio de las disimilitudes (o distancias) de la observación i con las demás observaciones del clúster al que pertenece i .
- b es la distancia mínima a otro clúster que no es el mismo en el que está la observación i . Ese clúster es la segunda mejor opción para i y se lo denomina vecindad de i .

El valor de $s(i)$ puede ser obtenido combinando los valores de a y b como se muestra a continuación en la fórmula (8):

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & \text{si } a(i) < b(i) \\ 0 & \text{si } a(i) = b(i). \\ \frac{b(i)}{a(i)} - 1 & \text{si } a(i) > b(i) \end{cases} \quad (8)$$

El coeficiente de Silueta varia de $-1 \leq s(i) \leq 1$, con ello:

- Para que el coeficiente de Silueta sea cercano a 1 el valor de b tiene que ser mayor al de a . Esto significa que la distancia de la observación i a los clústers vecinos es suficientemente grande para que su pertenencia al clúster actual sea la correcta. Es decir, no es similar a sus vecinos.
- Un valor de $s(i)$ que sea cercano a cero nos va a indicar que la observación i está en la frontera de dos clústeres.

- Y si el valor de $s(i)$ es negativo, entonces la observación i debería ser asignada al clúster más cercano.

Resumiendo:

- $s(i) \approx 1$, la observación i está bien asignada a su clúster.
- $s(i) \approx 0$, la observación i está entre dos clústeres.
- $s(i) \approx -1$, la observación i está mal asignada a su clúster.

Finalmente es posible calcular el coeficiente de Silueta como el promedio de todos los $s(i)$ para todas las observaciones del conjunto de datos.

3.5.8 Visualización

Dendrograma: se describe en detalle en la sección 2.3.2.

t-SNE (t-Distributed Stochastic Neighbor Embedding): Es una técnica no lineal de reducción de dimensionalidad que proyecta datos de alta dimensión en un espacio de menor dimensión, generalmente en dos o tres dimensiones, preservando las relaciones de proximidad entre los puntos. Esto facilita la visualización de estructuras complejas y patrones en los datos (van der Maaten, L., & Hinton, G. 2008).

PCA (Análisis de Componentes Principales): Es una técnica lineal de reducción de dimensionalidad que transforma las variables originales en un nuevo conjunto de variables ortogonales llamadas componentes principales. Estas componentes capturan la mayor parte de la varianza presente en los datos, permitiendo una representación más sencilla y eficiente de la información (Amat Rodrigo, J. 2017).

3.6 Códigos

Los códigos utilizados en el desarrollo de este trabajo están disponibles en el siguiente repositorio de GitHub: <https://github.com/JMezaV2020/CGRE-SEN-AST-2020-2023>

CAPÍTULO IV: Análisis de Clústeres e hiperparámetros

En este capítulo, se utilizarán los datos seleccionados y procesados conforme al criterio de trabajo que se centra únicamente en aquellas generadoras que registraron actividad durante este periodo. Inicialmente, se contaba con un total de 1.419 generadoras de energía; sin embargo, tras un análisis exhaustivo de los datos, se identificaron solo 623 generadoras que presentaron registros consistentes de generación de energía durante el periodo de tiempo seleccionado. Este enfoque permite garantizar que el análisis posterior se base en datos relevantes y representativos del funcionamiento real del sistema eléctrico, facilitando la comparación de los métodos de agrupamiento utilizados y optimizando el tiempo de cómputo.

4.1 Análisis de hiperparámetros

El análisis de hiperparámetros es un paso crítico en el proceso de entrenamiento de modelos de agrupamiento y Deep learning, ya que estos parámetros influyen directamente en la calidad y el rendimiento de los resultados obtenidos. La selección adecuada de hiperparámetros puede optimizar el ajuste del modelo a los datos y mejorar la generalización, evitando problemas como el sobreajuste o el subajuste (Bergstra & Bengio, 2012).

En este estudio, se emplearán diferentes métodos para seleccionar los hiperparámetros más adecuados para cada algoritmo. En general, se utilizará la métrica de Suma de Errores Cuadráticos dentro del Clúster (WSS), que permite determinar el número óptimo de clústeres al medir la compactación de los mismos. Por otro lado, La métrica de Fuerza de predicción evalúa la estabilidad del agrupamiento, es decir, la consistencia de los clústeres obtenidos para diferentes subconjuntos de datos. Un valor alto de Fuerza de predicción indica clústeres robustos y coherentes frente a variaciones en los datos (Tibshirani & Walther, 2005).

4.1.1 Selección de hiperparámetros para K -means

Para seleccionar el valor óptimo de K , se realizó un análisis visual mediante el método del codo. Este método consiste en graficar el WSS en función de diferentes valores de K y buscar el punto en el que la disminución del WSS comienza a estabilizarse, formando un "codo" en la gráfica. Este punto indica que agregar más clústeres no resulta en una mejora significativa en la compactación de los clústeres, lo que sugiere que el número de clústeres seleccionado es adecuado (Kodinariya & Makwana, 2013).

Mientras que Fuerza de predicción, A medida que el valor de K aumenta, los valores de Fuerza de predicción también incrementan, sugiriendo una mejora en el rendimiento, especialmente al utilizar DTW (Dynamic Time Warping) como métrica de distancia.

Para esta configuración en la Figura 7 con $K = 6$, se identifica como un punto óptimo desde una perspectiva de WSS, ya que proporciona una buena compactación de los clústeres con una reducción significativa de la variabilidad intraclúster.

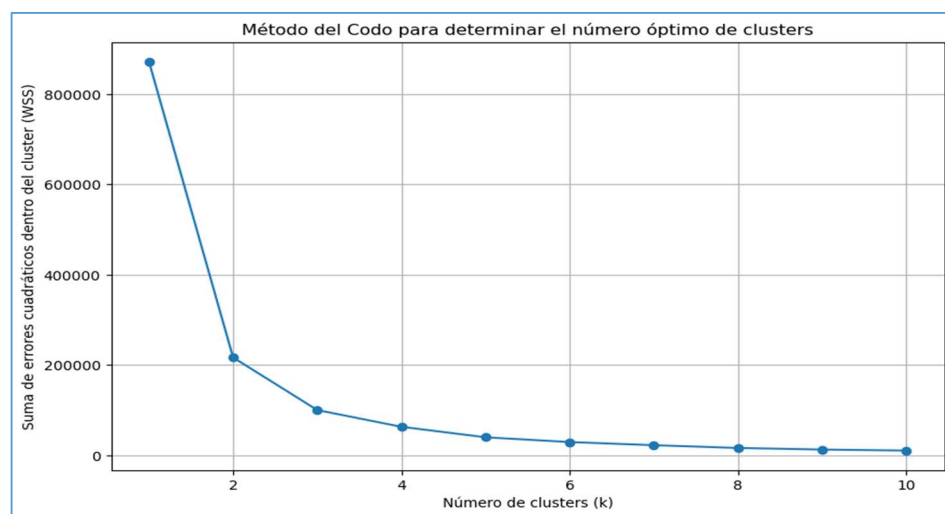


Figura 7: Resultados de hiperparámetros con WSS

Fuente: elaboración propia

Además, se evaluó la Fuerza de predicción, cuya evolución se ilustra en la Figura 8, como una métrica adicional para validar la estabilidad y consistencia de los clústeres formados. Este enfoque garantiza que la elección de K no solo sea óptima en términos de compactación, sino también en la capacidad de generalización del modelo. Esta Figura 8 muestra la relación entre la Fuerza de predicción y el número de clústeres para las métricas Euclidiana y DTW. Los resultados indican que, a medida que aumenta el número de clústeres, la Fuerza de predicción tiende a incrementarse, lo que sugiere una mejora en la robustez del agrupamiento. Sin embargo, es fundamental encontrar un equilibrio adecuado entre la Fuerza de predicción, el WSS y el tiempo de ejecución, ya que estos son factores determinantes al considerar el agrupamiento de patrones de consumo según tipo de energía. Además, dado el objetivo de obtener clústeres compactos y fácilmente interpretables, se busca seleccionar un valor de K lo más pequeño posible, manteniendo la robustez de los grupos formados.

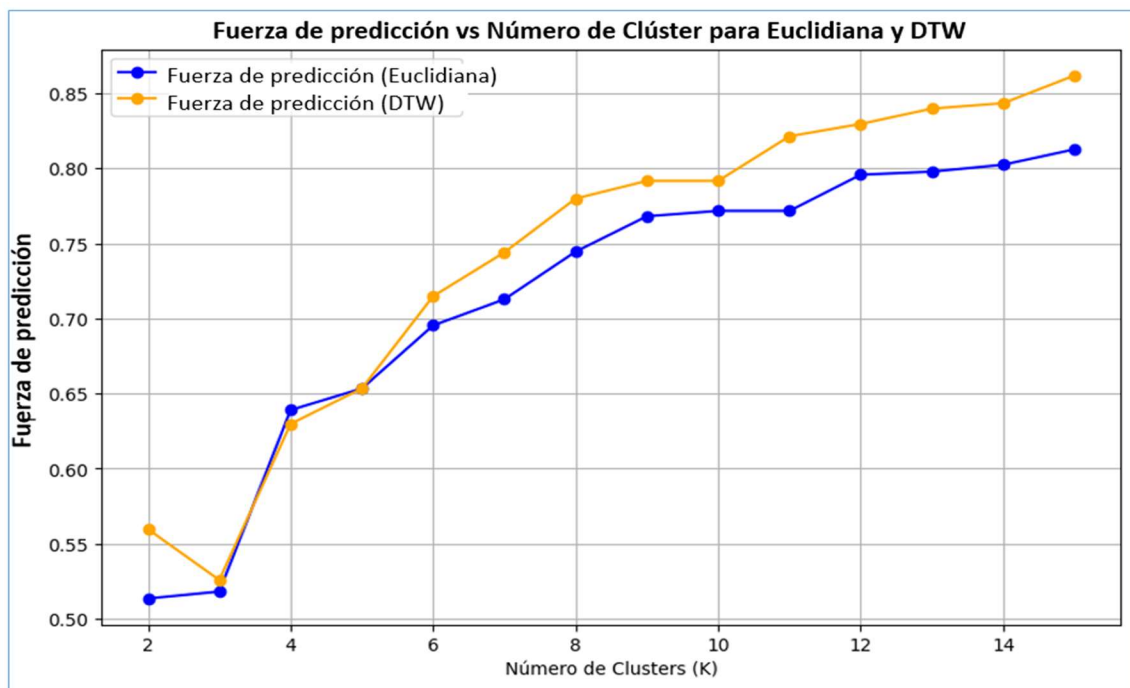


Figura 8: Resultados de hiperparámetros en métricas Euclidiana y DTW

Fuente: elaboración propia

A continuación, se presenta la Tabla 7, que resume los resultados de las métricas de distancia Euclidiana y DTW para diferentes valores de K en el análisis de agrupamiento. La tabla incluye dos métricas clave: la Fuerza de Predicción, que evalúa la estabilidad de los clústeres formados, y el Tiempo de Ejecución, que indica la eficiencia computacional del algoritmo.

