UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

DETECCIÓN DEL TRASTORNO DE DÉFICIT DE ATENCIÓN E HIPERACTIVIDAD UTILIZANDO ALGORITMOS DE CLASIFICACIÓN DE MACHINE LEARNING BASADO EN SEÑALES DE ELECTROENCEFALOGRAFÍA

POR

Aura Madeleyn Toledo Araneda

Memoria de Titulo presentada a la Facultad de Ingeniería de la Universidad de Concepción para optar al título profesional de Ingeniero/a Civil Biomédico

Profesor Guía
Esteban Pino

Enero 2025
Concepción
(Chile)

Resumen

El TDAH es uno de los trastornos psiquiátricos y neuroconductuales más comunes alrededor del mundo, afectando a millones de niños y adolescentes. Este proyecto de investigación implementó una metodología basada en aprendizaje automático que buscaba identificar y clasificar señales de EEG de sujetos con TDAH presentes en una base de datos de 121 participantes (61 TDAH, 60 control). Para desarrollar la metodología, se comenzó filtrando la señal y eliminando artefactos, luego se almacenaron y prepararon los datos para extraer las características en el dominio del tiempo (estadísticas, morfológicas y no lineales) y en el dominio de la frecuencia. La potencia de banda se calculó para 4 bandas de interés: delta, theta, alfa y beta aplicando el método Welch. Las características más relevantes fueron seleccionadas usando LASSO, lo que redujo significativamente la dimensión de la matriz de almacenamiento y además mejoró el rendimiento de los modelos de clasificación. También se realizó un análisis estadístico a las características más relevantes, evidenciando que predominaban en la zona frontal y parietal del lado derecho y además, presentaban más rítmos de ondas alfa y beta. Se evaluaron 4 clasificadores: SVM, Regresión Logística, Random Forest y Naive Bayes. Los resultados del desempeño de cada uno fue presentado en en términos de accuracy, precisión, recall y F1-score siendo el más robusto en sus resultados Regresión Logística con 85.83 % de accuracy en validación cruzada de 10 folds. La metodología fue efectiva en esta tarea de clasificación de señales para identificar individuos con TDAH y podría ser utilizada en investigaciones posteriores.

Abstract

ADHD is one of the most common psychiatric and neurodevelopmental disorders worldwide, affecting millions of children and adolescents. This research project implemented a methodology based on machine learning aimed at identifying and classifying EEG signals from subjects with ADHD in a dataset of 121 participants (61 ADHD, 60 control). To develop the methodology, the signals were first filtered and artifacts were removed, then the data was stored and prepared to extract features in both the time domain (statistical, morphological, and non-linear) and the frequency domain. Band power was calculated for 4 bands of interest: delta, theta, alpha, and beta using the Welch method. The most relevant features were selected using LASSO, which significantly reduced the storage matrix dimension and also improved the performance of the classification models. A statistical analysis was also performed on the most relevant features, showing that they predominated in the frontal and parietal areas of the right hemisphere and also exhibited more alpha and beta wave rhythms. Four classifiers were evaluated: SVM, Logistic Regression, Random Forest, and Naive Bayes. The performance results of each were presented in terms of accuracy, precision, recall, and F1-score, with Logistic Regression being the most robust in its results, achieving 85.83 % accuracy in 10-fold cross-validation. The methodology was effective in this task of classifying signals to identify individuals with ADHD and could be used in future research.

Tabla de contenidos

Índice de Tablas

Índice de Figuras

Capitulo 1. Introducción

1.1 Introducción General

El Trastorno de Déficit de Atención e Hiperactividad (TDAH) es un trastorno conductual del desarrollo neurológico que afecta al 11 % de los niños alrededor del mundo. Sus síntomas incluyen innatención, impulsividad e hiperactividad [5]. Los primeros síntomas se presentan en edad preescolar y suelen agudizarse en la etapa escolar afectando el desarrollo de sus actividades académicas, personales y sociales hasta la edad adulta, siendo más común en hombres [6]. La detección temprana es crucial para que los pacientes puedan recibir atención adecuada, tratamientos y herramientas que los ayude a desenvolverse y a ser conscientes de su trastorno [4]. El diagnóstico de TDAH es realizado por médicos especialistas, generalmente, bajo criterios descritos en el Manual Diagnóstico y Estadístico de Trastornos Mentales (DSM-5) acompañado de evaluación escolar y familiar [7] [8]. Diagnosticar TDAH toma bastante tiempo dada toda la información que se debe recopilar acerca del comportamiento del paciente en su entorno familiar y escolar. Además, requiere un alto grado de competencia clínica y, en ocasiones, puede ser inexacta [8]. En este contexto, se ha generado un gran interés por parte de los investigadores para desarrollar nuevos métodos de detección de TDAH que sean eficaces y complementen la confiabilidad profesional con datos cuantificables implemetando técnicas de aprendizaje automático (ML) a señales de electroencefalografía (EEG) [8]. El EEG es un método fiable que proporciona información sobre la actividad cerebral por lo que puede ser una herramienta útil para investigar y diagnosticar a los niños con TDAH [9].

El desarrollo de un método basado en EEG implica una serie de etapas que incluyen la adquisición de señales, su preprocesamiento para eliminar ruido y artefactos, la extracción de características significativas y su uso como entradas en algoritmos de clasificación. Estas técnicas no solo pueden mejorar la precisión del diagnóstico, sino también ofrecer información cuantitativa que complemente la evaluación clínica tradicional, fortaleciendo la toma de decisiones médicas. Con una base de datos adecuada de señales EEG y un diseño riguroso, se espera avanzar hacia un enfoque más objetivo y accesible para la identificación temprana de TDAH.

1.2 Objetivos

1.2.1 Objetivo General

Implementar un algoritmo de clasificación del Trastorno de Déficit de Atención e Hiperactividad basado en señales electroencefalogría, utilizando herramientas de aprendizaje automático para apoyo en el diagnóstico.

1.2.2 Objetivos Específicos

- Obtener una base de datos con señales EEG de un grupo control y un grupo diagnosticado con TDAH.
- Implementar algoritmos de aprendizaje automático en Python para clasificar las señales EEG.
- Validar el rendimiento del algoritmo de clasificación a través de métricas de evaluación.

1.3 Alcances y Limitaciones

- Implementación de algoritmo basado en una sola base de datos.
- Calidad de las señales EEG de la base de datos.

Capitulo 2. Marco Teórico

2.1 Introducción

Este capítulo busca proporcionar una comprensión teórica de las características clínicas del TDAH y su impacto en la población infantil. Se mencionarán las principales características del EEG, cómo es adquirida la señal, de dónde provinen los potenciales eléctricos y el equipo electrónico que se requiere para obtener digitalmente los resultados de la actividad cerebral. También se describirán los métodos comúnmente adoptados para procesar señales de EEG y ejecutar un clasificador desde el preprocesamiento hasta las métricas de evaluación de desempeño. Con esto se tendrá una visión teórica de todo lo que implica el procesamiento de señales y su utilización dentro del aprendizaje automático.

2.2 Trastorno de Déficit de Atención e Hiperactividad

El TDAH se conoce como un trastorno conductual en niños caracterizado por falta de atención, impulsividad e hiperactividad diagnosticado típicamente en la infancia, siendo más prevalente en niños que en niñas. Se han diagnosticado unos 6.4 millones de niños estadounidenses entre 4 y 17 años. Los síntomas suelen comenzar antes de los 12 años y pueden aparecer desde los 3 años dificultando el control sobre su atención, comportamiento y emociones, desencadenando dificultades en su desarrollo personal y académico [9].

Se ha observado que una lesión cerebral, factores genéticos, consumo de alcohol durante el embarazo, parto prematuro y bajo peso al nacer pueden asociarse con el TDAH, siendo la detección temprana relevante para ayudar a minimizar el impacto de los síntomas [6].

Cuando el TDAH persiste en la adultez, puede tener consecuencias graves a largo plazo, como bajo rendimiento académico y laboral, incluyendo riesgo de comportamiento antisocial, abuso de drogas y alcohol. Bajo estos antecedentes, la detección temprana es de gran valor para evitar mayores problemas en la edad adulta [9].

El TDAH se ha convertido en un tema de investigación importante al afectar a millones de niños en todo el mundo. Además de aumentar la comprensión sobre el TDAH, estos estudios pueden proporcionarnos un biomarcador confiable para el diagnóstico temprano [5].

2.3 EEG

Electroencefalografía es un tipo de estudio no invasivo realizado para proporcionar información valiosa sobre el funcionamiento cerebral en diversas condiciones y estados mentales. Su funcionamiento se basa en detectar la actividad eléctrica generada por la activación de neuronas dentro del cerebro por medio de electrodos instalados sobre el cuero cabelludo. Las señales de EEG adquiridas de un ser humano pueden utilizarse, por ejemplo, para investigar los siguientes problemas clínicos [2]:

- Monitoreo del estado de alerta, coma y muerte cerebral.
- Localización de áreas dañadas después de una lesión en la cabeza, accidente cerebrovascular y tumor.
- Monitoreo del compromiso cognitivo.
- Investigación de la epilepsia y localización del origen de las convulsiones.
- Evaluación de los efectos de los fármacos antiepilépticos.
- Monitoreo del desarrollo cerebral.
- Investigación de trastornos del sueño y la fisiología.
- Investigación de trastornos mentales.
- Reconocimiento de emociones en autistas.
- Monitoreo de la fatiga mental en pilotos y conductores.

2.3.1 Actividad neuronal

Las actividades en el Sistema Nervioso Central (SNC) están principalmente relacionadas con las corrientes sinápticas transferidas entre las uniones de los axones y dendritas, o dendritas y dendritas de células [2]. En la imagen 2.1 se pueden ver las principales estructuras de la neurona.

El potencial de acción es la información transmitida por un nervio. Estos potenciales son causados por un intercambio de iones a través de la membrana neuronal y un potencial de acción es un cambio temporal en el potencial de la membrana que se transmite a lo largo del axón. El potencial de membrana se

despolariza (se vuelve más positivo), produciendo un peak. Después del peak, la membrana se repolariza (se vuelve más negativa). El potencial se vuelve más negativo que el potencial de reposo y luego vuelve a la normalidad como se muestra en la imagen 2.2.

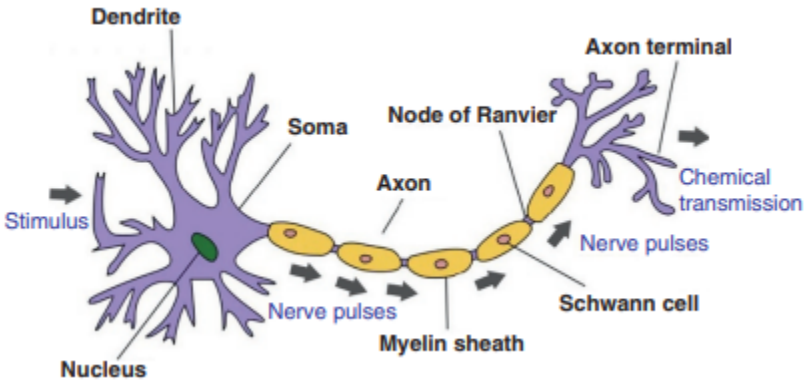

Fig. 2.1: Estructura de una Neurona [2]

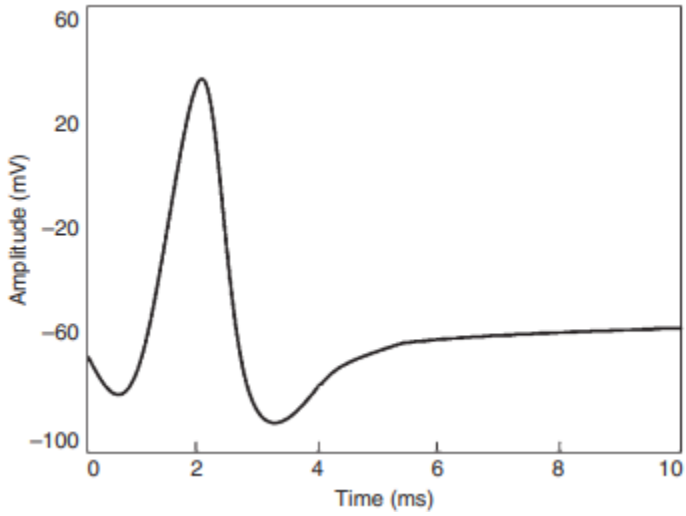

Fig. 2.2: Potencial de Acción [2]

2.3.2 Bandas cerebrales

Por medio del estudio de EEG se han podido distinguir 5 pricipales ondas cerebrales que se diferencian entre sí por sus rangos de frecuencia. Estas 5 bandas de frecuencia se denominan alfa (α), theta (θ), beta (β), delta (δ) y gamma (γ). Cada una de estas bandas, se asocia a diferentes estados del cerebro descritos en la Tabla 2.1 [1].

En la imagen 2.3 se puede observar cada banda y notar la diferencia de frecuencia y amplitud.

