Aplicación de algoritmos de aprendizaje automático, con énfasis en aprendizaje no supervisado, para la identificación y categorización de segmentos de interés en señales bioeléctricas para el estudio de la epilepsia - Fase V

Dylan Antonio Ixcayau Morán

UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería

Aplicación de algoritmos de aprendizaje automático, con énfasis en aprendizaje no supervisado, para la identificación y categorización de segmentos de interés en señales bioeléctricas para el estudio de la epilepsia - Fase V

Trabajo de graduación presentado por Dylan Antonio Ixcayau Morán
para optar al grado académico de Licenciado en Ingeniería Mecatrónica

Guatemala,

2024

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería

Aplicación de algoritmos de aprendizaje automático, con énfasis en aprendizaje no supervisado, para la identificación y categorización de segmentos de interés en señales bioeléctricas para el estudio de la epilepsia - Fase V

Trabajo de graduación presentado por Dylan Antonio Ixcayau Morán
para optar al grado académico de Licenciado en Ingeniería Mecatrónica

Guatemala,

2024

Vo.Bo.:

(f) _____

M. Sc. Carlos Esquit

Tribunal Examinador:

(f) _____

M.Sc. Carlos Esquit

(f) _____

M. Sc. Miguel Enrique Zea Arenales

(f) _____

Ing. Kurt Emmanuel Kellner

Fecha de aprobación: Guatemala, 13 de febrero de 2025.

Prefacio

Mi vida universitaria concluye con esta presentación de tesis, y es inevitable detenerse por un momento para mirar hacia atrás, reflexionar sobre todas las personas que conocí en el camino, los que se quedaron y los que se fueron, y todas las experiencias vividas que se convertirán en recuerdos imborrables.

En primer lugar, quiero expresar mi más profundo agradecimiento a mis padres, quienes han sido un pilar fundamental en mi vida. Siempre diré que no podría haber tenido mejores padres; me han apoyado desde el momento en que nací y nunca me han dejado caer. Este trabajo y mi graduación son, en gran medida, gracias a ellos. También quiero agradecer a mis hermanos, quienes, a pesar de las dificultades, siempre han estado ahí para apoyarme, brindándome consejos y su cariño incondicional.

Un agradecimiento especial a mi asesor de tesis, el Dr. Luis Rivera, cuya guía, conocimiento, experiencia y, sobre todo, paciencia fueron invaluables durante este proceso. No solo compartió su experiencia académica, sino que también me brindó inspiración y motivación. Su compromiso fue esencial para que esta investigación pudiera llevarse a cabo con éxito.

Esta tesis no solo representa la culminación de un proyecto, sino también el cierre de uno de los capítulos más importantes y memorables de mi vida, tanto para mí como para mi familia. Es el resultado del esfuerzo conjunto de todas las personas que me han apoyado hasta este momento.

Finalmente, gracias a todos los que formaron parte de este viaje y, por supuesto, a Cristiano Ronaldo, cuya pasión, disciplina y excelencia en cada gol me inspiraron a dar lo mejor de mí. SIUUUUUUU. También a Chris Bumstead, quien, con su constancia, disciplina y mentalidad inquebrantable, me motivó a mantenerme firme en mis objetivos dentro y fuera de los estudios. Ambos me enseñaron que la grandeza se construye día a día, con esfuerzo y dedicación.

Índice

Prefacio III

Lista de figuras VIII

Lista de cuadros IX

Resumen X

Abstract XI

1. Introducción 1

2. Antecedentes 2
2.1. Antecedentes fuera de la UVG . . . 2
2.2. Investigaciones previas en la UVG . . . 5

3. Justificación 13

4. Objetivos 14
4.1. Objetivo general . . . 14
4.2. Objetivos específicos . . . 14

5. Alcance 15

6. Marco teórico 16
6.1. Epilepsia . . . 16
6.1.1. Tipos de convulsiones . . . 16
6.1.2. Tipos de epilepsia . . . 17
6.2. Señales bioeléctricas . . . 18
6.3. Señales electroencefalográficas . . . 18
6.3.1. Ritmos y formas de onda del EEG . . . 19
6.4. Electromiografía . . . 19
6.5. Electrocardiograma . . . 20
6.6. Características en el dominio del tiempo . . . 20

6.7. Características en el dominio de la frecuencia . . . 21
6.8. Características en el dominio de tiempo-frecuencia . . . 21
6.9. Aprendizaje automático . . . 22
6.10. Tipos de aprendizaje automático . . . 22
6.10.1. Aprendizaje supervisado . . . 22
6.10.2. Aprendizaje no supervisado . . . 23
6.10.3. Aprendizaje reforzado . . . 26
6.11. VAT . . . 27
6.12. PCA . . . 28

7. Obtención y Procesamiento de Datos EEG **29**
7.1. Datos de HUMANA de pacientes con epilepsia . . . 29
7.2. Datos del TUH EEG Epilepsy Corpus . . . 30
7.3. Agrupamiento de datos . . . 30
7.4. Análisis de señales EEG . . . 31

8. Resultados Preliminares de Clustering **32**
8.1. VAT: Evaluación Visual de la Tendencia de Clustering . . . 32
8.2. Extracción de características . . . 33
8.3. Resultado inciales del clustering . . . 35
8.4. Prueba uno a uno . . . 36
8.5. Prueba sujeto a sujeto . . . 37
8.6. Evaluación General del Conjunto de Datos . . . 39
8.7. PCA en el contexto del estudio . . . 41
8.8. Prueba uno a uno con PCA . . . 42
8.9. Prueba sujeto a sujeto con PCA . . . 46
8.10. Evaluación General del Conjunto de Datos con PCA . . . 50

9. Normalización *Z-score* **53**
9.1. Resultados de pruebas anteriores . . . 55

10. *Fuzzy C-Means* **58**

11. Métodos Finales **60**

12. Actualización de la herramienta **62**

13. Conclusiones **71**

14. Recomendaciones **73**

15. Bibliografía **74**

16. Anexos **78**
16.1. Nombres de los sujetos en TUH EEG Epilepsy corpus . . . 78
16.2. Resultados sujeto a sujeto . . . 78
16.3. Resultados sin potencia y con PCA . . . 79
16.4. Resultados normalizando . . . 80

Lista de figuras

1. Ventana principal del toolbox [7]. 6
2. Ventana de selección y visualización [7]. 6
3. Ventana de extracción de caracteristicas [7]. 7
4. Ventana de implementación de red neuronal artificial [7]. 7
5. Ventana de implementación de máquina de vectores de soporte [7]. 8
6. Resumen de los resultados de los clasificadores generados: red neuronal (RNA) y máquina de vectores de soporte (SVM) [8]. 8
7. Ventana con la gráfica de una ventana del registro EEG del canal seleccionado [8]. 9
8. Resumen del rendimiento de la RNA para dos clases variando la cantidad de características utilizadas en el dominio de la frecuencia [9]. 9
9. Resúmen de resultados clústering jerárquico con Rand Index [9]. 10
10. Mejoras en el interfaz de usuario [10]. 10
11. Aviso descarga finalizada [10]. 11
12. Funcionamiento de ventana para visualizar señales [10]. 11
13. Funcionamiento de la ventana multigráfico [10]. 12
14. Instalador de la aplicación standalone en sistema de computo de HUMANA [10]. 12
15. Ventana de anotaciones automáticas [11]. 12

16. Tipos de convulsiones [17]. 17
17. Tipos de epilepsia [18]. 17
18. Ritmos cerebrales en un EEG [24] 20
19. Ejemplo de agrupación [34]. 24
20. Ejemplo de agrupación por medio de agrupación jerárquica [34]. 24
21. Ejemplo de agrupación por medio de agrupación k-medias [34]. 25
22. Ejemplo de agrupación por medio de agrupamiento difuso. 25
23. Ejemplo de visualización del VAT. 27

24. Grupo de datos ideal para VAT. 33
25. VAT ideal con grupos claros. 33
26. Grupo de datos ideal para VAT. 34
27. VAT ideal con grupos claros. 34

28. Dispersión de las características, enfocando la característica de potencia. . . . 41
29. PCA sesgada por la característica de potencia. . . . 42
30. Varianza características puras. . . . 43
31. Componentes principales. . . . 43
32. Datos agrupados de manera real. . . . 44
33. Datos agrupados por K-means. . . . 45
34. Datos agrupado por el cluster jerarquíco. . . . 45
35. Datos agrupados de manera real en prueba sujeto 3-4. . . . 47
36. Datos agrupados por K-means sujeto 3-4. . . . 48
37. Datos agrupado por el cluster jerarquíco sujetos 3-4. . . . 48
38. Datos agrupados de manera real en prueba sujeto 5-4. . . . 48
39. Datos agrupados por K-means sujeto 5-4. . . . 49
40. Datos agrupado por el cluster jerarquíco sujetos 5-4. . . . 49
41. Datos agrupados de manera real en prueba sujeto 3-4. . . . 51
42. Datos agrupados por K-means sujeto 3-4. . . . 51
43. Datos agrupado por el cluster jerarquíco sujetos 3-4. . . . 52

44. Varianza de características puras. . . . 53
45. PCA de características puras sin normalizar. . . . 54
46. Varianza de características puras normalizadas. . . . 55
47. PCA de características puras normalizadas. . . . 55

48. Fuzzy en una señal EEG. . . . 58

49. Estudio Gika con las secciones agrupadas, donde un color es un grupo y otro color es otro (señal obtenida de la herramienta actualizada). . . . 61
50. Estudio Al con las secciones agrupadas, donde un color es un grupo y otro color es otro (señal obtenida de la herramienta actualizada). . . . 61
51. Estudio CLEA con las secciones agrupadas, donde un color es un grupo y otro color es otro (señal obtenida de la herramienta actualizada). . . . 61
52. Estudio HCHC con las secciones agrupadas, donde un color es un grupo y otro color es otro (señal obtenida de la herramienta actualizada). . . . 61

53. Cierre de programa desde la ventana principal. . . . 63
54. Inicio de sesión amigable. . . . 64
55. Botones de tipos de anotaciones. . . . 65
56. Selección de aprendizaje a utilizar. . . . 65
57. Botón VAT 66
58. Selección de algoritmos y número de agrupaciones. . . . 66
59. Selección de PCA a utilizar. . . . 66
60. Selección de algoritmos. . . . 67
61. Opción de montaje para visualización de la señal. . . . 67
62. Visualización con el montaje en el que viene los canales. . . . 67
63. Ventana de selección de canales para realizar un montaje. . . . 68
64. Botón Observar anotaciones. . . . 69
65. Ventana de anotaciones no supervisadas. . . . 69
66. Selección de canal a visualizar. . . . 70
67. Ingreso de cantidad de segundo a visualizar en la ventana. . . . 70
68. Puntero para obtener información de un punto específico. . . . 70

Lista de cuadros

1. Información de grabaciones dadas por HUMANA 29

2. Prueba con un solo estudio para epilepsia y uno para no epilepsia en frecuencia. 36
3. Prueba con un solo estudio para epilepsia y uno para no epilepsia en tiempo continuo. 36
4. Prueba con un solo estudio para epilepsia y uno para no epilepsia usando todas las características. 36
5. Resultados de Rand Index para pruebas frecuencia con combinaciones de diferentes sujetos epilépticos y sanos (4-5, 4-3). 37
6. Resultados de Rand Index para pruebas tiempo continuo con combinaciones de diferentes sujetos epilépticos y sanos (4-5, 4-3). 38
7. Resultados de Rand Index para combinación de características con combinaciones de diferentes sujetos epilépticos y sanos (4-5, 4-3). 38
8. Prueba con 254 EDF sin epilepsia y 184 EDF con epilepsia en frecuencia. . . 40
9. Prueba con 254 EDF sin epilepsia y 184 EDF con epilepsia en tiempo continuo. 40
10. Prueba con 254 EDF sin epilepsia y 184 EDF con epilepsia combinando todas las características. 40
11. Rand Index usando combinación de características. 44
12. Rand Index usando PCA. 44
13. Resultados de Rand Index para pruebas combinando caracteristicas con combinaciones de diferentes sujetos epilépticos y sanos (4-5, 4-3). 46
14. Resultados de Rand Index para pruebas usando el PCA con combinaciones de diferentes sujetos epilépticos y sanos (4-5, 4-3). 46
15. Agrupamiento con 254 EDF sin epilepsia y 184 EDF con epilepsia combinando todas las características. 50
16. Agrupamiento con 254 EDF sin epilepsia y 184 EDF con epilepsia utilizando PCA. 50

17. Rand Index usando combinación de características en la prueba dato a dato. . 55
18. Rand Index usando PCA en la prueba uno a uno. 56
19. Resultados de Rand Index para pruebas combinando caracteristicas con combinaciones de diferentes sujetos epilépticos y sanos (4-5, 4-3). 56

20. Resultados de Rand Index para pruebas usando el PCA con combinaciones de diferentes sujetos epilépticos y sanos (4-5, 4-3). 56
21. Agrupamiento con 254 EDF sin epilepsia y 184 EDF con epilepsia combinando todas las características. 56
22. Agrupamiento con 254 EDF sin epilepsia y 184 EDF con epilepsia utilizando PCA. 57

23. Nombre de los sujetos dentro de TUH EEG Epilepsy corpus. 78
24. Rand Index para diferentes combinaciones de sujetos frecuencia. 79
25. Rand Index para diferentes combinaciones de sujetos tiempo continuo. 79
26. Rand Index para diferentes combinaciones de sujetos Todas las características. 79
27. Rand Index para diferentes combinaciones de sujetos todas las caracteristicas sin potencia. 80
28. Rand Index para diferentes combinaciones de sujetos PCA sin potencia. 80
29. Rand Index para diferentes combinaciones de sujetos todas las caracteristicas normalizadas. 80
30. Rand Index para diferentes combinaciones de sujetos PCA normalizada. 81

Resumen

En el trabajo presentado en este documento se investigarón e implementaron distintos métodos y algoritmos con el fin de mejorar la identificación y categorización de segmentos de interés en señales bioeléctricas para el estudio de la epilepsia. Este trabajo incluyó el procesamiento de señales provenientes del TUH EEG Epilepsy Corpus y pruebas adicionales con datos de HUMANA para evaluar su desempeño en diferentes escenarios.

Se implementaron técnicas como la normalización Z-score, que asegura una escala uniforme entre características, y el Análisis de Componentes Principales (PCA), utilizado para reducir la dimensionalidad, optimizar la carga computacional y mejorar la visualización de los agrupamientos generados. Asimismo, se evaluaron algoritmos de aprendizaje automático no supervisado, como K-Means, Clustering Jerárquico y Fuzzy C-Means. Estos algoritmos demostraron su efectividad para diferenciar segmentos epilépticos y no epilépticos, con resultados particularmente robustos al combinar características extraídas de los dominios del tiempo, frecuencia.

Una contribución destacada de este trabajo fue la integración de Fuzzy C-Means, que permitió la generación de anotaciones con gradientes, mejorando la interpretación visual de los resultados en la herramienta desarrollada. La herramienta también mostró consistencia y adaptabilidad con diferentes señales validando su capacidad para aplicaciones clínicas y de investigación.

Este proyecto concluyó con el desarrollo de una herramienta funcional que puede ser utilizada en el análisis y diagnóstico de la epilepsia, proporcionando una base para futuras investigaciones en aprendizaje automático y validaciones clínicas más amplias.

Abstract

In the work presented in this document, various methods and algorithms were investigated and implemented to improve the identification and categorization of segments of interest in bioelectrical signals for the study of epilepsy. This work included the processing of signals from the TUH EEG Epilepsy Corpus and additional tests with HUMANA data to evaluate their performance in different scenarios.

Techniques such as Z-score normalization, which ensures a uniform scale between features, and Principal Component Analysis (PCA), used to reduce dimensionality, optimize computational load, and improve the visualization of generated clusters, were implemented. Additionally, unsupervised machine learning algorithms, such as K-Means, Hierarchical Clustering, and Fuzzy C-Means, were evaluated. These algorithms demonstrated their effectiveness in differentiating epileptic and non-epileptic segments, with particularly robust results when combining features extracted from time and frequency domains.

A notable contribution of this work was the integration of Fuzzy C-Means, which enabled the generation of gradient annotations, improving the visual interpretation of the results in the developed tool. The tool also demonstrated consistency and adaptability with different signals, validating its suitability for clinical and research applications.

This project concluded with the development of a functional tool that can be used for the analysis and diagnosis of epilepsy, providing a foundation for future research in machine learning and broader clinical validations.

CAPÍTULO 1

Introducción

El análisis de señales bioeléctricas, en particular de las señales EEG (electroencefalográficas), juega un papel crucial en el estudio de la epilepsia. La detección precisa de eventos epilépticos y la identificación de patrones que caracterizan los episodios ictales y no ictales es fundamental para mejorar los diagnósticos y tratamientos. Sin embargo, debido a la gran cantidad de datos que estas señales generan y a la variabilidad inherente en los pacientes, el procesamiento manual se convierte en una tarea difícil y tediosa. Esta situación plantea la necesidad de desarrollar herramientas automatizadas que permitan analizar estas señales de manera eficiente.

El presente trabajo buscó implementar algoritmos de aprendizaje automático, con un enfoque particular en técnicas de aprendizaje no supervisado, para la identificación y categorización de segmentos relevantes en señales bioeléctricas. El objetivo fue mejorar la capacidad de detección automática de eventos epilépticos y generar anotaciones relevantes a partir de los datos obtenidos, cumpliendo con los parámetros establecidos por HUMANA.

Las señales EEG fueron procesadas mediante algoritmos de aprendizaje no supervisado previamente desarrollados, con el objetivo de optimizar la identificación de segmentos relevantes y mejorar la precisión en el agrupamiento de las señales. Sin embargo, en esta fase del proyecto, se buscó obtener nuevas señales bioeléctricas, adicionales a las que se utilizaron antes de este trabajo de graduación. Estas señales provinieron principalmente de la base de datos del Temple University Hospital (TUH EEG) [1], complementadas con algunas señales obtenidas de HUMANA en fases anteriores del proyecto. Este enfoque permitió abordar un análisis más completo y variado de los datos. Estas señales fueron procesadas utilizando algoritmos de aprendizaje automático previamente desarrollados, con el fin de optimizar la detección de segmentos de interés y mejorar la precisión de los agrupamientos.

Para alcanzar estos objetivos, se investigaron características adicionales de las señales bioeléctricas y se evaluaron nuevos algoritmos de aprendizaje no supervisado. A través de un análisis estadísticos, se determinaron las mejores características para describir los datos y se actualizó la herramienta de software existente, incorporando estas mejoras.

CAPÍTULO 2

Antecedentes

La epilepsia es una enfermedad cerebral crónica no transmisible que afecta a personas de todas las edades. En todo el mundo, unos 50 millones de personas padecen epilepsia, lo que la convierte en uno de los trastornos neurológicos más comunes. La epilepsia se caracteriza por convulsiones recurrentes, las convulsiones se deben a descargas eléctricas excesivas en un grupo de células cerebrales que pueden producirse en diferentes partes del cerebro [2].

Cerca del 80 % de los pacientes viven en países de ingresos bajos y medianos, donde el acceso a los tratamientos puede ser limitado o inaccesible. El riesgo de muerte prematura en personas con epilepsia es hasta tres veces mayor que en la población general. En muchas partes del mundo, los enfermos de epilepsia y sus familias sufren estigmatización y discriminación [2].

2.1. Antecedentes fuera de la UVG

En el artículo Predicción de Crisis Epilépticas Utilizando Señales electroencefalográficas (EEG) y Aprendizaje Automático se tuvo como objetivo principal desarrollar un modelo para predecir crisis epilépticas con el fin de mejorar la calidad de vida de los pacientes afectados. El problema a resolver se centró en la necesidad de anticipar las crisis epilépticas mediante el análisis de señales de EEG, utilizando técnicas de aprendizaje automático para lograr una predicción precisa y oportuna [3].

Para lograr el objetivo se utilizó el conjunto de datos CHB-MIT, que consta de señales de EEG de 22 sujetos, para desarrollar el modelo de predicción SVM, el cuál en este caso se empleó para distinguir entre el estado preictal y el estado inercial en las señales de EEG, lo que permitió predecir la ocurrencia de crisis epilépticas con una sensibilidad del 88.3 %. La metodología incluyó dos etapas de preprocesamiento de datos: la conversión de las señales de EEG de 23 canales en un canal sustituto para mejorar la relación señal-ruido y la aplicación de la descomposición en modos empíricos (EMD) para aumentar aún más la relación señal-ruido. Se exploraron diferentes métodos de filtrado para obtener el canal sustituto de la señal

de EEG los cuales fueron: Filtro de promediado, para suavizar las señales de EEG y reducir el ruido en los datos; Filtro Laplaciano Grande, para mejorar la calidad de las señales de EEG al resaltar las características relevantes; Filtro de patrón especial común para mejorar la separabilidad de las clases en las señales de EEG y aumentar la relación señal-ruido [3].

Se demostró que el modelo propuesto para predecir crisis epilépticas supera a otros en cuanto a sensibilidad y tiempo promedio de predicción con un tiempo promedio de 23.6 minutos, brindando a los pacientes más tiempo para recibir la medicación adecuada y prevenir la crisis epiléptica [3].

El alcance del artículo se centró en el desarrollo y evaluación del modelo de predicción de crisis epilépticas utilizando el conjunto de datos CHB-MIT. Sin embargo, se reconoció la limitación de no incluir la adquisición de datos de EEG directamente de los pacientes, ya que se utilizó un conjunto de datos disponible públicamente. Además, se identificó la posibilidad de mejorar el preprocesamiento de las señales de EEG para aumentar la sensibilidad de la predicción de crisis en futuras investigaciones [3].

El trabajo de detección de convulsiones epilépticas en registros largos de EEG utilizando un detector de anomalías con rechazo de artefactos abordó el desafío de detectar convulsiones epilépticas en registros prolongados de electroencefalograma (EEG) de manera automática. Se destaca en el trabajo la dificultad y la propensión a errores del proceso manual de detección de convulsiones, así como la necesidad de profesionales experimentados para llevar a cabo esta tarea. El objetivo principal fue proponer un método novedoso para detectar convulsiones epilépticas a partir de registros largos de EEG, utilizando detectores de anomalías de vanguardia y técnicas de rechazo de artefactos [4].

Se propuso un método de dos etapas para la detección de convulsiones epilépticas en registros largos de EEG. En la primera etapa, se preseleccionan los EEG para la detección potencial de convulsiones utilizando seis métodos de detección de anomalías diferentes los cuales fueron [4]:

- COPOD: Detección de anomalías basada en cópulas (*Copula-Based Outlier Detection*).
- ECOD: Detección de anomalías utilizando funciones de distribución acumulativa empírica (*Empirical Cumulative Distribution Functions*).
- LSCP: Combinación selectiva local de conjuntos paralelos de detectores de anomalías (*Locally Selective Combination of Parallel Outlier Ensembles*).
- LODA: Detector ligero en línea de anomalías (*Lightweight On-line Detector of Anomalies*).
- IForest: Bosque de aislamiento *(Isolation Forest*).
- OCSVM: Máquina de vectores de soporte de una clase (*One-Class Support Vector Machine*).

Estos métodos generaron un puntaje de anomalía para cada segmento de EEG, lo que permite identificar segmentos con un alto potencial de contener convulsiones. La segunda etapa implicó la detección de convulsiones mediante la eliminación de artefactos. Se aplicaron

técnicas específicas para abordar la pérdida de contacto de todos los electrodos, la pérdida de contacto de un solo electrodo y la eliminación de artefactos generales. Las cuales fueron [4]:

- *Bipolar Montage* (Montaje Bipolar).
- Cálculo de la tasa de potencia absoluta.
- Cálculo de la tasa de potencia absoluta promedio.

Estas técnicas se utilizaron para filtrar los segmentos de EEG que contienen convulsiones, eliminando los artefactos y preservando los eventos de convulsiones [4].

En este estudio, se compararon seis métodos de detección de anomalías utilizando el valor óptimo de δ para cada método. Se observó que los métodos COPOD, ECOD y IForest mostraron un rendimiento similar, mientras que los demás métodos tuvieron un rendimiento inferior. Además, se encontró que el desempeño del método COPOD está positivamente correlacionado con la especificidad y que la detección de eventos de convulsiones estuvo relacionada con la sensibilidad hasta cierto punto de δ. La evaluación del método propuesto con el dataset CHB-MIT de epilepsia mostró una sensibilidad promedio del 82.8 %, especificidad del 88.2 % y un ratio de detección del 88.4 %, lo que indica que el método propuesto sigue siendo prometedor a pesar del desafío de tener un conjunto de datos altamente desequilibrado [4].