Tabla 7: Resultados de las métricas Euclidiana y DTW

Métrica: EUCLIDIANA			Métrica: DTW		
K	Fuerza de predicción	Tiempo de ejecución (s)	K	Fuerza de predicción	Tiempo de ejecución (s)
2	0,51	0,51	2	0,56	2.877,92
3	0,52	1,70	3	0,53	1.238,63
4	0,64	0,46	4	0,63	1.789,04
5	0,65	0,46	5	0,65	1.700,53
6	0,70	0,50	6	0,71	1.596,36
7	0,71	0,45	7	0,74	1.758,30
8	0,74	0,94	8	0,78	1.904,82
9	0,77	0,64	9	0,79	2.498,27
10	0,77	0,64	10	0,79	2.453,86
11	0,77	0,76	11	0,82	4.338,88
12	0,80	0,70	12	0,83	3.302,93
13	0,80	0,64	13	0,84	3.124,68
14	0,80	0,60	14	0,84	2.823,97
15	0,81	0,61	15	0,86	2.946,85

Fuente: elaboración propia

El análisis de los resultados muestra que, a medida que aumenta el número de clústeres K , la Fuerza de Predicción tiende a incrementar, lo que sugiere una mejora en la robustez y consistencia de los clústeres obtenidos. Sin embargo, también se observa un incremento en el tiempo de ejecución, lo que pone de manifiesto la necesidad de balancear la complejidad del modelo y la calidad del

agrupamiento. Esta información es fundamental para decidir el número óptimo de clústeres que se utilizará en el análisis final, asegurando que se logre un agrupamiento eficiente y representativo.

Los resultados presentados en la Tabla 8 se correlacionan con la evolución de la Fuerza de predicción ilustrada en la Figura 6. Al analizar los datos y que resumimos para $K = 2, 6$ y 15 en la Tabla 8 es posible indicar:

Tabla 8: Comparativo de resultados de métricas con $K = 2, 6$ y 15

Métrica	K	Fuerza de predicción	Observaciones
Euclidiana	2	0,51	Estabilidad moderada en los clústeres formados.
	6	0,7	Mejora en la robustez del agrupamiento.
	15	0,81	Máximo valor, clústeres muy consistentes bajo variaciones.
DTW	2	0,56	Estabilidad inicial moderada.
	6	0,71	Incremento en la robustez del agrupamiento.
	15	0,86	Máximo valor, mayor robustez en comparación con Euclidiana.

Fuente: elaboración propia

Para la métrica Euclidiana:

Se observa que para $K = 2$, la Fuerza de predicción es de 0,51, lo que indica una estabilidad moderada en los clústeres formados. A medida que se incrementa el número de clústeres a 6, la Fuerza de predicción aumenta a 0,70, sugiriendo una mejora en la robustez del agrupamiento.

El valor máximo de Fuerza de predicción se alcanza con $K = 15$, donde se registra un 0,81, lo que implica que los clústeres son muy consistentes bajo variaciones de datos.

Para la métrica DTW:

El comportamiento es similar; comenzando con una Fuerza de predicción de 0,56 para $K = 2$, incrementa hasta 0,71 en $K = 6$. Este patrón de aumento continúa, alcanzando un valor máximo de 0,86 en $K = 15$. A lo largo de los valores de K , se puede notar que la métrica DTW presenta una mayor Fuerza de predicción en comparación con la métrica Euclidiana en los valores más altos de K .

Este análisis numérico resalta que tanto la métrica Euclidiana como la DTW muestran un incremento de la fuerza de predicción con el aumento de K , aunque la métrica DTW logra valores más altos de robustez en el agrupamiento a partir de $K = 10$ en comparación con Euclidiana, como se observa en la Tabla 8. Esto sugiere que la elección de la métrica de distancia puede influir significativamente en la calidad del agrupamiento y su estabilidad, lo que debe ser considerado al seleccionar el número óptimo de clústeres para el análisis. Debido a que la métrica DTW presenta una mejor estabilidad y robustez, se descartará la métrica Euclidiana para el análisis, priorizando la métrica DTW para asegurar resultados más consistentes y representativos en la identificación de patrones de generación de energía.

Para esta configuración, $K = 6$ se identifica como un punto óptimo desde la perspectiva de la Suma de Errores Cuadráticos dentro del Clúster (WSS), ya que proporciona una buena compactación de los clústeres y una reducción significativa de la variabilidad intra clúster. Aunque $K = 10$ también mostró un aumento en la Fuerza de predicción, el objetivo de este análisis es lograr clústeres compactos y fácilmente interpretables, lo que hace que la selección de un valor de K más pequeño, como $K = 6$, sea preferible. Esto es fundamental en el contexto del análisis de la generación de energía, donde la simplicidad y la claridad en la interpretación de los clústeres son esenciales para la toma de decisiones informadas sobre la gestión y optimización del sistema.

4.1.2 Resultados de hiperparámetros para AHC

En este apartado, se analizarán los hiperparámetros utilizados para el Agrupamiento Jerárquico Aglomerativo (AHC), detallando los métodos de enlace empleados y los tiempos de ejecución asociados. Se presentan los resultados en la Figura 9, que muestra la Fuerza de predicción en función del número de clústeres para los métodos Ward y Complete.

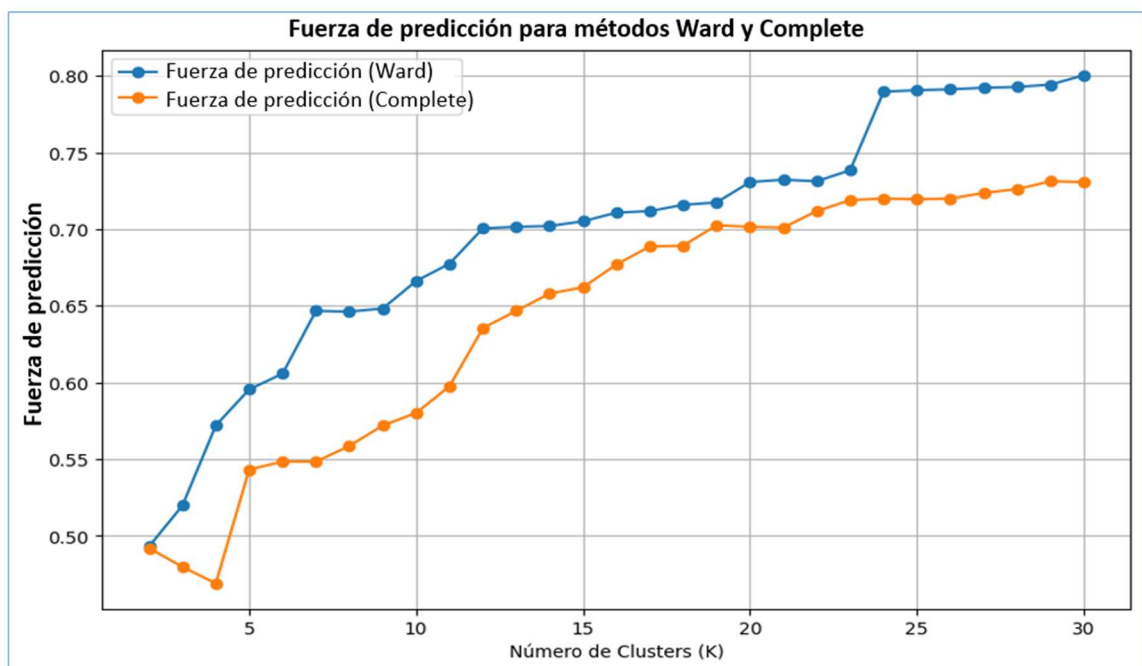


Figura 9: Resultados de Hiperparámetros en métricas Ward y Complete

Fuente: elaboración propia

La Tabla 9 a continuación proporciona una comparación de la Fuerza de Predicción y el tiempo de ejecución para ambos métodos a medida que se varía el número de clústeres K :

Tabla 9: Resultados de hiperparámetros para AHC

Método: Ward

<i>K</i>	Fuerza de Predicción	Tiempo de Ejecución (segundos)
2	0,49	0,20
3	0,52	0,10
4	0,57	0,18
5	0,60	0,18
6	0,61	0,18
7	0,65	0,18
8	0,65	0,18
9	0,65	0,18
10	0,67	0,19
11	0,68	0,18
12	0,70	0,18
13	0,70	0,18
14	0,70	0,18
15	0,70	0,18

Método: Complete

<i>K</i>	Fuerza de Predicción	Tiempo de Ejecución (segundos)
2	0,49	0,20
3	0,48	0,20
4	0,47	0,20
5	0,54	0,19
6	0,55	0,18
7	0,55	0,18
8	0,56	0,18
9	0,57	0,18
10	0,58	0,18
11	0,60	0,18
12	0,64	0,18
13	0,65	0,18
14	0,66	0,18
15	0,66	0,18

Fuente: elaboración propia

Los resultados muestran que tanto el método de enlace Ward como el método Complete presentan una fuerza de predicción similar en los valores más bajos de *K*, con un aumento progresivo a medida que el número de clústeres se incrementa. No obstante, el método Ward destaca por producir agrupamientos más compactos y mejor separados, proporcionando una mayor estabilidad en comparación con el método Complete, que no logra un desempeño satisfactorio y es descartado para este análisis.

A medida que se exploran valores de K mayores, la Fuerza de predicción continúa mejorando, alcanzando un nivel de 0.80 para $K = 30$ y sugiriendo que un mayor número de clústeres podría seguir incrementando la estabilidad del agrupamiento. Sin embargo, optar por un número de clústeres tan alto puede resultar en una representación excesivamente fragmentada y menos interpretable, lo que comprometería el objetivo de ofrecer una estructura generalizable y útil de los datos.

Además, al observar $K = 12$, se nota un valle en la Fuerza de predicción, indicando una posible reducción de estabilidad en este punto. Por este motivo, seleccionar un valor de $K = 10$ se presenta como una opción intermedia, proporcionando una Fuerza de predicción aceptable sin fragmentar en exceso los datos.

Por lo que, la elección de $K = 10$ balancea adecuadamente la calidad del agrupamiento y la interpretabilidad de los resultados. Aunque un valor de K mayor podría ofrecer agrupamientos aún más estables, la selección de $K = 10$ facilita un análisis claro y útil, alineado con el objetivo de obtener una estructura comprensible de los datos.

4.1.3 Resultados de hiperparámetros para Autoencoders LSTM

En esta sección, se analizan diferentes combinaciones de hiperparámetros para optimizar el rendimiento del autoencoder LSTM. La mejor configuración se estableció como:

- Tamaño de capa oculta: 128.
- Número de capas: 2.
- Tasa de aprendizaje: 0.001.
- Tasa de dropout: 0.3.

En la Figura 10 se ilustra la minimización de la función de pérdida durante el entrenamiento, junto con el comportamiento del autoencoder LSTM en relación con el número de epoch. En ella aparece el comportamiento del autoencoder LSTM, donde se evaluaron diferentes combinaciones de hiperparámetros para determinar la configuración óptima. Los hiperparámetros seleccionados, {'hidden_size': 128, 'num_layers': 2, 'learning_rate': 0.001, 'dropout_rate': 0.3}, demostraron un rendimiento superior en términos de minimización de la función de pérdida (loss).