Tabla 2.1: Clasificación de bandas cerebrales [1]

Banda	**Frecuencia (Hz)**	**Descripción**
Delta (δ)	< 4	Se estima que tienen su origen en regiones subcorticales y se asocian con estados de sueño profundo o encefalopatías graves.
Theta (θ)	4-8	Las ondas teta aparecen a medida que la conciencia se desliza hacia la somnolencia. Los cambios en el ritmo de las ondas teta se examinan para estudios de maduración y emocionales.
Alfa (α)	8-13	Se asocian con estados de relajación y alerta tranquila, sin atención o concentración específica. Tiene una amplitud más alta sobre las áreas occipitales.
Beta (β)	13-30	Esta onda está presente en el ritmo habitual de vigilia del cerebro asociado con el pensamiento activo, la atención activa, el enfoque o la resolución de problemas concretos. La actividad beta rítmica se encuentra principalmente sobre las regiones frontales y centrales.
Gamma (γ)	<30	La detección de estos ritmos puede usarse para la confirmación de ciertas enfermedades cerebrales. Las regiones de frecuencias EEG altas y los niveles más altos de flujo sanguíneo cerebral se encuentran en el área frontocentral.

2.3.3 Adquisición de la señal

El sistema de medición de EEG debe consistir de varios elementos necesarios, incluyendo los electrodos con medios conductivos, amplificadores con filtros, convertidor analógico a digital y dispositivo de grabación. Las señales de microvoltios registradas desde los electrodos en el cuero cabelludo son transformadas por los amplificadores en señales dentro de un rango adecuado de voltaje. Luego, las señales son transformadas por el convertidor de un formato analógico a uno digital y finalmente almacenadas a través del dispositivo de grabación [3]. Esquemáticamente se muestra en la Figura 2.4.

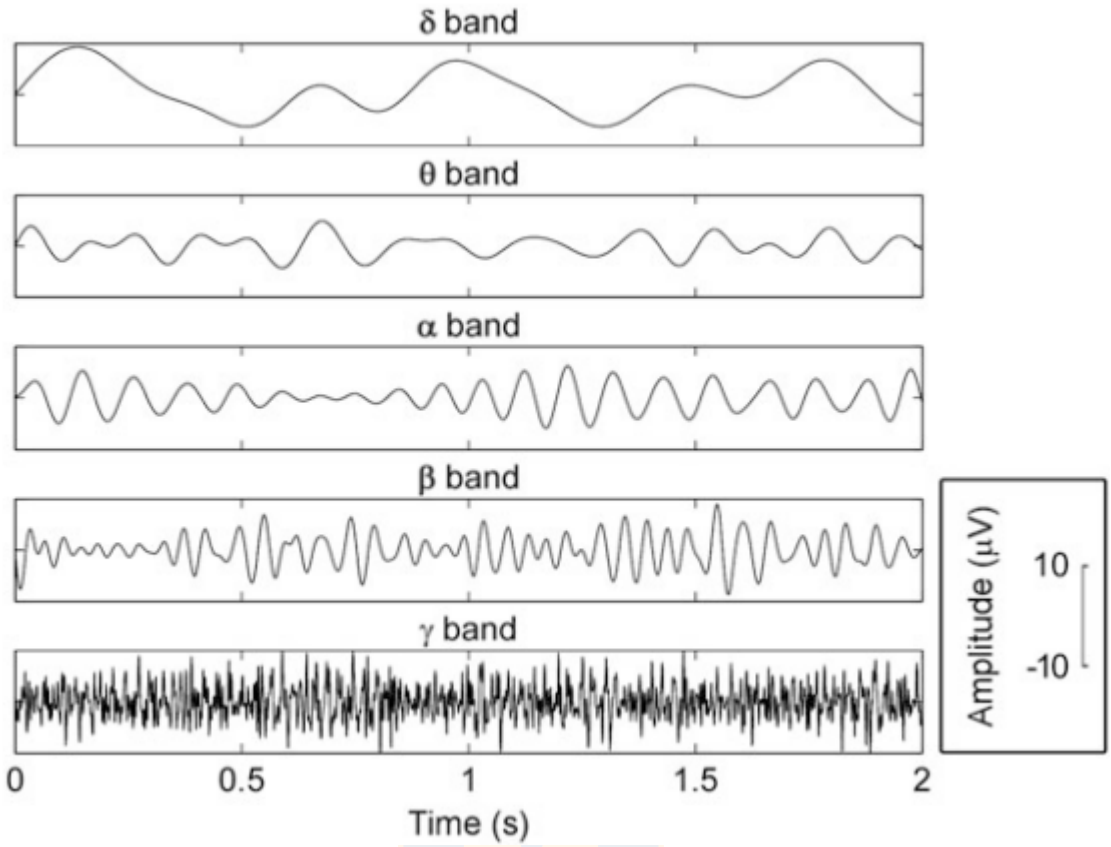

Fig. 2.3: Bandas Cerebrales [1]

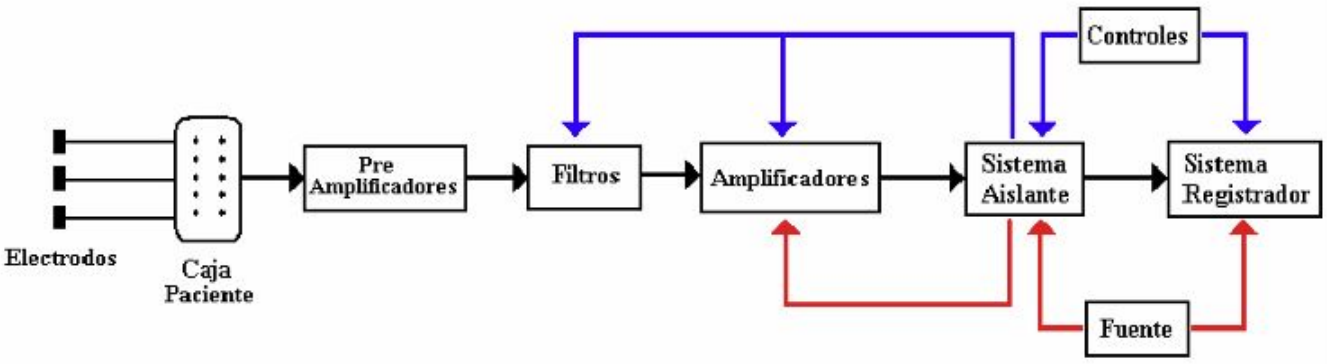

Fig. 2.4: Esquema Electroencefalógrafo [3]

2.3.4 Tipos de electrodos

Los electrodos comúnmente utilizados para EEG se pueden clasificar según sus cacterísticas electrónicas de adquisición [10]:

- Electrodo pasivo: Los cables de los electrodos pasivos se conectan a una caja de interruptores que

contiene todos los componentes electrónicos de conmutación.

- Electrodo activo: Los cables de los electrodos activos tienen componentes electrónicos de conmutación integrados.

También se pueden clasificar según cómo se adhieren a la cabeza para adquirir las señales cerebrales:

- Electrodo seco: No utiliza gel conductor y pueden ser electrodos de contacto, sin contacto y aislantes.
- Electrodo húmedo: Utiliza un gel conductor que está en contacto con el cabello.

2.3.5 Ubicación de los electrodos

La Federación Internacional de Sociedades de Electroencefalografía y Neurofisiología Clínica recomienda la disposición convencional de electrodos denominada 10-20 para 19 electrodos [2] (excluyendo los electrodos del lóbulo de la oreja). El '10' y el '20' se refieren al hecho de que las distancias reales entre electrodos adyacentes son el 10 o el 20 % de la distancia total adelante-atrás o derecha-izquierda del cráneo [1]. Cuando se registra un EEG usando más electrodos, se interpolan electrodos adicionales usando la división del 10 %, que llena los sitios intermedios a medio camino entre los del sistema 10-20 existente. La distribución 10-20 se muestra en la Figura 2.5.

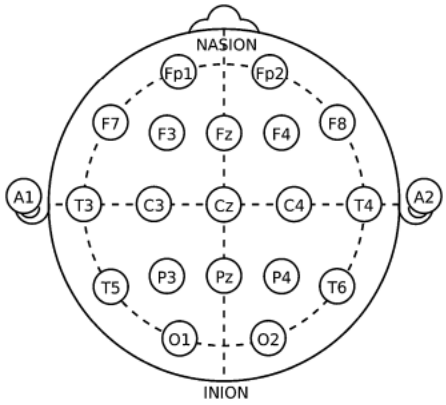

Fig. 2.5: Ubicación de Electrodos Sistema 10-20 [4]

2.4 Machine Learning

Machine Learning (aprendizaje automático) es una combinación de aprendizaje a partir de clasificación o predicción de datos o agrupamiento. En la mayoría de las aplicaciones, se combina con algunas técnicas de procesamiento de señales para extraer las mejores características discriminativas de los datos antes de tomar pasos de aprendizaje y clasificación [4].

Existen principalmente dos tipos de técnicas: aprendizaje supervisado y aprendizaje no supervisado. En la técnica de aprendizaje supervisado, el objetivo se conoce durante la fase de entrenamiento y el clasificador se entrena para minimizar la diferencia entre la salida real y los valores objetivo, es decir, ajusta sus parámetros para que las predicciones sean lo más precisas posible. Por otro lado, en el aprendizaje no supervisado, el clasificador encuentra patrones ocultos o estructuras intrínsecas en los datos de entrada [2].

Aplicar aprendizaje automático en el contexto de estudio de señales EEG, generalmente implica los siguientes pasos: [11]:

- Adquisición de la señal
- Preprocesamiento
- Selección de características relevantes
- Entrenamiento/Validación del modelo
- Optimización de parámetros

En la siguiente sección, se detallarán los principales pasos involucrados en el procesamiento de señales de EEG para algoritmos de clasificación.

2.5 Procesamiento de señales de EEG

2.5.1 Preprocesamiento

El preprocesamiento es un paso crucial para obtener resultados precisos y fiables en el análisis de señales de EEG. La adquisición de señales de EEG viene acompañada de ruidos de línea y artefactos. Los

artefactos son señales fisiológicas no deseadas que pueden distorsionar los datos. Entre los artefactos más comunes se encuentran el electromiograma (EMG), el electrocardiograma (ECG) y aquellos causados por movimientos corporales menores, movimientos oculares y sudoración [1].

La presencia de artefactos dificultan el análisis e interpretación de las señales de EEG, ya que introducen variabilidad y ruido que no están relacionados con la actividad cerebral que se desea estudiar. Es fundamental implementar técnicas de preprocesamiento que permitan mejorar la calidad de la señal antes de proceder con la extracción de características [1].

Este proceso implica varias etapas que ayudan a mitigar el ruido y los artefactos que pueden distorsionar los registros. A continuación, se mencionan los principales pasos involucrados en el preprocesamiento de las señales de EEG [1]:

- Filtrado
- Re-Referencia
- Extracción de Épocas
- Eliminación de Canales Malos
- Eliminación de Artefactos

A continuación, la sección se centrará en detallar dos de los procedimientos más relevantes realizados comúnmente durante el preprocesamiento de señales de EEG en contexto de detección de TDAH: Filtrado y Eliminación de Artefactos Fisiológicos.

Filtrado

El filtrado se utiliza para eliminar el ruido de línea y las frecuencias no deseadas que pueden estar presentes en las señales de EEG. Este paso ayuda a mejorar la relación señal-ruido y a aislar las frecuencias de interés que corresponden a la actividad cerebral relevante [1].

Una de las fuentes más comunes de artefactos no fisiológicos es la interferencia eléctrica. Esta proviene del suministro de energía alterna, que es de 60 Hz o 50 Hz dependiendo de dónde es adquirida la señal. La interferencia de la línea eléctrica puede eliminarse aplicando un filtro Notch [12] [9].

De acuerdo a las características señaladas en la Tabla 2.1, las bandas de frecuencia se asociasan a diferentes estados del cerebro. Tal como se señala en [9] la mayoría de los pacientes con TDAH tienen un patrón de ondas cerebrales común que consiste en una abundancia de ondas cerebrales lentas (delta o teta) y una escasez de ondas cerebrales rápidas (beta), por lo que se pueden utilizar además, filtros para seleccionar solo las ondas relevantes para el estudio.

Componentes de Artefactos Independientes (ICA)

Los artefactos, como los movimientos oculares, la actividad muscular y otros ruidos fisiológicos, pueden distorsionar significativamente las señales de EEG. La aplicación de técnicas como el Análisis de Componentes Independientes (ICA) permite separar y eliminar estos artefactos, mejorando las señales neuronales [2].

ICA es una técnica de separación de fuentes que intenta identificar fuentes independientes en los datos de EEG. Aplicar ICA a los datos de EEG implica la descomposición de las series temporales de EEG en un conjunto de componentes. Los datos de EEG se transforman en una colección de salidas simultáneas de filtros espaciales aplicados a todos los datos multicanal [1].