El alcance se centró en la detección automática de convulsiones epilépticas a partir de registros de EEG, utilizando métodos de detección de anomalías y técnicas de eliminación de artefactos. Sin embargo, se reconoce que el método propuesto aún presenta limitaciones, como una precisión relativamente baja y la necesidad de mejorar la especificidad. Además, se destaca que la evaluación del método se realizó en un conjunto de datos privado, lo que limita su generalización a otros conjuntos de datos [4].

El estudio de clasificación de crisis epilépticas y no epilépticas psicógenas mediante electroencefalografía y electrocardiografía se enfocó en desarrollar algoritmos de clasificación basados en técnicas de aprendizaje automático que permitan diferenciar de manera precisa entre estos dos tipos de crisis, lo cual presenta un desafío clínico significativo debido a sus similitudes en la presentación clínica [5].

Se recopilaron datos de EEG y ECG de los participantes, quienes proporcionaron su consentimiento informado. Se extrajeron características de desaceleración crítica de los datos preictales de EEG y ECG, y se utilizaron diferentes clasificadores como [5]:

- k-vecinos más cercanos (*k-nearest neighbors*): En el análisis preliminar, se observó que los parámetros en el clasificador tenían un efecto trivial en el redimiento de la clasificación. Aunque se obtuvieron resultados estables, el rendimiento fue insatisfactorio.
- Árbol de decisión (*decision tree*): Se utilizó para la clasificación de los datos preictales de EEG y ECG. Se observó que el rendimiento fue menor en comparación con el clasificador de bosque aleatorio.

- Bosque aleatorio (*Random forest*): Fue uno de los clasificadores más efectivos. Se obtuvo la precisión más alta (87.83 %) utilizando este clasificador en el intervalo preictal de 15-0 minutos de datos de EEG y ECG.
- *Naive Bayes* (NB): Su rendimiento fue menor.
- Máquina de vectores de soporte (SVM): Se empleó con el núcleo de RBF. Se observó que el rendimiento del clasificador SVM fue menor en comparación con el clasificador de bosque aleatorio.

Se evaluó el rendimiento de los clasificadores y se realizó un análisis estadístico para determinar la significancia de los resultados [5].

Se observó que la combinación de datos de EEG y ECG mejoró significativamente el rendimiento de la clasificación, con una precisión máxima del 87.83 % utilizando el clasificador de bosque aleatorio. Se identificó que las características de desaceleración crítica en el intervalo preictal de 15-0 minutos lograron el mejor rendimiento en la clasificación de crisis, lo que sugiere la posibilidad de predecir eventos hasta una hora antes de su inicio [5].

Aunque los resultados fueron prometedores, este estudio presentó limitaciones como el tamaño relativamente pequeño de la muestra de pacientes y crisis, lo que sugiere la necesidad de ampliar el estudio a una población más grande. Además, no se exploraron a fondo las diferencias en la fisiopatología subyacente entre las crisis epilépticas y no epilépticas psicogénicas, lo que plantea áreas para investigaciones futuras. Se destaca la importancia de investigar la selección de canales y la duración óptima de los datos para mejorar la precisión de la clasificación en futuros estudios [5].

2.2. Investigaciones previas en la UVG

La línea de investigación en la Universidad del Valle de Guatemala (UVG) comenzó con el trabajo de graduación de María Angulo [6] en el año 2020, quien se dedicó a implementar algoritmos de clasificación basados en máquinas de vectores de soporte (SVM) y redes neuronales en señales biomédicas de pacientes con epilepsia. El objetivo era identificar características y patrones relevantes de esta enfermedad. Angulo logró demostrar que es factible detectar crisis epilépticas mediante el aprendizaje automático, al caracterizar y clasificar correctamente los segmentos de las señales.

En ese mismo año, María Fernanda Pineda [7] inició el desarrollo e implementación de una base de datos de señales biomédicas de pacientes con epilepsia del Centro de Epilepsia y Neurocirugía Funcional (HUMANA), la cual organizó y clasificó meticulosamente. Para ello, empleó el software de código abierto *phpMyAdmin* para crear una base de datos relacional dentro del entorno de *MySQL*. Además, en colaboración con María Angulo, desarrolló un toolbox de enlace que permitía escribir, obtener y manipular datos pertenecientes a la base de datos mediante Matlab. Este toolbox facilitó la interacción de HUMANA con los datos, presentando campos amigables para el usuario. Sin embargo, se identificaron limitaciones en el almacenamiento de datos, por lo que se recomendó almacenar los datos de manera individual para evitar ralentizar el proceso al manejar grandes cantidades de información.

Adicionalmente, María Pineda trabajo una interfaz sencilla para conectar la base de datos con Matlab y poder hacer análisis de los mismos mediante los algoritmos desarrollados por María Angulo. En las Figuras 1, 2, 3, 4 y 5 puede observarse la interfaz.

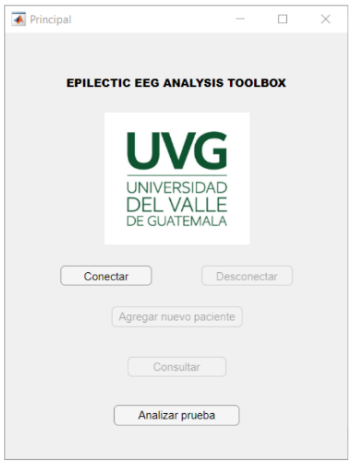

Figura 1: Ventana principal del toolbox [7].

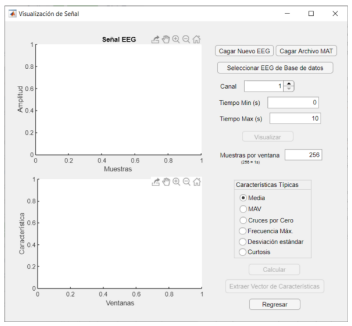

Figura 2: Ventana de selección y visualización [7].

Siguiendo el trabajo de María Angulo, en el año 2021 David Alejandro Vela [8] desarrolló y validó un proceso de reconocimiento y anotación de posibles episodios ictales en señales electroencefalográficas (EEG) de pacientes con epilepsia, por medio de técnicas de aprendizaje automático. Con respecto al trabajo anterior agregó una característica adicional para la clasificación, asi como otras dos clases a disciminar: preictales e interictales.

En este caso, se realizaron ajustes a la interfaz anterior utilizando algoritmos de aprendizaje automático supervisado para clasificar las señales EEG. El clasificador con mejor desempeño obtuvo una exactitud del 96.7 %, utilizando SVM con características en kernel gaussiano y tiempo continuo. Una de las limitaciones del proyecto fue que se trabajó más con el aprendizaje supervisado, lo cual dejaba un espacio abierto para explorar el aprendizaje no supervisado [8]. Los resultados del trabajo se pueden observar en la Figura 6 y en la Figura 7 se muestra la forma en que se clasifica un segmento de señal por color, según su estado ictal.

En el trabajo de investigación de Camila Lemus [9] el objetivo fue mejorar una herramienta de software para el estudio de la epilepsia mediante la incorporación de análisis de diversas señales bioeléctricas, características en el dominio de la frecuencia y técnicas

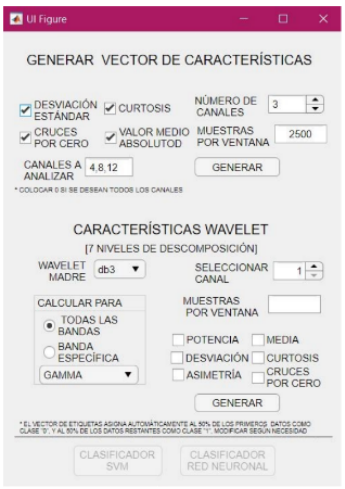

Figura 3: Ventana de extracción de caracteristicas [7].

Figura 4: Ventana de implementación de red neuronal artificial [7].

de aprendizaje automático no supervisado. Esto se lograría evaluando nuevas señales bioeléctricas relevantes, explorando características en el dominio de la frecuencia, integrando algoritmos de aprendizaje no supervisado y generando anotaciones automáticas para validar los resultados con especialistas.

Para alcanzar los objetivos establecidos, Camila Lemus llevó a cabo experimentos de aprendizaje automático utilizando diversas bases de datos, incluida la de la Universidad de Bonn en Alemania, que contenía registros de pacientes con epilepsia en diferentes estados. Realizó selecciones de características a diferentes razones y exploró múltiples combinaciones de ellas para determinar cuáles eran óptimas para las señales EEG en el dominio de la frecuencia. Posteriormente, entrenó un clasificador de redes neuronales y evaluó su desempeño para clasificar dos, tres y hasta seis clases de forma satisfactoria [9].

Los resultados de clasificación para los dos pares de clases fueron muy similares al utilizar dos o más características. El nivel de exactitud para el clasificador Ictal/Sano se encontró por encima del 99.99 % en estos casos, y el clasificador Interictal/Preictal tuvo una exactitud superior al 98.80 % de igual manera. Luego, realizó experimentos similares para identificar crisis en señales ECG, manteniendo la misma metodología pero cambiando el tipo de señal y las características. Posteriormente, exploró el aprendizaje no supervisado utilizando

Figura 5: Ventana de implementación de máquina de vectores de soporte [7].

Características	Tiempo Continuo			Wavelet		
Modelo	RNA	SVM		RNA	SVM	
Kernel	-	Gaussiano	Lineal	-	Gaussiano	Lineal
2 Clases	100.00%	99.80%	100.00%	97.70%	98.70%	97.90%
3 Clases	97.90%	98.90%	97.20%	98.20%	98.30%	97.20%
4 Clases	88.00%	91.30%	88.30%	81.20%	83.30%	77.10%
Promedio	95.30%	96.70%	95.20%	92.40%	93.40%	90.70%
Desv. Estándar	5.23%	3.81%	4.99%	7.90%	7.17%	9.64%

Figura 6: Resumen de los resultados de los clasificadores generados: red neuronal (RNA) y máquina de vectores de soporte (SVM) [8].

técnicas como *K-means* y *Clustering Jerárquico* para agrupar datos en diferentes niveles, validando los resultados utilizando el método *Rand Index*. Este enfoque permitió identificar las características más precisas y adecuadas para generar resultados satisfactorios con un menor costo computacional, así como determinar el nivel óptimo de agrupamiento para la clasificación jerárquica [9].

Los resultados destacaron la eficacia de las redes neuronales para clasificar señales EEG, con una precisión promedio del 99.60 % y una mínima desviación estándar del 0.61 %. Aunque la extracción de características en el dominio de la frecuencia mostró menor eficiencia que otras técnicas, como Wavelet o en el dominio del tiempo, demostró mayor precisión en el clasificador neuronal. Se desarrolló una nueva sección en la herramienta de software para el análisis de señales ECG, logrando una precisión del 90.3 % en la clasificación binaria mediante el análisis no lineal (Poincaré). La combinación de características específicas, como la razón θ/α y β/α, resultó ventajosa, superando el 98.90 % de precisión en menos tiempo. Aunque los algoritmos de aprendizaje no supervisado fueron más rápidos, la precisión de la clasificación se vio afectada, resaltando la importancia de equilibrar precisión y tiempo de procesamiento en la selección de algoritmos [9]. ——————————
==============================————-==========–

El alcance de este proyecto fue ampliar las funciones de la herramienta de análisis de epilepsia desarrollada previamente en la Universidad del Valle de Guatemala en 2020 y 2021. Se exploró el rendimiento de distintos clasificadores de aprendizaje automático supervisado para señales EEG, utilizando características en el dominio de la frecuencia. Además, se incluyó el análisis de señales de electrocardiograma (ECG) mediante técnicas de aprendizaje no supervisado. Entre las limitaciones del proyecto se encontraron la falta de una validación profunda de los resultados mediante asesoría médica y la imposibilidad de realizar pruebas

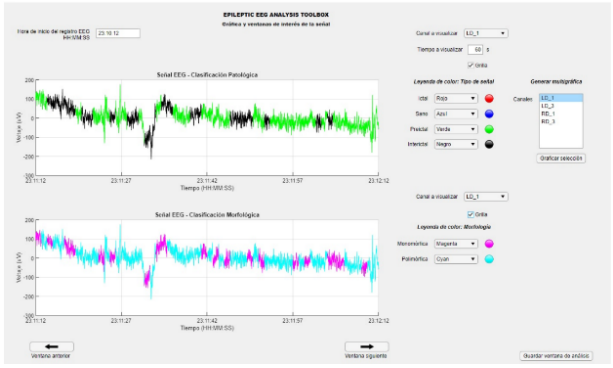

Figura 7: Ventana con la gráfica de una ventana del registro EEG del canal seleccionado [8].

Cantidad de características	Características utilizadas (razones)	Exactitud RNA	
		Ubonn - Sano/Ictal	Kaggle - Interictal/Preictal
1	1	91.55%	90.50%
2	1, 2	100%	98.90%
3	1, 2, 3	99.99%	98.80%
4	1, 2, 3, 4	99.99%	98.80%
5	1, 2, 3, 4, 5	99.99%	98.80%
6	1, 2, 3, 4, 5, std	100%	99.40%

Figura 8: Resumen del rendimiento de la RNA para dos clases variando la cantidad de características utilizadas en el dominio de la frecuencia [9].

con los datos de pacientes del hospital HUMANA debido a restricciones de tiempo y recursos disponibles [9].

El objetivo de Diego Méndez en su trabajo de graduación del 2023 fue extender, validar y migrar la herramienta de software para el estudió de la epilepsia desarrollada en fases anteriores para su uso en HUMANA [10].

Como parte de la mejora de la herramienta se realizó una modificación en el interfaz de usuario de la ventana de consulta para hacerla más intuitiva y amigable como se observa en la Figura 10. Se añadió un sistema de descarga para las señales y su metadata, facilitando su análisis como se ve en la Figura 11, Además, se diseñó un procedimiento de borrado de datos y se ajustó la visualización del eje x para mostrarlo en función del tiempo como se ve en Figura 12. Se agregaron ventanas que permiten visualizar canales en simultáneo mejorando la experiencia del usuario como se ve en la Figura 13 [10].

Se buscó migrar la herramienta a sistemas de cómputo fuera de la universidad, utilizando Matlab Compiler para crear una versión compatible con diferentes sistemas. Se desarrollaron manuales de instalación y usuario para facilitar su implementación y su uso externo. Para evaluar el funcionamiento adecuado de la herramienta, se realizaron pruebas completas de funcionalidades y una visita a HUMANA para verificar la factibilidad de su uso en las computadoras de la organización [10].

Los resultados del proyecto incluye mejoras significativas en la interfaz de usuario, la actualización exitosa de los comandos de conexión entre la herramienta y la base de datos, y

Clustering Jerárquico	
No. de clases	Exactitud
2	98.08%
3	82.25%
4	76.20%
Promedio	85.51%
Desv. Estándar	11.30%

Figura 9: Resúmen de resultados clústering jerárquico con Rand Index [9].

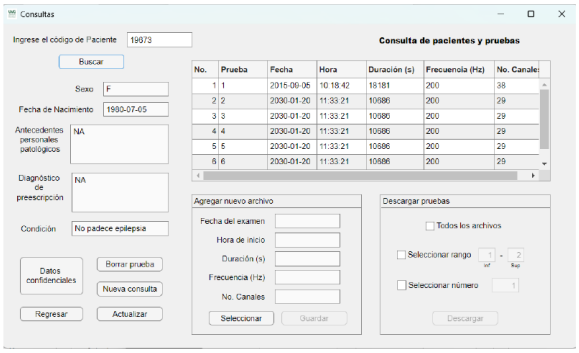

Figura 10: Mejoras en el interfaz de usuario [10].

la implementación de una ventana para visualizar varías gráficas para un análisis simultáneo. Se logró reducir el tiempo de carga de los datos y se mejoró el código para la creación y generación de relaciones en la base de datos [10].

El alcance del proyecto se enfocó en la mejora continua y la optimización de la herramienta para el estudio de la epilepsia, con el objetivo de permitir su ejecución en computadoras sin licencia activa de Matlab. Sin embargo, al finalizar el proyecto, se identificó una limitación significativa: la falta de una licencia activa de MATLAB Compiler. Esta carencia impidió que la última versión autónoma compilada de la herramienta incorporara todas las modificaciones y mejoras realizadas durante el proyecto. Como consecuencia, esta restricción afectó la capacidad de despliegue y distribución de la herramienta desarrollada para el estudio de la epilepsia en HUMANA [10].

El trabajo de investigación de Christopher Patzán del año 2023 tuvo como objetivo aplicar algoritmos de aprendizaje automático a una mayor cantidad de señales bioeléctricas para mejorar el proceso de detección de segmentos de interés en el estudio de la epilepsia [11].

Se utilizaron señales EEG y electromiográficas (EMG) obtenidas con el equipo Biopac de la UVG y proporcionadas por HUMANA para implementar y evaluar algoritmos de aprendizaje automático. Se extrajeron características en el dominio del tiempo, frecuencia y wavelets, optimizando el tiempo de entrenamiento de los algoritmos y mejorando la clasificación de las señales [11]. Se clasificaron las señales en categorías como “Ictal”, “Sano”, “Interictal”, “Preictal”, validando la efectividad de las redes neuronales y las SVM [11]. Esto puede observarse en la Figura 15

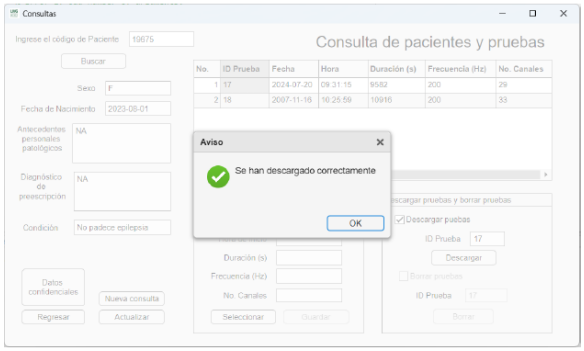

Figura 11: Aviso descarga finalizada [10].

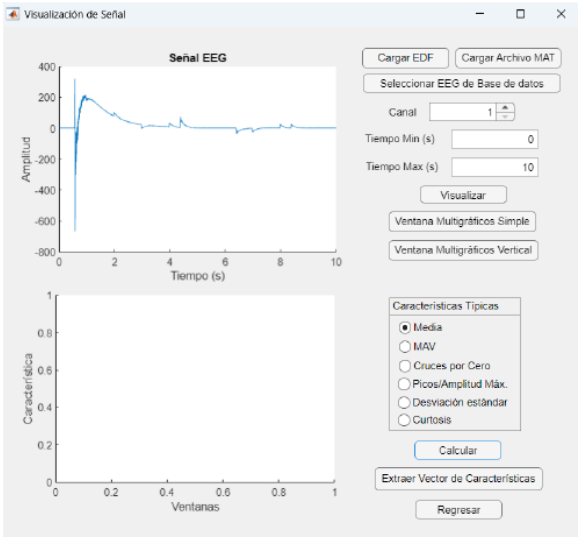

Figura 12: Funcionamiento de ventana para visualizar señales [10].

La actualización de la herramienta de software implicó la corrección de errores y la optimización del procesamiento de señales, con la recolección de 187 grabaciones de señales EEG y EMG en la UVG y el uso de algoritmos de agrupamiento K-means y jerárquico [11]. Se identificaron características eficaces en el dominio del tiempo-frecuencia, especialmente utilizando transformadas wavelets mejorando la clasificación de segmentos de interés en señales bioeléctricas [11].

El proyecto se enfocó principalmente en su aplicación en Guatemala y entornos similares. Sin embargo, enfrentó limitaciones, como la restricción geográfica y la necesidad de adaptar técnicas existentes en lugar de desarrollar nuevas desde cero. Además, se encontró una limitada cantidad de datos de señales bioeléctricas disponibles, lo que sugiere que aumentar esta cantidad podría enriquecer y mejorar la capacidad de detección de patrones. También se observó que hubo pocas oportunidades para interactuar con especialistas y aprovechar su experiencia, lo cual podría haber contribuido a mejorar el reconocimiento de los segmentos de interés y hacerlos más precisos y útiles para ellos [11].

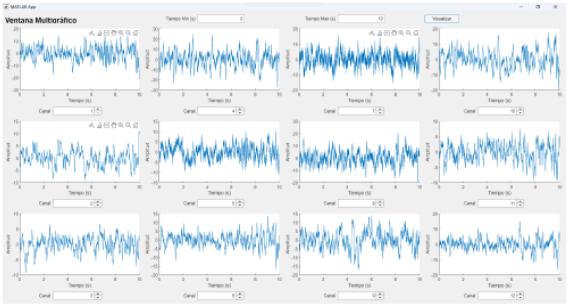

Figura 13: Funcionamiento de la ventana multigráfico [10].

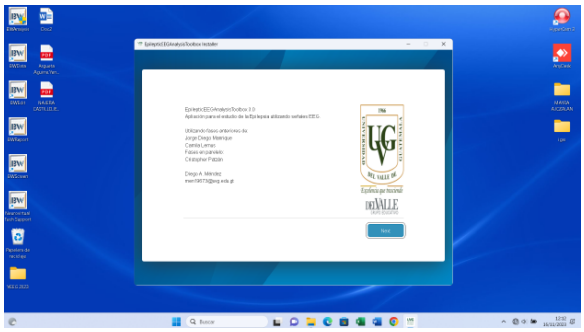

Figura 14: Instalador de la aplicación standalone en sistema de computo de HUMANA [10].

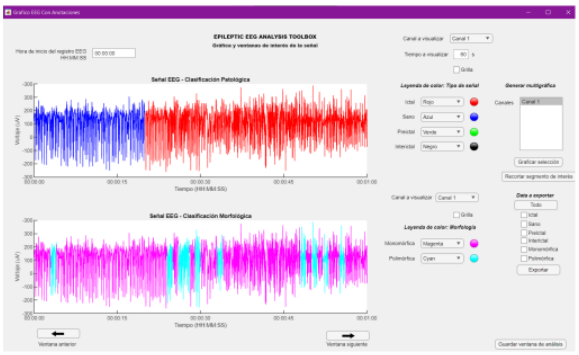

Figura 15: Ventana de anotaciones automáticas [11].

CAPÍTULO 3

Justificación

La epilepsia no es un problema eventual, es un problema que social que afecta alrededor de 325,000 personas en Guatemala y tiene alto impacto emocional, personal, familiar y social; en muchas ocasiones el problema no son las crisis convulsivas, si no el momento y lugar donde se presentan, ya que puede poner en riesgo la vida de los pacientes y la de las personas que lo rodean [12]. A pesar de la alta prevalencia de la epilepsia en el país, hay una escasa cantidad de investigaciones sobre esta afección, lo que hace necesario llenar este vacío en el conocimiento y obtener información importante sobre la epilepsia en esta población [13].

En la actualidad, en el Centro de Epilepsia y Neurocirigía Funcional (HUMANA), se recolectan datos de pacientes con epilepsia y luego analiza de forma manual. Este método para analizar las señales resulta ser ineficiente para los especialistas, ya que tardan mucho tiempo. Debido a esto, la automatización en el análisis y clasificación de los datos es una prioridad.

En las fases anteriores se desarrollaron diferentes algoritmos de aprendizaje, tanto supervisado como no supervisado. Sin embargo, la cantidad de datos disponibles para el análisis era limitada y no se contaba con un recurso estructurado que permitiera optimizar el desarrollo de los algoritmos no supervisados para captar correctamente los segmentos de interés. En esta fase, se hizo uso de un libro de la institución HUMANA, el cual contiene información clave sobre epilepsia, señales EEG, tipos de montaje y directrices para la interpretación de las señales, sirviendo como base para mejorar la comprensión y análisis de los datos.

En esta fase se contó con acceso a una de las bases más grandes del mundo utilizadas para este tipo de estudios, TUH EEG, lo que permitió trabajar con un conjunto de datos más completo y representativo. Esto facilitó el desarrollo de algoritmos de aprendizaje no supervisado más precisos y efectivos en la identificación de segmentos de interés dentro de los datos analizados. Además, se utilizó información contenida en un libro de HUMANA sobre epilepsia y señales EEG, lo que ayudó a identificar características clave que debían ser consideradas en el proceso de análisis.

CAPÍTULO 4

Objetivos

4.1. Objetivo general

Implementar algoritmos de aprendizaje automático, con énfasis en técnicas de aprendizaje no supervisado, para la identificación y categorización de segmentos de interés en señales bioeléctricas para el estudio de la epilepsia.