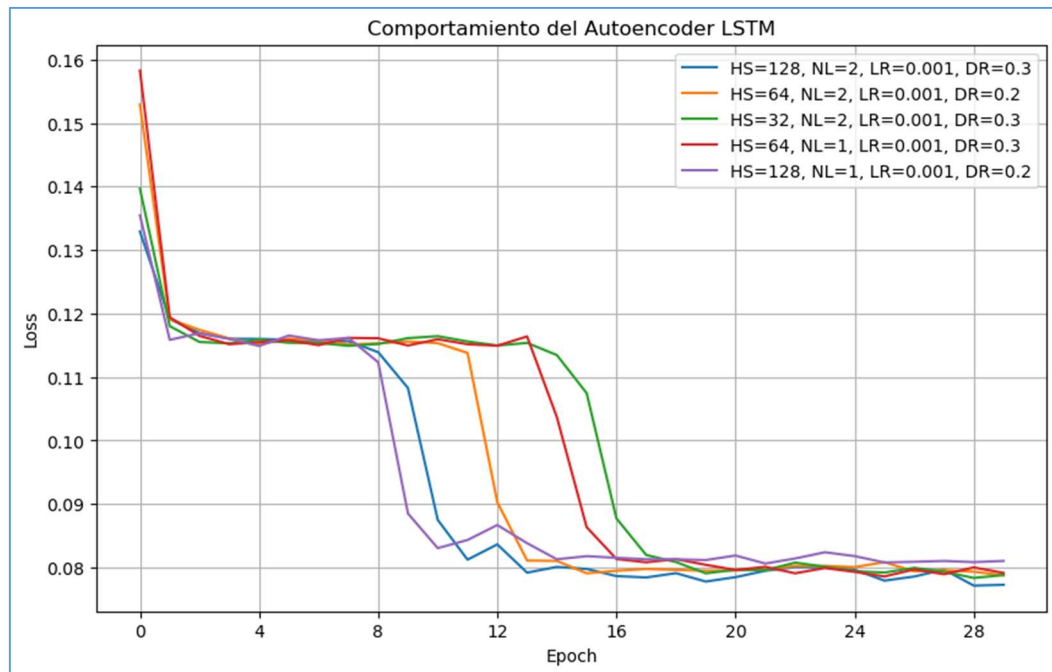


Figura 10: Comportamiento del autoencoder LSTM vs Epoch

Fuente: elaboración propia

En el gráfico de la Figura 10 se observa que esta configuración (línea azul) logra una disminución rápida y consistente de la pérdida, alcanzando los valores más bajos de loss en un número razonable de épocas sin mostrar oscilaciones o sobreajuste evidentes. La combinación de un tamaño de capa oculta de 128, con 2 capas y una tasa de dropout de 0.3, parece ser adecuada para capturar las

características de la serie temporal, manteniendo al mismo tiempo una estabilidad en el aprendizaje.

La elección adecuada de hiperparámetros, como el tamaño de la capa oculta y la tasa de dropout, es fundamental para evitar el sobreajuste y mejorar la capacidad de generalización en redes neuronales. Según Srivastava et al. (2014), el uso de dropout durante el entrenamiento de redes neuronales profundas previene el sobreajuste y mejora la generalización del modelo.

A continuación, se presentan los resultados de los hiperparámetros para el K -means aplicado en el espacio latente del LSTM. En la Figura 11 se muestra la relación entre la Fuerza de Predicción y los clústeres K utilizando la métrica Euclidiana, mientras que en la Figura 12 se ilustra esta misma relación empleando la métrica DTW, ambos en el contexto del autoencoder LSTM. En estas figuras es posible observar que la fuerza de predicción es relativamente alta para valores bajos de K ($K = 2$ a $K = 6$), manteniéndose alrededor de 0.75 - 0.80, lo cual sugiere que el modelo está generando clústeres bien definidos y estables en esta región. Sin embargo, al llegar a $K = 8$, la Fuerza de predicción cae drásticamente a aproximadamente 0.50, indicando que en este punto el modelo pierde estabilidad en el agrupamiento. Esto puede suceder porque el incremento en el número de clústeres sobrepasa la capacidad del modelo para encontrar patrones claros, lo que genera una mayor fragmentación y menor consistencia en los clústeres.

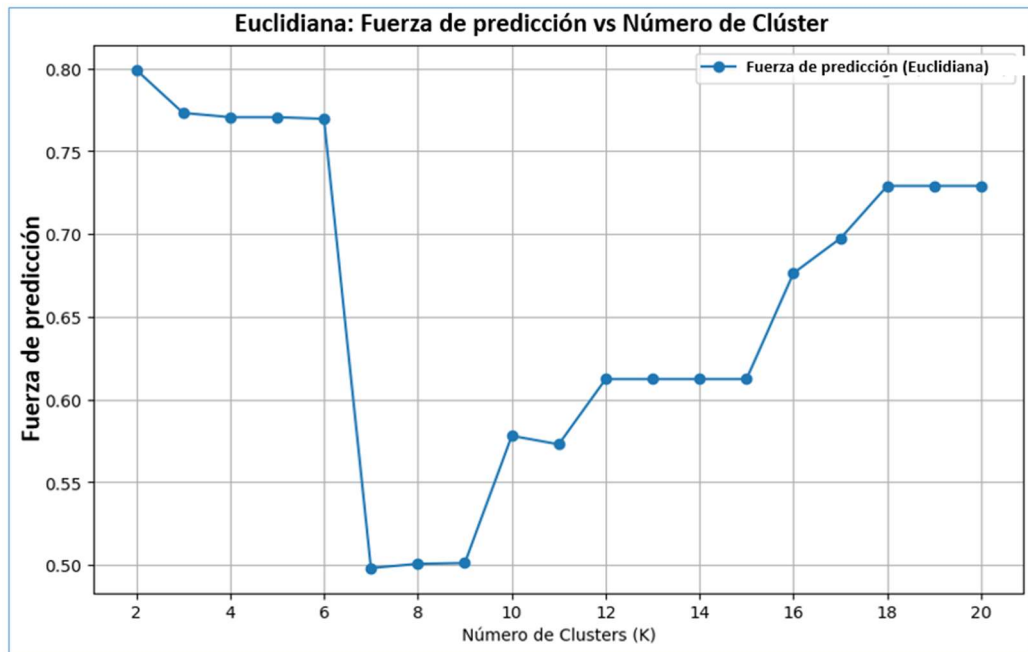


Figura 11: Fuerza de predicción vs clúster en autoencoder LSTM (Euclidiana)

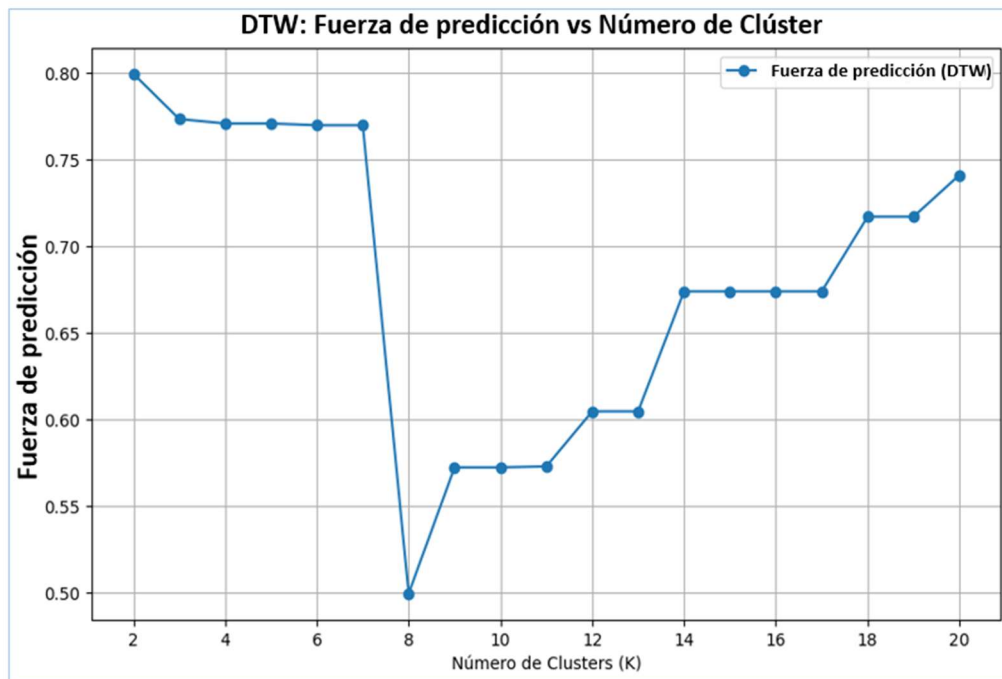


Figura 12: Fuerza de predicción vs clúster en autoencoder LSTM (DTW)

Después de este punto, se observa un aumento gradual y constante en la Fuerza de predicción conforme K aumenta de 10 a 20. Esto indica que, a medida que se

agregan más clústeres, el modelo recupera parte de la estabilidad en los agrupamientos. Aunque no alcanza los niveles más altos observados al inicio, esta tendencia ascendente sugiere que un mayor número de clústeres podría proporcionar agrupamientos más estables, pero también introduce una mayor complejidad en la interpretación y el riesgo de sobreajuste.

Es importante señalar que, al comparar las métricas de distancia, la métrica Euclidiana se descartó para este análisis debido a su comportamiento más abrupto: la caída en la Fuerza de predicción es mucho más rápida y la recuperación mucho más lenta que la observada con DTW. Esta diferencia en la dinámica de recuperación llevó a la elección de DTW como métrica de distancia más adecuada, ya que presenta una mayor estabilidad y un mejor desempeño en la identificación de patrones a medida que aumenta el número de clústeres.

Elegir un valor de $K = 6$ permite evitar esta zona de inestabilidad, aprovechando el rango estable de K sin caer en el comportamiento inconsistente observado en valores de K superiores.

4.2 Resultados del análisis de clústeres

El análisis de clústeres es una técnica fundamental en la minería de datos que permite agrupar un conjunto de objetos en grupos o clústeres, donde los objetos dentro de un mismo clúster son más similares entre sí que aquellos de diferentes clústeres. El análisis de clústeres es una técnica esencial en la generación de energía, ya que permite identificar patrones y agrupaciones significativas en los datos de consumo y producción. Según Hodge y Austin (2004), esta metodología facilita la exploración y comprensión de la estructura subyacente de los datos, lo cual es fundamental para la toma de decisiones informadas en diversas aplicaciones.

En este estudio, se aplicaron tres métodos de agrupamiento: *K*-means, Agrupamiento Jerárquico Aglomerativo (AHC) y LTSM. Se evaluaron las características de los clústeres resultantes, así como su interpretación y robustez. A continuación, se presentan los resultados del análisis utilizando el método *K*-means, seguido de los resultados del AHC y LTSM, lo que permite una comparación de la efectividad de cada técnica en la identificación de patrones en el Sistema Eléctrico Nacional (SEN).