Después de la descomposición por ICA, cada fila de la matriz de datos transformada representa la evolución temporal de la actividad de un Componente Independiente (IC) que ha sido filtrado espacialmente a partir de los datos del canal. Las salidas del procedimiento de ICA son formas de onda de IC, así como una matriz que transforma los datos de EEG a datos de IC, y su matriz inversa para transformar los datos de IC de nuevo a datos de EEG. Estas salidas proporcionan información sobre las propiedades temporales y espaciales de un IC [1].

Los datos de EEG pueden considerarse como sumas de señales reales de EEG y artefactos, que son independientes entre sí. Por lo tanto, ICA es potencialmente una metodología útil para separar artefactos de señales de EEG. Para la eliminación de artefactos en los registros de EEG, los ICs calculados se clasifican primero como componentes artefactuales o relacionados con el sistema nervioso. Si se detectan y se identifican como ICs relacionados con artefactos, pueden restarse de los datos registrados y los datos restantes pueden remezclarse. ICA separa los componentes con el fin de identificar artefactos relacionados con movimientos oculares, ECG, EMG y ruido de línea. Estos ICs relevantes para artefactos tienen formas características (topografías, evoluciones temporales y espectros de frecuencia) y pueden identificarse automáticamente mediante Software [1].

Esta técnica de separación de fuentes ha sido utilizada en estudios relacionados con el procesamiento

de señales EEG [13], [14]. Es muy probable que a lo largo de las señales EEG se encuentren artefactos y este método ha sido perfeccionado y estudiado arrojando resultados favorables en la etapa de preprocesamiento.

2.5.2 Extracción de Características

Para un entrenamiento eficiente del algoritmo de clasificación, es necesaria la extracción de características [9]. Esta es una de las etapas más determinantes de los resultados de la clasificación. Realizar una extracción eficaz de características de los datos de EEG puede utilizarse para identificar y pronosticar enfermedades cerebrales [15]. A continuación, se presentan las características comúnmente extraídas que han sido estudiadas en investigaciones relacionadas con la clasificación del TDAH en el Dominio del Tiempo y en el Dominio de la Frecuencia [16] [9], [4], [5], [15].

Características en el Dominio del Tiempo

Las señales de EEG no son estacionarias, por lo que sus propiedades varían sustancialmente con el tiempo. Estas características variables en el tiempo, podrían ser moduladas por condiciones experimentales o estados mentales y, por lo tanto, transmiten información importante [1]. Las características estudiadas en el dominio del tiempo pueden ser estadísticas, morfológicas y no lineales y se han utilizado en varios estudios de señales de EEG [6], [16], [9], [4], [5]. Las características estadísticas consideran modelar los datos en términos de rasgos estadísticos como: media, mediana, rango, varianza, curtosis y asimetría [17]. Para especificar más detalladamente, se presenta la fórmula de cada característica y una pequeña descripción.

- **Características Estadísticas**

  - **Media Aritmética o Promedio**: Estadística más comúnmente usada, la cual se puede obtener sumando todos los valores de la muestra y dividiendo por el tamaño de esta. Representa la tendencia central.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.5.1}$$

    donde:

    - $\overline{x}$ es la media aritmética.
    - n es el tamaño de la muestra.

- x_i es el valor de la muestra i-ésima.

- **Mediana**: Valor en el medio de los valores ordenados de la muestra, lo cual divide los datos de la muestra en dos mitades iguales.

  Si el tamaño n es impar, la mediana es:

$$\text{Me} = x_{\left(\frac{n+1}{2}\right)} \tag{2.5.2}$$

  Si el tamaño n es par, la mediana es:

$$\text{Me} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} \tag{2.5.3}$$

  donde:

  - $x_{(k)}$ representa el k-ésimo valor ordenado de la muestra.

- **Varianza**: Utilizada para medir la diferencia de cada dato de muestra respecto al valor medio. La varianza de un conjunto de datos finito con tamaño n, se puede calcular mediante la siguiente fórmula:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{2.5.4}$$

  donde:

  - σ^2 es la varianza.
  - $\overline{x}$ es la media aritmética de la muestra.

- **Desviación Estándar de la Muestra**: Medida estadística que indica la dispersión o variabilidad de un conjunto de datos respecto a su media aritmética.

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{2.5.5}$$

  donde:

  - σ es la desviación estándar muestral.

- **Rango Intercuartílico (IQR):** Es una medida de la dispersión de los datos alrededor de la mediana $Q2$, y se calcula restando el primer cuartil $Q1$ al tercer cuartil $Q3$:

$$\text{IQR} = Q3 - Q1$$

  donde:

- $Q1$ es el primer cuartil (el valor que divide el 25 % inferior de los datos).
- $Q3$ es el tercer cuartil (el valor que divide el 25 % superior de los datos).
- $Q2$ es la mediana.

- **Sesgo (Skewness) y Curtosis (Kurtosis)**: Son medidas que describen la forma de la distribución de los datos.

  El sesgo mide la asimetría de la distribución de los datos respecto a su media. Se calcula como:

$$\text{Skew} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma}\right)^3 \tag{2.5.6}$$

  La curtosis mide la forma de la cola de la distribución en comparación con una distribución normal. Se calcula como:

$$\text{Kurt} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma}\right)^4 \tag{2.5.7}$$

  donde:

  - $\overline{x}$ es la media de la muestra.
  - σ es la desviación estándar de la muestra.

- **Parámetros de Hjorth**

  Los parámetros de Hjorth son una serie de métricas que describen la forma de una señal en términos de tres parámetros: actividad, movilidad y complejidad.

  - **Actividad**

    La actividad es una medida de la varianza de la señal, lo cual representa la potencia del espectro de la señal. Se define como:

$$\text{Actividad} = \sigma_x^2 \tag{2.5.8}$$

    donde:

    - σ_x^2 es la varianza de la señal x.

  - **Movilidad** La movilidad es una medida de la variación de la primera derivada de la señal en comparación con la variación de la señal en sí. Se calcula como:

$$\text{Movilidad} = \sqrt{\frac{\sigma_{x'}^2}{\sigma_x^2}} \tag{2.5.9}$$

    donde:

- $\sigma^2_{x'}$ es la varianza de la primera derivada de la señal.
- σ^2_x es la varianza de la señal original x.

donde $\sigma^2_{x'}$ es la varianza de la primera derivada de la señal.

- **Complejidad** La complejidad es una medida de la variación de la segunda derivada de la señal en comparación con la primera derivada. Indica cuán compleja es la forma de la señal en relación a una onda sinusoidal pura. Se define como:

$$\text{Complejidad} = \frac{\text{Movilidad de } x'}{\text{Movilidad de } x} \tag{2.5.10}$$

donde la movilidad de x' es la movilidad de la primera derivada de la señal.

donde:

- La **movilidad de** x' es la movilidad de la primera derivada de la señal, calculada como:

$$\text{Movilidad de } x' = \sqrt{\frac{\sigma^2_{x''}}{\sigma^2_{x'}}}$$

- La **movilidad de** x es la movilidad de la señal original.
- $\sigma^2_{x''}$ es la varianza de la segunda derivada de la señal.

Estos parámetros son útiles para caracterizar las señales EEG, proporcionando información sobre su estructura y dinámica.

- **Características Morfológicas**

  - **Amplitud Absoluta**

    El valor absoluto de la señal está definido por:

$$\text{AA} = \max |x(t)| \tag{2.5.11}$$

    donde:

    - $x(t)$ es la señal en el tiempo t.
    - $\max |x(t)|$ representa el valor máximo absoluto de la señal en todo el intervalo de tiempo.

  - **Área Positiva**

    La suma de los valores positivos de una señal:

$$\text{AP} = \sum_t 0.5 \left(x(t) + |x(t)| \right) \tag{2.5.12}$$

    donde:

- $x(t)$ es la señal en el tiempo t.
- $\sum_t$ denota la suma a lo largo de todos los instantes de tiempo t.

- **Área Negativa**

  La suma de los valores negativos de una señal:

$$\text{AN} = \sum_t 0.5\,(x(t) - |x(t)|) \tag{2.5.13}$$

  donde:

  - $x(t)$ es la señal en el tiempo t.
  - $\sum_t$ denota la suma a lo largo de todos los instantes de tiempo t.

- **Área Total**

  La suma de los valores positivos y negativos de una señal:

$$\text{AT} = \text{AP} + \text{AN} \tag{2.5.14}$$

  donde:

  - AP es el área positiva de la señal.
  - AN es el área negativa de la señal.

- **Peak to Peak**

  La diferencia entre el máximo y el mínimo de una señal:

$$\text{PP} = X_{\max} - X_{\min} \tag{2.5.15}$$

  donde:

  - $X_{\max}$ es el valor máximo de la señal $x(t)$.
  - $X_{\min}$ es el valor mínimo de la señal $x(t)$.

- **No Lineales**

  Las características no lineales pueden interpretarse como la complejidad de un sistema. En el contexto del cerebro y las señales EEG, estas características no lineales reflejan la complejidad de la actividad cerebral, que es altamente dinámica y cambia en escalas de tiempo. La atención fortalece un sentido mientras debilita otros y reduce la complejidad del cerebro, por lo que las características no lineales del EEG pueden ser herramientas adecuadas para investigar la atención [18].

  La dimensión fractal es la medida de la complejidad de patrones fractales, expresando su complejidad como una proporción definida por los cambios. Los fractales son estructuras que exhiben

autosimilitud, lo que significa que se repiten a diferentes escalas de observación. La dimensión fractal cuantifica cómo cambia la estructura de un fractal a medida que se examina a diferentes niveles de detalle o escala. Si la estructura cambia significativamente con la escala, la dimensión fractal será mayor, indicando mayor complejidad. El análisis de estas características a menudo se ha realizado con estos tres algoritmos [5]:

- **Dimensión Fractal de Katz**

$$FD_{\text{Katz}} = \frac{\ln(N-1)}{\ln(N-1) - \ln\left(\frac{d}{L}\right)}$$

  - N: Número total de puntos en la serie temporal.
  - d: La distancia entre los puntos más alejados en la serie.
  - L: La longitud de la serie temporal.

- **Dimensión Fractal de Petrosian**

$$D_{\text{Petrosian}} = \frac{\log_{10}(n)}{\log_{10}(n) + \log_{10}\left(\frac{n}{n+0.4N_{\Delta}}\right)}$$

  - n: Número de segmentos en la serie temporal que son mayores que el promedio.
  - N_{Δ}: Número total de cambios en la serie temporal.

- **Dimensión Fractal de Higuchi**

$$D_{\text{Higuchi}} = \frac{\log_{10}(N)}{\log_{10}(k)} - 1$$

  - N: Número total de puntos en la serie temporal.
  - k: Escala utilizada para el cálculo de la dimensión fractal.

Características en el Dominio de la Frecuencia

La frecuencia es un parámetro fundamental e importante para describir una señal. El análisis espectral puede transformar las señales de EEG del dominio del tiempo al dominio de la frecuencia, lo que puede revelar cómo se distribuye la potencia de las señales de EEG a lo largo de las frecuencias. La estimulación sensorial o las tareas cognitivas podrían aumentar o disminuir las actividades rítmicas del EEG en ciertas bandas de frecuencia. Se utiliza comúnmente en una amplia gama de señales que exhiben patrones oscilatorios y rítmicos, y es de particular importancia en el análisis de EEG. En el contexto del análisis de EEG, la estimación espectral se aplica normalmente a grabaciones continuas de EEG durante un período de tiempo para calcular la potencia de varios

ritmos específicos: theta, delta, alfa, beta y gamma. El propósito de la estimación espectral es detectar periodicidades en los datos, observando peaks en las frecuencias correspondientes a estas periodicidades. La base de la estimación espectral es la Transformada de Fourier (FT) [1]:

$$X(f) = \int_{-\infty}^{\infty} x(t)\, e^{-j2\pi ft}\, dt \tag{2.5.16}$$

donde $X(f)$ es la representación de la señal en el dominio de la frecuencia, $x(t)$ es la señal en el dominio del tiempo, f es la frecuencia y j es la unidad imaginaria.

Existen otros métodos a partir de esta transformada que permiten el estudio de las señales en frecuencia como lo es el método Welch. Este método divide la señal en segmentos superpuestos, calcula la Densidad Espectral de Potencia (PSD) en cada segmento y promedia los resultados para obtener una estimación más precisa y suave de la potencia en el dominio de la frecuencia. Puede manejar señales no estacionarias al dividirlas en pequeños segmentos, lo que permite una mayor resolución en el análisis de la frecuencia sin introducir ruido adicional. Además, permite ajustar el solapamiento entre segmentos y aplicar ventanas. Este método es ampliamente utilizado para analizar señales biomédicas como el EEG, donde es crucial estimar la distribución de potencia en las diferentes bandas de frecuencia [1].

Como se mencionaba anteriormente, en el análisis espectral se puede calcular la PSD, que describe cómo se distribuye la potencia de una señal aleatoria a lo largo de diferentes frecuencias y la Potencia de Banda (PD) para evaluar la cantidad de potencia contenida en una banda específica. La PD se puede calcular integrando la PSD sobre el rango de frecuencias de interés [1].

$$P_{\text{banda}} = \int_{f_1}^{f_2} \mathrm{PSD}(f)\, df \tag{2.5.17}$$

donde:

- f_1 y f_2 son los límites inferior y superior de la banda de frecuencias de interés.