4.2. Objetivos específicos

- Obtener señales bioeléctricas adicionales a las de fases anteriores y de pacientes con epilepsia de HUMANA.
- Aplicar algoritmos de aprendizaje automático desarrollados en fases anteriores a las señales EEG con sus múltiples canales junto con otros tipos de señales bioeléctricas que se hayan obtenido.
- Mejorar el proceso de detección de segmentos de interés en las señales y la generación automática de anotaciones relevantes, según los parámetros de HUMANA.
- Investigar y evaluar características de las señales bioeléctricas, así como algoritmos de aprendizaje automático, con énfasis en aprendizaje no supervisado, con el fin de mejorar el rendimiento de los agrupamientos y análisis de cúmulos desarrollados.
- Realizar análisis estadísticos para evaluar el rendimiento de los algoritmos y determinar las mejores características para describir a las señales bioeléctricas.
- Actualizar la herramienta de software para el estudio de la epilepsia desarrollada en fases anteriores, incorporando las mejoras a los algoritmos de agrupamiento y detección de segmentos de interés de las señales bioeléctricas.

CAPÍTULO 5

Alcance

En esta fase del proyecto, se lograron avances significativos en la implementación y validación de algoritmos de aprendizaje no supervisado para la identificación y categorización de segmentos de interés en señales EEG. La incorporación de un conjunto de datos más amplio y representativo permitió una validación más robusta de los algoritmos, utilizando el Rand Index como métrica de evaluación principal para determinar la consistencia de los agrupamientos en diferentes configuraciones.

Se implementó un nuevo método de agrupamiento, Fuzzy C-Means, y se incluyó dentro de la herramienta de software una visualización mejorada que permite representar los resultados de este algoritmo con un gradiente de colores, en lugar de los colores sólidos utilizados previamente. Esta funcionalidad brinda una representación más detallada de los valores de pertenencia generados por el algoritmo, facilitando una interpretación más clara de los datos.

Además, se inspeccionaron diversas características para determinar cuáles ofrecían mejores resultados en la clasificación de las señales, empleando técnicas de preprocesamiento como la normalización z-score y métodos de reducción de dimensionalidad como el Análisis de Componentes Principales (PCA). Estas técnicas optimizaron la calidad del agrupamiento y redujeron la variabilidad en los resultados.

En cuanto a la herramienta de software, se ampliaron las funcionalidades para permitir la selección de diferentes algoritmos de agrupamiento y el número de grupos deseados, además de incluir opciones de visualización de montajes monopolares y bipolares, brindando mayor flexibilidad en el análisis de las señales EEG.

Aunque se lograron avances significativos, una limitación fue la falta de señales intrasujeto más largas y con etiquetas, lo que habría permitido una validación más específica de las agrupaciones generadas por los algoritmos. Este aspecto representa un área de oportunidad para futuros estudios.

CAPÍTULO 6

Marco teórico

6.1. Epilepsia

La epilepsia es uno de los primero transtornos documentados en la historia de la neurología [14]. Estudios epidemiológicos indican que entre 0.5 y 1 % de la población mundial padece epilepsia y se considera que entre 1 y 3 % de la población tendrá epilepsia durante su vida [15]. La epilepsia no es un problema eventual, es un problema social que afecta alrededor de 325,000 personas en Guatemala [12].

La epilepsia es una alteración del sistema nervioso central caracterizada un incremento y sincronización anormales de la actividad eléctrica neuronal, que se manifiesta con crisis recurrentes y espontáneas así como por cambios electroencefalográficos [15]. Las crisis epilépticas se dividen en dos tipos principales: crisis generalizadas y crisis parciales o focales. En las crisis generalizadas, la descarga epiléptica afecta simultáneamente a toda la superficie del cerebro, mientras que en las crisis parciales o focales, la descarga epiléptica se origina en una parte específica del cerebro [16].

6.1.1. Tipos de convulsiones

Nuevos términos para describir y clasificar convulsiones han sido desarrollados por la liga internacional contra la epilepsia. Esto se hizo para hacer que los nombres de las convulsiones sean más precisos, menos confusos y más descriptivos de lo que esta sucediendo. Los nuevos términos son [17]:

- Focal: También conocidas como convulsiones parciales, comienzan en una parte específica del cerebro. Los síntomas pueden variar dependiendo de la parte del cerebro afectada, pero pueden incluir movimientos involuntarios de un solo lado del cuerpo, cambios en la percepción sensorial, emociones intensas o cambios en la conciencia.

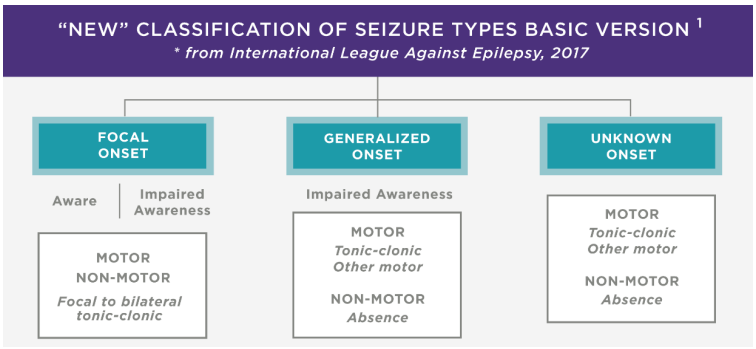

Figura 16: Tipos de convulsiones [17].

Estas convulsiones pueden ser simples (sin pérdida de conciencia) o complejas (con alteración de la conciencia).

- Generalizada: Estas convulsiones involucran ambos hemisferios cerebrales desde el inicio de la actividad. Pueden incluir subtipos como convulsiones tónico-clónicas (anteriormente conocidas como convulsiones gran mal), que se caracterizan por rigidez muscular seguida de sacudidas musculares rítmicas y pérdida de conciencia, así como convulsiones de ausencia, que implican breves periodos de ausencia o desconexión sin movimiento notable.
- Desconocida: Algunas convulsiones no pueden clasificarse claramente como focales o generalizadas debido a la falta de información sobre su inicio o propagación en el cerebro. Estas convulsiones pueden requerir más evaluación para determinar su tipo específico.

6.1.2. Tipos de epilepsia

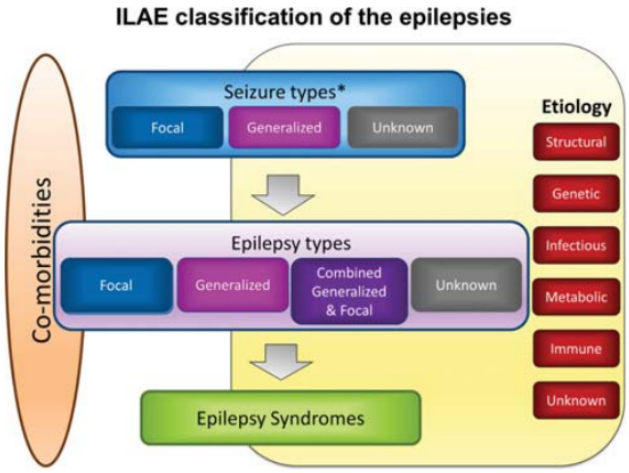

Figura 17: Tipos de epilepsia [18].

- Focal: se caracteriza por convulsiones que comienzan en una parte específica del cere-

bro. Las convulsiones pueden tener diversos síntomas dependiendo de la región cerebral afectada, como movimientos involuntarios, sensaciones anormales o cambios en la percepción.

- Generalizada: las convulsiones involucran ambos hemisferios cerebrales desde el inicio de la actividad epiléptica. Las convulsiones generalizadas pueden incluir convulsiones tónico-clónicas, convulsiones de ausencia y otros tipos de convulsiones que afectan a todo el cerebro desde el principio.
- Algunas personas pueden experimentar una combinación de convulsiones focales y generalizadas. Esto significa que las convulsiones pueden comenzar en una parte específica del cerebro pero luego propagarse y afectar a todo el cerebro. Se considera una forma mixta de epilepsia.
- Desconocida: En algunos casos, no es posible determinar claramente si las convulsiones son focales o generalizadas debido a la falta de información sobre su origen o propagación en el cerebro. En tales casos, se puede diagnosticar epilepsia de origen desconocido.

6.2. Señales bioeléctricas

Las señales bioeléctricas son señales generadas por nervios y células musculares que pueden medirse y controlarse continuamente [18]. Las señales bioeléctricas tienen características únicas que las hacen útiles en una amplia gama de aplicaciones, como la cardiología, la neurología y la ingeniería biomédica [19].

6.3. Señales electroencefalográficas

Un electroencefalograma (EEG) es una prueba utilizada para diagnosticar la epilepsia mediante el registro de la actividad eléctrica en el cerebro [20].

El electroencefalograma es un estudio de la función cerebral que recoge la actividad eléctrica del cerebro en situación basal y con métodos de activación, como la hiperventilación y la fotoestimulación. Es conveniente también registrar durante el sueño [21].

La señal eléctrica recogida se amplifica y representa en forma de líneas, interpretándose la actividad de las distintas áreas cerebrales a lo largo del tiempo, estas serían las señales electroencefalográficas (Señales EEG) [21].

Existen patrones normales y patrones anormales que hacen sospechar lesiones o enfermedades características. Es, por tanto, un medio de diagnóstico funcional de enfermedades cerebrales complementario a otros estudios, especialmente los radiológicos (TAC, resonancia magnética). Es necesario realizar un EEG al momento en que los pacientes muestren síntomas de deterioro en memoria, facultades intelectuales, conciencia o que exista la sospecha de haber sufrido una crisis epiléptica [21].

6.3.1. Ritmos y formas de onda del EEG

Los ritmos cerebrales, que son los patrones más comunes observados en las señales EEG y que describen la actividad cerebral, varían según la edad, el estado de vigilia o sueño, y la presencia de alguna enfermedad [22]. Como se ve en la figura 18

- El ritmo delta (δ) se caracteriza por una frecuencia menor a 4 Hz y una amplitud significativa. Se observa comúnmente durante el sueño profundo y en algunos casos anormales. Es predominante en lactantes hasta el año de edad y se manifiesta durante las etapas 3 y 4 del sueño. La producción de estas ondas coincide con la regeneración y restauración del Sistema Nervioso Central [23].
- El ritmo theta (θ) se caracteriza por una frecuencia entre 4 a 7 Hz, se observa durante la somnolencia, en ciertas fases del sueño y durante el enfoque interno, como en la meditación. Reflejan el estado entre la vigilia y el sueño, y están asociados con la mente subconsciente. Son anormales en adultos despiertos pero normales en niños hasta los 13 años. Cuando funcionan adecuadamente, pueden facilitar conductas adaptativas complejas como el aprendizaje y la memoria [23].
- Ritmos alpha (α) se caracteriza por una frecuencia entre 8 y 13 Hz, predominante en un electroencefalograma normal, se manifiesta cuando el Sistema Nervioso Central está en reposo pero la persona está despierta y atenta. Es común a lo largo de la vida, especialmente después de los trece años, cuando se convierte en el ritmo principal del reposo. Este estado cerebral se presenta cuando una persona está alerta pero no está procesando información activamente, es decir, cuando está tranquila pero consciente. Su prominencia es mayor en sujetos normales que están relajados y despiertos con los ojos cerrados, y la actividad se reduce cuando los ojos se abren, siendo su amplitud más marcada en las regiones occipitales [23].
- Ritmos beta (β) se caracteriza por una frecuencia entre 14 y 30 Hz, predominan cuando estamos despiertos y concentrados en tareas externas. Son rápidas y están asociadas a la actividad cortical, principalmente en las áreas frontal y central. Este ritmo refleja la atención enfocada y se observa durante la resolución de problemas [23].
- Ritmos gamma (γ) se caracteriza por una frecuencia mayor a 30 Hz, son las más rápidas y se producen en ráfagas cortas, relacionándose con el procesamiento simultáneo de información en diferentes áreas del Sistema Nervioso Central. Son el único grupo de frecuencia presente en todas las partes del cerebro y se observan durante estados de alta resolución mental. El ritmo gamma, también con frecuencias mayores a 30 Hz, está asociado con un procesamiento activo de la información en el córtex. Se puede observar utilizando un electrodo sobre el área sensoriomotora y técnicas de registro de alta sensibilidad [23].

6.4. Electromiografía

La electromiografía (EMG) mide la respuesta muscular o la actividad eléctrica en respuesta a una estimulación nerviosa del músculo. La prueba se utiliza para ayudar a detectar

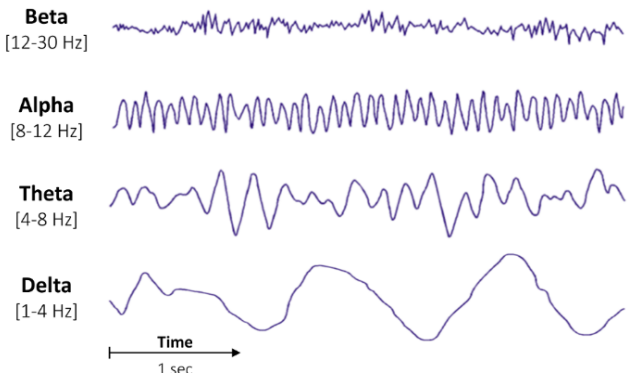

Figura 18: Ritmos cerebrales en un EEG [24]

anomalías neuromusculares. Durante la prueba, una o más agujas pequeñas (también llamadas electrodos) se insertan a través de la piel en el músculo. La actividad eléctrica captada por los electrodos se muestra en un osciloscopio. EMG mide la actividad eléctrica del músculo durante el descanso, la contracción leve y la contracción contundente. El tejido muscular normalmente no produce señales eléctricas durante el descanso. Cuando se inserta un electrodo, se puede ver un breve período de actividad en el osciloscopio, pero después de eso, no debe haber señal [25].

6.5. Electrocardiograma

Un electrocardiograma (ECG) es un procedimiento simple, indoloro y rápido que registra la actividad eléctrica de su corazón. Cada vez que el corazón late, una señal eléctrica circula a través de él. La señal activa las cuatro cámaras de su corazón para que se contraigan (aprieten) al ritmo correcto para que su corazón pueda bombear sangre a su cuerpo [26]. Esto se realiza por medio de las variaciones de voltaje en relación al tiempo. A diferencia de los EEGs que son de larga duración, los ECGs usualmente no exceden los 30 segundos [27].

6.6. Características en el dominio del tiempo

Las características en el dominio del tiempo (TDF) son aquellas calculadas a partir de señales EEG sin procesar o de señales preprocesadas realizadas en el dominio del tiempo, como la descomposición en modo empírico (EMD) [28]. Algunas características en el dominio del tiempo son:

- Media: Valor promedio de la señal durante un intervalo de tiempo específico.
- Desviación estandar: Mide la variabilidad de los señal dentro de la ventana de tiempo.
- Cruces por cero: Número de veces que la señal cruza el eje cero dentro de la ventana, lo que indica la frecuencia de oscilación de la señal.

- Asimetría: Mide la falta de simetría en la distribución de los valores de la señal alrededor de su media. Indica si la distribución de los valores está sesgada hacia la izquierda (negativa) o hacia la derecha (positiva) con respecto a la media.
- Curtosis: Es una medida de cuán propensa es una distribución a tener valores atípicos. La curtosis de la distribución normal es 3. Las distribuciones más propensas a valores atípicos que la distribución normal tienen una curtosis mayor a 3; las distribuciones menos propensas tienen una curtosis menor a 3. Algunas definiciones de curtosis restan 3 al valor calculado, de modo que la distribución normal tiene una curtosis de 0.

6.7. Características en el dominio de la frecuencia

El análisis del dominio de la frecuencia es crucial, ya que una representación de frecuencia de una señal de EEG proporciona información útil sobre los patrones de la señal. La PSD (Power Spectral Density o Densidad espectral de potencia en español) y la PSD normalizada (por la potencia total) se utilizan principalmente para extraer características que representan la partición de potencia en cada frecuencia [28]. Algunas características en el dominio de la frecuencia son:

- Potencia: Es la cantidad total de energía que una señal transporta dentro de una banda específica de frecuencia en el dominio de la frecuencia.
- Energía: Representa la cantidad total de energía contenida en la señal en una banda de frecuencia específica.
- Densidad Espectral de potencia (PSD): Es una medida de la distribución de potencia de una señal en función de la frecuencia. Indica cómo se distribuye la potencia de la señal en diferentes frecuencias y es útil para caracterizar la naturaleza estocástica de la señal, especialmente en aplicaciones de comunicaciones y procesamiento de señales.
- Frecuencia Centroide: Medida de la ubicación promedio de la energía espectral de la señal.
- Propagación: Se refiere a cómo las diferentes componentes de frecuencia de una señal se transmiten a través de un medio o canal. Esto puede incluir efectos como la atenuación y dispersión de diferentes frecuencias, afectando el espectro de la señal a medida que viaja.

6.8. Características en el dominio de tiempo-frecuencia

Proporciona una representación conjunta de la energía o intensidad de una señal en función del tiempo y la frecuencia. En este dominio, se pueden observar características únicas que no son evidentes en el dominio de la frecuencia única, como la evolución espectral de una señal a lo largo del tiempo y la frecuencia instantánea en cada instante [29]. Algunas de las características clave del dominio tiempo-frecuencia incluyen:

- Representación detallada de la evolución espectral: Permite analizar cómo varía el contenido espectral de una señal en diferentes momentos, lo que es crucial para entender cambios temporales en la señal, como en el caso de señales no estacionarias.
- Análisis de la frecuencia instantánea: Proporciona información sobre la frecuencia dominante en cada instante de tiempo, lo que es esencial para comprender fenómenos como la modulación de frecuencia y la propagación de señales en medios variables.
- Caracterización de eventos transitorios: Permite identificar y analizar eventos transitorios en una señal, que pueden no ser evidentes en un análisis de frecuencia única, lo que resulta útil en aplicaciones como el procesamiento de señales sísmicas o de radar.
- Interpretación intuitiva de la energía de la señal: Facilita la visualización y comprensión de cómo se distribuye la energía de una señal en diferentes componentes de frecuencia a lo largo del tiempo, lo que puede revelar información importante sobre el comportamiento de la señal.

6.9. Aprendizaje automático

El aprendizaje automático (*machine learning* en inglés) es una subrama de la inteligencia artificial, que es el proceso por el que los ordenadores aprovechan las redes neuronales para reconocer patrones y mejorar su capacidad para identificarlos. Con los suficientes ajustes y datos, un algoritmo de aprendizaje automático puede predecir nuevos patrones e información [30].

El aprendizaje automático describe la capacidad de los sistemas para aprender de datos de entrenamiento específicos de problemas para automatizar el proceso de construcción de modelos analíticos y resolver tareas asociadas [31].

6.10. Tipos de aprendizaje automático

6.10.1. Aprendizaje supervisado

El aprendizaje automático supervisado es un tipo de aprendizaje automático en el que un algoritmo aprende a partir de un conjunto de datos etiquetado. Los datos etiquetados consisten en ejemplos de entrada y salida deseados. El algoritmo utiliza estos ejemplos para aprender a mapear las entradas a las salidas deseadas [32].

El aprendizaje supervisado se puede dividir en dos tipos de problemas cuando se extrae información: clasificación y regresión [32]:

- La clasificación utiliza un algoritmo para asignar con precisión los datos de prueba en categorías específicas. Reconoce entidades específicas dentro del conjunto de datos e intenta sacar algunas conclusiones sobre cómo se deben etiquetar o definir esas entidades.

- La regresión se utiliza para comprender la relación entre variables dependientes e independientes. Se utiliza comúnmente para realizar proyecciones, como los ingresos por ventas de una empresa determinada

Algunos de los métodos de aprendizaje supervisado más utilizados son [32]:

- Regresión lineal: se utiliza para identificar la relación entre una variable dependiente y una o más variables independientes y, por lo general, se aprovecha para hacer predicciones sobre resultados futuros. Cuando solo hay una variable independiente y una variable dependiente, se le conoce como regresión lineal simple. A medida que aumenta el número de variables independientes, se denomina regresión lineal múltiple.
- Regresión logística: se selecciona cuando la variable dependiente es categórica, lo que significa que tienen resultados binarios, como "verdadero" y "falso" o "sí" y "no". La regresión logística se utiliza principalmente para resolver problemas de clasificación binaria, como la identificación de spam.
- Máquinas de vectores de soporte (SVM):se utiliza tanto para la clasificación como para la regresión de datos. Normalmente se aprovecha para problemas de clasificación, construyendo un hiperplano donde la distancia entre dos clases de puntos de datos es máxima. Este hiperplano se conoce como límite de decisión y separa las clases de puntos de datos (por ejemplo, naranjas frente a manzanas) a cada lado del plano.
- K-vecino más cercano: Clasifica puntos de datos basándose en su proximidad a otros datos. Asume que los puntos similares están cerca entre sí y usa la distancia, generalmente euclidiana, para determinar las categorías. Asigna la categoría más frecuente o el promedio. Es fácil de usar y tiene bajo tiempo de cálculo, lo que lo hace popular entre los científicos de datos. Sin embargo, su tiempo de procesamiento aumenta con conjuntos de datos grandes, lo que limita su uso en clasificación. KNN se aplica en motores de recomendación y reconocimiento de imágenes.
- Bosque aleatorio: el bosque aleatorio es otro algoritmo flexible de aprendizaje automático supervisado que se utiliza tanto para fines de clasificación como de regresión. El "bosque" hace referencia a una colección de árboles de decisión no correlacionados, que luego se fusionan para reducir la variación y crear predicciones de datos más precisas.

6.10.2. Aprendizaje no supervisado

El aprendizaje no supervisado descubren patrones ocultos o agrupaciones de datos sin necesidad de intervención humana [33].

La capacidad del aprendizaje no supervisado para descubrir similitudes y diferencias en la información lo convierte en la solución ideal para el análisis exploratorio de datos, estrategias de venta cruzada, segmentación de clientes y reconocimiento de imágenes [33].

Los modelos de aprendizaje no supervisado se utilizan para tres tareas principales: agrupación, asociación y reducción de dimensionalidad [33]. Algunos métodos de aprendizaje no supervisado son:

- Agrupación (*clustering* en inglés): Es una técnica de minería de datos que agrupa datos sin etiquetar en función de sus similitudes o diferencias. Los algoritmos de agrupamiento se utilizan para procesar objetos de datos sin procesar y no clasificados en grupos representados por estructuras o patrones en la información. Los algoritmos de agrupamiento se pueden clasificar en algunos tipos, específicamente exclusivos, superpuestos, jerárquicos y probabilísticos.

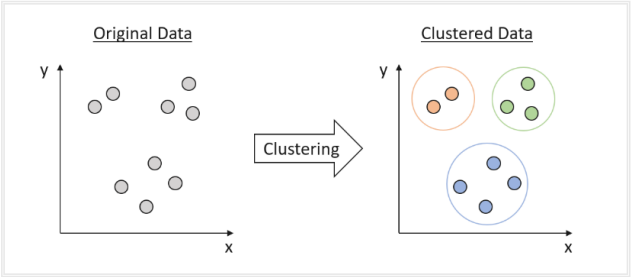

Figura 19: Ejemplo de agrupación [34].

  - Agrupamiento jerárquico (HCA por sus siglas en inglés): Se puede categorizar de dos formas: pueden ser aglomerados o divisivos:
    - Aglomerados: Sus puntos de datos se aíslan inicialmente como agrupaciones separadas y luego se fusionan de forma iterativa según la similitud hasta que se logra crear un grupo. Normalmente se utilizan cuatro métodos diferentes para medir la similitud: Método Ward, enclace promedio, enlace completo (o vecino más lejano), enlace simple (o vecino más cercano).
    - Divisiva: Un solo grupo (*cluster* en inglés) se divide en función de las diferencias entre los puntos de datos.

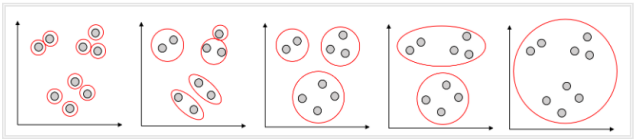

Figura 20: Ejemplo de agrupación por medio de agrupación jerárquica [34].