4.2.1 Resultados de *K*-means

En la Tabla 10 se presenta la participación porcentual por tipo de generadora de energía en el Sistema Eléctrico Nacional (SEN). Esta información es fundamental para comprender la distribución de las fuentes de energía, lo que influye en el análisis de clústeres.

Tabla 10: Participación *porcentual por tipo de generadora*

Tipo	Total	Proporción
Solar	222	35,63%
Térmica	215	34,51%
Hidráulica	152	24,40%
Eólica	34	5,46%
	623	

Se observa que la generación solar representa el 35,63% del total, seguida por la generación térmica con un 34,51%. Las generadoras hidráulicas y eólicas tienen proporciones más bajas, con un 24,40% y un 5,46% respectivamente. Estos datos permitirán una evaluación más precisa del agrupamiento realizado y de los patrones de consumo asociados.

4.2.1.1 Parámetros y Ejecución DTW

- Métrica Utilizada: DTW.
- Tiempo de Ejecución: 4.737,20 segundos.
- Índice de Silueta: 0,26.

La Tabla 11 a continuación presenta la distribución de generadoras por clúster en cantidad y porcentaje con el método DTW

Tabla 11: Generadoras *por* clúster en cantidad y *porcentaje* DTW

Clúster	Generadoras	Térmica	Hidráulica	Solar	Eólica
0	95	27	34	34	0
1	175	155	14	5	1
2	105	10	23	72	0
3	107	1	2	104	0
4	46	10	0	3	33
5	95	12	79	4	0
	623	215	152	222	34
Clúster	Generadoras	Térmica	Hidráulica	Solar	Eólica
0	15,25%	28,42%	35,79%	35,79%	0,00%
1	28,09%	88,57%	8,00%	2,86%	0,57%
2	16,85%	9,52%	21,90%	68,57%	0,00%
3	17,17%	0,93%	1,87%	97,20%	0,00%
4	7,38%	21,74%	0,00%	6,52%	71,74%
5	15,25%	12,63%	83,16%	4,21%	0,00%
		34,51%	24,40%	35,63%	5,46%

Las Figuras 13 y 14 permiten visualizar los clústeres usando PCA y t-SNE respectivamente en métrica DTW.

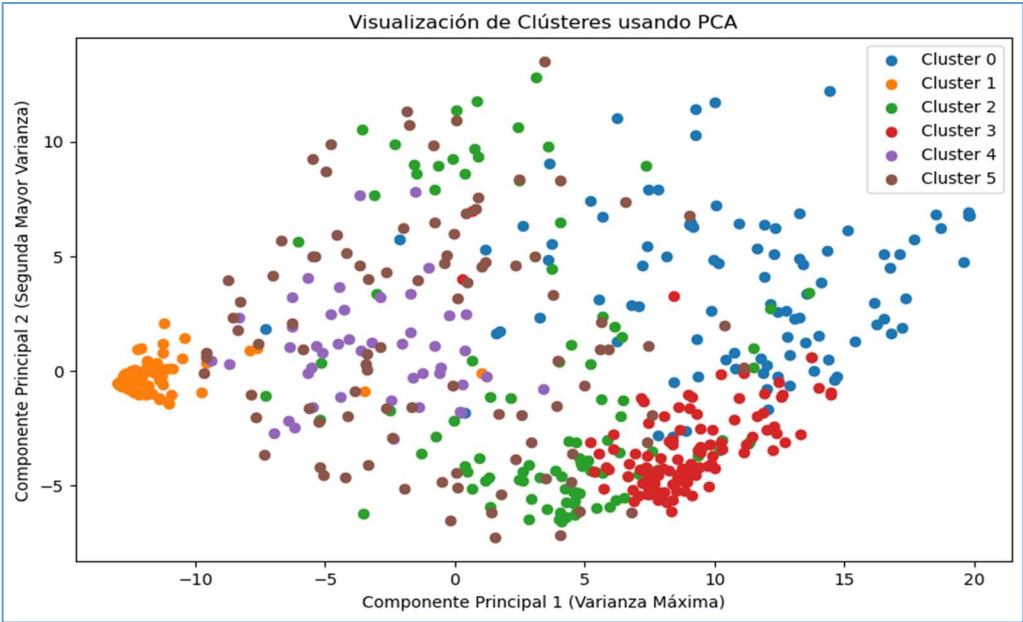


Figura 13: Generadoras por clúster con métrica DTW usando PCA

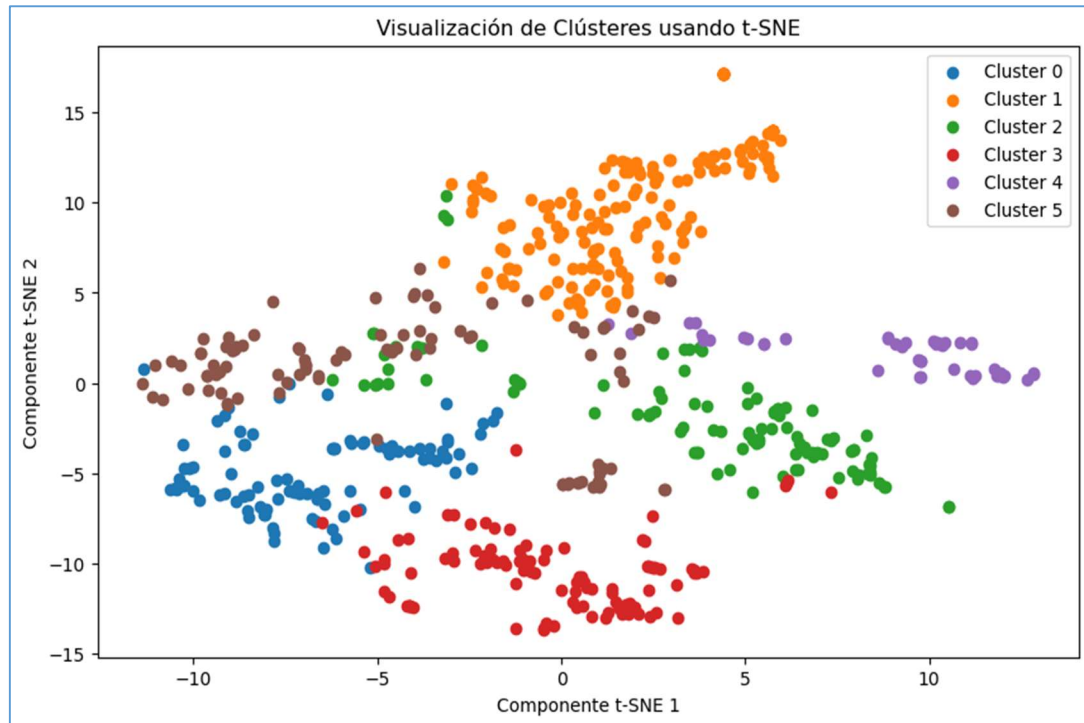


Figura 14: Generadoras por clúster con métrica DTW usando t-SNE

4.2.1 Resultados del agrupamiento jerárquico aglomerativo (AHC)

- Método Utilizado: Ward.
- Tiempo de Ejecución: 0,15 segundos.
- Índice de Silueta: 0,31.

Las Figura 15 muestra un dendrograma que representa el resultado de un agrupamiento jerárquico

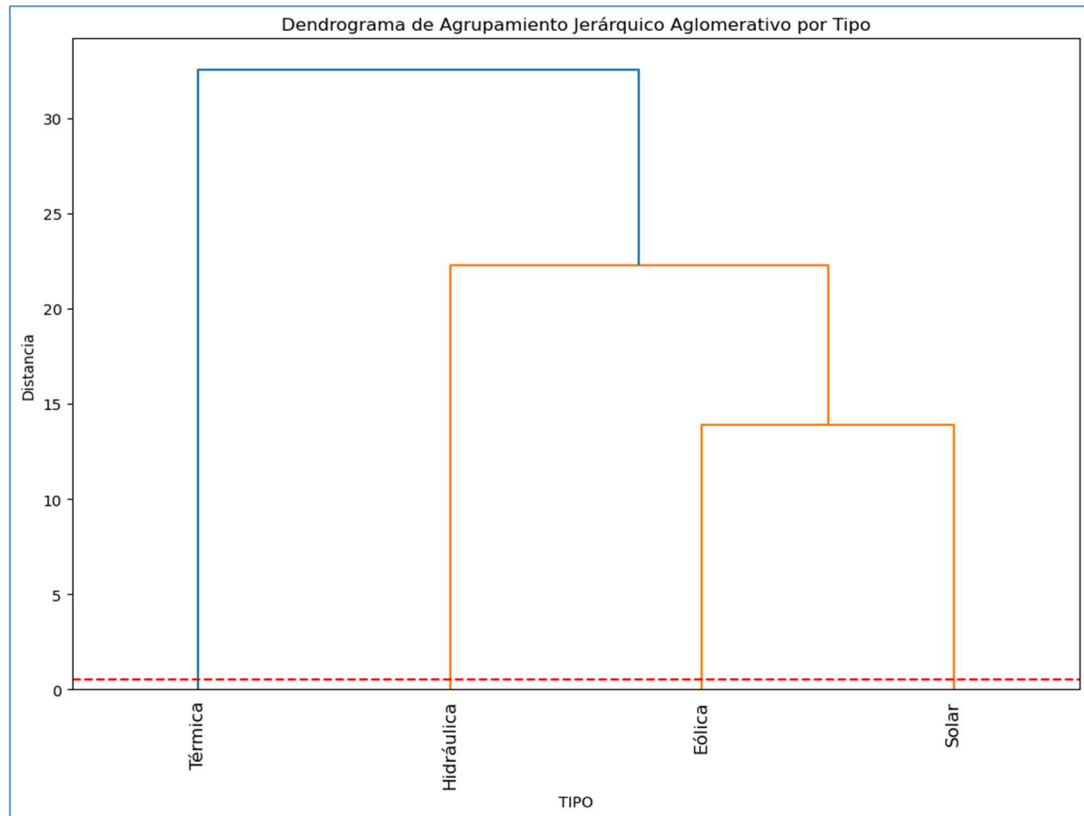


Figura 15: Dendrograma para agrupamiento jerárquico aglomerativo

La Tabla 12 muestra la distribución de generadoras por clúster en cantidad y porcentaje con AHC utilizando método Ward.

Tabla 12: Generadoras por clúster en cantidad y porcentaje AHC

Clúster	Generadoras	Térmica	Hidráulica	Solar	Eólica
0	12	8	3	1	0
1	63	22	15	26	0
2	67	2	4	61	0
3	37	3	33	1	0
4	122	0	4	118	0
5	172	156	13	3	0
6	63	2	61	0	0
7	18	7	6	5	0
8	55	15	13	7	20
9	14	0	0	0	14
	623	215	152	222	34

Clúster	Generadoras	Térmica	Hidráulica	Solar	Eólica
0	1,93%	66,67%	25,00%	8,33%	0,00%
1	10,11%	34,92%	23,81%	41,27%	0,00%
2	10,75%	2,99%	5,97%	91,04%	0,00%
3	5,94%	8,11%	89,19%	2,70%	0,00%
4	19,58%	0,00%	3,28%	96,72%	0,00%
5	27,61%	90,70%	7,56%	1,74%	0,00%
6	10,11%	3,17%	96,83%	0,00%	0,00%
7	2,89%	38,89%	33,33%	27,78%	0,00%
8	8,83%	27,27%	23,64%	12,73%	36,36%
9	2,25%	0,00%	0,00%	0,00%	100,00%
		34,51%	24,40%	35,63%	5,46%

Las Figuras 16 y 17 permiten visualizar los clústeres usando PCA y t-SNE respectivamente en métrica WARD.