La unidad de la PSD es V^2/Hz y se puede expresar en decibelios (dB) como $10\log_{10}(\mathrm{V}^2/\mathrm{Hz})$. La potencia de banda, al integrar la PSD sobre un rango de frecuencias, tiene unidades de V^2.

2.5.3 Selección de Características Relevantes

Usar todas las características extraídas dará como resultado la inclusión de datos irrelevantes o duplicados, lo que reducirá la precisión de la clasificación [19]. Esta etapa es el procedimiento para

seleccionar características cruciales entre una amplia cantidad de ellas. El beneficio de la selección de características es que reduce el sobreajuste y aumenta la precisión de la clasificación [16].

LASSO (Least Absolute Shrinkage and Selection Operator) es una técnica de regularización utilizada en la selección de características mediante el método de contracción. En esta técnica, los coeficientes determinados en el modelo lineal se contraen hacia cero, introduciendo un factor de penalización basado en la norma L_1. El parámetro de penalización α denota la cantidad de contracción (o restricción) que se aplicará en la ecuación. Puede ser cualquier número real entre cero e infinito; cuanto mayor sea el valor, más agresiva será la penalización. Como resultado de este encogimiento, LASSO tiende a llevar algunos coeficientes a exactamente cero. Esto implica que el modelo puede efectivamente excluir ciertas características del conjunto de datos, lo que resulta en un subconjunto más manejable de variables relevantes. De esta manera, LASSO proporciona un mecanismo eficiente para la selección de características [20].

.

$$L_{\text{lasso}}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\hat{\beta})^2 + \alpha\sum_{j=1}^{m}|\hat{\beta}_j| \tag{2.5.18}$$

- $\sum_{i=1}^{n}$: Representa la suma de los cuadrados de los residuos, calculados como las diferencias entre los valores observados y_i y los valores predichos por el modelo $\mathbf{x}_i^T\hat{\beta}$, para cada una de las n observaciones. Este término mide qué tan bien el modelo ajusta los datos.
- $\mathbf{x}_i^T\hat{\beta}$: Producto escalar entre el vector de predictores $\mathbf{x}_i$ (transpuesto) y el vector de coeficientes $\hat{\beta}$. Este término calcula la predicción del modelo para una observación específica i, considerando los valores de las variables predictoras y sus respectivos coeficientes.
- $\alpha\sum_{j=1}^{m}|\hat{\beta}_j|$: Penalización basada en la norma L_1, que suma los valores absolutos de los coeficientes $\hat{\beta}_j$ del modelo, escalada por el parámetro de regularización α. Este término controla la complejidad del modelo: al aumentar α, se incentiva que algunos coeficientes se reduzcan a cero, promoviendo la selección de características más relevantes.

Algunos estudios han demostrado la efectividad de LASSO en la selección de características para tareas de clasificación enfocadas en la detección de TDAH. En [19] se aplicó esta técnica a características no lineales seleccionando 38 de un total de 58, obteniéndose la mejor métrica de accuracy para el clasificador RUS Boosted Trees (94.6 %). En [6] se utilizó LASSO-LR donde el método seleccionó 47 de 342 características dando como mejor métrica en el clasificador SVM accuracy 94.2 %. También se utilizó LASSO-LR en [4] donde se indentificaron 28 de 132 características más relevantes en base a 6 de 19 canales seleccionados anteriormente, como resultado se obtuvo 97.53 % de accuracy en el clasificador Gaussian Process (GP). Esto demuestra que LASSO

puede ser una herramienta poderosa para reducir el conjunto de variables a las más importantes, mejorando la precisión de los modelos.

2.5.4 Clasificadores

- **Support Vector Machine**

  Probablemente el clasificador más popular para muchas aplicaciones. Corresponde a un tipo de algoritmo supervisado de clasificación. Puede clasificar datos encontrando la frontera de decisión lineal (hiperplano) que separa todos los puntos de datos de una clase de los de la otra clase. El mejor hiperplano para un SVM es aquel con el mayor margen entre las dos clases, cuando los datos son linealmente separables. Es muy utilizado en datos que tienen exactamente dos clases [2].

- **Naive Bayes**

  Este clasificador es un modelo probabilístico basado en el teorema de Bayes, que asume que las características son independientes entre sí. Aunque esta suposición rara vez es cierta en la práctica, Naive Bayes ha demostrado ser efectivo en problemas con datos de alta dimensionalidad y conjuntos de datos pequeños. El clasificador asigna una etiqueta de clase basándose en la probabilidad posterior calculada a partir de las probabilidades de las características observadas, es sencillo de implementar y eficiente computacionalmente. [21].

- **Random Forest**

  Es un algoritmo de aprendizaje automático basado en ensamblajes que construye múltiples árboles de decisión y combina sus predicciones para mejorar la precisión y reducir el riesgo de sobreajuste. Cada árbol en el bosque se construye utilizando una muestra aleatoria del conjunto de datos de entrenamiento, y en cada nodo del árbol, se selecciona aleatoriamente un subconjunto de características para determinar la mejor división. Random Forest es conocido por su capacidad de manejar datos de alta dimensionalidad, tolerar datos faltantes y reducir el sobreajuste al promediar las predicciones de múltiples árboles [22].

- **Regresión Logística**

  Este modelo se basa en una representación de las entradas como un vector de características y utiliza funciones como la sigmoide (para clasificación binaria) o softmax (para clasificación multiclase) para estimar $p(y|x)$, donde y es la clase objetivo y x son las características de entrada. Durante el entrenamiento, el modelo ajusta sus parámetros minimizando una función de pérdida. En la fase de prueba, la regresión logística clasifica una nueva entrada asignándola

a la clase con la probabilidad más alta. Este método es ampliamente utilizado debido a su simplicidad y efectividad en problemas linealmente separables [23].

2.5.5 Entrenamiento/Validación del Modelo

El objetivo de un clasificador de aprendizaje automático es estimar la clase correcta correspondiente a un vector de características dado, basado en algún conocimiento previo obtenido a través del entrenamiento. El entrenamiento es el proceso donde un clasificador aprende la relación de mapeo entre los vectores de características y sus etiquetas de clase correspondientes. Esta relación entre los vectores de características y las etiquetas de clase correspondientes forma un límite de decisión en el espacio de características que separa los patrones de diferentes clases entre sí. Antes de entrenar el clasificador, se deben organizar las muestras [1]:

- Conjunto de entrenamiento (train set): para entrenar un clasificador, con etiquetas de clase conocidas.
- Conjunto de prueba (test set): para evaluar el rendimiento de un clasificador, con etiquetas de clase conocidas.

En hold-out (validación simple) usualmente el 80 % de los datos se utiliza para entrenar el modelo y el 20 % restante, se utiliza como conjunto de prueba para evaluar el rendimiento del modelo con datos que no se han utilizado en el entrenamiento. Este es un enfoque básico para separar los datos y medir la capacidad de un modelo para trabajar con datos que no ha visto antes. Sin embargo, los resultados pueden variar significativamente dependiendo de cómo se dividen los datos [24].

Adicionalmente, se han desarrollado otros enfoques como la validación cruzada para utilizar todos los datos disponibles tanto para entrenamiento como para prueba. Esto ayuda a aprovechar al máximo los datos y a evaluar la capacidad del clasificador para generalizar. A continuación se describen dos de ellos ampliamente utilizados [1]:

- LOOCV (leave-one-out cross validation): Utilizando este método con N muestras, se puede dividir en N-1 muestras de entrenamiento y una muestra de prueba, y el mismo procedimiento puede repetirse N veces para asegurar que cada muestra se utilice una vez como muestra de prueba.
- K-fold: En esta técnica K es el número de partes en las que se divide el conjunto de datos (por ejemplo, 5-fold o 10-fold). Para cada uno de los K folds, el modelo se entrena utilizando K-1 folds y se valida en el fold restante. Este proceso se repite K veces, de modo que cada fold se utiliza una vez como conjunto de validación y K-1 veces como conjunto de entrenamiento.

En comparación con LOOCV, la validación cruzada K-fold puede ser más eficiente desde el punto de vista computacional y es menos probable que el clasificador entrenado esté sobreajustado.

2.5.6 Métricas de Evaluación de Desempeño

La evaluación del desempeño de los modelos de clasificación es crucial para comprender la efectividad y la precisión de los métodos aplicados. Las métricas de evaluación proporcionan una manera cuantitativa de medir la capacidad de un modelo para realizar predicciones correctas y generalizar a datos no vistos. Utilizar una combinación de métricas permite obtener una visión completa del rendimiento del modelo [1].

A continuación, se describen las principales métricas utilizadas para evaluar el desempeño de los modelos de clasificación:

Matriz de Confusión

Proporciona una visión general de cómo se comporta el modelo, mostrando las predicciones correctas y los errores cometidos [1]. A continuación, se muestra la matriz de confusión generada por la función `confusion_matrix` de la librería `sklearn.metrics`, que ordena la matriz en base a las etiquetas de clase presentes en los vectores de predicciones y valores reales. En este caso, la fila 1 (Real = 0) corresponde a la clase negativa y la fila 2 (Real = 1) corresponde a la clase positiva.

		Predicción	
		0	**1**
Real	**0**	**Verdaderos Negativos (VN)**	**Falsos Positivos (FP)**
	1	**Falsos Negativos (FN)**	**Verdaderos Positivos (VP)**

Fig. 2.6: Matriz de Confusión (Elaboración propia)

- **VN**: cantidad de negativos que fueron clasificados correctamente como negativos por el modelo.
- **FN**: cantidad de positivos que fueron clasificados incorrectamente como negativos.
- **FP**: cantidad de negativos que fueron clasificados incorrectamente como positivos.

- **VP**: cantidad de positivos que fueron clasificados correctamente como positivos por el modelo.

De la matriz de confusión, se pueden calcular las siguientes métricas [4]:

- Precisión Global (Accuracy): es la proporción de casos clasificados correctamente, lo que incluye los verdaderos positivos (VP) y los verdaderos negativos (VN), respecto al número total de casos. Representa qué tan bien el modelo predice correctamente en general y se define matemáticamente de la siguiente manera:

$$\text{ACC}\ (\%) = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}} \times 100 \tag{2.5.19}$$

- Precisión: es la proporción de casos positivos correctamente clasificados (VP) respecto al total de casos predichos como positivos. Esta métrica mide la exactitud de las predicciones positivas realizadas por el modelo y se define matemáticamente de la siguiente manera:

$$\text{Precision}\ (\%) = \frac{\text{VP}}{\text{VP} + \text{FP}} \times 100 \tag{2.5.20}$$

- Recall: es la proporción de casos positivos correctamente clasificados (VP) respecto al total de casos que realmente son positivos. Esta métrica evalúa la capacidad del modelo para identificar correctamente las instancias positivas y se define matemáticamente de la siguiente manera:

$$\text{Recall}\ (\%) = \frac{\text{VP}}{\text{VP} + \text{FN}} \times 100 \tag{2.5.21}$$

- F1-Score: es la media armónica de la precisión (Precision) y el recall (Sensibilidad). Esta métrica proporciona un balance entre ambas. Se define como:

$$\text{F1-Score}\ (\%) = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100 \tag{2.5.22}$$

Capitulo 3. Metodología

3.1 Introducción

En este capítulo se detallan los pasos seguidos desde la adquisición de los datos hasta la evaluación de los clasificadores, permitiendo comprender cómo se procesaron los datos y cómo se validaron los modelos de clasificación. La investigación se basa en el análisis de señales EEG, con el objetivo de identificar sujetos con TDAH. Se describe la procedencia del dataset y sus características, las técnicas de preprocesamiento utilizadas para preparar las señales EEG, así como el proceso de extracción de características. También se detallarán las técnicas de validación utilizadas para posteriormente presentar los resultados.

3.2 Base de Datos

Para el desarrollo de este proyecto de investigación, se utilizó el dataset "EEG DATA FOR ADHD / CONTROL CHILDREN" [25]. Es de acceso público y fue obtenido desde IEEEDataPort. El dataset cuenta con los registros de EEG de 121 niños entre 7 y 12 años. Del conjunto de datos, 61 niños (M: 48 vs. F: 13) fueron diagnosticados con TDAH por un psiquiatra experimentado según los criterios del DSM-5 y han tomado Ritalin por hasta 6 meses. El grupo control está compuesto por 60 niños (M: 50 vs. F: 10) donde ninguno tiene antecedentes de trastornos psiquiátricos, epilepsia ni ningún informe de conductas de alto riesgo.

El registro EEG se realizó según el estándar de distribución de electrodos 10-20 mediante 19 canales: Fz, Cz, Pz, C3, T3, C4, T4, Fp1, Fp2, F3, F4, F7, F8, P3, P4, T5, T6, O1, O2 (ver Anexo A) a una frecuencia de muestreo de 128 Hz. Los electrodos A1 y A2 fueron las referencias ubicadas en los lóbulos de las orejas.