  - K-medias (*k-means* en inglés): Es un método de agrupación exclusivo o "hard". funciona categorizando puntos de datos en clústeres mediante el uso de una medida de distancia matemática, generalmente euclidiana, desde el centro del clúster. El objetivo es minimizar la suma de distancias entre los puntos de datos y sus clústeres asignados. Los puntos de datos más cercanos a un centroide se agrupan dentro de la misma categoría. Un valor k más alto, o el número de grupos, significa grupos más pequeños con mayor detalle, mientras que un valor k más bajo da como resultado grupos más grandes con menos detalle [35].
  - Agrupamiento difuso (*Fuzzy c-means (FCM)* en inglés) es una técnica de agrupamiento de datos donde cada uno el punto de datos pertenece a un clúster en un grado especificado por una membresía grado. El algoritmo FCM comienza con una suposición inicial para los centros de clúster, que representar la ubicación media de cada grupo. La suposición inicial para este grupo los centros son

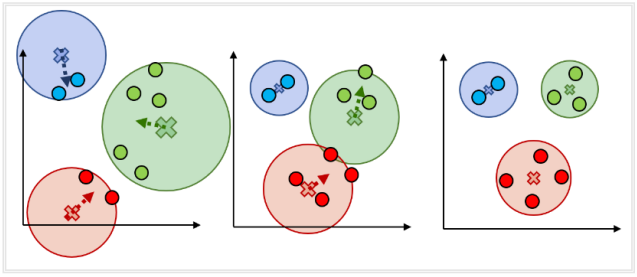

Figura 21: Ejemplo de agrupación por medio de agrupación k-medias [34].

probablemente incorrectos. Adicionalmente, FCM asigna cada punto de datos a grado de membresía para cada clúster. Actualizando iterativamente los centros de clúster y las calificaciones de membresía para cada punto de datos, el algoritmo mueve iterativamente el el clúster se centra en la ubicación óptima dentro de un conjunto de datos. Esta iteración se basa minimizar una función objetiva que represente la distancia desde cualquier dato dado apunte a un centro de clúster ponderado por el grado de membresía del punto de datos [36].

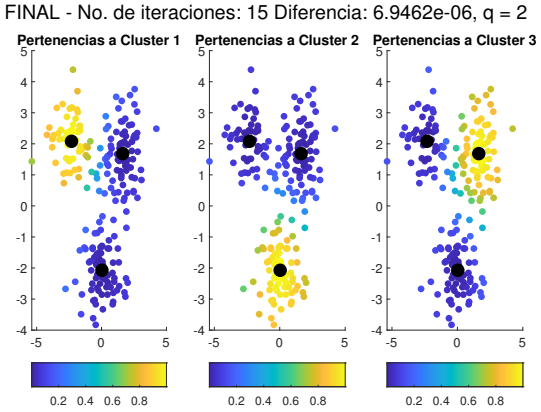

Figura 22: Ejemplo de agrupación por medio de agrupamiento difuso.

- Reglas de asociación: Es un método basado en reglas para encontrar relaciones entre variables en un conjunto de datos determinado. Estos métodos se utilizan con frecuencia para el análisis de la cesta de la compra, lo que permite a las empresas comprender mejor las relaciones entre los diferentes productos.
- Reducción de dimensionalidad: Es una técnica que se utiliza cuando la cantidad de características o dimensiones en un conjunto de datos determinado es demasiado alta. Reduce la cantidad de entradas de datos a un tamaño manejable y al mismo tiempo preserva la integridad del conjunto de datos tanto como sea posible

Validación de agrupamientos

La validez de los agrupamientos consiste en un conjunto de técnicas diseñadas para identificar el conjunto de clústeres que mejor se ajuste a las particiones naturales de los datos, sin contar con información de clases previa. El proceso de agrupamiento es evaluado mediante índices de validez de clústeres, los cuales permiten medir la calidad y coherencia de los agrupamientos generados [37].

Algunos de estos indices de validación son:

- Indice Rand: es una métrica de validación externa utilizada para medir la similitud entre dos particiones de un conjunto de datos. Evalúa la proporción de pares de elementos que están agrupados de manera consistente (en el mismo clúster o en clústeres diferentes) en ambas particiones [38]. Su valor varía entre 0 y 1, donde 1 indica una coincidencia perfecta. Esta métrica es ampliamente utilizada en el análisis de agrupamientos para evaluar la calidad de los algoritmos en función de su capacidad para replicar particiones de referencia conocidas.
- Indice Dunn: es una métrica utilizada para evaluar la calidad de los agrupamientos generados por un algoritmo de clustering. Se calcula como el cociente entre la menor distancia entre dos clústeres diferentes (indicando separación) y la mayor distancia entre los puntos de un mismo clúster (indicando cohesión). Su objetivo es maximizar la separación entre clústeres y minimizar la dispersión dentro de ellos. Un valor alto del índice indica que los clústeres están bien separados y son compactos [39].
- Indice Xie-Beni: es una métrica ampliamente utilizada para validar agrupamientos generados mediante el algoritmo *Fuzzy C-Means* (FCM). Este índice evalúa tanto la compactación dentro de los clústeres como la separación entre ellos, proporcionando una métrica combinada que ayuda a determinar la calidad de los agrupamientos. Se define como la relación entre la distancia promedio entre los puntos y sus centros de clústeres y la distancia mínima entre los centros de clústeres. El valor óptimo del índice indica el número de clústeres que mejor representan la estructura subyacente de los datos [40].

6.10.3. Aprendizaje reforzado

Es una técnica del aprendizaje automático en la que un agente informático aprende a realizar una tarea a través de repetidas interacciones de prueba y error con un entorno dinámico. Este enfoque de aprendizaje permite que el agente tome una serie de decisiones que amplían al máximo una métrica de recompensa por la tarea hecha, sin intervención humana y sin estar programado explícitamente para completar la tarea [41].

Los programas de inteligencia artificial entrenados con aprendizaje reforzado superan a los humanos en juegos de mesa tales como Go y el ajedrez, así como en videojuegos. Aunque el aprendizaje reforzado no es de ninguna manera un concepto nuevo, los recientes avances en cuanto a Deep Learning y potencia informática han permitido lograr algunos resultados notables en el área de la inteligencia artificial [41].

6.11. VAT

El VAT (*Visual Assessment of (Cluster) Tendency*) es una técnica visual que evalúa la predisposición de un conjunto de datos para formar agrupamientos o clústeres de manera natural. Esta tendencia a la agrupación mide hasta qué punto los datos pueden dividirse en subconjuntos homogéneos sin necesidad de realizar una segmentación previa. El VAT es especialmente útil como paso exploratorio inicial para determinar si la aplicación de algoritmos de agrupamiento es viable y, de ser así, cuán estructurados están los datos [42].

El método VAT consta de dos pasos principales:

- Reordenación de los objetos: Utilizando una matriz de disimilaridad, los objetos se reorganizan para que aquellos con mayor similitud se ubiquen cercanos entre sí. Este paso permite resaltar patrones de relación en los datos.
- Generación de una imagen de intensidad: La matriz reordenada se transforma en una representación visual donde los valores de disimilaridad se expresan mediante niveles de intensidad. Los clústeres emergen como bloques oscuros alineados a lo largo de la diagonal principal de la imagen, ofreciendo una representación intuitiva de las tendencias de agrupamiento.

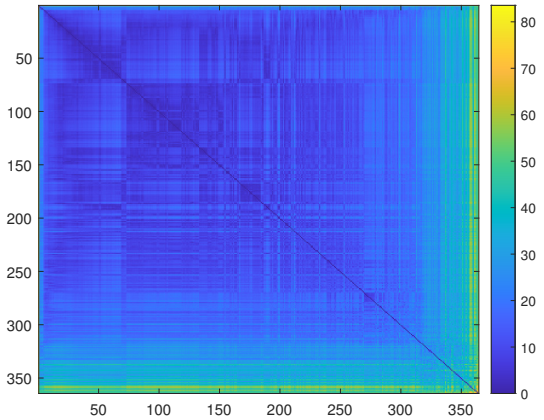

Figura 23: Ejemplo de visualización del VAT.

El VAT destaca por su simplicidad y versatilidad, siendo aplicable a cualquier conjunto de datos que pueda representarse como vectores de objetos o valores de disimilaridad por pares. Además, esta técnica es robusta frente a ruido y valores atípicos, lo que la convierte en una herramienta poderosa para detectar conglomerados antes de aplicar métodos de agrupamiento más complejos [42].

6.12. PCA

El Análisis de Componentes Principales (PCA) es una técnica estadística utilizada para reducir la dimensionalidad de un conjunto de datos, conservando la mayor cantidad posible de información original. Este método transforma un conjunto de variables posiblemente correlacionadas en un conjunto más pequeño de variables no correlacionadas, denominadas componentes principales. Estas nuevas variables están ordenadas según la cantidad de varianza que explican dentro del conjunto de datos [43].

El PCA se basa en los siguientes pasos fundamentales:

- Estandarización de los datos: Garantiza que todas las variables contribuyan de manera equitativa al análisis, independientemente de sus escalas.
- Cálculo de la matriz de covarianza: Evalúa las relaciones entre las variables del conjunto de datos.
- Obtención de valores y vectores propios: Los vectores propios indican las direcciones de máxima varianza, mientras que los valores propios reflejan la importancia relativa de cada dirección.
- Selección de componentes principales: Se eligen los componentes que explican la mayor parte de la varianza acumulada.

El PCA es una herramienta versátil que se utiliza en una amplia gama de aplicaciones, como la visualización de datos, el preprocesamiento para aprendizaje automático y la reducción de ruido en datos complejos. Es particularmente útil en situaciones donde los datos de alta dimensión pueden dificultar el análisis, facilitando la identificación de patrones y estructuras subyacentes en el conjunto de datos [43].

CAPÍTULO 7

Obtención y Procesamiento de Datos EEG

En este capítulo se describen las fuentes de datos utilizadas en el proyecto, incluyendo las señales EEG obtenidas de HUMANA y la base de datos TUH EEG Epilepsy Corpus. Se detallan los aspectos generales de cada conjunto de datos, como su origen, composición y relevancia para el análisis automático de señales EEG en el contexto de la detección de epilepsia.

7.1. Datos de HUMANA de pacientes con epilepsia

HUMANA ha compartido 6 grabaciones de señales EEG. La información de dichas señales se puede ver en el Cuadro 1. Estos datos son valiosos, ya que permiten evaluar la forma en que la institución captura sus señales EEG, ofreciendo un conjunto diverso para realizar pruebas con los algoritmos desarrollados en este proyecto.

Nombre	**Canales**	**Duración** (hh:mm:ss)	**Frecuencia de muestreo**
AL.edf	33	03:01:56	200Hz
CLEA.edf	29	02:39:42	200Hz
GIKA.edf	29	02:58:06	200Hz
HCHC.edf	29	23:02:44	200Hz
Ajczalar M.edf	30	15:57:13	300Hz
GUADRON T.edf	30	03:02:01	300Hz

Cuadro 1: Información de grabaciones de señales bioeléctricas de pacientes con epilepsia brindadas por HUMANA [11].

7.2. Datos del TUH EEG Epilepsy Corpus

El Temple University Hospital (TUH) EEG Corpus es una de las bases de datos de electroencefalografía (EEG) más grandes y completas disponibles públicamente para la investigación en aprendizaje automático. Esta base de datos contiene más de 25,000 estudios de EEG y datos de más de 14,000 pacientes, ofreciendo una gran cantidad de información clínica, incluyendo interpretaciones de neurólogos, historial médico y datos demográficos de los pacientes. La base de datos fue desarrollada por el Departamento de Neurología y el Neural Engineering Data Consortium de la Universidad de Temple, con el objetivo de facilitar la investigación y el desarrollo de algoritmos de aprendizaje automático para el análisis automático de EEGs [1].

En el presente proyecto se utilizó el subconjunto TUH EEG Epilepsy Corpus, que fue creado específicamente para el análisis automático de datos de pacientes con epilepsia. Este subconjunto contiene archivos en formato EDF (European Data Format) y reportes médicos correspondientes a 1,799 archivos en 570 sesiones de 200 pacientes, de los cuales 1,473 archivos de 436 sesiones pertenecen a pacientes diagnosticados con epilepsia, y los 326 archivos restantes de 134 sesiones corresponden a pacientes sin epilepsia [44].

El TUH EEG Epilepsy Corpus proporciona una amplia variedad de señales EEG con diferentes configuraciones de canales y características clínicas, lo que lo convierte en un recurso valioso para la evaluación y desarrollo de los algoritmos implementados en este proyecto. Estas señales, al igual que las proporcionadas por HUMANA, se someten a técnicas de agrupamiento para identificar y agrupar automáticamente los segmentos de interés relacionados con eventos epilépticos.

- Total de archivos: 1,799
- Número de sesiones: 570
- Pacientes con epilepsia: 1,473 archivos, 436 sesiones
- Pacientes sin epilepsia: 326 archivos, 134 sesiones
- Formato: EDF (European Data Format)

Los datos obtenidos del TUH EEG Epilepsy Corpus han sido utilizados para probar la robustez de los algoritmos de aprendizaje no supervisado desarrollados en este proyecto, permitiendo una evaluación exhaustiva de su capacidad para identificar y agrupar eventos epilépticos de forma automática.

7.3. Agrupamiento de datos

Dado que se utiliza la base de datos TUH EEG Epilepsy Corpus, los estudios están divididos en archivos EDF, donde cada uno contiene registros que detectan exclusivamente eventos epilépticos o no epilépticos. Con el fin de validar la efectividad de los algoritmos, se decidió agrupar los datos de manera intersujeto, estudio entre sujetos diferentes. Esto

significa que se combinaron estudios de un paciente con epilepsia junto con estudios de un paciente sano en un único análisis, asignando etiquetas a cada uno de estos estudios.

Posteriormente, los segmentos se mezclaron de forma aleatoria junto con sus etiquetas correspondientes, generando un conjunto de datos mixto que permitió evaluar el desempeño de los algoritmos de agrupamiento. A partir de este conjunto, se analizó la efectividad de los algoritmos comparando los agrupamientos generados con las etiquetas reales y evaluando, además, la similitud entre los agrupamientos producidos por los diferentes algoritmos.

Este enfoque permite medir de manera robusta la capacidad de los algoritmos para identificar correctamente los eventos epilépticos y diferenciarlos de los no epilépticos en un contexto variado, lo que resulta crucial para validar su uso en entornos clínicos.

Una vez organizados los datos, es importante considerar las propiedades intrínsecas de las señales EEG, que juegan un papel clave en la elección de los métodos analíticos y de agrupamiento, como se detalla a continuación.

7.4. Análisis de señales EEG

Las señales EEG representan la actividad eléctrica del cerebro, la cual es altamente compleja y puede variar significativamente entre personas debido a factores como la edad, la disposición de los electrodos, y otras características individuales. En muchos estudios, se tiende a realizar un análisis intrasujeto, que implica analizar las señales de un mismo individuo para identificar patrones específicos de esa persona. Este enfoque es útil en situaciones donde se busca monitorear la evolución de la actividad cerebral en un solo paciente.

En este proyecto, se optó por un análisis intersujeto debido a la falta de señales extensas de una sola persona en la base de datos que cuenten con etiquetas de segmentos sanos e ictales, lo que imposibilita validar las agrupaciones de manera intrasujeto. Este enfoque tiene como objetivo comprobar de forma robusta si los algoritmos son capaces de clasificar correctamente eventos epilépticos y no epilépticos en diferentes individuos. Aunque las señales EEG son altamente variables entre personas, lo que puede dificultar el desempeño de los algoritmos de agrupamiento, este tipo de análisis permite identificar patrones generales en grupos de individuos. Esto es clave para desarrollar modelos diagnósticos aplicables a una población más amplia, y resulta particularmente relevante en la detección de trastornos neurológicos como la epilepsia.

CAPÍTULO 8

Resultados Preliminares de Clustering

En este capítulo se detallan los resultados iniciales obtenidos en el proceso de agrupamiento de señales EEG mediante algoritmos de aprendizaje no supervisado. Estos resultados corresponden a un primer acercamiento al análisis y validación de los algoritmos, empleando diferentes configuraciones de características. Además, se presenta el uso preliminar de herramientas clave, como el *Visual Assessment of Tendency* (VAT), para evaluar la tendencia de agrupación de los datos. Este análisis preliminar no solo busca evaluar la capacidad de los algoritmos para identificar patrones, sino también establecer una base sólida para comprender las dinámicas del aprendizaje automático en el contexto del estudio.

8.1. VAT: Evaluación Visual de la Tendencia de Clustering

El método VAT es una herramienta diseñada para evaluar visualmente la tendencia de agrupación en un conjunto de datos, proporcionando una representación gráfica que permite determinar si los datos presentan una estructura agrupada natural. Este método es particularmente efectivo cuando los grupos son de forma aproximadamente circular o compacta, lo que facilita la identificación de patrones claros.

En el caso ideal, como se muestra en la Figura 24, los datos presentan grupos bien definidos, compactos y con una separación clara entre ellos. Esta estructura simplificada permite que el VAT sea útil para indicar la cantidad de grupos presentes y su distribución como se muestra en la Figura 25.

Sin embargo, en nuestro análisis, como se ilustra en la Figura 26, las muestras presentan una mayor dispersión y desorganización. Este tipo de distribución dificulta la eficacia del VAT, ya que los grupos no tienen una forma definida ni una separación evidente, lo que complica la interpretación visual del resultado como se observa en la Figura 27.

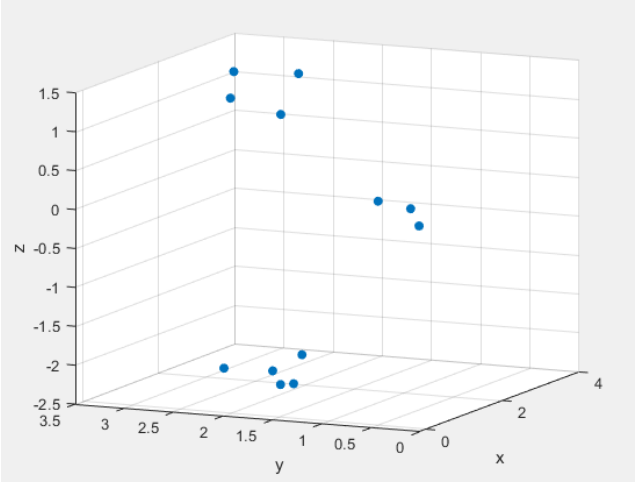

Figura 24: Grupo de datos ideal para VAT.

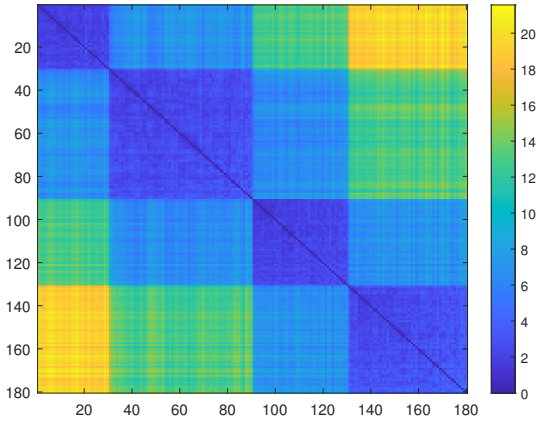

Figura 25: VAT ideal con grupos claros.

A pesar de estas limitaciones, el VAT sigue siendo útil como una herramienta exploratoria preliminar, permitiendo identificar datos que podrían no ser apropiados para ciertos métodos de agrupamiento y motivando el uso de técnicas adicionales, como PCA o normalización, para mejorar la calidad de las agrupaciones.

8.2. Extracción de características

Para realizar un análisis efectivo de las señales EEG, se ha implementado un proceso de extracción de características. Esto es necesario, ya que los algoritmos de aprendizaje no supervisado funcionan mediante vectores de características que describen la señal.

Se extrajeron características en en tres dominios diferentes: tiempo, frecuencia y wavelets. Cada uno de estos enfoques proporciona información que ayuda a los algoritmos de aprendizaje automático no supervisado a identificar patrones significativos dentro de las

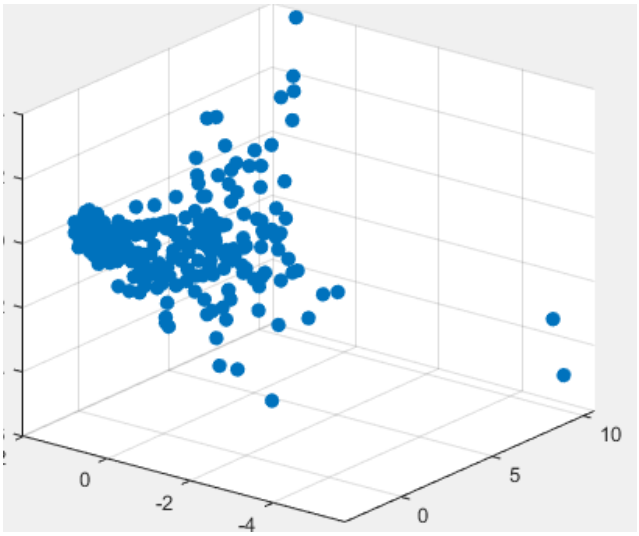

Figura 26: Grupo de datos ideal para VAT.

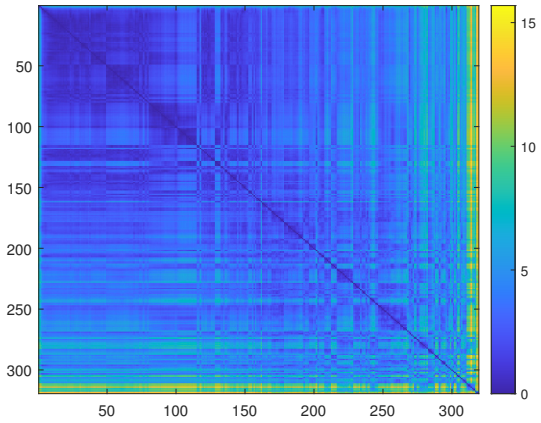

Figura 27: VAT ideal con grupos claros.

señales.

- Características en el dominio del tiempo: Las siguientes características fueron extraídas para capturar información relevante de la señal en el tiempo
  - Desviación estándar (STD).
  - Valor Medio Absoluto (MAV).
  - Cruces por cero (ZC).
  - Curtosis.
  - Asimetría.
- Características en el dominio de la frecuencia: Mediante la transformada de Fourier, se analizaron las siguientes características de frecuencia de la señal EEG:

- Relación Theta/Alpha: Relación entre las potencias en las bandas de frecuencia Theta (4-7 Hz) y Alpha (8-13 Hz).
- Relación Beta/Alpha: Relación entre las potencias en las bandas de frecuencia Beta (14-30 Hz) y Alpha.
- Relación Theta/Beta: Relación entre las potencias en las bandas de frecuencia Theta y Beta.
- Relación (Theta + Alpha)/Beta: Suma de las potencias en las bandas Theta y Alpha en relación con la potencia en la banda Beta.
- Relación (Theta + Alpha)/(Alpha + Beta): Suma de las potencias en Theta y Alpha en comparación con la suma de las potencias en Alpha y Beta.

- Características mediante transformadas wavelet: Las señales EEG también fueron analizadas mediante la Transformada Wavelet Discreta (DWT), utilizando la wavelet madre Daubechies 3 (db3) con un nivel de descomposición de 2. Esto permitió descomponer la señal en diferentes niveles de resolución, capturando tanto información en alta como en baja frecuencia. La característica extraída a partir de esta descomposición fue:

  - Potencia.

8.3. Resultado inciales del clustering

Las pruebas se realizaron utilizando datos de la base TUH EEG Epilepsy Corpus, agrupados de manera intersujeto. En algunas pruebas, se combinaron estudios de epilepsia y no epilepsia en grupos de diferentes tamaños. Se buscó que los sujetos de las pruebas estuvieran en un mismo rango de edad, ya que las señales EEG varían significativamente dependiendo de la edad. Para estas pruebas, se seleccionó un rango de edad de 40 a 60 años, debido a que hay más personas dentro de este rango de edad dentro del TUH EEG Epilepsy Corpus, lo que permite obtener estudios más similares y consistentes.

Una vez que los métodos de agrupamiento y validación fueron seleccionados y ajustados utilizando los datos del TUH EEG Corpus, las señales de HUMANA se utilizaron como una prueba adicional. Este enfoque permitió evaluar la generalización de los algoritmos desarrollados y analizar su comportamiento en un conjunto de datos diferente, contribuyendo a una validación más amplia de los métodos.

En este documento, los sujetos incluidos en las pruebas han sido referidos como “Sujeto 1”, “Sujeto 2”, etc., para facilitar la lectura y comprensión de los resultados. Sin embargo, los nombres originales asignados en la base de datos (por ejemplo, aaaaajrh, aaaaajrm) se encuentran documentados en los anexos de este trabajo para su consulta.

Cuando se trabaja con conjuntos de datos que contienen más de tres características, la visualización directa se vuelve inviable debido a las limitaciones impuestas por la dimensionalidad. En un espacio de más de tres dimensiones, los datos no pueden ser representados de forma gráfica en un entorno tridimensional, lo que dificulta la interpretación intuitiva de las relaciones y patrones entre las variables.