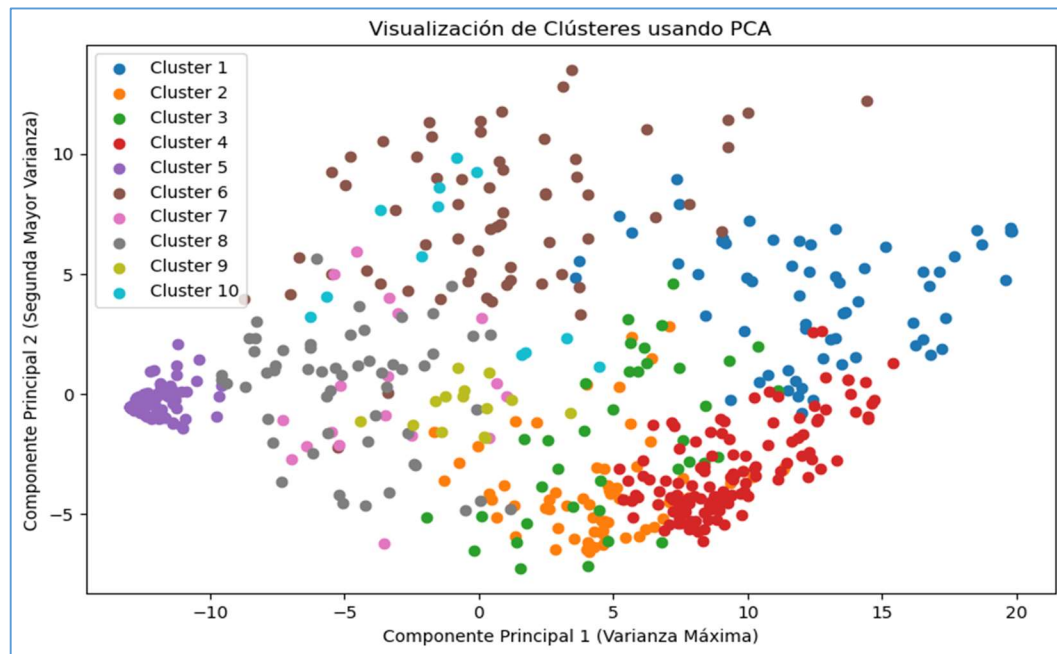


Figura 16: Generadoras por clúster con método AHC usando PCA

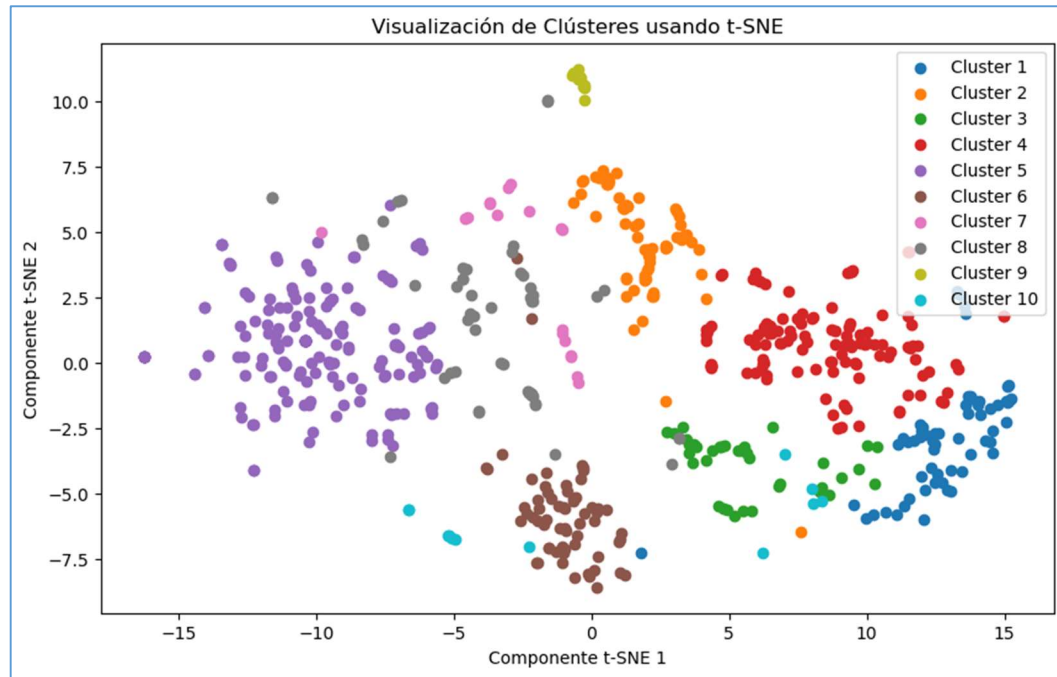


Figura 17: Generadoras por clúster con método AHC usando t-SNE

4.2.3 Resultados del Autocodificador LSTM DTW

- Tiempo de entrenamiento del Autocodificador LSTM: 32,13 segundos.
- Tiempo de ejecución para *K*-means (DTW): 17,82 segundos.
- Índice de Silueta con DTW: 0,22.

En la Tabla 13 se presenta la distribución de generadoras por clúster en cantidad y porcentaje con Autocodificador LSTM DTW.

Tabla 13: Generadoras por clúster en cantidad y porcentaje LSTM – DTW

Clúster	Generadoras	Térmica	Hidráulica	Solar	Eólica
0	125	21	34	65	5
1	161	139	8	4	10
2	109	17	30	55	7
3	84	12	29	37	6
4	98	20	32	42	4
5	46	6	19	19	2
	623	215	152	222	34

Clúster	Generadoras	Térmica	Hidráulica	Solar	Eólica
0	20,06%	16,80%	27,20%	52,00%	4,00%
1	25,84%	86,34%	4,97%	2,48%	6,21%
2	17,50%	15,60%	27,52%	50,46%	6,42%
3	13,48%	14,29%	34,52%	44,05%	7,14%
4	15,73%	20,41%	32,65%	42,86%	4,08%
5	7,38%	13,04%	41,30%	41,30%	4,35%
		34,51%	24,40%	35,63%	5,46%

Las Figuras 18 y 19 permiten visualizar los clústeres usando PCA y t-SNE respectivamente con LSTM DTW.

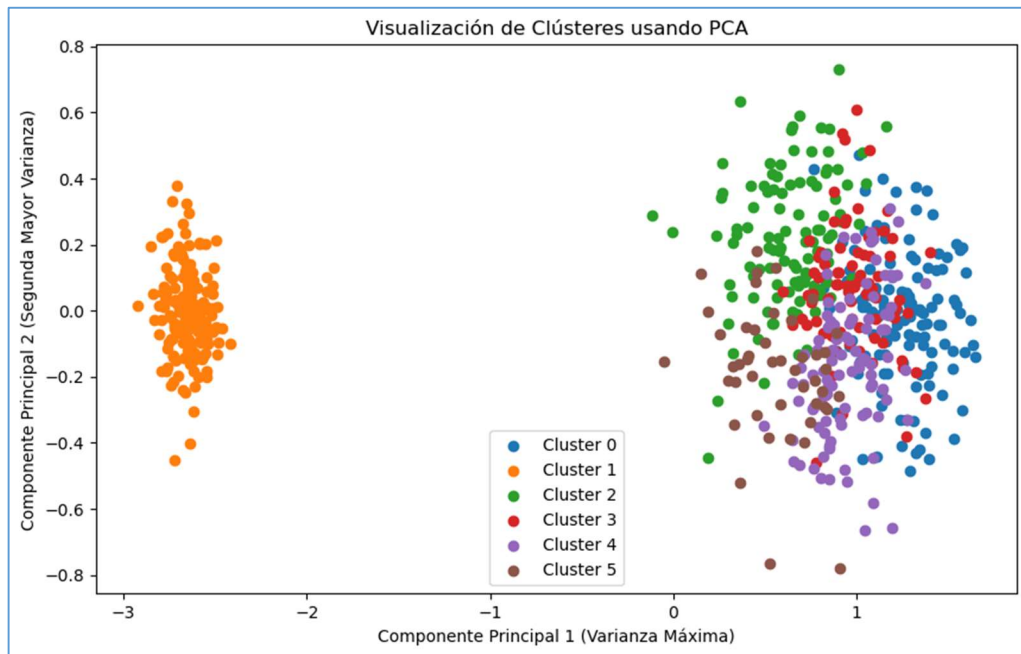


Figura 18: Generadoras por clúster, método LSTM DTW usando PCA

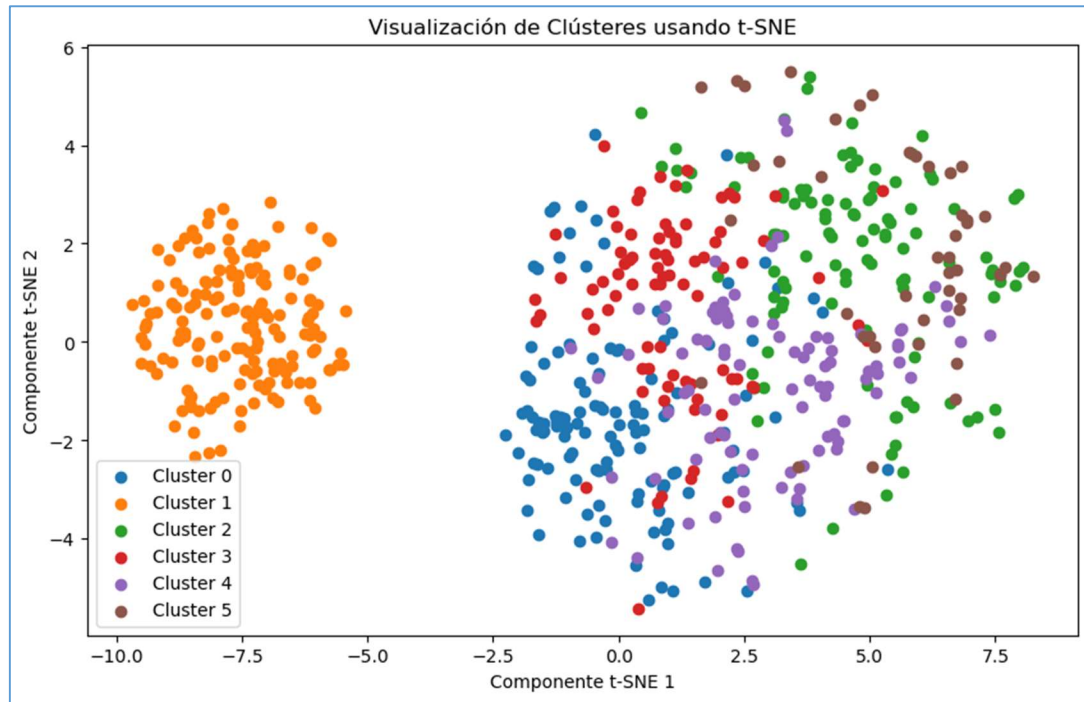


Figura 19: Generadoras por clúster, método LSTM DTW usando t-SNE

4.3 Análisis de Resultados

En esta sección se presentan los hallazgos derivados del análisis de clústeres aplicados a los datos de generación de energía del Sistema Eléctrico Nacional (SEN) de Chile. Se utilizaron tres métodos principales: *K*-means, Agrupamiento Jerárquico Aglomerativo (AHC) y Autoencoder LSTM, aplicados a los datos de generación de energía del periodo 2020-2023. Estos métodos se evaluaron en términos de cohesión y estabilidad de los clústeres, utilizando el Índice de Silueta y la Fuerza de Predicción para comparar y validar los resultados.