El protocolo de registro del EEG se basó en tareas de atención visual. En la tarea, se mostró a los niños una serie de imágenes de personajes de dibujos animados y se les pidió que los contaran. El número de personajes en cada imagen se seleccionó al azar entre 5 y 16, y el tamaño de las imágenes era lo suficientemente grande como para que los niños pudieran verlos y contarlos fácilmente. Para tener un estímulo continuo durante la grabación de la señal, cada imagen se mostraba inmediatamente y de forma ininterrumpida después de la respuesta del niño. Por lo tanto, la duración del registro EEG a lo largo de esta tarea visual cognitiva dependía del desempeño del niño.

3.3 Diagrama del Proyecto

En la Figura 3.7, se muestra un diagrama que describe los pasos a seguir en la metodología del proyecto. El flujo de trabajo comienza con la preparación y preprocesamiento de las señales EEG, seguido de la extracción de características y su posterior selección. La fase siguiente incluye el entrenamiento del clasificador, con la incorporación de validación cruzada. Finalmente, se evalúa el desempeño de los clasificadores.

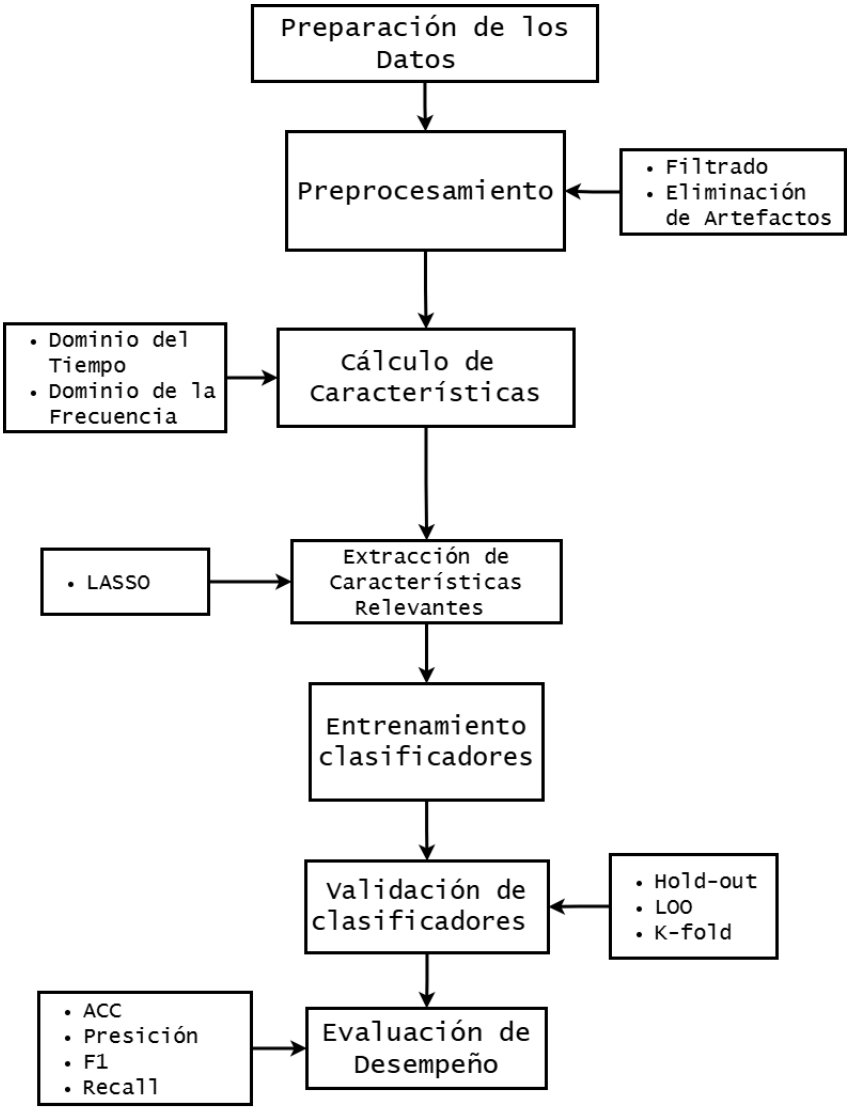

Fig. 3.7: Diagrama del Proyecto (Elaboración propia)

3.4 Software

Para el desarrollo de esta metodología se utilizó:

- MATLAB R2024b y Toolbox EEGLAB 2024.2.1
- Python 3.12.7 (dentro del entorno Spyder 5.5.1 de Anaconda Navigator 2.6.3)

3.5 Preparación de los Datos

La base de datos seleccionada presenta las señales en formato `.mat`. Este formato no es admitido directamente en EEGLAB [26], por lo que se deben cargar los datos en formato `.set`. Mediante un script en MATLAB, se realizó un cambio de formato a los 121 registros de señales.

Una vez listo el formato, las señales se cargaron en el ambiente de EEGLAB. Para identificar correctamente cada canal, se cargó un archivo `.ced` que contenía el nombre de cada electrodo según su posición. En la Figura 3.8 se puede observar la distribución de los canales de adquisición según el protocolo internacional 10-20.

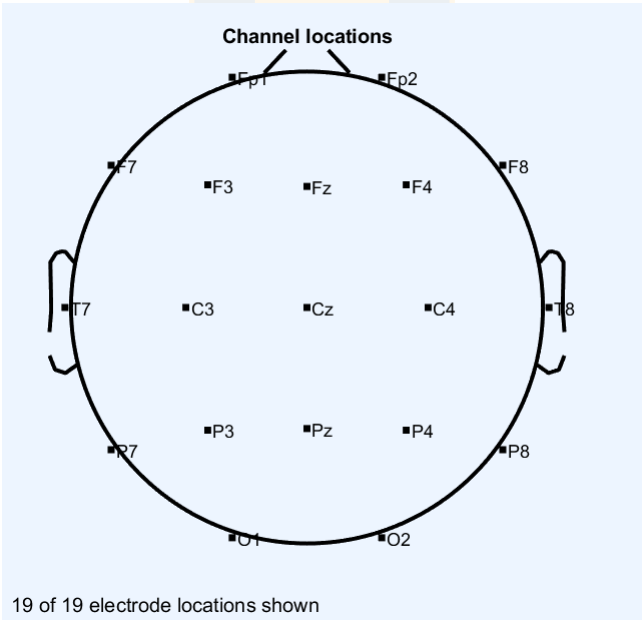

Fig. 3.8: Distribución 10-20 electrodos del dataset

3.6 Preprocesamiento

Se utilizó un filtro pasa alto de 0.5 Hz y un filtro pasa bajo de 30 Hz, lo que permitió eliminar el ruido de línea y filtrar las frecuencias relevantes asociadas al TDAH, que abarcan desde las ondas delta hasta las ondas beta. Según [9], la mayoría de los niños con TDAH presentan una abundancia de ondas lentas (delta-theta) y una disminución de ondas beta, lo que resulta en una alta proporción theta-beta. Por otro lado, en [27] se investigaron tanto grupos con este perfil típico como un perfil atípico caracterizado por un exceso de ondas beta y una baja proporción theta-beta, con el objetivo de determinar si el exceso de beta está relacionado con un estado de hiperactivación o hiperalerta del SNC. En la Figura 3.9 se puede observar el espectro en frecuencia de la señal antes y después de ser filtrada. Luego de filtrar las señales, se realizó la descomposición en fuentes mediante ICA. Este algoritmo se aplica de manera automática dentro de la interfaz de EEGLAB. Cuenta con una herramienta de etiquetado de artefactos 'ICLabel' que cumple la labor de nombrar cada fuente como: 'Brain', 'Muscle' 'Eye','Heart','Line Noise', 'Channel Noise' y 'Other'. La probabilidad de que componentes de la señal pertenezcan a artefactos se estableció en 60 %. El número de artefactos detectados y rechazados, varió de sujeto a sujeto. Finalmente se guardaron 121 matrices `.mat` que contenían canales x muestras. Estos datos se utilizaron en los procesamientos posteriores durante la extracción de características.

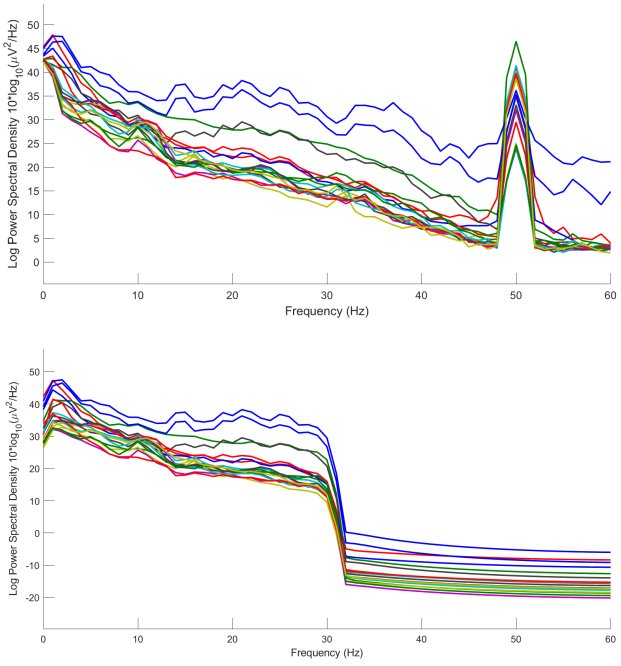

Fig. 3.9: Espectro en frecuencia antes y después de filtrar la señal

3.7 Extracción de Características

Si bien se tenían 121 sujetos en la base de datos, se eliminó la señal con menor duración del grupo TDAH para trabajar con las clases totalmente balanceadas.

- **Características en el Dominio del Tiempo**

  Todas las características extraídas en el dominio del tiempo se muestran a continuación. Cada característica fue calculada para 19 canales, 60 sujetos Control y 60 sujetos TDAH por lo que se almacenaron en matrices 60x19. En la Tabla 3.2 se resumen las dimensiones de cada matriz por grupo de característica. Como resultado, se obtuvieron 18 matrices para el grupo Control y 18 matrices para el grupo TDAH.

  1. Media Aritmética ($\overline{x}$)
  2. Mediana (Me)
  3. Varianza (σ^2)
  4. Desviación Estándar de la Muestra (σ)
  5. IQR
  6. Sesgo (Skew)
  7. Curtosis (Kurt)
  8. Hjorth Actividad
  9. Hjorth Movilidad
  10. Hjorth Complejidad
  11. Amplitud Absoluta (AA)
  12. Área Positiva (AP)
  13. Área Negativa (AN)
  14. Área Total (AT)
  15. Peak to Peak (PP)
  16. Dimensión Fractal de Katz
  17. Dimensión Fractal de Petrosian
  18. Dimensión Fractal de Higuchi

Tabla 3.2: Dimensiones de almacenamiento características extraídas en el Dominio del Tiempo

Características	TDAH (sub x ch)	CTRL (sub x ch)
Estadísticas (7): $\overline{x}$, Me, σ^2, σ, IQR, Skew, Kurt	60x19	60x19
Morfológicas (5): AA, AP, AN, AT, PP	60x19	60x19
Hjorth (3): Actividad, Movilidad, Complejidad	60x19	60x19
Dimensión Fractal (3): Katz, Petrosian, Higuchi	60x19	60x19

sub = sujetos; ch = canales

- **Características en el Dominio de la Frecuencia**

  En el caso de la Potencia de Banda, se calculó para las bandas cerebrales de interés: delta, teta, alfa y beta en los 19 canales de cada sujeto mediante el método Welch. Al aplicar este método, se establecieron ventanas de 4s y 50 % de solapamiento. La matriz resultante de cada

banda se almacenó con dimensiones 60x19 para grupo Control y lo mismo para grupo TDAH como se muestra en la Tabla 3.3.

Tabla 3.3: Dimensiones de almacenamiento características extraídas en el Dominio de la Frecuencia

Potencia de Banda	TDAH (sub x ch)	CTRL (sub x ch)
Delta 0.5-4 Hz	60x19	60x19
Theta 4-8 Hz	60x19	60x19
Alfa 8-13 Hz	60x19	60x19
Beta 13-30 Hz	60x19	60x19

sub = sujetos; ch = canales

3.8 Carga y Organización de los datos en Python

Luego, todas las matrices se cargaron en Python y, para garantizar la homogeneidad en las escalas de las características, los datos fueron normalizados utilizando `StandardScaler` de la librería `scikit-learn`, aplicándose una transformación estándar (media 0 y desviación estándar 1).

$$X_{\text{normalizada}} = \frac{X - \mu}{\sigma} \tag{3.8.1}$$

Donde:

- X es el valor original de la característica.
- μ es la media de la característica.
- σ es la desviación estándar de la característica.

Se definieron 2 listas: `carac_adhd` y `carac_ctrl` para almacenar las rutas a los archivos correspondientes. Luego, mediante un bucle, se recorrió cada archivo de las listas, extrayendo las matrices de datos asociadas a cada característica. Estas matrices se almacenaron temporalmente en las listas `data_adhd_list` y `data_ctrl_list`.