8.4. Prueba uno a uno

Esta prueba tuvo un propósito exploratorio, diseñada para evaluar preliminarmente el comportamiento de los algoritmos de agrupamiento con un conjunto de datos pequeño y controlado. Se incluyó un solo estudio de una persona con epilepsia y un solo estudio de una persona sana, asegurando que ambos sujetos tuvieran edades similares. Los sujetos fueron seleccionados aleatoriamente de la base de datos TUH EEG Corpus, bajo los nombres de "sujeto 1" y "sujeto 2", para facilitar su identificación en el análisis. Los resultados obtenidos, basados en la métrica de evaluación Rand Index, se muestran en los Cuadros 2, 3, 4.

Prueba frecuencia	
Algoritmos	**Resultado Rand Index**
K-means y etiquetas reales	0.8813
Clustering Jerárquico y etiquetas reales	0.83829
K-means y Clustering Jerárquico	0.95063

Cuadro 2: Prueba con un solo estudio para epilepsia y uno para no epilepsia en frecuencia.

Prueba tiempo continuo	
Algoritmos	**Resultado Rand Index**
K-means y etiquetas reales	1
Clustering Jerárquico y etiquetas reales	1
K-means y Clustering Jerárquico	1

Cuadro 3: Prueba con un solo estudio para epilepsia y uno para no epilepsia en tiempo continuo.

Prueba combinando características	
Algoritmos	**Resultado Rand Index**
K-means y etiquetas reales	1
Clustering Jerárquico y etiquetas reales	1
K-means y Clustering Jerárquico	1

Cuadro 4: Prueba con un solo estudio para epilepsia y uno para no epilepsia usando todas las características.

Dado que en el análisis basado en wavelets se utilizó únicamente una característica, se decidió omitir la prueba que involucraba exclusivamente esta categoría. Sin embargo, dicha característica fue incluida en las pruebas que combinaron todas las características juntas, permitiendo evaluar su aporte al desempeño general de los algoritmos de clustering en un contexto más amplio.

En cuanto a las características de frecuencia (Cuadro 2), se observa un rendimiento aceptable pero inferior comparado con las demás características. El Rand Index alcanzó valores de 0.8813 para K-Means y de 0.83829 para Clustering Jerárquico, lo cual refleja una buena pero no perfecta separación entre las señales ictales y no ictales. Estos resultados sugieren que las características de frecuencia, aunque útiles, no son completamente efectivas en la diferenciación de los eventos en este contexto.

Por otro lado, las características de tiempo continuo (Cuadro 3) mostraron un rendi-

miento sobresaliente, alcanzando un Rand Index perfecto de 1 en todos los casos, tanto al comparar con etiquetas reales como entre los dos algoritmos de agrupamiento. Esto indica que las características en el dominio del tiempo son altamente efectivas para capturar las diferencias entre las señales ictales y no ictales en este tipo de prueba. Este rendimiento sugiere que las pruebas intrasujeto, donde se analizan datos de un solo individuo, podrían generar resultados consistentemente altos en escenarios clínicos reales.

Finalmente, al combinar todas las características (Cuadro 4), el Rand Index también alcanzó un valor perfecto de 1, mostrando que el uso de múltiples características fortalece la robustez del análisis y confirma la capacidad de los algoritmos para identificar patrones relevantes en los datos. Este resultado reafirma la importancia de emplear un enfoque combinado en aplicaciones prácticas, donde se puedan integrar múltiples dimensiones de información.

8.5. Prueba sujeto a sujeto

Para esta prueba, se seleccionó un conjunto de estudios de un sujeto con epilepsia y otro conjunto de estudios de un sujeto sano, asegurando que ambos tuvieran edades similares y un número comparable de estudios. Los sujetos denominados sujeto 5 (23 estudios), sujeto 4 (19 estudios) y sujeto 3 (17 estudios) fueron escogidos para este análisis. Los estudios se agruparon y analizaron utilizando los algoritmos de agrupamiento no supervisado K-Means y Clustering Jerárquico. Esta prueba sigue un enfoque similar al de la prueba uno a uno, pero incorpora un mayor número de datos, garantizando que provengan de las mismas personas para mantener la coherencia en el análisis. Los resultados obtenidos, basados en la métrica de evaluación Rand Index, se muestran en las tablas 5, 6, 7.

Prueba frecuencia		
Algoritmos	**Rand Index sujeto 4 y 5**	**Rand Index sujeto 4 y 3**
K-means y etiquetas reales	0.60265	0.86225
Clustering Jerárquico y etiquetas reales	0.61021	0.82594
K-means y Clustering Jerárquico	0.98376	0.95719

Cuadro 5: Resultados de Rand Index para pruebas frecuencia con combinaciones de diferentes sujetos epilépticos y sanos (4-5, 4-3).

Prueba tiempo continuo		
Algoritmos	**Rand Index sujeto 4 y 5**	**Rand Index sujeto 4 y 3**
K-means y etiquetas reales	0.86346	0.78489
Clustering Jerárquico y etiquetas reales	0.87224	0.84023
K-means y Clustering Jerárquico	0.98982	0.93242

Cuadro 6: Resultados de Rand Index para pruebas tiempo continuo con combinaciones de diferentes sujetos epilépticos y sanos (4-5, 4-3).

Prueba combinando características		
Algoritmos	**Rand Index sujeto 4 y 5**	**Rand Index sujeto 4 y 3**
K-means y etiquetas reales	0.62513	0.91629
Clustering Jerárquico y etiquetas reales	0.62309	0.55123
K-means y Clustering Jerárquico	0.99591	0.52667

Cuadro 7: Resultados de Rand Index para combinación de características con combinaciones de diferentes sujetos epilépticos y sanos (4-5, 4-3).

En el Cuadro 5 se observa que el Rand Index entre K-Means y Clustering Jerárquico es alto (0.98376 y 0.95719), lo que indica una alta consistencia entre ambos algoritmos al utilizar características basadas en frecuencia. Sin embargo, al comparar los algoritmos con las etiquetas reales, los valores del Rand Index para la combinación de sujetos 4-5 son más bajos (0.60265 y 0.61021), lo que sugiere que las características de frecuencia no son suficientes para separar completamente los segmentos epilépticos y no epilépticos en este conjunto de datos. En contraste, los valores del Rand Index para la combinación de sujetos 5 y 3 son más altos (0.86225 y 0.82594), lo que indica que, en este caso, las características de frecuencia sí resultan efectivas para diferenciar los segmentos.

En el Cuadro 6 los resultados muestran una mejora notable tanto para la combinación de sujetos 4-5 y resultados un poco menores para la combinación de sujetos 3-5. En el caso de los sujetos 4-5, el Rand Index entre K-Means y Clustering Jerárquico es alto (0.98982), reflejando una fuerte concordancia entre los algoritmos al utilizar características del dominio del tiempo. Los valores del Rand Index en comparación con las etiquetas reales también son elevados (0.86346 para K-Means y 0.87224 para Clustering Jerárquico), superando significativamente los resultados obtenidos con las características de frecuencia.

Por otro lado, para los sujetos 3-5, aunque los valores del Rand Index entre los algoritmos (0.93242) y respecto a las etiquetas reales (0.78489 para K-Means y 0.84023 para Clustering Jerárquico) son algo menores en comparación con los sujetos 4-5, siguen mostrando un desempeño considerablemente bueno. Esto indica que las características en el dominio del

tiempo son efectivas para capturar información relevante en ambos casos, aunque es evidente que la selección de sujetos influye en los resultados.

En el Cuadro 7 los resultados para las pruebas con la combinación de características muestran tendencias interesantes. Para la combinación de sujetos 4-5, el Rand Index entre K-Means y Clustering Jerárquico es muy alto (0.99591), indicando una fuerte concordancia entre los dos algoritmos. Sin embargo, al compararlos con las etiquetas reales, los valores son más bajos (0.62513 para K-Means y 0.62309 para Clustering Jerárquico), lo que sugiere que, aunque la combinación de características proporciona una visión más completa de los datos, no siempre logra una separación ideal entre segmentos epilépticos y no epilépticos en este contexto.

En el caso de la combinación de sujetos 5-3, los valores del Rand Index al compararlos con las etiquetas reales (0.91629 para K-Means y 0.55123 para Clustering Jerárquico) son considerablemente variables. Esto podría estar relacionado con las diferencias inherentes en las características de los datos de los sujetos seleccionados, lo que afecta la capacidad de los algoritmos para capturar patrones consistentes. Además, el Rand Index entre los algoritmos (0.52667) muestra una menor concordancia, lo que podría reflejar cómo las características combinadas interactúan de manera diferente en este caso.

Es importante señalar que los resultados del Rand Index también pueden depender significativamente de la inicialización de los algoritmos de agrupamiento, especialmente en el caso de K-Means, donde la elección de los centroides iniciales puede influir en la convergencia final del algoritmo. Por lo tanto, estas variaciones subrayan la importancia de realizar múltiples inicializaciones o experimentos para obtener un panorama más robusto de los resultados y su estabilidad.

Estos resultados resaltan la importancia de contar con señales intrasujeto para estos análisis, ya que la variabilidad de las señales entre diferentes personas puede afectar considerablemente los resultados del agrupamiento. Utilizar señales de un solo sujeto elimina la influencia de las diferencias individuales, permitiendo evaluar de manera más precisa la capacidad de las características seleccionadas para separar los segmentos epilépticos y no epilépticos. Por ello, es fundamental priorizar el análisis intrasujeto cuando se busca obtener resultados más consistentes y clínicamente relevantes.

8.6. Evaluación General del Conjunto de Datos

Para esta prueba, se incluyeron un total de 37 personas sin epilepsia con 254 estudios y 17 personas con epilepsia con 184 estudios. Los estudios se agruparon y analizaron utilizando dos algoritmos de agrupamiento no supervisado: K-Means y Clustering Jerárquico. Los resultados obtenidos para la métrica de evaluación Rand Index se muestran en los Cuadros 8 y 9.

En el Cuadro 8 se observa que los algoritmos K-Means y Clustering Jerárquico generan agrupamientos altamente similares entre ellos, ya que el Rand Index entre ambos es de 0.99751, lo que indica una casi perfecta coincidencia en los agrupamientos generados con las características utilizadas en frecuencia.

Prueba con muchos EDFs frecuencia	
Algoritmos	**Resultado Rand Index**
K-means y etiquetas reales	0.63049
Clustering Jerárquico y etiquetas reales	0.62998
K-means y Clustering Jerárquico	0.99751

Cuadro 8: Prueba con 254 EDF sin epilepsia y 184 EDF con epilepsia en frecuencia.

Sin embargo, al comparar ambos algoritmos con las etiquetas reales, los valores del Rand Index muestran valores un poco bajos, pero aceptables teniendo en cuenta la cantidad de personas que se incluyeron en esta prueba.

Prueba con muchos EDFs tiempo continuo	
Algoritmos	**Resultado Rand Index**
K-means y etiquetas reales	0.61709
Clustering Jerárquico y etiquetas reales	0.64697
K-means y Clustering Jerárquico	0.90637

Cuadro 9: Prueba con 254 EDF sin epilepsia y 184 EDF con epilepsia en tiempo continuo.

En el Cuadro 9 se observa que los algoritmos K-Means y Clustering Jerárquico generan agrupamientos altamente similares entre ellos, ya que el Rand Index entre ambos es de 0.90637, lo que refleja una similitud alta al utilizar características en el dominio de tiempo continuo. Este resultado es consistente con la tendencia observada en la prueba de frecuencia, lo que sugiere que ambos algoritmos son capaces de encontrar patrones similares en estos tipos de características.

Al igual que en los resultados obtenidos para el análisis en frecuencia, los valores del Rand Index en comparación con las etiquetas reales son más bajos, pero siguen siendo aceptables considerando la diversidad de personas incluidas en el agrupamiento. La variabilidad de las señales EEG entre individuos podría explicar la dificultad de alcanzar una correspondencia más alta con las etiquetas reales.

Prueba con muchos EDFs combinando características	
Algoritmos	**Resultado Rand Index**
K-means y etiquetas reales	0.5199
Clustering Jerárquico y etiquetas reales	0.51599
K-means y Clustering Jerárquico	0.97823

Cuadro 10: Prueba con 254 EDF sin epilepsia y 184 EDF con epilepsia combinando todas las características.

En el Cuadro 10 se observa que la combinación de todas las características permitió evaluar el desempeño global de los algoritmos de clustering en un contexto más amplio. Los resultados muestran que, al combinar las características, los algoritmos K-Means y Clustering Jerárquico siguen produciendo agrupamientos altamente similares entre ellos, con un Rand Index de 0.97823. Esto reafirma la consistencia de los algoritmos al trabajar con datos diversos y multidimensionales.

Sin embargo, al comparar los agrupamientos con las etiquetas reales, los valores del Rand Index disminuyen considerablemente en comparación con las pruebas específicas de características, lo que podría deberse al incremento de la complejidad de los datos al incluir características adicionales junto a la alta variación en las señales al ser tantos estudios de diferentes sujetos y diferentes tiempos.

A pesar de ello, esta combinación de características representa una herramienta valiosa para explorar de manera más general las capacidades de los algoritmos y su respuesta ante conjuntos de datos más heterogéneos. Este análisis global proporciona un punto de partida para futuras optimizaciones en la selección de características y en la mejora de los algoritmos utilizados.

8.7. PCA en el contexto del estudio

En este proyecto, el Análisis de Componentes Principales (PCA) se empleó para reducir la dimensionalidad del conjunto de datos, originalmente compuesto por 10 características extraídas de las señales EEG. Esta técnica permitió transformar las características en un espacio reducido, optimizando la representación de los datos al retener la mayor proporción posible de la varianza explicada.

Al inicio de esta fase, al analizar las características puras (Figura 28), se identificó que la característica "Potencia" tenía una varianza considerablemente superior al resto, lo que generaba un sesgo significativo en los resultados del PCA. Este sesgo influía en la distribución de las componentes principales (Figura 29), concentrando la varianza principalmente en la primera componente principal y dificultando la interpretación del resto de las dimensiones. Por esta razón, se decidió excluir la característica "Potencia" en las pruebas subsecuentes, mejorando así la distribución de los datos y reduciendo la influencia de características dominantes.

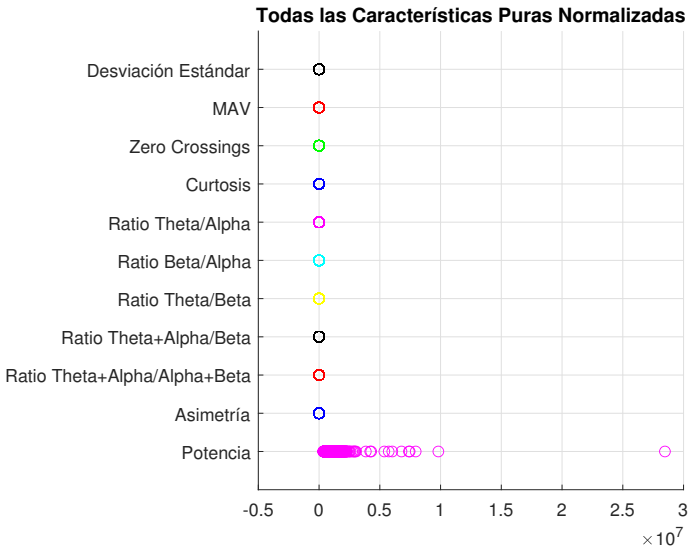

Figura 28: Dispersión de las características, enfocando la característica de potencia.

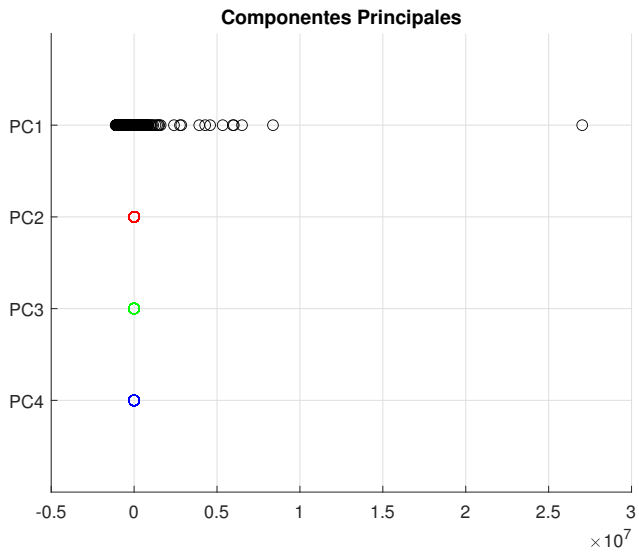

Figura 29: PCA sesgada por la característica de potencia.

Además, el uso de PCA tuvo beneficios en términos de eficiencia y rendimiento. Al reducir la cantidad de dimensiones, se disminuyó significativamente la carga computacional del proceso de agrupamiento, lo que facilitó la implementación de los algoritmos K-Means y Clustering Jerárquico. Comparando los resultados con y sin PCA, se observó una mejora consistente en el Rand Index al validar los agrupamientos generados con las etiquetas reales. Esto evidencia que PCA no solo optimizó el tiempo de procesamiento, sino que también contribuyó a una mayor precisión en la validación de los clusters generados.

No necesariamente el primer componente principal proporciona la mejor separación entre los datos, ya que esta representa la mayor varianza de los datos en general, pero no siempre captura las características más relevantes para la clasificación. Por ello, se consideraron también la segunda y tercera componentes principales, así como combinaciones de estas, para determinar cuál proporcionaba el mejor resultado en términos de agrupamiento.

8.8. Prueba uno a uno con PCA

Siguiendo con la línea de las pruebas realizadas anteriormente, se llevaron a cabo pruebas uno a uno utilizando el PCA. Estas pruebas emplearon los mismos algoritmos de agrupamiento, K-Means y Clustering Jerárquico, con el objetivo de evaluar el impacto del PCA en la calidad de los agrupamientos generados.

Para la comparación, se utilizó la combinación de todas las características originales como referencia, ya que proporciona una descripción amplia de la señal EEG. Esta se contrastó con los resultados obtenidos al trabajar con los primeros tres componentes principales derivados del PCA, que retienen la mayor parte de la varianza explicada en los datos. Además, se identificará y mencionará en los casos donde el segundo o tercer componente principal haya proporcionado mejores resultados en comparación con la combinación de los primeros tres componentes principales.

En la Figura 30 se puede observar la varianza de las características y en la Figura 31 se puede observar la varianza del PCA. Estas visualizaciones pueden darnos una idea de como los algoritmos buscarán los grupos.

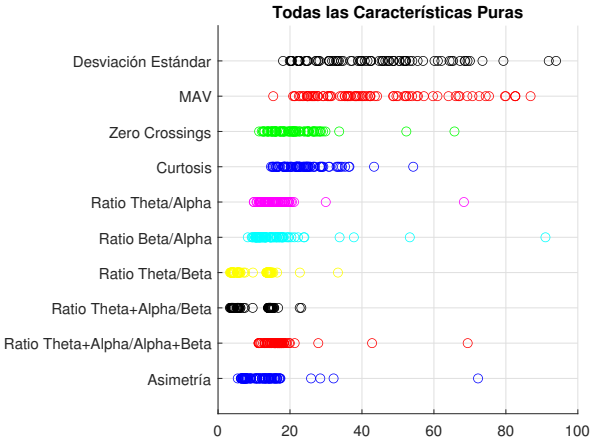

Figura 30: Varianza características puras.

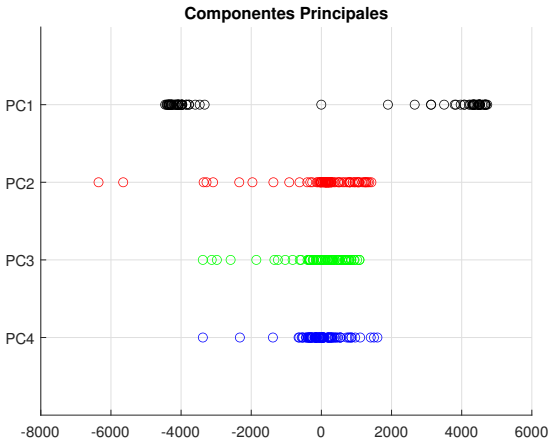

Figura 31: Componentes principales.

En el Cuadro 11, se observa que la combinación de características, incluso excluyendo la potencia, permitió obtener una identificación perfecta de los segmentos, logrando un Rand Index de 1 en todos los casos. Esto evidencia la efectividad de las características seleccionadas para capturar patrones relevantes en las señales y diferenciar claramente entre segmentos epilépticos y no epilépticos.

Por otro lado, en el Cuadro 12, los resultados muestran un desempeño ligeramente inferior al utilizar PCA. Aunque K-means mantiene un Rand Index de 1 al ser comparado con las etiquetas reales, los valores obtenidos para Clustering Jerárquico muestran una ligera disminución, alcanzando un Rand Index de 0.97561 tanto al compararse con las etiquetas reales como con K-means. Sin embargo, esta diferencia mínima en el desempeño es aceptable, especialmente considerando que el uso de PCA reduce significativamente la carga computacional, facilitando el análisis y el procesamiento de los datos.

Combinación de características	
Algoritmos	**Resultado Rand Index**
K-means y etiquetas reales	1
Clustering Jerárquico y etiquetas reales	1
K-means y Clustering Jerárquico	1

Cuadro 11: Rand Index usando combinación de características.

PCA dato a dato	
Algoritmos	**Resultado Rand Index**
K-means y etiquetas reales	1
Clustering Jerárquico y etiquetas reales	0.97561
K-means y Clustering Jerárquico	0.97561

Cuadro 12: Rand Index usando PCA.

Además, gracias a la incorporación del Análisis de Componentes Principales (PCA), fue posible visualizar los datos y los agrupamientos en un espacio tridimensional, lo cual facilitó la interpretación de los resultados obtenidos. Las Figuras 32, 33 y 34 presentan ejemplos de estas visualizaciones, donde se observa cómo los datos proyectados en los primeros tres componentes principales permiten identificar de manera más clara los patrones y la estructura de los clústeres generados. Esta capacidad de representación gráfica en 3D no solo aporta una mejor comprensión de los datos, sino que también refuerza la utilidad del PCA como herramienta clave en el análisis exploratorio y en la validación de los algoritmos de agrupamiento.

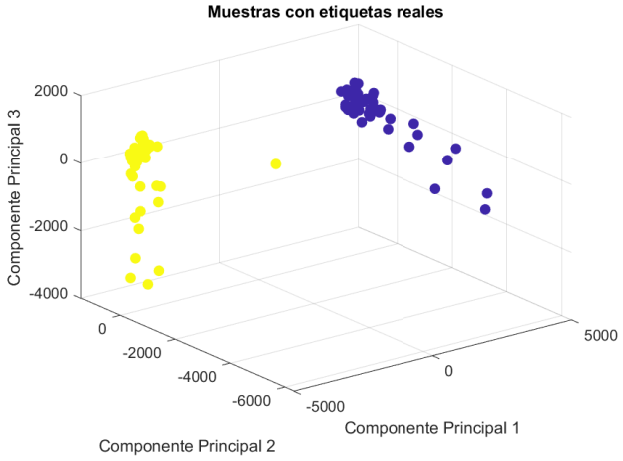

Figura 32: Datos agrupados de manera real.

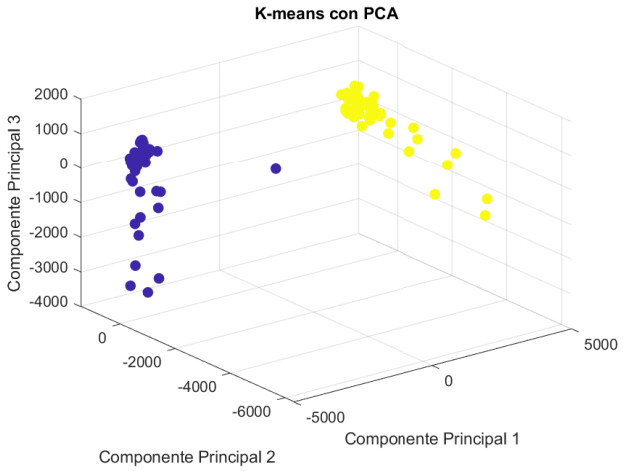

Figura 33: Datos agrupados por K-means.

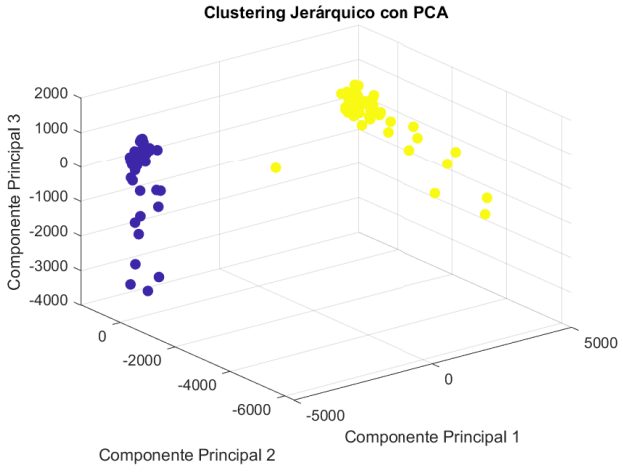

Figura 34: Datos agrupado por el cluster jerarquíco.