4.3.1 Resultados de *K*-means

El método *K*-means con métrica DTW, su Fuerza de Predicción alcanzo un valor de 0.86 con $K = 15$, lo cual sugiere una mayor robustez del agrupamiento bajo esta métrica. En la Tabla 14 se presenta la distribución porcentual de generadoras por clúster con métrica DTW, en ella se observa que la aglomeración porcentual por clúster de las generadoras para el tipo térmica es

de 72,09% en el clúster 1, mientras que la eólica en el clúster 4 es de 97,06% y la hidráulica en el clúster 5 es de 51,97%, lo que indica que son valores representativos, por otra parte, la solar se separa en dos clústeres representativos 2 y 3 con 32,43% y 46,85% respectivamente.

Tabla 14: Distribución porcentual de generadoras por clúster con *DTW*

Clúster	Térmica	Hidráulica	Solar	Eólica
0	12,56%	22,37%	15,32%	0,00%
1	72,09%	9,21%	2,25%	2,94%
2	4,65%	15,13%	32,43%	0,00%
3	0,47%	1,32%	46,85%	0,00%
4	4,65%	0,00%	1,35%	97,06%
5	5,58%	51,97%	1,80%	0,00%

Es importante señalar que el valor del eje *Y* no tiene una unidad específica y debe interpretarse como el grado de disimilitud o la "fuerza" que refleja la diferencia entre los clústeres. Este valor indica cuán distintos son los clústeres entre sí en función de sus características, mostrando el nivel de separación en el comportamiento de las series temporales. A medida que el valor aumenta, la diferencia entre los grupos se hace más pronunciada, lo que implica que los clústeres son más distintos, mientras que valores bajos indican que los grupos son más similares entre sí, lo que se observa en las Figuras 20 a la 24.