A continuación, se concatenaron las características de cada grupo a lo largo de las columnas usando `np.concatenate` de la librería `Numpy`. Como resultado, se obtuvieron 2 matrices que combinaban el número de canales (19) y la cantidad de características por cada grupo (22). Las dimensiones obtenidas como resultado se muestran en la Tabla 3.4. En la Tabla 3.5 se puede ver un resultado simplificado del procedimiento.

Tabla 3.4: Dimensiones de matrices al concatenar horizontalmente

Matriz	sub x (ch x feat)
TDAH	60x418
CONTROL	60x418

sub = sujetos; ch = canales; feat = características.

Tabla 3.5: Concatenación horizontal de canales y características

Sub	ch 1 (feat 1)	ch 2 (feat 1)	...	ch 1 (feat 2)	ch 2 (feat 2)	...
1	0.53	0.67	...	0.82	0.74	...
2	0.41	0.58	...	0.76	0.63	...
...	...	...	...	...	...	...
60	0.89	0.81	...	0.61	0.72	...

sub = sujetos; ch = canales; feat = característica

Luego ambas matrices se concatenaron verticalmente en una sola estructura de datos mediante la función `vstack` de la librería `Numpy`, lo que resultó en una matriz final X de dimensiones 120×418. En la Tabla 3.6 se muestra el resultado de forma simplificada.

Tabla 3.6: Concatenación vertical de la matriz grupo TDAH y grupo Control

Sub	ch 1 (feat 1)	ch 2 (feat 1)	...	ch 1 (feat 2)	ch 2 (feat 2)	...
1	0.53	0.67	...	0.82	0.74	...
2	0.41	0.58	...	0.76	0.63	...
...	...	...	...	...	...	...
60	0.89	0.81	...	0.61	0.72	...
61	0.90	0.79	...	0.65	0.68	...
62	0.88	0.82	...	0.64	0.70	...
...	...	...	...	...	...	...
120	0.87	0.84	...	0.63	0.73	...

sub = sujetos; ch = canales; feat = característica

Las etiquetas para los grupos se generaron como un vector y, asignando el valor de 1 para los sujetos TDAH y 0 para los sujetos del grupo Control.

3.9 Selección de características más relevantes

Para seleccionar las características más relevantes para la clasificación, se aplicó la técnica de selección de características LASSO. El parámetro α fue ajustado manualmente probando diferentes valores: $10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}$ y 1. Este rango permitió evaluar cómo la regularización influye en el número de características seleccionadas y en el rendimiento de los clasificadores. El valor óptimo de $\alpha = 10^{-2}$ resultó en la selección de 70 características de un total de 418, generando una matriz de dimensiones 120x70.

En la Figura 3.10 se muestran gráficamente las primeras características más relevantes de la lista, ordenadas en orden descendente según el valor absoluto de su coeficiente. La lista completa se puede revisar en el Anexo B.

Las características seleccionadas se utilizaron para entrenar 4 modelos de clasificación: Support Vector Machine, Regresión Logística, Random Forest y Naive Bayes. Estos modelos se ajustaron con hiperparametros que se pueden revisar en el Anexo C.

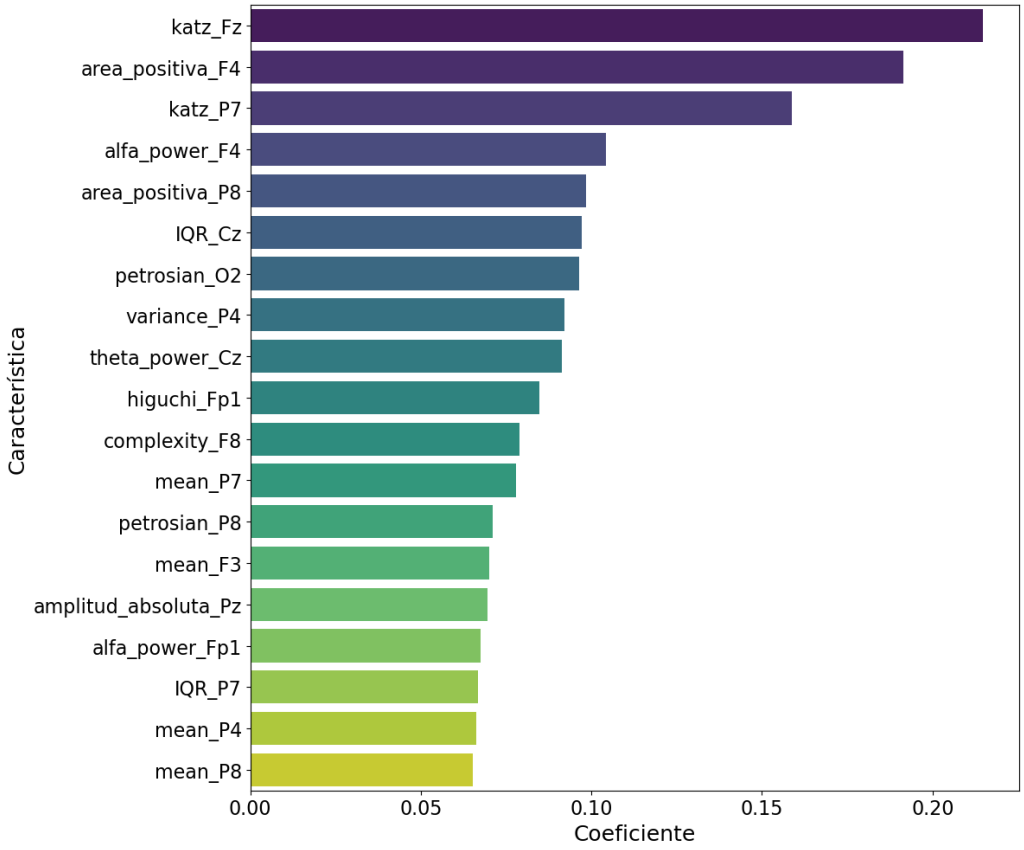

Fig. 3.10: Características más relevantes seleccionadas por LASSO

3.10 Conjuntos de Entrenamiento/Validación

Se ejecutaron 4 formas de entrenamiento y validación de algoritmos:

1. Hold-out: Se dividieron los datos en conjuntos de entrenamiento (80 %), con dimensiones 96x70 y prueba (20 %), con dimensiones 24x70.
2. Leave-One-Out: Cada muestra del conjunto de datos fue utilizada como conjunto de prueba una vez, mientras que las restantes se utilizaron como conjunto de entrenamiento, repitiéndose este proceso un total de 120 veces, correspondiente al número de sujetos en el conjunto de datos.
3. K-Folds: Se empleó este método de validación cruzada para dividir los datos en 5, 10 y 15 folds. En cada iteración, uno de los folds fue utilizado como conjunto de prueba, mientras que los restantes son usados para entrenamiento. Los datos fueron seleccionados de manera aleatoria, lo que asegura que las particiones sean representativas del conjunto de datos, aunque no garantiza una distribución equilibrada de las clases en cada fold.
4. StratifiedKFold: Este método es una variación de Kfold que asegura que cada fold tenga una proporción representativa de muestras de cada clase evitando que alguna clase esté subrepresentada o ausente en los conjuntos de entrenamiento o prueba. Al igual que en K-Folds, se realizaron validaciones cruzadas con 5, 10 y 15 folds. Esta variación de Kfold mejora la estabilidad y confiabilidad de los resultados al preservar las proporciones originales de las clases en cada partición por lo que los resultados finales se presentarán con esta validación.

El desempeño de cada clasificador en estas validaciones, se presentan en el siguiente capítulo de Resultados.

Capitulo 4. Resultados

4.1 Introducción

Este capítulo detalla el desempeño de los 4 modelos utilizados en la clasificación, en términos de accuracy, presición, recall, f1-score y sus respectivas matrices de confusión. También se presenta un análisis estadístico que mostrará las zonas del cerebro donde se ubicaron las características más relevantes seleccionadas por LASSO. La comprensión de estos resultados permitirá tener una visión clara de cuán bien funcionó la metodología empleada y cuál ha sido el mejor modelo para esta tarea de clasificación.

4.2 Resultados ubicación caracteristicas más relevantes

En la Figura 4.11 el gráfico de barras presenta la distribución de las características más relevantes en las diferentes zonas del cerebro. Como se puede observar, en la zona Frontal (F) se presenta el mayor número de características relevantes. Esto se puede atribuir al impacto del tamaño del lóbulo frontal en la actividad cerebral y cómo está involucrado en funciones del lenguaje expresivo, decisiones e interacciones sociales. Le sigue en número, la zona Parietal (P) que se relaciona con la percepción, sensación e integración de la información sensorial con el sistema visual [28].

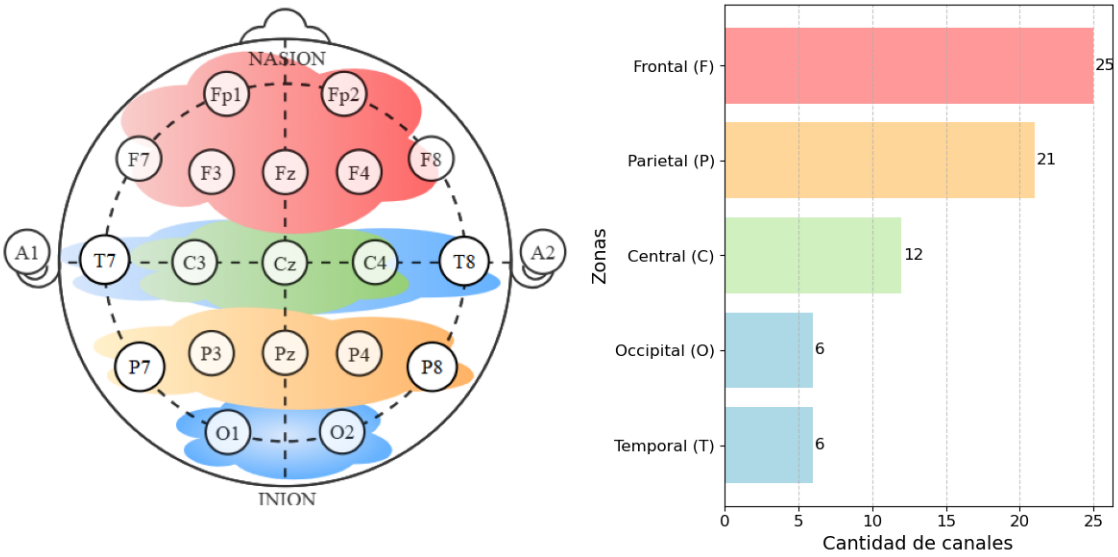

Fig. 4.11: Distribución de características más relevantes por zonas del cerebro

En el segundo gráfico (Figura 4.12), se presenta una comparación del número de características más relevantes entre el hemisferio izquierdo y derecho del cerebro. Se observa que las características predominan en el lado derecho en todas las zonas a excepción de la zona Occipital (canales O). Esta distribución podría estar asociada a individuos con funciones lateralizadas atribuidas al lado derecho, como el análisis global de imágenes o conceptos y localización de objetos [29].

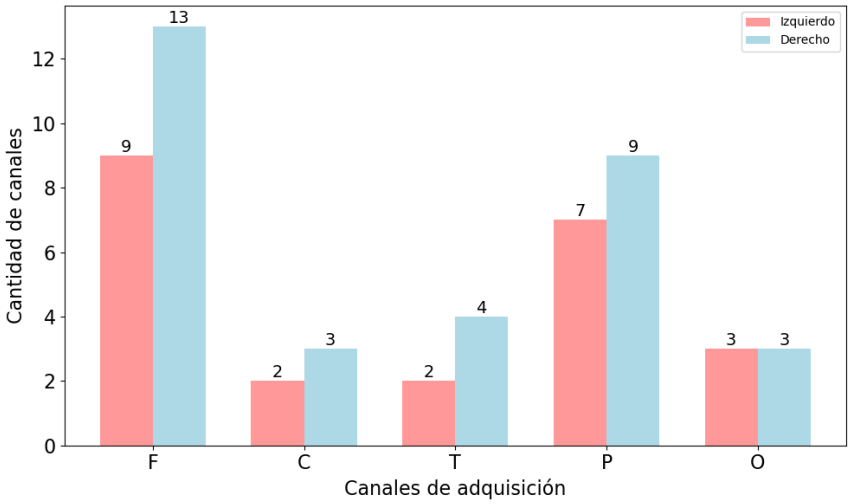

Fig. 4.12: Distribución de características más relevantes en hemisferios izquierdo y derecho

Con respecto al análisis realizado sobre las frecuencias cerebrales presentes en las características más relevantes, predominan alfa y beta, seguido de un menor número de ondas theta y sin presencia de delta. Alfa está relacionada con la alerta tranquila y beta está presente en el ritmo habitual de vigilia, asociado con el pensamiento activo, la atención activa y el enfoque. Theta se asocia a la transición entre la vigilia y la somnolencia y en estados de relajación profunda [1].