8.9. Prueba sujeto a sujeto con PCA

Igual que en la sección anterior se siguió con la línea de las pruebas, en este caso se llevo a cabo la prueba sujeto a sujeto utilizando el PCA. Estas pruebas se hicieron seleccionando un conjunto de estudio de un sujeto con epilepsia y otro conjunto de estudios de un sujeto sano, asegurando que los sujetos tuvieran edades similares y un número comparable de estudios. Los sujetos elegidos para el estudió fueron los sujetos 3, 4 y 5.

Combinación de características		
Algoritmos	**Rand Index sujeto 4 y 5**	**Rand Index sujeto 4 y 3**
K-means y etiquetas reales	0.86346	0.82948
Clustering Jerárquico y etiquetas reales	0.87224	0.84023
K-means y Clustering Jerárquico	0.98982	0.70821

Cuadro 13: Resultados de Rand Index para pruebas combinando caracteristicas con combinaciones de diferentes sujetos epilépticos y sanos (4-5, 4-3).

PCA sujeto a sujeto		
Algoritmos	**Rand Index sujeto 4 y 5**	**Rand Index sujeto 4 y 3**
K-means y etiquetas reales	0.86346	0.75905
Clustering Jerárquico y etiquetas reales	0.86872	0.84023
K-means y Clustering Jerárquico	0.98982	0.90047

Cuadro 14: Resultados de Rand Index para pruebas usando el PCA con combinaciones de diferentes sujetos epilépticos y sanos (4-5, 4-3).

En los Cuadros 13 y 14 se presentan los resultados del Rand Index para las pruebas sujeto a sujeto, utilizando tanto la combinación de características originales (sin la potencia) como el Análisis de Componentes Principales (PCA). Los resultados permiten evaluar la efectividad de los algoritmos de agrupamiento en ambas configuraciones y para diferentes combinaciones de sujetos.

Para la combinación de características originales (Cuadro 13), los valores del Rand Index son consistentemente altos en las combinaciones de los sujetos 4 y 5. El Rand Index alcanza valores de 0.98982 para la combinación de K-Means y Clustering Jerárquico, lo que refleja una gran concordancia entre los algoritmos y las etiquetas reales. Sin embargo, al pasar a la combinación de los sujetos 4 y 3, se observa una disminución considerable en el Rand Index, especialmente para K-Means y Clustering Jerárquico (0.70821). Esto sugiere que la variabilidad introducida por las diferencias individuales entre los sujetos tiene un impacto

significativo en los resultados de agrupamiento.

En el caso de PCA sujeto a sujeto (Cuadro 14), los resultados muestran un comportamiento similar. Para los sujetos 4 y 5, el Rand Index es igualmente alto (0.98982) para la combinación de K-Means y Clustering Jerárquico, indicando que PCA logra preservar la estructura de los datos y produce agrupamientos efectivos. Sin embargo, para los sujetos 4 y 3, se observa una mejora en comparación con las características originales: el Rand Index para K-Means y Clustering Jerárquico sube a 0.90047. Esto demuestra que PCA puede mitigar parcialmente el impacto de la variabilidad intersujeto, permitiendo una mejor agrupación en escenarios más complejos.

Los resultados entre las clasificaciones realizadas mediante PCA y la combinación de características puras son bastante similares. Entre todos los valores de Rand Index obtenidos con las etiquetas reales, se calcula una desviación estándar de 0.034, lo que indica una variabilidad baja en los resultados. Esto es un aspecto muy valioso para la integración del PCA, ya que, aunque no se observe una mejora sustancial en los índices de agrupamiento, se logra mantener un rendimiento comparable mientras se reduce significativamente la dimensionalidad de los datos. Esto, a su vez, disminuye la carga computacional, haciendo que el proceso sea más eficiente sin comprometer la precisión del modelo.

En las Figuras 35, 36 y 37 podemos ver el grupo de muestra con las etiquetas reales y las agrupaciones de los clusters de la combinación sujetos 3-4. En las Figuras 38, 39 y 40 podemos ver el grupo de muestra con las etiquetas reales y las agrupaciones de los clusters de la combinación sujetos 5-4.

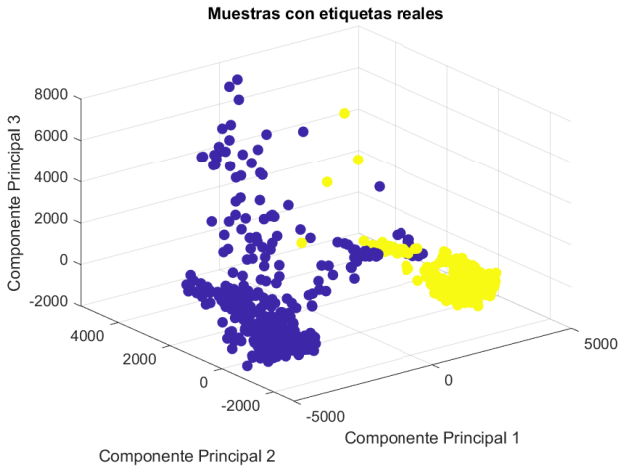

Figura 35: Datos agrupados de manera real en prueba sujeto 3-4.

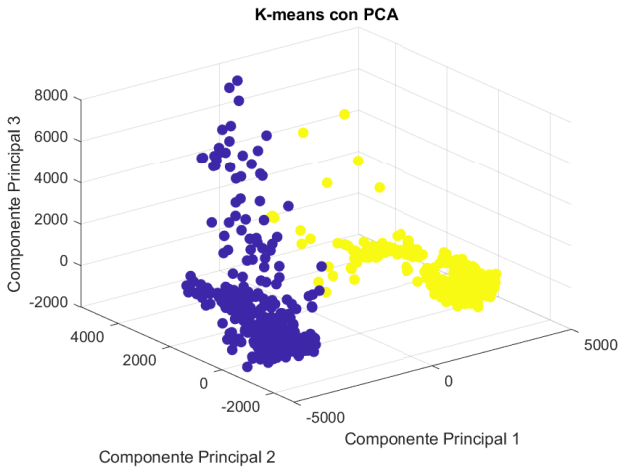

Figura 36: Datos agrupados por K-means sujeto 3-4.

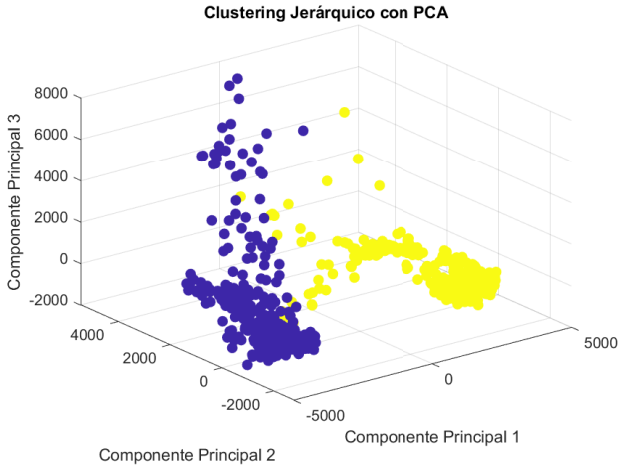

Figura 37: Datos agrupado por el cluster jerarquíco sujetos 3-4.

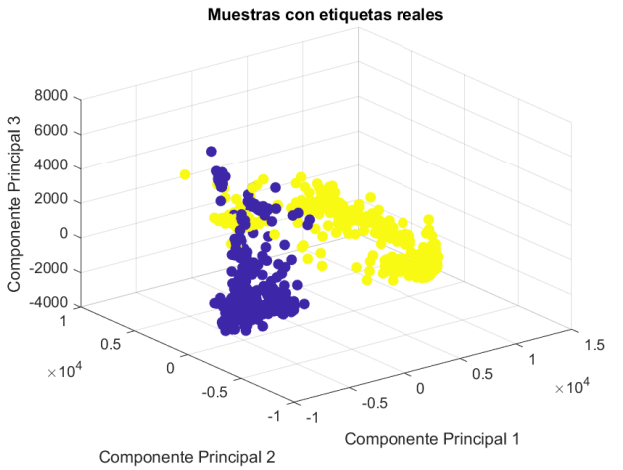

Figura 38: Datos agrupados de manera real en prueba sujeto 5-4.

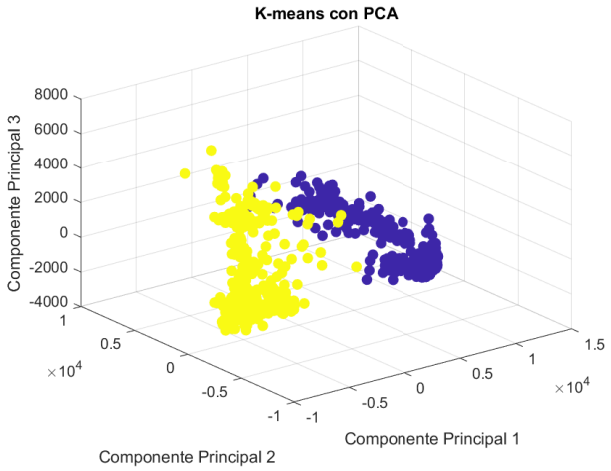

Figura 39: Datos agrupados por K-means sujeto 5-4.

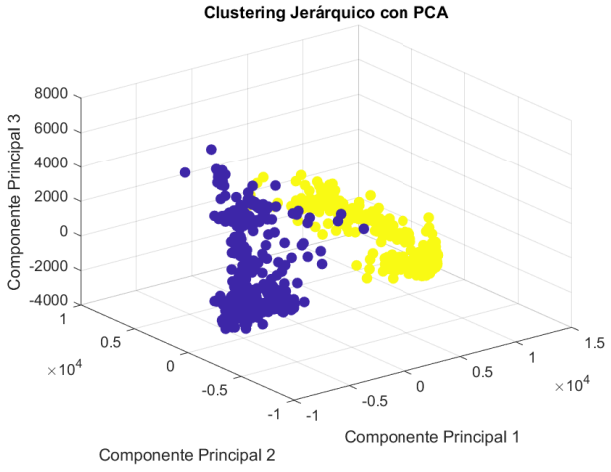

Figura 40: Datos agrupado por el cluster jerarquíco sujetos 5-4.

8.10. Evaluación General del Conjunto de Datos con PCA

En línea con las pruebas realizadas anteriormente, se llevó a cabo una evaluación general integrando un mayor número de señales EEG. Esta prueba incluyó estudios de múltiples personas, lo que añade una mayor versatilidad a los datos al considerar la variabilidad inherente entre los sujetos. Al utilizar PCA en este escenario, el objetivo principal fue analizar si la reducción de dimensionalidad permite mantener un rendimiento consistente en los algoritmos de agrupamiento al enfrentar un conjunto de datos más diverso.

En este análisis, se compararon los resultados obtenidos mediante la combinación de características originales y los componentes principales seleccionados con PCA, evaluando cómo estas configuraciones impactan en la capacidad de los algoritmos para identificar patrones y agrupar correctamente los segmentos epilépticos y no epilépticos en un conjunto de datos más amplio y complejo.

Agrupamiento con muchos EDFs combinando características	
Algoritmos	**Resultado Rand Index**
K-means y etiquetas reales	0.61709
Clustering Jerárquico y etiquetas reales	0.64
K-means y Clustering Jerárquico	0.90702

Cuadro 15: Agrupamiento con 254 EDF sin epilepsia y 184 EDF con epilepsia combinando todas las características.

Agrupamiento con muchos EDFs utilizando PCA	
Algoritmos	**Resultado Rand Index**
K-means y etiquetas reales	0.61595
Clustering Jerárquico y etiquetas reales	0.63479
K-means y Clustering Jerárquico	0.93041

Cuadro 16: Agrupamiento con 254 EDF sin epilepsia y 184 EDF con epilepsia utilizando PCA.

En el Cuadro 15 y el Cuadro 16, se presentan los resultados del Rand Index obtenidos al realizar pruebas de agrupamiento con un conjunto extenso de datos, utilizando tanto la combinación de características originales como los componentes principales generados con PCA.

Para ambos enfoques, los resultados del Rand Index al comparar K-means y etiquetas reales son muy similares, alcanzando valores de 0.61709 para la combinación de características y 0.61595 al utilizar PCA. De manera similar, al comparar Clustering Jerárquico y etiquetas reales, se obtienen valores de 0.64 con la combinación de características y 0.63479 con PCA. Estas métricas reflejan que la integración del PCA no afecta significativamente el rendimiento de los algoritmos cuando se comparan con las etiquetas reales, demostrando que el PCA preserva la información relevante incluso después de reducir la dimensionalidad.

Sin embargo, al observar la métrica de K-means y Clustering Jerárquico, se nota una ligera mejora al utilizar PCA, con un Rand Index de 0.93041 frente a 0.90702 al emplear la combinación de características. Esto indica una mayor concordancia entre los algoritmos de agrupamiento al usar PCA, lo que refuerza su capacidad para capturar patrones consistentes

en un conjunto de datos más amplio y diverso.

Un aspecto clave es que el uso de PCA no solo mantiene la calidad de los agrupamientos generados, sino que también aporta beneficios significativos al reducir la complejidad computacional y facilitar la interpretación de los datos. En escenarios con grandes volúmenes de datos, como el presente estudio, esta reducción de dimensionalidad es especialmente valiosa, permitiendo un análisis más eficiente sin comprometer el rendimiento.

En las figuras 41, 42 y 43, se presentan las visualizaciones de los datos etiquetados y los resultados de los algoritmos de agrupamiento. La Figura 41 muestra la distribución de los datos de acuerdo con sus etiquetas reales, lo que permite observar la separación teórica de los segmentos epilépticos y no epilépticos. Por otro lado, las Figuras 42 y 43 representan los agrupamientos generados por los algoritmos K-Means y Clustering Jerárquico, respectivamente, utilizando PCA para reducir la dimensionalidad de los datos. Estas visualizaciones son fundamentales para evaluar cómo los algoritmos identifican patrones y cuán similares son sus resultados respecto a las etiquetas reales.

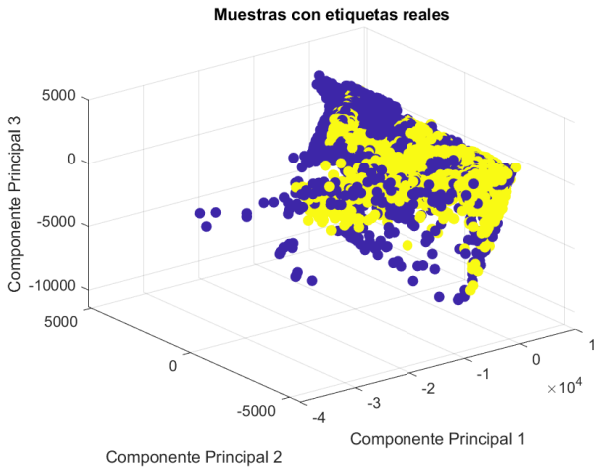

Figura 41: Datos agrupados de manera real en prueba sujeto 3-4.

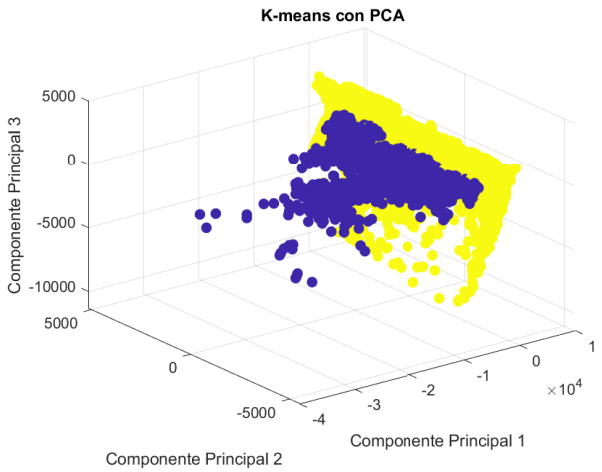

Figura 42: Datos agrupados por K-means sujeto 3-4.

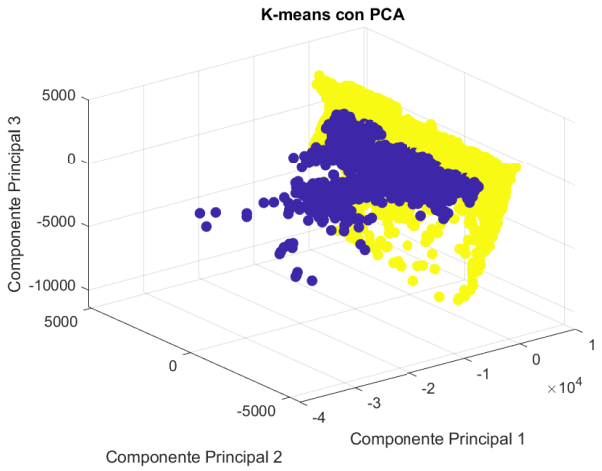

Figura 43: Datos agrupado por el cluster jerarquíco sujetos 3-4.

CAPÍTULO 9

Normalización *Z-score*

En las pruebas intrasujeto, se observó que la varianza de los datos se sesgaba debido a que algunas características presentaban valores significativamente más altos que otras como se observa en la Figura 44 afectando al PCA como se ve en la Figura 45. Este sesgo podía influir negativamente en los resultados de los algoritmos de agrupamiento, dificultando la identificación precisa de patrones en los datos. Para solucionar este problema, se decidió implementar la normalización *Z-score*.

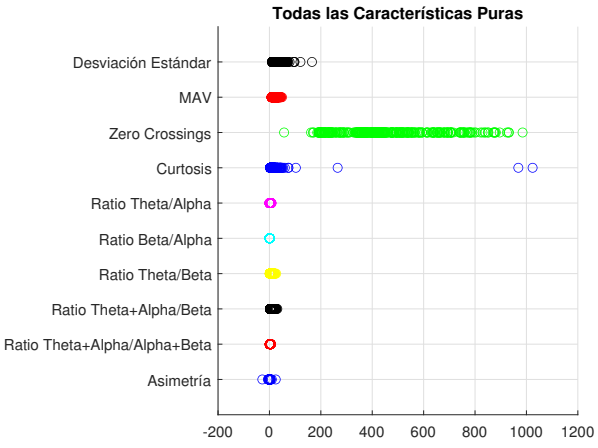

Figura 44: Varianza de características puras.

La normalización *Z-score* transforma los datos de un conjunto para que tengan una media de 0 y una desviación estándar de 1. Esta técnica garantiza que todas las características contribuyan de manera equitativa al análisis, eliminando los efectos de las diferencias en escalas o magnitudes. La fórmula para calcular el *Z-score* de una variable X es la siguiente:

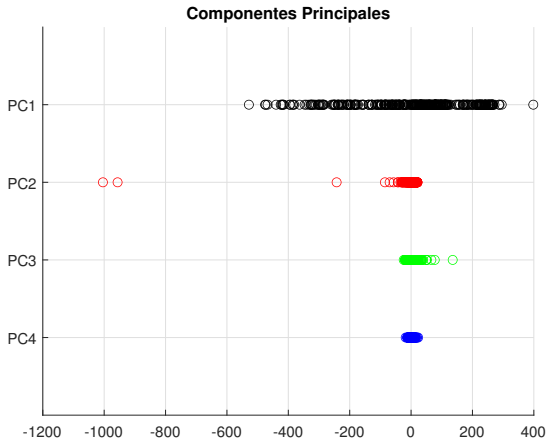

Figura 45: PCA de características puras sin normalizar.

$$Z = \frac{X - \mu}{\sigma}$$

- Z: Valor normalizado.
- X: Valor original de la variable.
- μ: Media de los datos.
- σ: Desviación estándar.

El uso de esta normalización permitió igualar la escala de todas las características, asegurando que ninguna característica con mayor varianza dominara el análisis como se observa en la Figura 46, esta normalización también afecta de manera positiva al PCA como se ve en la Figura 47, ya que antes se sesgaba junto con las características. Esto fue especialmente relevante para las pruebas intrasujeto, donde las características deben reflejar con precisión las diferencias entre segmentos epilépticos y no epilépticos, sin introducir sesgos debido a escalas desbalanceadas.

Sin embargo, cabe destacar que las características intersujeto no parecían verse afectadas de manera significativa por las diferencias de escala y varianza, como ocurrió en las pruebas intrasujeto. Esto se debe a que, en un contexto intersujeto, las variaciones inherentes entre diferentes individuos dominan el análisis, lo que hace que el impacto de características con varianza desbalanceada sea menos crítico. A pesar de esto, la normalización *Z-score* fue implementada como una medida general para garantizar consistencia y equidad en las contribuciones de las características en todos los escenarios.

Es importante mencionar que la implementación de esta técnica podría afectar ligeramente los resultados de las pruebas intersujeto, pero se consideró esencial para las pruebas intrasujeto, ya que en un contexto clínico real, la aplicación práctica de los algoritmos se centrará principalmente en analizar datos de un solo individuo (intrasujeto). Esto refuerza la relevancia de priorizar técnicas como la normalización *Z-score*, que aseguren un análisis confiable y robusto en el entorno clínico.

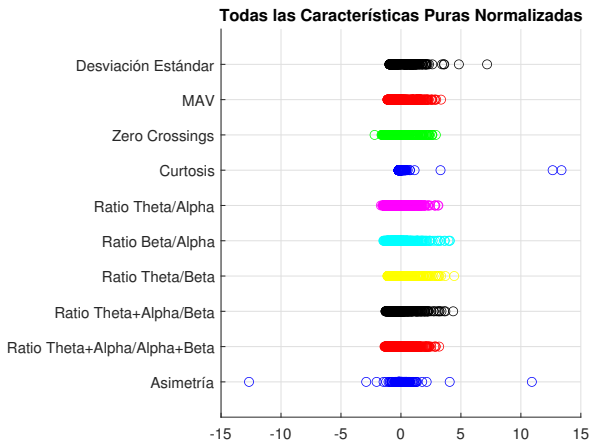

Figura 46: Varianza de características puras normalizadas.

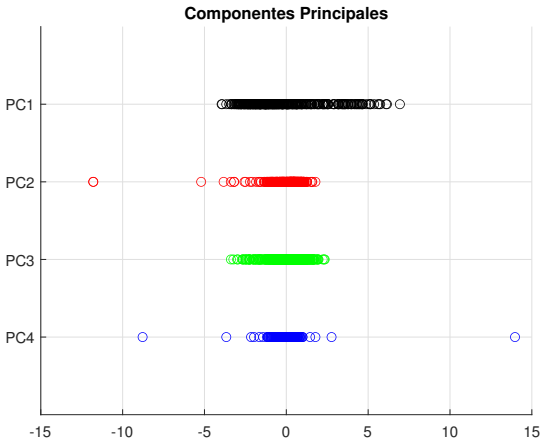

Figura 47: PCA de características puras normalizadas.

9.1. Resultados de pruebas anteriores

En esta sección se hará un breve resumen de los resultados de las mismas pruebas hechas anteriormente, pero ahora con la normalización *Z-score*, luego se discutirá sobre los resultados.

Combinación de características	
Algoritmos	**Resultado Rand Index**
K-means y etiquetas reales	0.9756
Clustering Jerárquico y etiquetas reales	1
K-means y Clustering Jerárquico	0.9756

Cuadro 17: Rand Index usando combinación de características en la prueba dato a dato.

PCA	
Algoritmos	**Resultado Rand Index**
K-means y etiquetas reales	1
Clustering Jerárquico y etiquetas reales	1
K-means y Clustering Jerárquico	1

Cuadro 18: Rand Index usando PCA en la prueba uno a uno.

Combinación de características sujeto a sujeto		
Algoritmos	**Rand Index sujeto 4 y 5**	**Rand Index sujeto 4 y 3**
K-means y etiquetas reales	0.64666	0.9729
Clustering Jerárquico y etiquetas reales	0.64666	0.99771
K-means y Clustering Jerárquico	0.71372	0.72779

Cuadro 19: Resultados de Rand Index para pruebas combinando caracteristicas con combinaciones de diferentes sujetos epilépticos y sanos (4-5, 4-3).

PCA sujeto a sujeto		
Algoritmos	**Rand Index sujeto 4 y 5**	**Rand Index sujeto 4 y 3**
K-means y etiquetas reales	0.65642	0.9751
Clustering Jerárquico y etiquetas reales	0.92443	0.99543
K-means y Clustering Jerárquico	0.65863	0.72471

Cuadro 20: Resultados de Rand Index para pruebas usando el PCA con combinaciones de diferentes sujetos epilépticos y sanos (4-5, 4-3).