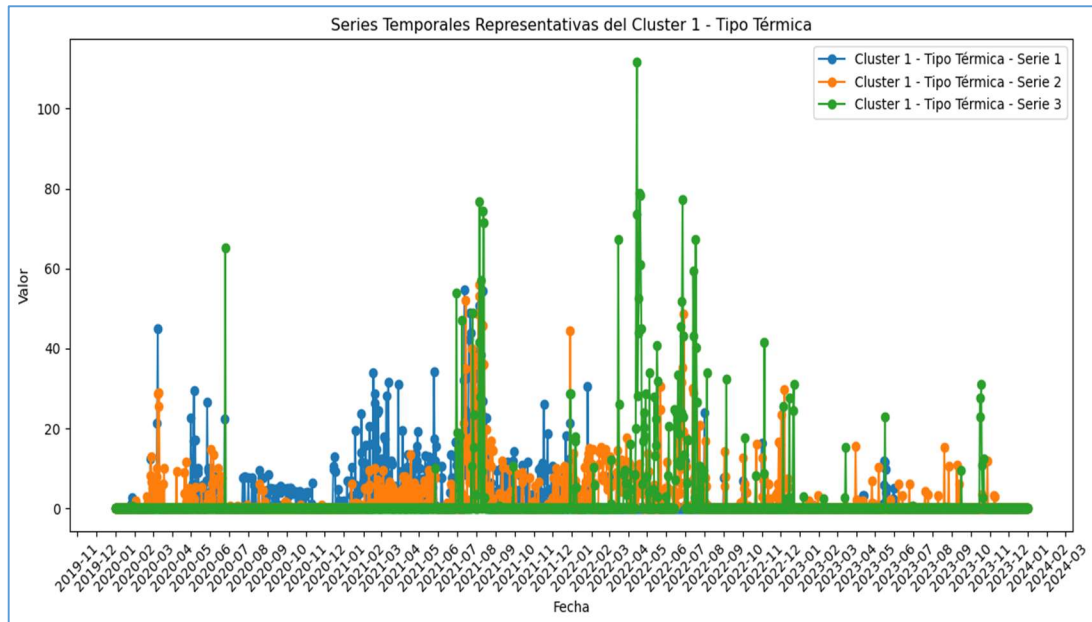


Figura 20: Curvas representativas del Clúster 1 para tipo Térmica

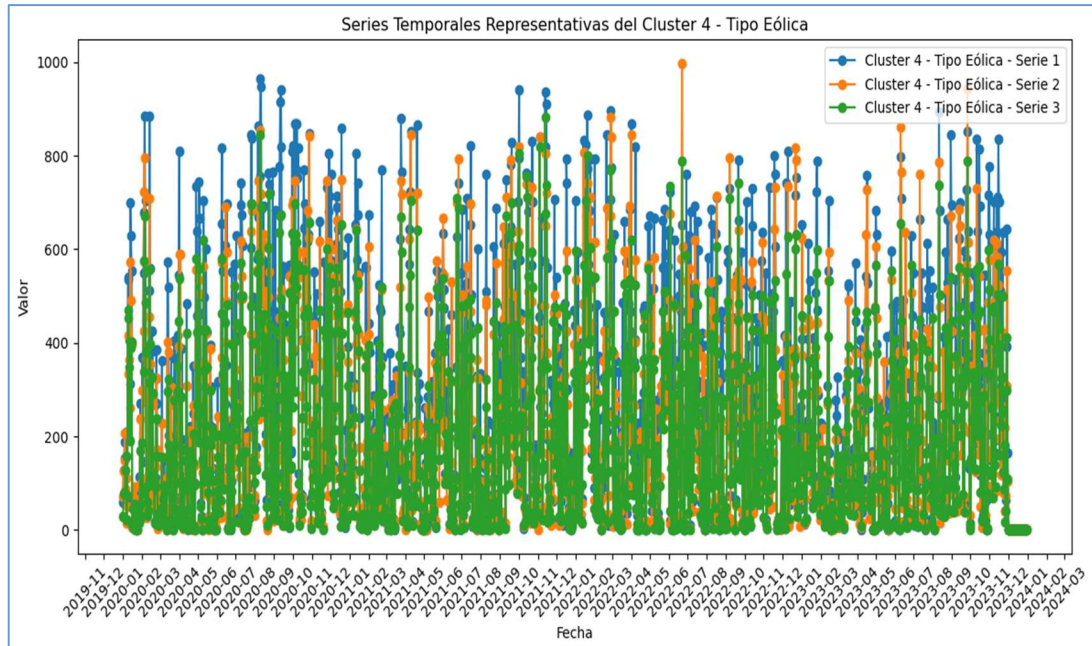


Figura 21: Curvas representativas del Clúster 4 para tipo Eólica

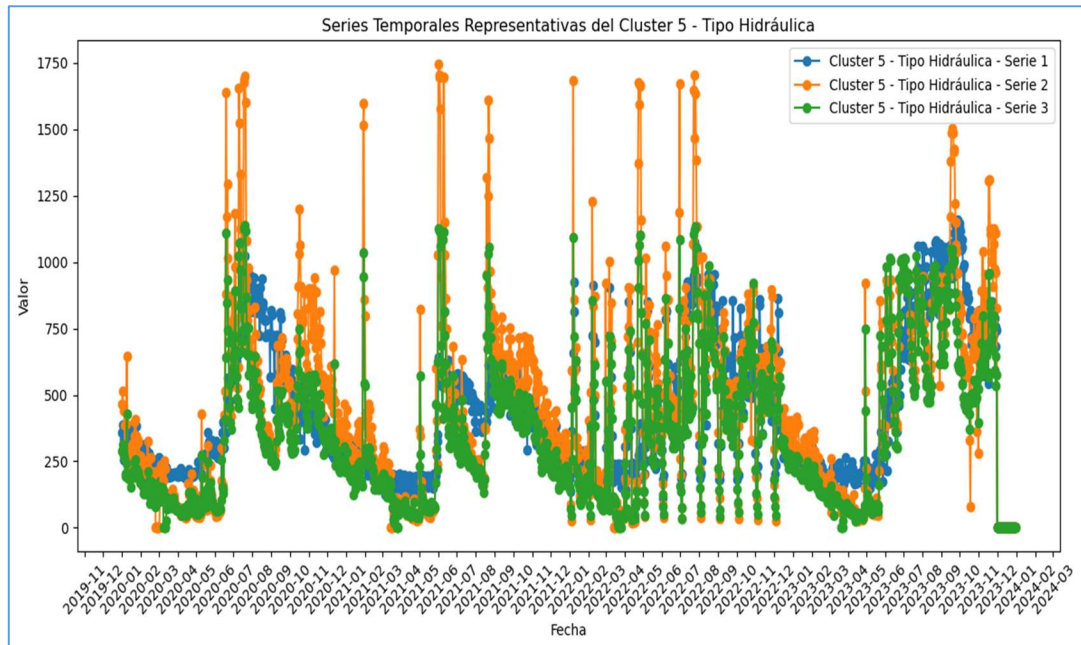


Figura 22: Curvas representativas del Clúster 5 para tipo Hidráulica

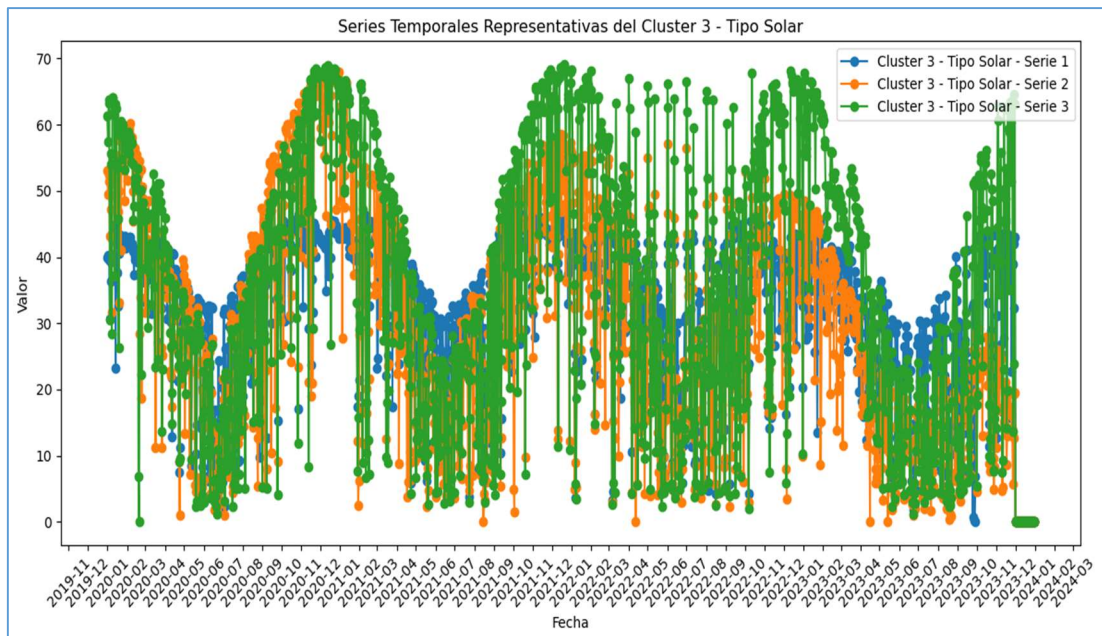


Figura 23: Curvas representativas del Clúster 3 para tipo Solar

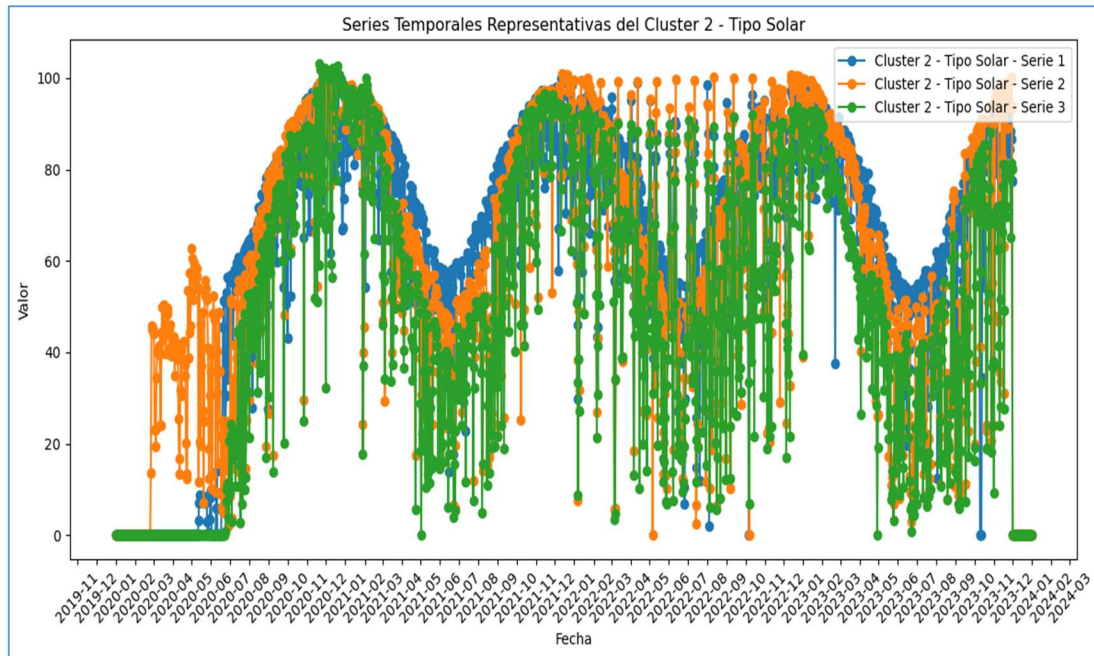


Figura 24: Curvas representativas del Clúster 2 para tipo Solar

4.3.2 Resultados del Agrupamiento Jerárquico Aglomerativo (AHC)

Los resultados mostraron que el método Ward produjo clústeres más compactos y con mejor cohesión, en comparación con el enlace Complete. La Fuerza de Predicción para el método Ward fue superior a la del método Complete, con una tendencia a estabilizarse a partir de $K = 10$, sugiriendo que los clústeres eran robustos y coherentes. El índice de Silueta también respaldó estos resultados, mostrando valores más altos en comparación con K -means.

En la Tabla 15 se presentan resultados con métrica Ward respectivamente, en ella se observa que la aglomeración porcentual por clúster de las generadoras para el tipo térmica es de 72,56% en el clúster 5, mientras que la eólica en el clúster 8 y 9 es de 58,82% y 41,18% respectivamente, la hidráulica en el clúster 6 y 3 es de 40,13%, por otra parte, la solar se separa en los clústeres 4 y 2 con 53,15% y 27,48% debido a esta gran cantidad no se mostraron las curvas representativas.

Tabla 15: Distribución en % de generadoras por clúster con Ward

Clúster	Térmica	Hidráulica	Solar	Eólica
0	3,72%	1,97%	0,45%	0,00%
1	10,23%	9,87%	11,71%	0,00%
2	0,93%	2,63%	27,48%	0,00%
3	1,40%	21,71%	0,45%	0,00%
4	0,00%	2,63%	53,15%	0,00%
5	72,56%	8,55%	1,35%	0,00%
6	0,93%	40,13%	0,00%	0,00%
7	3,26%	3,95%	2,25%	0,00%
8	6,98%	8,55%	3,15%	58,82%
9	0,00%	0,00%	0,00%	41,18%

4.3.3 Resultados del Autoencoder LSTM

En la Tabla 16 para LSTM evaluado con las métricas DTW, mostró una Fuerza de Predicción relativamente alta para valores bajos de K (de 0.75 a 0.80 en el rango de $K = 2$ a $K = 6$), lo que sugiere que el modelo generaba clústeres bien definidos y estables en esta configuración. Sin embargo, al incrementar k a valores entre 8 y 9, la Fuerza de Predicción disminuyó, indicando que el modelo perdía capacidad de generalización y estabilidad en los agrupamientos, lo cual podría deberse a la complejidad adicional introducida con un mayor número de clústeres, en valores superiores se fue recuperando poco a poco. La métrica DTW en el LSTM ofreció una robustez ligeramente mejor en comparación con Euclidiana.

En los resultados obtenidos con LSTM en DTW, solo se logró aglomerar las generadoras de tipo térmica en el clúster 1 con un **64,65%**. En cuanto a las generadoras eólicas, estas se presentaron como perturbaciones en los otros clústeres, quedando en el mencionado con un 29,41%.

Tabla 16: Distribución en % de generadoras por clúster métrica DTW

Clúster	Térmica	Hidráulica	Solar	Eólica
0	9,77%	22,37%	29,28%	14,71%
1	64,65%	5,26%	1,80%	29,41%
2	7,91%	19,74%	24,77%	20,59%
3	5,58%	19,08%	16,67%	17,65%
4	9,30%	21,05%	18,92%	11,76%
5	2,79%	12,50%	8,56%	5,88%

En resumen, cada uno de los métodos utilizados para el análisis de agrupamiento de series temporales presenta características y ventajas particulares. Dynamic Time Warping (DTW) se destaca por su capacidad para comparar dos series temporales permitiendo alineaciones no lineales. Esto le permite examinar toda la serie completa, buscando la mejor alineación temporal entre las series, independientemente de desajustes en el tiempo o variaciones en su duración. Su enfoque en las diferencias globales a lo largo de la serie completa lo hace especialmente útil para series con desajustes temporales o variabilidad en su duración.

El Agrupamiento Jerárquico Aglomerativo (AHC) construye una jerarquía de clústeres mediante una fusión iterativa de los puntos más similares. Si bien AHC puede usar diversas métricas de distancia, incluyendo Euclidiana o DTW, su principal enfoque está en cómo se agrupan los elementos en un dendrograma. Aunque AHC captura la estructura jerárquica de las relaciones entre las series, no está diseñado para manejar explícitamente las variaciones temporales de manera tan eficiente como otros métodos.

Finalmente, LSTM (Long Short-Term Memory), una red neuronal recurrente, es eficaz para aprender patrones en secuencias temporales, particularmente cuando se utiliza en un autoencoder LSTM. No obstante, su proceso de reducción de dimensiones en la fase de codificación puede resultar en una pérdida de

detalles, especialmente si la dimensión se reduce excesivamente siendo así en nuestro caso. Aunque LSTM captura las dependencias temporales en las series, la compresión de las representaciones latentes puede eliminar ciertos detalles complejos o menos frecuentes, a cambio de una representación más general y compacta.

CAPITULO V: Conclusiones

Este capítulo presenta las conclusiones obtenidas a partir del análisis de los patrones de generación de energía del Sistema Eléctrico Nacional (SEN) de Chile durante el período 2020-2023. Las conclusiones se organizan en función del objetivo general, los objetivos específicos y las preguntas de investigación planteadas en este estudio.

5.1 Identificación de Patrones

El uso de K -means con la métrica DTW permitió identificar patrones claros en las generadoras térmicas y solares, logrando agrupamientos compactos y diferenciados que facilitan el análisis de tendencias específicas en estos tipos de generación, se mejoró la identificación de patrones en las generadoras hidráulicas, permitiendo capturar las variaciones temporales y las características intermitentes propias de este tipo de energía. Por otro lado, AHC proporcionó una representación visual de las relaciones jerárquicas entre los tipos de generadoras, destacando la separación estructural entre térmicas y renovables, así como las diferencias internas entre las diversas fuentes renovables (eólicas, solares e hidráulicas). En cuanto a LSTM, aunque también mostró una clara separación entre las generadoras renovables y las térmicas, los métodos anteriores (K -means y AHC) lograron una distinción más precisa y consistente en la identificación de los patrones de generación energética. Esto es visible en las métricas de visualización como PCA y t-SNE

5.2 Comparación de Métodos

Los resultados y los tiempos de ejecución pueden variar según el entorno en el que se ejecuten los métodos, por lo que se procuró mantener una equivalencia en las condiciones para todos los algoritmos. Para ello, se probaron dos métricas en cada método y se simplificó la base de datos, asegurando que todas las métricas pudieran procesarla de manera eficiente. En este caso, AHC con la

métrica Ward mostró el mejor desempeño en términos de eficiencia, gracias a su bajo costo computacional y tiempos de ejecución reducidos, proporcionando resultados efectivos para clasificar los distintos tipos de generadoras según sus características de generación de energía. En cambio, DTW podría haber ofrecido un mejor desempeño en algunos casos, especialmente en la identificación de clústeres, pero su complejidad computacional resultó en tiempos de ejecución significativamente mayores en comparación con Ward. Por otro lado. En el caso de LSTM, aunque puede considerarse poco eficiente durante el proceso de entrenamiento debido a su alta demanda computacional, una vez entrenado, su eficiencia mejora significativamente, pero en el caso de estudio no fue capaz de aglomerar otras generadoras aparte de las térmicas.

5.3 Análisis de Tendencias y Ciclos

K-means, en general, fue capaz de capturar los comportamientos de los distintos tipos de generadoras, como se muestra en las curvas, destacando cómo la energía solar e hidráulica varía a lo largo del año, permitiendo identificar claramente los aumentos y descensos en la generación debido a la estacionalidad.