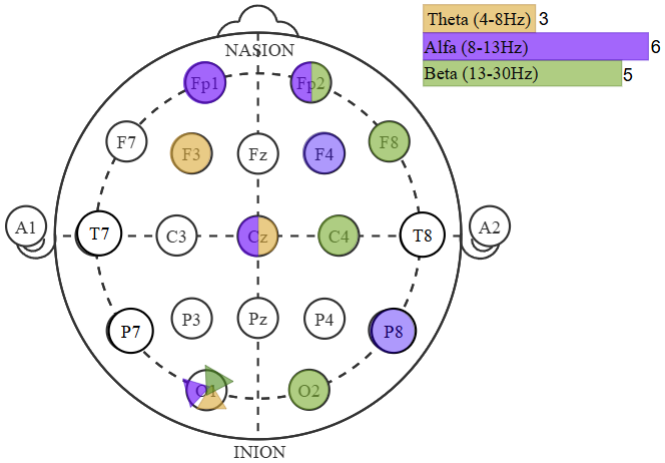

Fig. 4.13: Distribución de características más relevantes en bandas cerebrales

4.3 Resultados Support Vector Machine

Los resultados obtenidos muestran un desempeño variable según la validación empleada. En la Tabla 4.7, se presentan las métricas de evaluación para diferentes enfoques de validación cruzada y el método Hold-out. Este último alcanzó el mejor desempeño general, con un accuracy de 87.50 %. Al utilizar validaciones cruzadas, como Leave-One-Out, el modelo mostró el accuracy más bajo de todos los resultados (73.33 %) y para validación cruzada Kfold, el desempeño disminuyó a medida que se incrementó el número de folds. Esto puede deberse a que el tamaño de los conjuntos de entrenamiento en esquemas con más divisiones, dificulta el aprendizaje del modelo. Las matrices de confusión mostradas en la Figura 4.14 dejan en evidencia la disminución del desempeño, pues al aumentar los folds, aumentan los FP y disminuyen los VP.

Tabla 4.7: Resultados de validación para SVM

Train/Val	Accuracy (%)	Presicion (%)	Recall (%)	F1 (%)
Hold-out	87.50	85.71	92.31	88.89
LOOCV	73.33	71.88	76.67	74.19
5-Folds	81.67	80.91	83.33	81.97
10-Folds	80.83	85.02	78.33	79.22
15-Folds	75.83	78.03	78.33	75.49

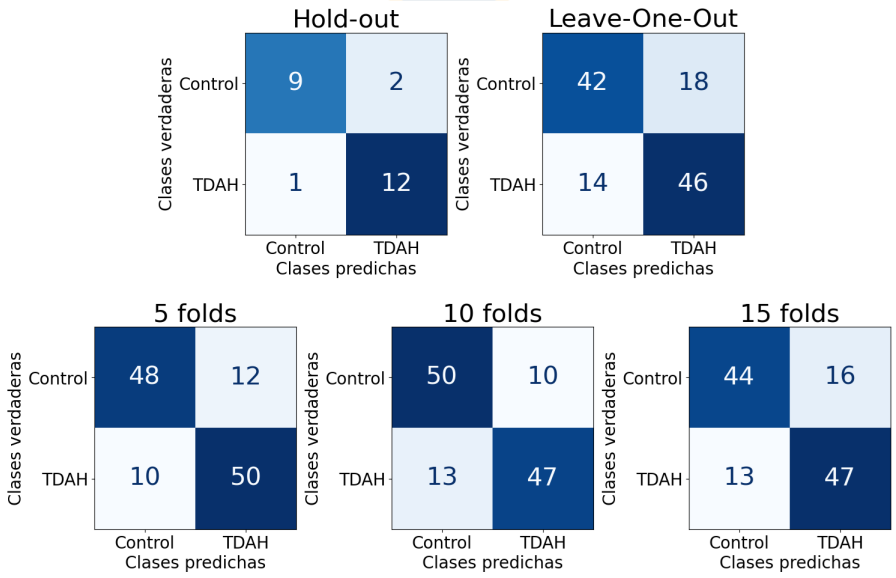

Fig. 4.14: Resultados Matrices de Confusión - SVM

4.4 Resultados Regresión Logística

Los resultados obtenidos muestran un buen desempeño general, destacándose en Hold-out donde alcanzó un 91.67 % de accuracy y luego presentó pequeñas variaciones entre las diferentes validaciones cruzadas. En la Tabla 4.8 se puede ver el detalle de las métricas. En Leave-One-Out, el desempeño fue algo más modesto con un accuracy de 84.17 %, mientras que en el resto de validaciones cruzadas obtuvo 85.00 %, 85.83 % y 84.17 % de accuracy para 5, 10 y 15 folds respectivamente. Esto último indica que no varió tanto con respecto a SVM, por lo que Regresión Logística podría ser menos sensible a grupos más pequeños de entrenamiento. Las matrices de confusión acumuladas para 5, 10 y 15 folds (Figura 4.15) muestran un comportamiento consistente en cuanto a la clasificación de la clase positiva y negativa, con una ligera variabilidad en la cantidad de falsos positivos y falsos negativos a medida que se aumenta el número de folds. También presentó buenos resultados en términos de precisión, recall y F1, convirtiéndose en el modelo más eficaz para esta clasificación.

Tabla 4.8: Resultados de validación para Regresión Logística

Train/Val	Accuracy (%)	Precisión (%)	Recall (%)	F1 (%)
Hold-out	91.67	92.31	92.31	92.31
LOOCV	84.17	84.75	83.33	84.03
5-Folds	85.00	87.26	83.33	84.26
10-Folds	85.83	89.29	85.00	85.34
15-Folds	84.17	86.89	83.33	83.38

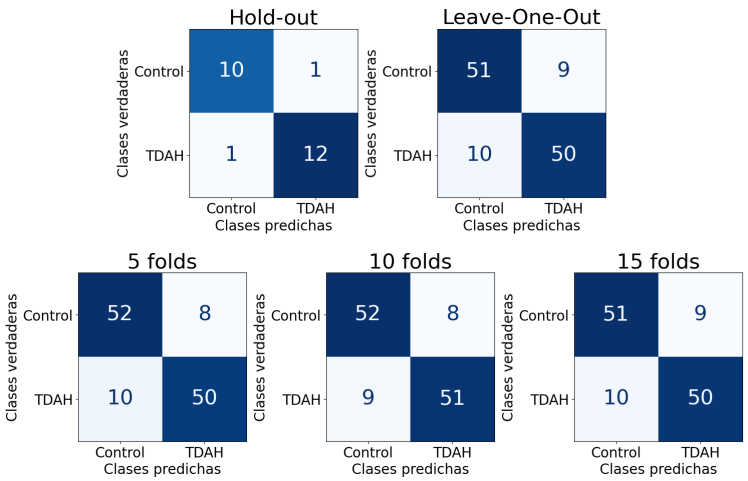

Fig. 4.15: Resultados Matrices de Confusión - Regresión Logística

4.5 Resultados Random Forest

Random Forest obtuvo un desempeño regular en comparación a SVM y RL. En la Tabla 4.9 se puede ver que el mejor desempeño fue en 10 folds, alcanzando un 80.83 % de accuracy. En términos de precisión, recall y F1, los resultados también fueron más consistentes en este caso, lo que sugiere que el modelo se benefició de un balance adecuado entre los datos de entrenamiento y validación. Con respecto a Leave-One-Out, el modelo alcanzó un accuracy de 78.33 %, mismo valor que 5 folds. Esto podría indicar que el clasificador no es particularmente sensible a cambios significativos en el tamaño del conjunto de entrenamiento, manteniendo resultados similares entre las distintas validaciones sin destacar por sobre los clasificadores mencionados anteriormente.

Tabla 4.9: Resultados de validación para Random Forest

Train/Val	Accuracy (%)	Precisión (%)	Recall (%)	F1 (%)
Hold-out	79.17	75.00	92.31	82.76
LOOCV	78.33	76.56	81.67	79.03
5-Folds	78.33	76.89	81.67	78.89
10-Folds	80.83	80.67	83.33	80.68
15-Folds	76.67	76.89	78.33	76.88

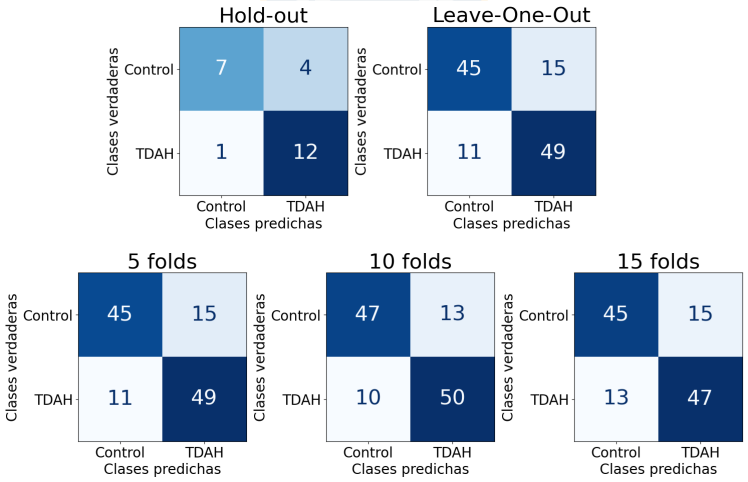

Fig. 4.16: Resultados Matrices de Confusión- Random Forest

4.6 Resultados Naive Bayes

Este modelo presenta un desempeño inferior en comparación con los clasificadores anteriores. En la Tabla 4.10 se observa que el mejor resultado se obtuvo con Hold-out, alcanzando un accuracy de 75.00 %, mientras que en Leave-One-Out y K-fold el accuracy se mantuvo constante alrededor del 70.00 %. Esto indica que el modelo no varía significativamente al variar los tamaños de los conjuntos de entrenamiento y validación. Las matrices de confusión (Figura 4.17) reflejan una tendencia recurrente a errores de clasificación, dando como resultado un recall entre 80-84.62 % y una precisión más baja (66.22-73.33 %) debido a la elevada cantidad de falsos positivos. El valor de F1, entre 71.81 % y 78.57 %, muestra un balance moderado entre precisión y recall, pero inferior al alcanzado por SVM o RL.

Tabla 4.10: Resultados de validación para Naive Bayes

Train/Val	Accuracy (%)	Precisión (%)	Recall (%)	F1 (%)
Hold-out	75.00	73.33	84.62	78.57
LOOCV	70.00	66.22	81.67	73.13
5-Folds	70.00	67.05	81.67	73.08
10-Folds	70.00	67.26	81.67	72.81
15-Folds	69.17	68.03	80.00	71.81

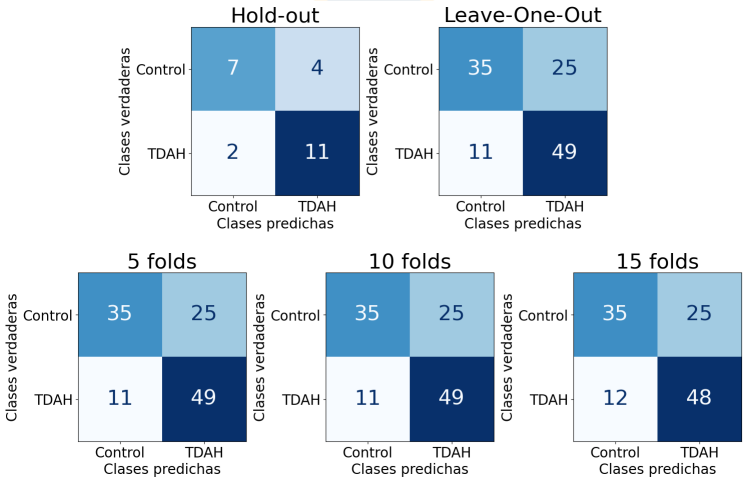

Fig. 4.17: Resultados Matrices de Confusión - Naive Bayes

Capitulo 5. Discusión y Conclusiones

5.1 Discusión de los resultados

En los resultados se obtuvo que las características más relevantes se concentraron en las zonas frontal y parietal del hemisferio derecho. Esta distribución se vincula a la activación de estas áreas durante las tareas de atención realizadas, como también a una posible lateralización funcional en el procesamiento cerebral. Además, se observó mayor presencia de ondas alfa y beta en las características más relevantes, lo que refleja un perfil coherente con actividades cognitivas asociadas al ritmo habitual de vigilia y la atención activa. Comparado con los patrones observados en señales adquiridas en reposo de niños con TDAH [30], donde predomina una mayor presencia de ondas theta y una disminución de las ondas beta, los resultados de esta investigación mostraron una activación cerebral diferente pero coherente como se mencionó anteriormente. Aunque los estudios en reposo ofrecen ventajas importantes, como la reducción de artefactos y una mayor facilidad en la recolección de datos, es fundamental considerar sus limitaciones. El análisis de señales en reposo no refleja completamente los patrones de activación asociados con tareas cognitivas específicas. Estudiar e idenficar perfiles típicos y atípicos de activación cerebral en sujetos con TDAH durante tareas de atención, creará una base de conocimiento para una mejor comprensión de los mecanismos cerebrales.