Agrupamiento con muchos EDFs combinando características	
Algoritmos	**Resultado Rand Index**
K-means y etiquetas reales	0.5021
Clustering Jerárquico y etiquetas reales	0.5019
K-means y Clustering Jerárquico	0.9352

Cuadro 21: Agrupamiento con 254 EDF sin epilepsia y 184 EDF con epilepsia combinando todas las características.

En los Cuadros 17 a 22 se presentan los resultados obtenidos al aplicar la normalización Z-score en las diferentes pruebas de agrupamiento. En la prueba dato a dato, utilizando la combinación de características originales, los resultados del Rand Index fueron altos, alcanzando 0.9766 para las métricas evaluadas con K-means y Clustering Jerárquico. La

Agrupamiento con muchos EDFs utilizando PCA	
Algoritmos	**Resultado Rand Index**
K-means y etiquetas reales	0.5021
Clustering Jerárquico y etiquetas reales	0.50104
K-means y Clustering Jerárquico	0.89646

Cuadro 22: Agrupamiento con 254 EDF sin epilepsia y 184 EDF con epilepsia utilizando PCA.

incorporación de la normalización Z-score no afectó significativamente la capacidad de los algoritmos para identificar patrones en esta configuración. Por otro lado, al utilizar PCA en esta misma prueba, los valores del Rand Index fueron perfectos (1), lo que indica que la combinación de Z-score y PCA permitió una agrupación más precisa y equitativa, aprovechando de manera más eficiente las características de los datos.

En las pruebas sujeto a sujeto, los resultados mostraron diferencias notables. Al trabajar con la combinación de características originales, los valores de Rand Index para la combinación de sujetos 4-5 alcanzaron 0.6666 para K-means y etiquetas reales, mientras que para sujetos 4-3 los valores fueron más altos, llegando hasta 0.9971. Esto evidencia que las características originales presentan una variabilidad que puede depender de la selección específica de los sujetos en análisis. Con la incorporación de PCA, los resultados mejoraron para los sujetos 4-5, alcanzando 0.9243 para Clustering Jerárquico y etiquetas reales. Esto subraya que la normalización Z-score, junto con PCA, es particularmente útil para pruebas intrasujeto, donde la variabilidad debe ser controlada para obtener agrupamientos más precisos.

En las pruebas masivas, los valores del Rand Index en la combinación de características originales mostraron cierta capacidad de agrupamiento entre K-means y Clustering Jerárquico, alcanzando un máximo de 0.9352. Sin embargo, los valores al comparar con etiquetas reales fueron más bajos, llegando a 0.521 para K-means y 0.501 para Clustering Jerárquico. Esto indica que, en pruebas intersujeto, las diferencias inherentes entre los sujetos complican la clasificación. Al integrar PCA en estas pruebas con múltiples sujetos, los resultados entre K-means y Clustering Jerárquico se mantuvieron altos, alcanzando 0.89646, lo que demuestra que PCA conserva la capacidad de los algoritmos para capturar patrones relevantes en datos más complejos. No obstante, los valores comparados con las etiquetas reales disminuyeron ligeramente, posiblemente debido a la pérdida de detalles específicos de los sujetos individuales.

En general, la normalización Z-score demostró ser una herramienta esencial para las pruebas intrasujeto, eliminando el sesgo causado por características con mayor varianza y permitiendo una representación más equitativa de los datos. Al combinarse con PCA, esta técnica mejoró significativamente los resultados de agrupamiento en escenarios donde la variabilidad entre los segmentos es menor, como en pruebas sujeto a sujeto. Aunque en pruebas intersujeto la normalización y PCA pueden mostrar ligeras reducciones en los resultados, la mejora en términos de eficiencia computacional y equidad en el tratamiento de las características justifica su uso, especialmente para aplicaciones clínicas intrasujeto donde se requiere un análisis más preciso y consistente.

CAPÍTULO 10

Fuzzy C-Means

Fuzzy es un algoritmo de agrupamiento que no se había explorado en las fases previas de este proyecto y cuya integración en esta etapa ha permitido ampliar las herramientas para el análisis de las señales EEG. A diferencia de otros métodos de agrupamiento, como K-Means o Clustering Jerárquico, FCM se basa en la asignación difusa de pertenencia de los datos a múltiples clusters. Esto significa que, en lugar de asignar cada punto de datos exclusivamente a un solo grupo, FCM calcula un grado de pertenencia para cada punto en relación con todos los clusters, proporcionando una representación más matizada de los datos.

En este trabajo, FCM se utilizó para analizar las señales EEG y evaluar su efectividad en la identificación y categorización de segmentos de interés, como se observa en la Figura 48. La gráfica muestra la señal EEG de un canal específico con una codificación de colores que representa el grado de pertenencia de los segmentos a diferentes clusters. Este enfoque visual proporciona información valiosa sobre las transiciones y patrones en la señal que no siempre son evidentes con métodos de agrupamiento tradicionales.

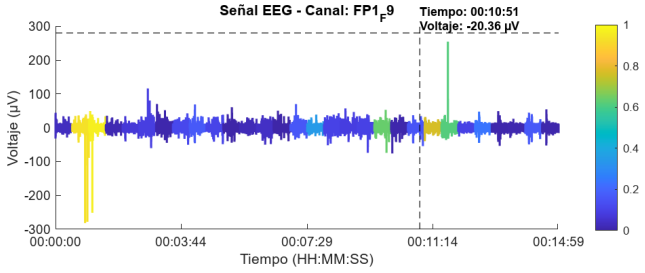

Figura 48: Fuzzy en una señal EEG.

La implementación de FCM no solo amplía las posibilidades de análisis, sino que también ofrece una nueva perspectiva para estudiar las características de las señales EEG. Su

incorporación permite comparar directamente su desempeño con otros algoritmos utilizados previamente, proporcionando un marco más robusto para la evaluación y validación de los agrupamientos en este proyecto.

Empleando un nuevo algoritmo, Fuzzy C-Means, se agrego un nueva funcionalidad. A diferencia de la categorización estricta a una fase u otra, este algoritmo permite visualizar la pertenencia de los segmentos de la señal mediante un gradiente de colores. Este gradiente representa el porcentaje de pertenencia a cada grupo o fase, brindando una visualización más flexible y precisa de las transiciones entre estados.

El uso del gradiente es especialmente útil para evitar clasificaciones estrictas, ya que las señales EEG suelen tener transiciones graduales entre fases. Esta visualización facilita la interpretación de los resultados, permitiendo a los usuarios observar de manera más intuitiva cómo varía la pertenencia de la señal de una fase a otra a lo largo del tiempo. Esto puede ser de gran valor en la detección de eventos neurológicos complejos, donde los límites entre fases no siempre son claros.

CAPÍTULO 11

Métodos Finales

Tras los análisis realizados en capítulos previos, se definieron los métodos finales que conforman el enfoque propuesto en este trabajo. Estos métodos se seleccionaron cuidadosamente con base en su desempeño y su capacidad para abordar los desafíos específicos del análisis de señales EEG. Los componentes principales del algoritmo final son los siguientes:

- Extracción de características: Se seleccionaron un total de 10 características descritas previamente en la Sección 8.2, "Extracción de características", excluyendo la potencia debido a su impacto desproporcionado en la varianza y en los algoritmos de agrupamiento.
- Reducción de dimensionalidad con PCA: Se utilizó el Análisis de Componentes Principales (PCA) para optimizar la carga computacional del modelo y facilitar la visualización de los clusters en espacios reducidos, sin comprometer la información relevante de los datos.
- Normalización Z-score: Esta técnica fue incorporada para igualar la escala de las características y evitar que las diferencias en magnitudes afectaran el rendimiento de los algoritmos de agrupamiento, especialmente en pruebas intrasujeto, donde este ajuste es crucial.

A continuación se presentan Figuras de la aplicación de los métodos en las señales proporcionadas por HUMANA.

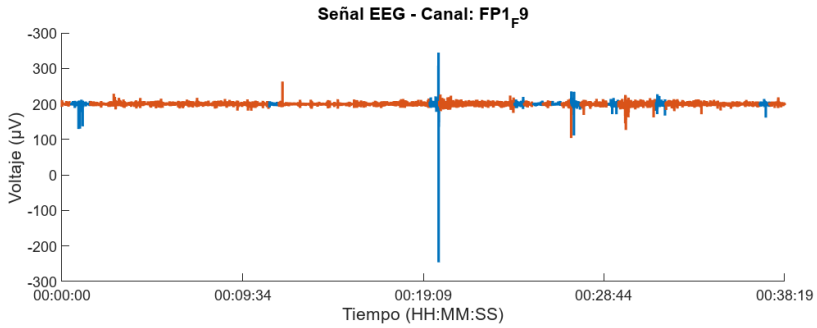

Figura 49: Estudio Gika con las secciones agrupadas, donde un color es un grupo y otro color es otro (señal obtenida de la herramienta actualizada).

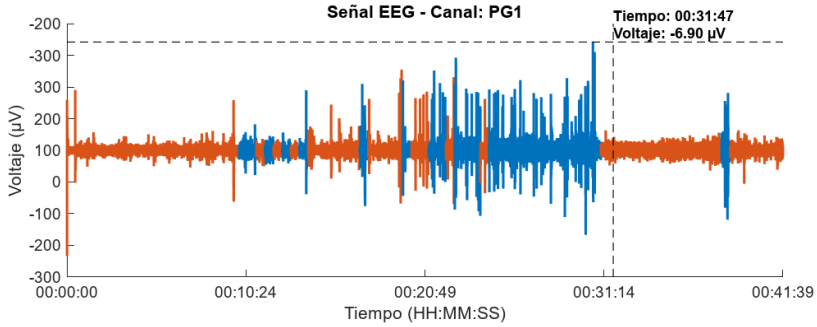

Figura 50: Estudio Al con las secciones agrupadas, donde un color es un grupo y otro color es otro (señal obtenida de la herramienta actualizada).

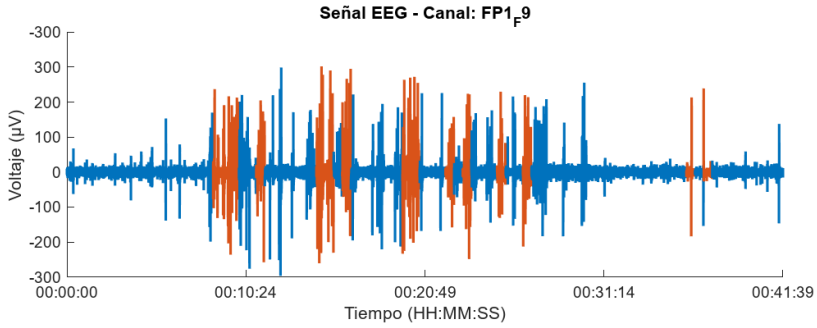

Figura 51: Estudio CLEA con las secciones agrupadas, donde un color es un grupo y otro color es otro (señal obtenida de la herramienta actualizada).

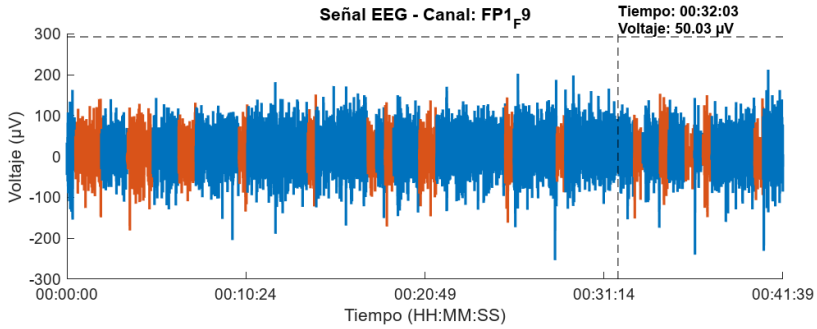

Figura 52: Estudio HCHC con las secciones agrupadas, donde un color es un grupo y otro color es otro (señal obtenida de la herramienta actualizada).

CAPÍTULO 12

Actualización de la herramienta

Como parte del desarrollo y mejora continua de la herramienta de análisis de EEG, se han implementado diversas actualizaciones tanto en la interfaz como en las funcionalidades, con el objetivo de optimizar la experiencia del usuario y facilitar el análisis automático de señales. Estas actualizaciones han sido desarrolladas de manera conjunta entre Javier Pérez, en el marco de su trabajo de graduación "Aplicación de algoritmos de aprendizaje automático, con énfasis en aprendizaje supervisado, para la identificación y categorización de segmentos de interés en señales bioeléctricas para el estudio de la epilepsia - Fase V" [45], y el trabajo de graduación presentado en este documento, centrado en el aprendizaje no supervisado. Cada uno ha contribuido con áreas de enfoque específicas. A continuación, se describen los cambios más relevantes.

Una de las mejoras conjuntas implementadas ha sido la incorporación de una alerta de cierre para la ventana principal, la cual se puede observar en la Figura 53. Esta funcionalidad fue diseñada para asegurar que, al intentar cerrar la ventana principal, se notifique al usuario que todas las ventanas secundarias también serán cerradas, evitando la pérdida de trabajo no guardado. Anteriormente, si la ventana principal se cerraba, las ventanas secundarias permanecían abiertas, lo que provocaba que MATLAB se bloqueara y fuera necesario un cierre forzoso mediante el administrador de tareas. Con esta nueva alerta, se garantiza que todas las ventanas se cierren de manera ordenada, mejorando la estabilidad de la herramienta.

Javier Pérez se ha encargado de la migración de la herramienta a la versión 2024 de MATLAB y ha implementado diversas mejoras en la interfaz de usuario, como:

- Interfaz de inicio de sesión: Se agregó un botón para visualizar la contraseña y la funcionalidad de ingresar automáticamente al presionar "Enter", sin necesidad de hacer clic en "Conectar" (ver Figura 54).
- Migración de la herramienta a MATLAB 2024: Ya que el programa viene de MATLAB

Figura 53: Cierre de programa desde la ventana principal.

2023, es necesario actualizarlo a la versión más reciente. Este cambio incluyó la actualización de funciones como str2double, que fue reemplazada por double(string('XX')) para mejorar la compatibilidad con la nueva versión.

Javier Pérez también agregó un botón llamo "Anotaciones por montaje" dentro de la ventana principal en el cuál se encuentra su trabajo dentro de la herramienta. Mientras que el trabajo presentado en este documento se encuentra dentro del botón "Anotaciones por canal" (ver Figura 55).

Las mejoras implementadas en este trabajo de graduación están enfocadas en optimizar las funcionalidades relacionadas con el aprendizaje no supervisado. Entre los principales avances se encuentran:

- Selección Aprendizaje: Se agregarón checkbox para individualizar los aprendizajes y poder escoger solo el no supervisado o solo el supervisado (ver Figura 56).
- Botón de VAT: Que muestra la matriz de disimilitud, permitirá evaluar visualmente la tendencia de los datos a agruparse (Figura **??**). Aunque este método no ofrece resultados óptimos para este tipo de datos se deja en la herramienta como otro tipo de visualización.
- Selección de algoritmos de agrupamiento: Se ha agregado una opción para elegir entre diferentes algoritmos de agrupamiento, como K-Means, Jerárquico y Fuzzy C-Means (ver Figura 58).
- Selección de PCA: Se ha agregado una opción para elegir entre que componente principal se desea usar, como el primero, segundo, tercero o combinación de los primero tres (ver Figura 59).

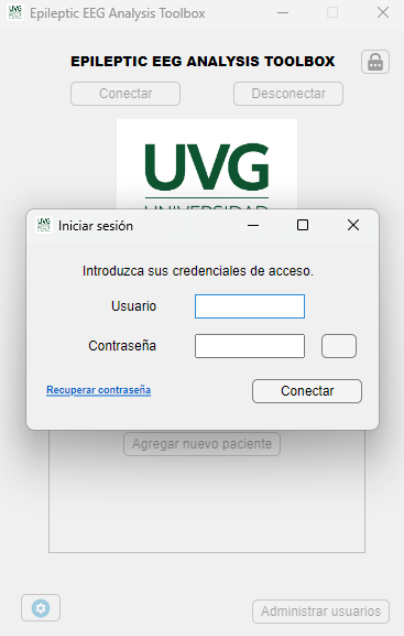

Figura 54: Inicio de sesión amigable.

- Selección de grupos: Se ha incorporado una opción que permite al usuario seleccionar la cantidad de grupos deseados para el agrupamiento. Si se elige la opción "Sano-Epilepsia", el sistema automáticamente seleccionará dos grupos. En caso de requerir más de dos grupos, es necesario activar el checkbox correspondiente e ingresar manualmente el número deseado de grupos (60).

- Configuración del montaje: Se implementó la opción de visualizar el montaje de las señales EEG mediante la función stackedplot (Figura 61). Esta función evalúa si la señal cargada está configurada en modo de montaje o como canales individuales. Si la señal está en modo montaje, se mostrará como en la Figura 62. Por otro lado, si está configurada en canales individuales, se abrirá la ventana ilustrada en la Figura 63, permitiendo su visualización en el formato correspondiente. Esta mejora proporciona una mayor flexibilidad para analizar las señales EEG según el formato de carga.

Figura 55: Botones de tipos de anotaciones.

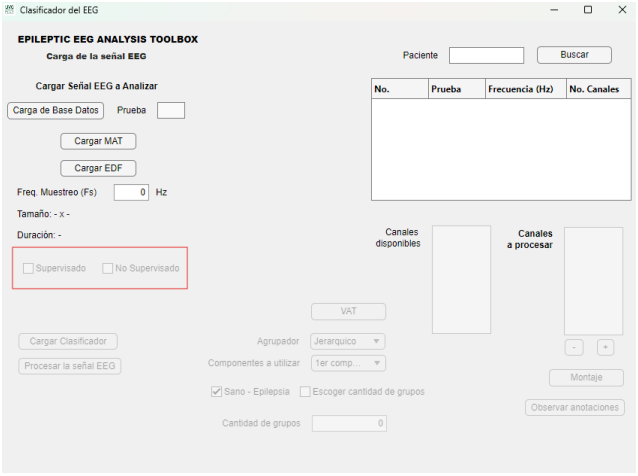

Figura 56: Selección de aprendizaje a utilizar.

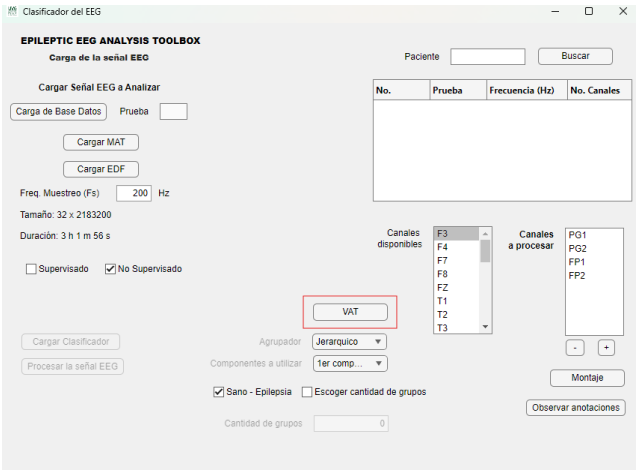

Figura 57: Botón VAT

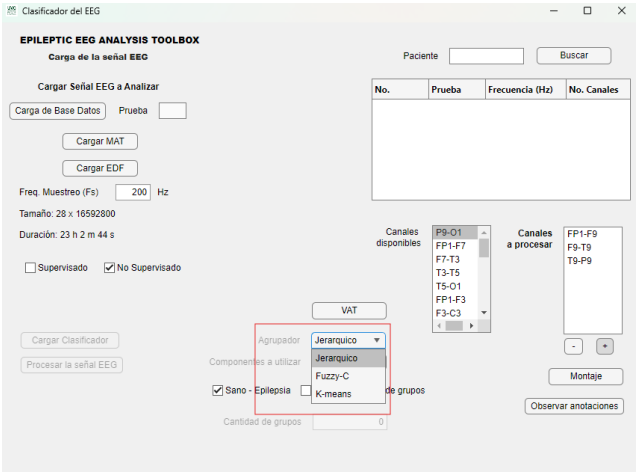

Figura 58: Selección de algoritmos y número de agrupaciones.

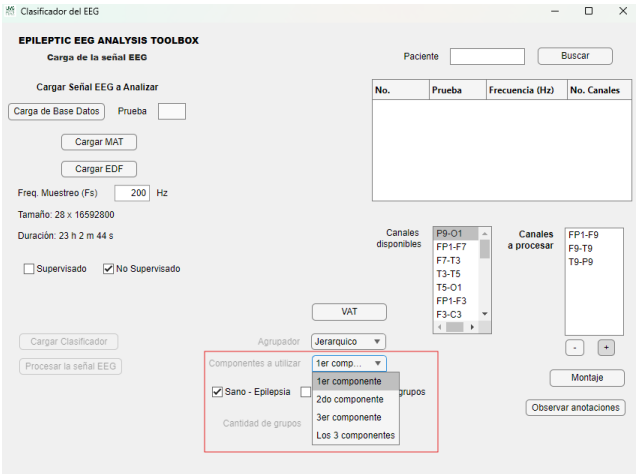

Figura 59: Selección de PCA a utilizar.

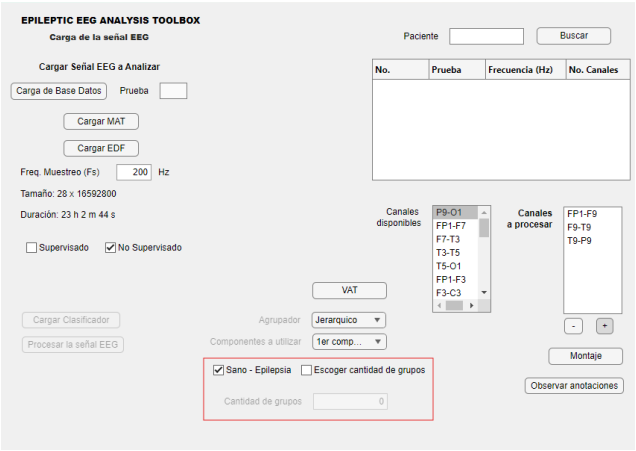

Figura 60: Selección de algoritmos.

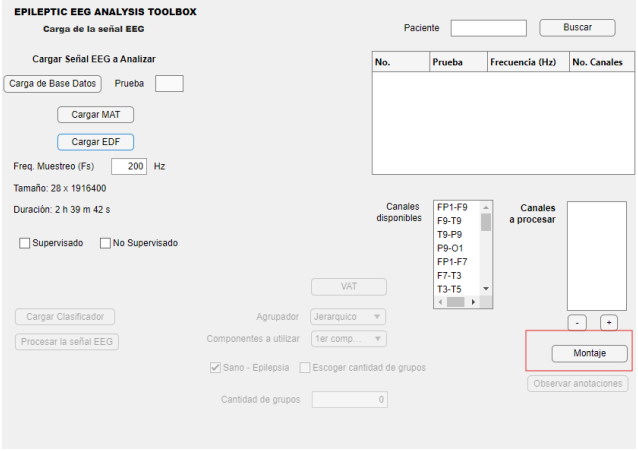

Figura 61: Opción de montaje para visualización de la señal.

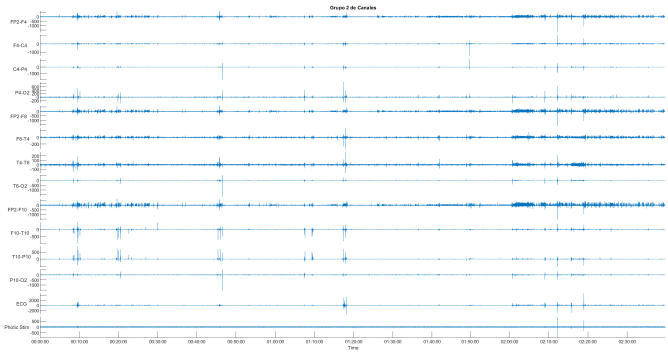

Figura 62: Visualización con el montaje en el que viene los canales.