AHC con Ward presentó una robustez moderada, pero fue útil para identificar subgrupos dentro de los tipos de generadoras, lo que permitió observar relaciones y jerarquías dentro de estos subgrupos, aportando una visión más detallada de las interacciones entre las generadoras.

LSTM presentó una calidad de agrupamiento baja, ya que solo pudo identificar de manera efectiva los grupos que se separaban claramente del resto, es decir, las generadoras térmicas. Esto limitó su capacidad para capturar otros patrones complejos presentes en las generadoras de otros tipos.

Al revisar las métricas de validación, la fuerza de predicción fue aceptable en los tres métodos, alcanzando 0,86 en *K*-means y 0,70 en AHC. El único que presentó

problemas con esta métrica fue LSTM, que experimentó una caída en la fuerza de predicción entre K 7 y 9, antes de volver a aumentar en los K siguientes. Esto sugiere que LSTM tuvo dificultades para mantener la coherencia en la predicción en ciertos clústeres, aunque se recuperó posteriormente.

El índice de Silueta para cada método fue el siguiente: K -means con DTW alcanzó 0,26, AHC con Ward obtuvo 0,31, y LSTM con DTW tuvo un valor de 0,22.

Estos valores indican que AHC con Ward tuvo la mayor cohesión y separación entre clústeres, lo que sugiere que los grupos formados son más compactos y bien definidos. Por otro lado, LSTM mostró un índice bajo, indicando que los clústeres generados no son tan bien definidos o son menos coherentes. DTW presentó un valor intermedio, lo que sugiere que, aunque mejora la identificación de patrones, su complejidad computacional puede afectar la calidad del agrupamiento.

5.4 Recomendaciones para futuras investigaciones.

Una recomendación importante para futuras investigaciones es ampliar el conjunto de datos no solo en cuanto a la cantidad generada diaria, sino también en términos de la cantidad de variables. Incluir más columnas, como la cantidad generada por hora en formato 24 horas para cada generadora, permitirá capturar de manera más precisa los patrones estacionales y los comportamientos diarios de la generación de energía. Esto enriquecería significativamente el análisis, permitiendo que los modelos de agrupamiento detecten variaciones más sutiles a lo largo del día y durante distintas estaciones del año. Además, mejorar el hardware con equipos más potentes, como procesadores de múltiples núcleos o GPUs, aceleraría los tiempos de procesamiento, permitiendo manejar grandes volúmenes de datos de manera más eficiente. Complementariamente, dividir el dataset en lotes optimizaría el uso de recursos computacionales, facilitando el

análisis de series temporales grandes y mejorando el rendimiento del modelo. Estas acciones no solo mejorarían la precisión del agrupamiento, sino que también permitirían una escalabilidad adecuada para manejar datos en tiempo real o más completos en el futuro.

5.5 Respuesta a las Preguntas de Investigación

1) ¿Cómo se pueden aplicar técnicas de agrupamiento de series temporales para caracterizar de manera efectiva los patrones de generación de energía real del SEN?

Existen diversas técnicas de agrupamiento, y la elección de cuál utilizar dependerá de los objetivos específicos del análisis y de cómo funcionan cada una de ellas. A continuación, se describe una metodología estructurada para su análisis.

- i. En primer lugar, es fundamental pre procesar los datos para minimizar posibles errores que puedan afectar los resultados de los modelos.
- ii. Luego, se debe seleccionar el algoritmo de agrupamiento más adecuado, como *K*-means, jerárquico, DBSCAN, o modelos basados en densidad, entre otros.
- iii. Posteriormente, es importante escoger las métricas que mejor se ajusten al método seleccionado.
- iv. Finalmente, se evalúan los clústeres generados mediante medidas de cohesión, separación y visualización, para interpretar la calidad del agrupamiento obtenido.

2) ¿Cuál de estas tres técnicas es más adecuada para identificar patrones y agrupaciones relevantes en la generación de energía en el SEN de Chile?

Entregar una respuesta definitiva es complejo, ya que depende de las agrupaciones que se deseen obtener y del hardware disponible. En el caso particular de este estudio, se recomienda optar por *K*-means con la métrica

DTW, debido a su consistencia en la caracterización de los datos y su capacidad para separar los diferentes tipos de energía. Este enfoque fue elegido buscando minimizar el valor de K . Por otro lado, si el objetivo es separar los datos según características propias dentro de las mismas aglomeraciones, se podría optar por un K más grande o utilizar AHC con la métrica Ward.

3) ¿Cómo se comparan las herramientas en términos de eficiencia computacional y calidad de los agrupamientos generados? ¿Qué prioriza cada uno a la hora de agrupar?

K -means ha mostrado un buen rendimiento ya que DTW evalúa las diferencias temporales entre las series, lo que es útil para identificar patrones similares en series temporales con desajustes temporales pero su complejidad computacional hace difícil el utilizar grandes volúmenes de datos. AHC, produjo clústeres relativamente coherentes. El método de Ward es conocido por su capacidad para minimizar la varianza interna de los clústeres durante el proceso de fusión, lo que genera grupos más compactos y homogéneos, fue capaz de mostrar comportamientos propios dentro de los mismos tipos de generadoras siendo eficiente a nivel de cálculo.

Los resultados obtenidos con otras métricas no fueron tan satisfactorios, lo que sugiere que, al utilizar diferentes métricas o al reducir demasiado las dimensiones, el agrupamiento no se adapta bien a la naturaleza de los datos. Esto puede ocurrir si las métricas utilizadas no capturan correctamente las relaciones temporales en las series. En el caso de LSTM, aunque puede considerarse poco eficiente durante el proceso de entrenamiento debido a su alta demanda computacional, una vez entrenado, su eficiencia mejora significativamente, siempre y cuando se cuente con suficientes datos para capturar adecuadamente los patrones y el comportamiento de las series temporales.

BIBLIOGRAFIA

CEN (2023). Reporte de Sostenibilidad - Cuenta Pública 2023. Recuperado de <https://www.coordinador.cl/wp-content/uploads/2024/05/CUENTA-PUBLICA-CEN-2023-3.pdf>

CEN (2022). Hoja de ruta para una transición energética acelerada: Visión del Coordinador Eléctrico Nacional. Versión para observaciones, octubre de 2024. Recuperado de <https://www.coordinador.cl/wp-content/uploads/2024/10/HOJA-DE-RUTA-2024-V1.pdf>

Torres, R., & García, N. (2021). Matriz energética de Chile. Elaborado para Comisión de Medioambiente de la Cámara de Diputados, N° SUP: 132210.

Ramírez-Murillo, H., Torres-Pinzón, C. A., & Forero-García, E. F. (2019). Estimación del potencial fotovoltaico mediante minería de datos en cuatro ciudades de Colombia. *TecnoLógicas*, 22(46), 65-85.

Xu, Y., Wang, Q., & Chen, L. (2020). Supervised MPC control of large-scale electricity networks via clustering methods. *Energy and AI*, 1(2), 100012. <https://doi.org/10.1016/j.egyai.2020.100012>

Zhou, Z., Li, H., & Wei, X. (2022). A Model-Adaptive Clustering Method for Low-Carbon Energy System Optimization. *Energy Reports*, 8(4), 284-301. <https://doi.org/10.1016/j.energy.2022.04.12467>

Marrero, Lester, Carrizo, Dante, García-Santander, Luis, & Ulloa-Vásquez, Fernando. (2021). Uso de algoritmo *K*-means para clasificar perfiles de clientes con datos de medidores inteligentes de consumo eléctrico: Un caso de estudio. *Ingeniare. Revista chilena de ingeniería*, 29(4), 778-787. <https://dx.doi.org/10.4067/S0718-33052021000400778>

Yajure-Ramírez, C. A. (2022). Uso de algoritmos de aprendizaje automático para analizar datos de energía eléctrica facturada. Caso: Chile 2015–2021. *I+D*

Tecnológico, 18(2), 17-31. <https://revistas.utp.ac.pa/index.php/id-tecnologico/article/view/3678>

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). *K*-means agrupamiento algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622, 178-210.

Gironés Roig *et al* (2017, p 117). *Minería de datos: Modelos y algoritmos*. Editorial UOC (Oberta UOC Publishing, SL). RambladelPoblenou, 156 08018 Barcelona. ISBN: 978-84-9116-904-8.

Murtagh, F., & Contreras, P. (2017). Algorithms for hierarchical agrupamiento: an overview, II. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 7(6), e1219.

Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data* (pp. 94-105).

Gironés Roig *et al* (2017, p 40). *Minería de datos: Modelos y algoritmos*. Editorial UOC (Oberta UOC Publishing, SL). RambladelPoblenou, 156 08018 Barcelona. ISBN: 978-84-9116-904-8.

Zaki, M. J., & Meira, W., Jr. (2014). Clustering jerárquico. En *Data Mining and Machine Learning: Fundamental Concepts and Algorithms* (Cap. 14). Cambridge University Press

Heaton, J. (2018). Ian Goodfellow, Yoshua Bengio, and Aaron Courville: *Deep learning: The mit press*, 2016, 800 pp, isbn: 0262035618. *Genetic programming and evolvable machines*, 19(1), 305-307.

Chollet, F. (2021). *Deep learning with Python*. Manning Publications.

Tavakoli et al. (2020). Agrupamiento de datos de series temporales mediante modelos de aprendizaje profundo basados en autocodificadores. *Aplica SN Ciencia*. 2 , 937, Springer.

Berzal, F. (2018). *Redes Neuronales & Deeplearning*. Granada, España: Independently published. ISBN-10: 1731314337.

McKinney, W. (2022). *Python for data analysis: Data wrangling with pandas, NumPy, and Jupyter* (3^a ed.). O'Reilly Media.

Liberti, L., & Lavor, C. (2017). *Euclidean distance geometry: An introduction*. Springer.

Müller, M. (2007). Dynamic time warping. En *Information retrieval for music and motion* (pp. 69–84). Springer. https://doi.org/10.1007/978-3-540-74048-3_4

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301), 236-244

Defays, D. (1977). An efficient algorithm for a complete-link method. *The Computer Journal*, 20(4), 364–366. <https://doi.org/10.1093/comjnl/20.4.364>

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in *K*-means Clustering. *International Journal*, 1(6), 90-95.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the Gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), 411–423. <https://doi.org/10.1111/1467-9868.00293>.

Tibshirani, R., & Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3), 511-528.

Van Der Maaten, L., Postma, E. O., & Van Den Herik, H. J. (2009). Dimensionality reduction: A comparative review. *Journal of machine learning research*, 10(66-71), 13.

Tan, P.-N., Steinbach, M., & Kumar, V. (2014). *Introduction to data mining*. Pearson Education Limited. ISBN 10: 1-292-02615-4.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Amat Rodrigo, J. (2017). Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE. *Ciencia de Datos*. Disponible bajo licencia CC BY-NC-SA 4.0 en https://www.cienciadedatos.net/documentos/35_principal_component_analysis.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2).

Tibshirani, R., & Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3), 511-528.

Kodinariya, T. M., & Makwana, P. R. (2013). Review on determining number of Cluster in *K*-means Clustering. *International Journal*, 1(6), 90-95.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929-1958.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial intelligence review*, 22, 85-126.