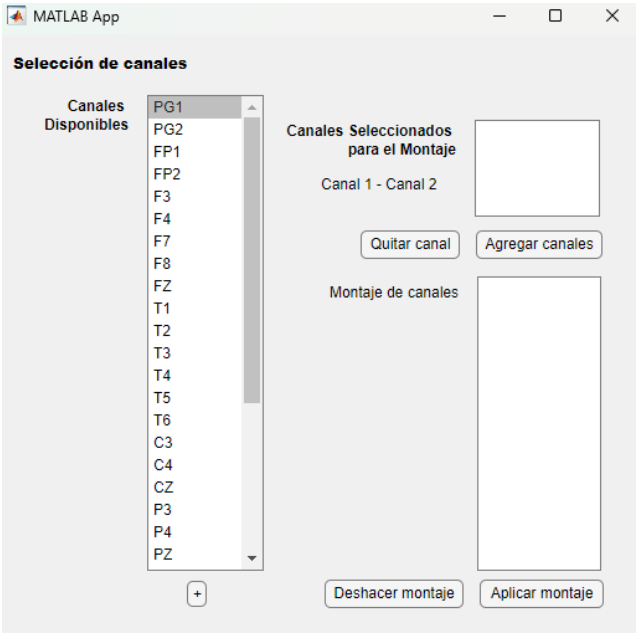

Figura 63: Ventana de selección de canales para realizar un montaje.

Después de seleccionar los canales a procesar y seleccionar los parámetros, se debe presionar el botón de "Observar anotaciones" (Figura 64) nos llevará a la ventana ilustrada en la Figura 65.

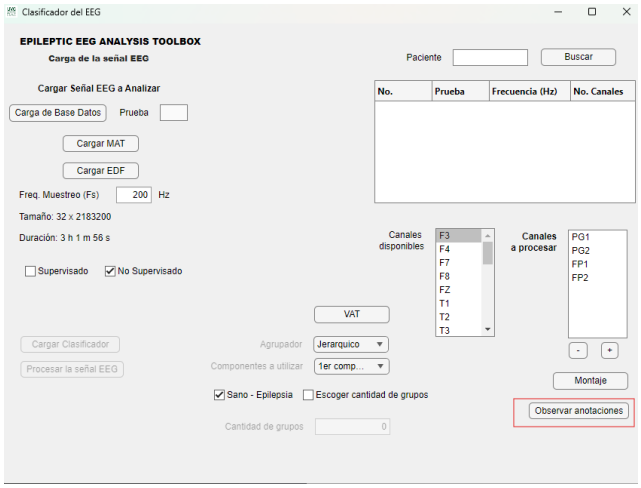

Figura 64: Botón Observar anotaciones.

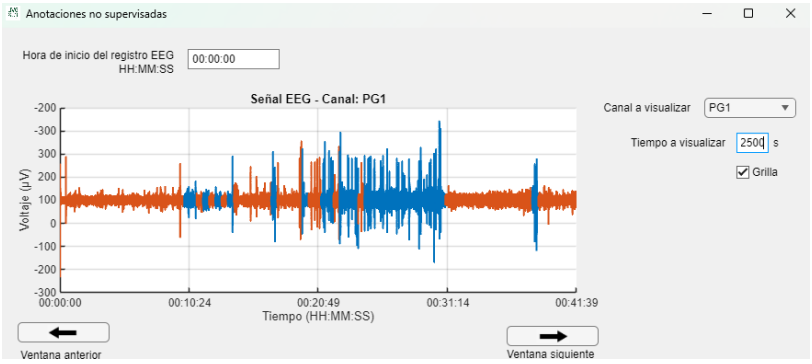

Figura 65: Ventana de anotaciones no supervisadas.

En la ventana de anotaciones no supervisadas es donde se ve gráficamente el resultado de todos los análisis previos. Aquí se presentan las agrupaciones realizadas por los algoritmos implementados, las etiquetas asignadas a los diferentes segmentos de la señal y los patrones identificados en la misma. Esto incluye tanto los clústeres generados como las características empleadas en el análisis.

Además, esta ventana proporciona herramientas interactivas para que el usuario explore los datos. Estos incluyen:

- Permite a los usuarios seleccionar el canal que desean visualizar agrupado (ver Figura 66).
- Ingreso de la cantidad de segundo que se desea ver de la señal (ver Figura 67).
- Puntero para obtener información de un punto específico de la señal (ver Figura 68).
- Botones para avanzar y retroceder en la señal.

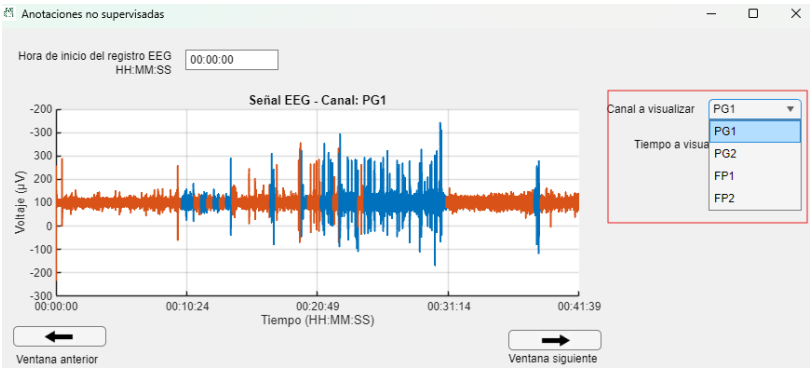

Figura 66: Selección de canal a visualizar.

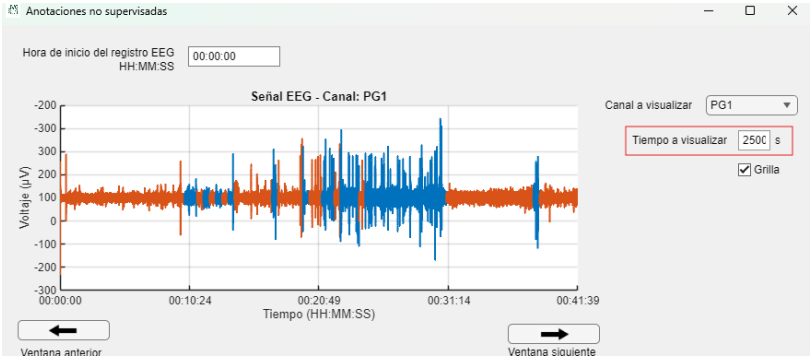

Figura 67: Ingreso de cantidad de segundo a visualizar en la ventana.

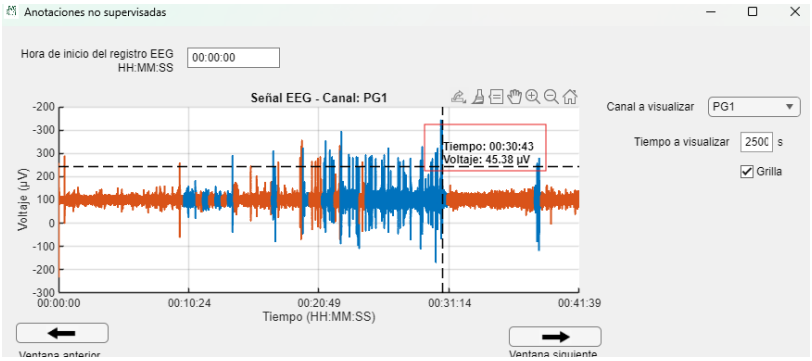

Figura 68: Puntero para obtener información de un punto específico.

CAPÍTULO 13

Conclusiones

- Las señales adicionales obtenidas por medio de TUH EEG permitieron ampliar el conjunto de datos analizados, lo que contribuyó a realizar evaluaciones más robustas de los algoritmos de agrupamiento. Sin embargo, la falta de una señal intrasujeto amplia etiquetada limitó la capacidad de validación directa.
- Las señales obtenidas de HUMANA permitieron realizar pruebas intrasujeto para la visualización en la herramienta.
- La aplicación de algoritmos como K-Means, Clustering Jerárquico y Fuzzy C-Means demostró que las características extraídas de las señales EEG permiten diferenciar segmentos epilépticos y no epilépticos de manera efectiva.
- Aunque K-Means es efectivo para identificar patrones, su desempeño depende significativamente de la inicialización, lo que puede llevar a agrupaciones erróneas en algunos casos. Por ello, se considera más confiable el uso de el Clustering Jerárquico, ya que este último ofrece resultados más robustos al no depender de una inicialización aleatoria.
- La incorporación de la normalización Z-score y la selección de características más relevantes mejoraron significativamente la precisión en la identificación de segmentos de interés, especialmente en pruebas intrasujeto.
- Los resultados obtenidos al comparar diferentes características (frecuencia, tiempo y wavelets) mostraron que las características combinadas ofrecen el mejor desempeño, mientras que la implementación de PCA ayudó a optimizar el análisis, al reducir la dimensionalidad y facilitar la identificación de patrones con menor carga computacional.
- La integración de Fuzzy C-means con su capacidad de generar anotaciones con gradientes y visualizaciones en la herramienta software permitió interpretar de manera intuitiva los resultados de los algoritmos, cumpliendo con los parámetros de HUMANA.
- Rand Index confirmó que las características normalizadas y el uso de PCA contribuyen a un mejor desempeño en agrupamientos tanto intersujeto como intrasujeto, con valores consistentes y robustos.

- La evaluación de características demostró que la variabilidad de los datos se gestiona mejor al emplear técnicas de normalización, lo cual será crucial para futuros estudios clínicos.
- La actualización de la herramienta a MATLAB 2024 y la incorporación de nuevas funcionalidades como visualizaciones de gradiente con Fuzzy, selección de grupos y montajes, optimizó su usabilidad y adaptabilidad a diferentes conjuntos de datos.
- La herramienta actualizada demostró ser funcional para aplicaciones tanto de investigación como clínicas, ofreciendo resultados consistentes y visualizaciones intuitivas para la interpretación de las señales EEG.

CAPÍTULO 14

Recomendaciones

- Se recomienda la recopilación de señales extensas de una sola persona, que pueda tener anotaciones sobre las diferentes secciones para poder validar de manera intrasujeto los clústers
- Se sugiere desarrollar dos versiones separadas de la herramienta: una enfocada en el uso clínico, optimizada para técnicos y profesionales de la salud, y otra diseñada específicamente para desarrolladores, que incluya configuraciones avanzadas. Esto evitará la sobrecarga de funcionalidades innecesarias en la versión clínica y garantizará una experiencia más intuitiva para sus usuarios.
- Se recomienda continuar investigando técnicas de reducción y análisis de dimensionalidad, como el Análisis de Componentes Independientes (ICA), para complementar los resultados obtenidos con PCA. Estas técnicas podrían ofrecer perspectivas adicionales sobre la separación de características relevantes y mejorar aún más el rendimiento de los algoritmos de agrupamiento, especialmente en aplicaciones con datos más complejos o heterogéneos.
- Se recomienda llevar a cabo una validación clínica más extensa de la herramienta de software en entornos médicos reales. La colaboración constante con instituciones médicas, podría facilitar la implementación de la metodología en la práctica clínica, permitiendo evaluar su eficacia en el diagnóstico y manejo de pacientes con epilepsia.
- Se sugiere seguir explorando e investigando otros algoritmos de aprendizaje automático no supervisado, ya que existe una amplia variedad de técnicas que podrían adaptarse mejor a las características específicas de los datos utilizados. Esta exploración podría contribuir a identificar métodos más eficaces y robustos para el análisis y agrupamiento de señales bioeléctricas.

CAPÍTULO 15

Bibliografía

[1] I. Obeid y J. Picone, "The Temple University Hospital EEG Data Corpus," *Frontiers in Neuroscience, Section Neural Technology*, vol. 10, pág. 196, 2016. DOI: 10.3389/fnins.2016.00196.

[2] W. H. Organization, *Epilepsy*, https://www.who.int/es/news-room/fact-sheets/detail/epilepsy, Consultado el 18 de abril de 2024, feb. de 2024.

[3] S. Usman, M. Usman y S. Fong, "Epileptic Seizures Prediction Using Machine Learning Methods," *Hindawi*, vol. 2017, págs. 1-10, 2017. DOI: 10.1155/2017/9074759.

[4] K. Mahmudul, X. Zhao, H. Sugano y T. Tanaka, "Detection of Epileptic Seizures in Long EEG Recordings Using an Anomaly Detector with Artifact Rejection," *IEEE*, págs. 2230-2234, 2024. DOI: 10.1109/ICASSP48485.2024.10447376.

[5] W. Xiong, E. Nurse, E. Lambert, M. Cook y T. Kameneva, "Classification of Epileptic and Psychogenic Non-Epileptic Seizures Using Electroencephalography and Electrocardiography," *IEEE*, vol. 31, págs. 2831-2838, 2023. DOI: 10.1109/TNSRE.2023.3288138.

[6] M. Angulo, "Análisis y Reconocimiento de Patrones de Señales Biomédicas de Pacientes con Epilepsia," Tesis de licenciatura, Universidad Del Valle de Guatemala, 2020.

[7] M. Pineda, "Diseño e Implementación de una Base de Datos de Señales Biomédicas de Pacientes con Epilepsia," Tesis de licenciatura, Universidad Del Valle de Guatemala, 2020.

[8] D. Vela, "Automatización del Proceso de Anotación de Señales EEG de Pacientes con Epilepsia por Medio de Técnicas de Aprendizaje Automático," Tesis de licenciatura, Universidad Del Valle de Guatemala, 2021.

[9] C. Lemus, "Análisis y anotación de señales bioeléctricas de pacientes con epilepsia utilizando técnicas de aprendizaje automático supervisado y no supervisado," Tesis de licenciatura, Universidad Del Valle de Guatemala, 2022.

[10] D. Méndez, "Extensión, validación y migración de una herramienta de software para el estudio de la epilepsia para su uso en el Centro de Epilepsia y Neurocirugía Funcional (HUMANA)," Tesis de licenciatura, Universidad Del Valle de Guatemala, 2023.

[11] C. Patzán, "Aplicación sistemática de algoritmos de aprendizaje automático para el estudio de la epilepsia y la detección de segmentos de interés en señales bioeléctricas," Tesis de licenciatura, Universidad Del Valle de Guatemala, 2023.

[12] Humana, *HUMANA: Centro de Epilepsia y Neurocirugía Funcional*, Sitio web, Accedido el fecha, mar. de 2024. dirección: https://humanagt.org/epilepsia/.

[13] J. E. Mendizabal y L. F. Salguero, "Prevalence of epilepsy in a rural community of Guatemala," *Epilepsia*, vol. 37, n.º 4, págs. 373-376, 1996.

[14] E. Wyllie, A. Gupta y D. Lachhwani, *The Treatment of Epilepsy: Principles & Practice*. Lippincott Williams & Wilkins, 2006.

[15] L. Meraz, M. Lourdes, L. Martín y M. Martín, "Conceptos Básicos de la Epilepsia," *Revista Médica de la Universidad Veracruzana*, vol. 9, n.º 2, págs. 31-37, 2009.

[16] R. García-Ramos, A. G. Pastor, J. Masjuan, C. Sánchez y A. Gil, "FEEN: Informe sociosanitario FEEN sobre la epilepsia en España," *Neurología*, vol. 26, n.º 9, págs. 548-555, 2011.

[17] E. Kiriakopoulos, R. Fisher y E. Wirrell, *Types of Seizures*, Sitio web, 2022. dirección: https://www.epilepsy.com/what-is-epilepsy/seizure-types.

[18] W. Fang, D. Wu, P. E y L. Ding, "Physiological computing for occupational health and safety in construction: Review, challenges and implications for future research," *ScienceDirect*, vol. 54, pág. 101 729, 2022. DOI: doi.org/10.1016/j.aei.2022.101729.

[19] D. Tinoco y D. Gudiño, *Redes neuronales en la Caracterización de Señales bioeléctricas*, https://virtual.cuautitlan.unam.mx/intar/?page_id=977, 2018.

[20] C. Okwaraji, A. Colasuonno, M. Davydova y E. Ambizas, "An Overview of Epilepsy," *U.S. Pharmacist*, vol. 47, n.º 11, págs. 5-12, 2022. dirección: https://www.uspharmacist.com/article/an-overview-of-epilepsy.

[21] Universidad de Navarra. "Electroencefalograma (EEG)."dirección: https://www.cun.es/diccionario-medico/terminos/electroencefalograma-eeg.

[22] C. Bazan, M. Blanco, J. Cardenas y F. Cruz, "Compresión de señales electroencefalográficas epilépticas y normales," *Ingeniería Electrónica, Automática y Comunicaciones*, vol. 33, n.º 1, págs. 25-32, 2012.

[23] N. Associates. "Brain Wave Frequencies."dirección: https://nhahealth.com/brainwaves-the-language/#:~:text=The%20raw%20EEG%20has%20usually%2C)%20for%20%E2%80%9Cactive%E2%80%9D%20intelligence..

[24] R. Vallat, *Compute the average bandpower of an EEG signal*, https://raphaelvallat.com/bandpower.html, 2018.

[25] "Electromyography (EMG)."dirección: https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/electromyography-emg.

[26] "Electrocardiograma."dirección: https://medlineplus.gov/spanish/pruebas-de-laboratorio/electrocardiograma/.

[27] L. Azcona, *El electrocardiograma*, https://www.fbbva.es/microsites/salud_cardio/mult/fbbva_libroCorazon_cap4.pdf, 2023.

[28] P. Boonyakitanont, A. Lek-uthai, K. Chomtho y J. Songsiri, "A review of feature extraction and performance evaluation in epileptic seizure detection using EEG," 2019. arXiv: 1908.00492 [eess.SP].

[29] L. Cohen, "Time-frequency distributions-a review," *Proceedings of the IEEE*, vol. 77, n.º 7, págs. 941-981, 1989. DOI: 10.1109/5.30749.

[30] "Aprendizaje automático," Hewlett Packard Enterprise Development LP. dirección: https://www.hpe.com/lamerica/es/what-is/machine-learning.html.

[31] C. Janiesch, P. Zschech y K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 20, n.º 1, pág. 685, 2021.

[32] "What Is Supervised Learning? | IBM," 2023. dirección: https://www.ibm.com/cloud/learn/supervised-learning.

[33] "What is unsupervised learning? | IBM," 2023. dirección: https://www.ibm.com/topics/unsupervised-learning.

[34] T. Amestoy, *Clustering basics and a demonstration in clustering infrastructure pathways*, https://waterprogramming.wordpress.com/2022/03/16/clustering-basics-and-a-demonstration-in-clustering-infrastructure-pathways/, mar. de 2022.

[35] E. Kavlakoglu y V. Winland, "What is k-means clustering? | IBM," 2024. dirección: https://www.ibm.com/topics/k-means-clustering.

[36] MathWorks, "Fuzzy Clustering," 2024. dirección: https://es.mathworks.com/help/fuzzy/fuzzy-clustering.html.

[37] M. Vazirgiannis, "Clustering Validity," en *Encyclopedia of Database Systems*, L. Liu y M. T. Özsu, eds., Boston, MA: Springer, 2009. DOI: 10.1007/978-0-387-39940-9_616. dirección: https://doi.org/10.1007/978-0-387-39940-9_616.

[38] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, n.º 336, págs. 846-850, 1971. DOI: 10.2307/2284239.

[39] R. Todeschini, D. Ballabio, V. Termopoli y V. Consonni, "Extended multivariate comparison of 68 cluster validity indices," *Chemometrics and Intelligent Laboratory Systems*, vol. 251, pág. 105 117, 2024. DOI: 10.1016/j.chemolab.2024.105117. dirección: https://doi.org/10.1016/j.chemolab.2024.105117.

[40] Y. Tang, F. Sun y Z. Sun, "Improved Validation Index for Fuzzy Clustering," *2005 American Control Conference*, págs. 1118-1125, 2005. DOI: 10.1109/ACC.2005.1470147. dirección: https://doi.org/10.1109/ACC.2005.1470147.

[41] MathWorks, "Introducción a Reinforcement Learning - MATLABSimulink," 2024. dirección: https://www.mathworks.com/help/reinforcement-learning/index.html.

[42] J. C. Bezdek y R. J. Hathaway, "VAT: A tool for visual assessment of (cluster) tendency," en *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, IEEE, vol. 3, 2002, págs. 2225-2230.

[43] "What is principal component analysis (PCA)? | IBM," 2023. dirección: https://www.ibm.com/topics/principal-component-analysis.

[44] L. Veloso, J. R. McHugh, E. von Weltin, I. Obeid y J. Picone, "Big Data Resources for EEGs: Enabling Deep Learning Research," en *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, I. Obeid y J. Picone, eds., Philadelphia, Pennsylvania, USA: IEEE, 2017, pág. 1.

[45] J. Pérez, "Aplicación de algoritmos de aprendizaje automático, con énfasis en aprendizaje supervisado, para la identificación y categorización de segmentos de interés en señales bioeléctricas para el estudio de la epilepsia - Fase V," Tesis de licenciatura, Universidad Del Valle de Guatemala, 2024.

CAPÍTULO 16

Anexos

16.1. Nombres de los sujetos en TUH EEG Epilepsy corpus

A continuación se presenta el Cuadro 23 con los nombres de los sujetos en la base de datos TUH EEG Epilepsy corpus.

Nombre de sujetos	
Sujetos	**Nombre**
Sujeto 1	aaaaaebo
Sujeto 2	aaaaaanr
Sujeto 3	aaaaangg
Sujeto 4	aaaaacrz
Sujeto 5	aaaaaoie
Sujeto 6	aaaaapmu
Sujeto 7	aaaaajgj
Sujeto 8	aaaaaeqq
Sujeto 9	aaaaajrh
Sujeto 10	aaaaadkv
Sujeto 11	aaaaakgy
Sujeto 12	aaaaammn

Cuadro 23: Nombre de los sujetos dentro de TUH EEG Epilepsy corpus.

16.2. Resultados sujeto a sujeto

A continuación se presentan algunos resultado extras a los que se mencionaron en los resultados.

Sujetos	Rand Index		
	K-means - reales	Jerárquico - reales	K-means - Jerárquico
Sujeto 6-4	0.50196	0.50357	0.78216
Sujeto 7-8	0.59167	0.59528	0.99177
Sujeto 9-10	0.49845	0.49789	0.98817
Sujeto 11-12	0.49383	0.49383	1

Cuadro 24: Rand Index para diferentes combinaciones de sujetos frecuencia.

Sujetos	Rand Index		
	K-means - reales	Jerárquico - reales	K-means - Jerárquico
Sujeto 6-4	0.67206	0.70533	0.9473
Sujeto 7-8	0.58814	0.59895	0.97551
Sujeto 9-10	0.49704	0.49704	1
Sujeto 11-12	0.65	0.65	1

Cuadro 25: Rand Index para diferentes combinaciones de sujetos tiempo continuo.

Sujetos	Rand Index		
	K-means - reales	Jerárquico - reales	K-means - Jerárquico
Sujeto 6-4	0.53613	0.54841	0.95876
Sujeto 7-8	0.6065	0.62657	0.95953
Sujeto 9-10	0.60693	0.64934	0.92012
Sujeto 11-12	0.49383	0.49444	0.97531

Cuadro 26: Rand Index para diferentes combinaciones de sujetos Todas las características.

16.3. Resultados sin potencia y con PCA

A continuación se presentan algunos resultado extras a los que se mencionaron en los resultados.

Sujetos	Rand Index		
	K-means - reales	Jerárquico - reales	K-means - Jerárquico
Sujeto 6-4	0.8163	0.70533	0.59678
Sujeto 7-8	0.54835	0.59895	0.53168
Sujeto 9-10	0.49704	0.49704	1
Sujeto 11-12	0.65	0.65	1

Cuadro 27: Rand Index para diferentes combinaciones de sujetos todas las caracteristicas sin potencia.

Sujetos	Rand Index		
	K-means - reales	Jerárquico - reales	K-means - Jerárquico
Sujeto 6-4	0.67088	0.66276	0.98606
Sujeto 7-8	0.6065	0.59528	0.50243
Sujeto 9-10	0.49704	0.49704	1
Sujeto 11-12	0.65	0.65	1

Cuadro 28: Rand Index para diferentes combinaciones de sujetos PCA sin potencia.

16.4. Resultados normalizando

A continuación se presentan algunos resultado extras a los que se mencionaron en los resultados.

Sujetos	Rand Index		
	K-means - reales	Jerárquico - reales	K-means - Jerárquico
Sujeto 6-4	0.50093	0.6329	0.57286
Sujeto 7-8	0.58548	0.52268	0.53168
Sujeto 9-10	0.69462	0.69462	1
Sujeto 11-12	0.63642	1	0.63642

Cuadro 29: Rand Index para diferentes combinaciones de sujetos todas las caracteristicas normalizadas.

Sujetos	**Rand Index**		
	K-means - reales	**Jerárquico - reales**	**K-means - Jerárquico**
Sujeto 6-4	0.59089	0.6329	0.85316
Sujeto 7-8	0.57461	0.53234	0.70155
Sujeto 9-10	0.69462	0.69462	1
Sujeto 11-12	0.50679	1	0.50679

Cuadro 30: Rand Index para diferentes combinaciones de sujetos PCA normalizada.