En relación a las métricas de desempeño de los modelos de clasificación, el modelo de Regresión Logística demostró ser el más eficaz, alcanzando un accuracy de 85.83 % en la validación cruzada de 10 folds. Aunque el modelo obtuvo un 91.67 % de accuracy en la validación hold-out, este resultado es menos representativo del desempeño general del modelo, ya que depende de una única división de los datos (80-20), lo que no garantiza su capacidad para generalizar a nuevas subdivisiones. Las validaciones cruzadas, al dividir los datos en múltiples subconjuntos, proporcionan un análisis más robusto y representativo de la capacidad de generalización del modelo, lo que hace que los resultados de la validación cruzada sean más confiables y aplicables en escenarios reales. La variabilidad observada en el rendimiento de SVM, indica que se deben explorar enfoques adicionales para optimizar su desempeño.

Para mejorar el resultado general de las métricas, será necesario ajustar la metodología desde la selección de características relevantes, considerando el análisis estadístico realizado en relación a estas y cómo se distribuyen por ciertas zonas y bandas cerebrales. También será necesario realizar ajustes a los hiperparámetros de los clasificadores para optimizar su desempeño.

Los resultados de los clasificadores abren la puerta a futuras investigaciones que exploren técnicas adicionales para mejorar la clasificación de señales EEG de sujetos con TDAH. Las métricas desmuestran la importancia de elegir un clasificador que logre un balance entre la capacidad de aprendizaje y la generalización. Regresión Logística se posiciona como la mejor opción en este contexto, dado su desempeño robusto y consistente. SVM es adecuado en escenarios donde los datos estén menos fragmentados, mientras que Random Forest podría ser una alternativa en casos donde se priorice la estabilidad. Naive Bayes, resultó menos competitivo en esta tarea de clasificación binaria.

Tabla 5.11: Resumen de métricas en base al mejor desempeño en 10 folds

Clasificador	**Accuracy (%)**	**Precisión (%)**	**Recall (%)**	**F1 (%)**
Regresión Logística	85.83	89.29	85.00	85.34
SVM	80.83	85.02	78.33	79.22
Random Forest	80.83	80.67	83.33	80.68
Naive Bayes	70.00	67.26	81.67	72.81

5.2 Conclusiones y trabajo futuro

Los resultados de la metodología aplicada demuestran que se implementó con éxito un algoritmo de clasificación para diferenciar sujetos con TDAH de un grupo control, utilizando señales de EEG y herramientas de aprendizaje automático. Se trabajó con una base de datos previamente procesada con señales EEG de ambos grupos, lo cual permitió identificar patrones relevantes en bandas de frecuencia y zonas frontal y parietal del hemisferio derecho. A través de este análisis, se observó una mayor presencia de ondas alfa y beta en las características relevantes, lo cual es coherente con procesos de atención activa y pensamiento enfocado, en contraste con los patrones predominantes en estudios en reposo, donde se observa un incremento de theta y una disminución de beta en sujetos con TDAH.

Para clasificar estas señales, se implementaron y evaluaron diferentes algoritmos de aprendizaje automático. Regresión Logística demostró ser el modelo más robusto, alcanzando un accuracy promedio del 85.83 %. Los resultados develan la importancia de las validaciones cruzadas para garantizar un análisis confiable y minimizar el riesgo de sobreajuste en los modelos.

La metodología de esta investigación resultó ser efectiva y podría ser la base para futuras investigaciones. Sería valioso explorar técnicas avanzadas de selección y extracción de características,

como aquellas basadas en redes neuronales profundas, que permitan identificar representaciones más complejas y potencialmente más discriminativas, como también considerar bases de datos más amplias. El estudio demuestra el potencial de investigar con señales biomédicas en conjunto con el aprendizaje automático como una herramienta prometedora para abordar desafíos clínicos complejos, abriendo el camino hacia soluciones más precisas y accesibles, contribuyendo significativamente al avance en la evaluación y tratamiento del TDAH.

Capitulo 6. Glosario

DSM-5	Manual Diagnóstico y Estadístico de Trastornos Mentales , en inglés, Diagnostic and Statistical Manual of Mental Disorders
ECG	Electrocardiografía
EEG	Electroencefalografía
EMG	Electromiografía
FT	Fourier Transform , en inglés, Transformada de Fourier
IC	Componente Independiente , en inglés, Independent Component
ICA	Análisis de Componentes Independientes , en inglés, Independent Component Analysis
ML	Aprendizaje Automático , en inglés, Machine Learning
PSD	Densidad Espectral de Potencia , en inglés, Power Spectral Density
SNC	Sistema Nervioso Central
TDAH	Trastorno de Déficit de Atención e Hiperactividad

Capitulo 7. Referencias

[1] L. Hu and Z. Zhang, *EEG Signal Processing and Feature Extraction*, first edition ed. Springer Singapore, 2019. [Online]. Available: https://doi.org/10.1007/978-981-13-9113-2

[2] S. Sanei and J. A. Chambers, *EEG Signal Processing and Machine Learning*, second edition ed. John Wiley & Sons, Ltd, 2021. [Online]. Available: https://www.wiley.com/en-jp/EEG+Signal+Processing+and+Machine+Learning%2C+2nd+Edition-p-9781119386957

[3] M. E. A. Perona and P. F. Diez, "Bioinstrumentación ii: Electroencefalografía," *Facultad de Ingeniería, Universidad Nacional de San Juan, Argentina*, 2006. [Online]. Available: http://dea.unsj.edu.ar/bioinstrumentacion2/apunteseeg.pdf

[4] N. A. MD. MANIRUZZAMAN, MD. AL MEHEDI HASAN and J. SHIN, "Optimal channels and features selection based adhd detection from eeg signal using statistical and machine learning techniques," *IEEE*, vol. 11, pp. 33 570 – 33 583, 04 2023. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3264266

[5] M. F. F. Ali Khaleghi, Pari Moradi Birgani and M. R. Mohammadi, "Applicable features of electroencephalogram for adhd diagnosis," *Research on Biomedical Engineering*, vol. 36, pp. 1–11, 01 2020. [Online]. Available: https://doi.org/10.1007/s42600-019-00036-9

[6] M. A. M. H. Md. Maniruzzaman1, Jungpil Shin1 and A. Yasumura, "Efficient feature selection and machine learning based adhd detection using eeg signal," *Computers, Materials and Continua*, vol. 72, pp. 5179–5195, 03 2022. [Online]. Available: https://doi.org/10.32604/cmc.2022.028339

[7] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5-TR)*, fifth edition ed. American Psychiatric Publishing, 2022. [Online]. Available: https://www.psychiatry.org/Psychiatrists/Practice/DSM

[8] O. Attallah, "Adhd-aid: Aiding tool for detecting children's attention deficit hyperactivity disorder via eeg-based multi-resolution analysis and feature selection," *Biomimetics*, vol. 9, 03 2024. [Online]. Available: https://doi.org/10.3390/biomimetics9030188

[9] A. Alim and M. H. Imtiaz, "Automatic identification of children with adhd from eeg brain waves," *MDPI*, vol. 13, pp. 193–205, 04 2023. [Online]. Available: http://doi.org/10.3390/signals4010010

[10] M. M. B. B. M.G. Sumithra, Rajesh Kumar Dhanaraj and C. Venkatesan, *Brain-Computer Interface: Using Deep Learning Applications*. John Wiley & Sons, Inc. and Scrivener Publishing LLC, 2023. [Online]. Available: https://doi.org/10.1002/9781119857655

[11] MathWorks, *Mastering Machine Learning A Step-by-Step Guide with MATLAB*. MathWorks, 2019, accessed: 2024-11-27. [Online]. Available: https://la.mathworks.com/campaigns/offers/mastering-machine-learning-with-matlab.html

[12] B. K. S. T. A. Özçelik Hakan Uyanık Hüseyin Üzen Abdülkadir Şengür, "Advanced eeg-based analysis for adhd identification utilizing convmixer and continuous wavelet transform," *Turkish Journal of Nature and Science*, vol. 13, pp. 19–25, 03 2024. [Online]. Available: https://doi.org/10.46810/tdfd.1388893

[13] G. C. V. A. G.-Y. Valery A. Ponomarev, Andreas Mueller and J. D. Kropotov, "Group independent component analysis (gica) and current source density (csd) in the study of eeg in adhd adults," *Clinical Neurophysiology*, vol. 125, pp. 83–97, 01 2014. [Online]. Available: https://doi.org/10.1016/j.clinph.2013.06.015

[14] A. M. N. Mehdi Samavati and M. R. Mohammadi, "Automatic minimization of eye blink artifacts using fractal dimension of independent components of multichannel eeg," in *20th Iranian Conference on Electrical Engineering*, 2012.

[15] V. Harpale and V. Bairagi, "An adaptive method for feature selection and extraction for classification of epileptic eeg signal in significant states," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, pp. 668–676, 07 2021. [Online]. Available: https://doi.org/10.1016/j.jksuci.2018.04.014

[16] A. T. Swadi and F. S. Miften, "Adhd detection using machine learning algorithms and eeg brain signals," *Journal of Education for Pure Science- University of Thi-Qar*, vol. 13, 03 2023. [Online]. Available: https://jceps.utq.edu.iq/index.php/main/article/view/237

[17] W. Zgallai, *Biomedical Signal Processing and Artificial Intelligence in Healthcare*. Elsevier, 2020. [Online]. Available: https://doi.org/10.1016/C2018-0-04775-1

[18] M. R. M. Armin Allahverdy, Alireza Khorrami and A. M. Nasrabadi, "Detecting adhd children using the attention continuity as nonlinear feature of eeg," *Frontiers in Biomedical Technologies*, vol. 3, pp. 1–2, 06 2016. [Online]. Available: https://www.researchgate.net/publication/327261963_Detecting_ADHD_Children_using_the_Attention_Continuity_as_Nonlinear_Feature_of_EEG

[19] N. A. T. Pham Thi Viet Huong and T. A. Vu, "Ensemble learning in detecting adhd children by utilizing the non-linear features of eeg signal," in *2nd International Conference on Human-*

centered Artificial Intelligence (Computing4Human 2021). CEUR Workshop Proceedings, October 28 2021. [Online]. Available: https://ceur-ws.org/Vol-3026/paper14.pdf

[20] DataCamp, "Tutorial: Lasso and ridge regression," last updated 2024 May 03. [Online]. Available: https://www.datacamp.com/es/tutorial/tutorial-lasso-ridge-regression

[21] O. Peretz, M. Koren, and O. Koren, "Naive bayes classifier – an ensemble procedure for recall and precision enrichment," *Engineering Applications of Artificial Intelligence*, vol. 136, 2024. [Online]. Available: https://doi.org/10.1016/j.engappai.2024.108972

[22] D. Thakur and S. Biswas, "An integration of feature extraction and guided regularized random forest feature selection for smartphone based human activity recognition," *Journal of Network and Computer Applications*, vol. 204, 2022. [Online]. Available: https://doi.org/10.1016/j.jnca.2022.103417

[23] D. Jurafsky and J. H. Martin, "Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition with language models, 3rd edition," *Stanford University and University of Colorado at Boulder*, 2024, online manuscript released August 20, 2024. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3

[24] J. Allibhai, "Holdout vs cross-validation in machine learning," 2019. [Online]. Available: https://medium.com/@jaz1/holdout-vs-cross-validation-in-machine-learning-7637112d3f8f

[25] A. Motie Nasrabadi, A. Allahverdy, M. Samavati, and M. R. Mohammadi, "Eeg data for adhd / control children," 2020. [Online]. Available: https://dx.doi.org/10.21227/rzfh-zn36

[26] D. A. . M. S, "Eeglab: an open-source toolbox for analysis of single-trial eeg dynamics, journal of neuroscience methods," 2004. [Online]. Available: https://eeglab.org/

[27] A. R. Clarke, R. J. Barry, F. E. Dupuy, R. McCarthy, M. Selikowitz, and S. J. Johnstone, "Excess beta activity in the eeg of children with attention-deficit/hyperactivity disorder: A disorder of arousal?" *International Journal of Psychophysiology*, vol. 89, no. 3, pp. 314–319, 2013. [Online]. Available: https://doi.org/10.1016/j.ijpsycho.2013.04.009

[28] K. Javed, V. Reddy, and F. Lui, "Neuroanatomy, cerebral cortex," *StatPearls [Internet]*, last updated 2023 Jul 25. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK537247/

[29] B. E. Morton, "Brain executive laterality and hemisity," *Personality Neuroscience*, vol. 3, 07 2020. [Online]. Available: https://doi.org/10.1017/pen.2020.6

[30] A. R. Clarke, R. J. Barry, R. McCarthy, M. Selikowitz, S. J. Johnstone, C.-I. Hsu, C. A. Magee, C. A. Lawrence, and R. J. Croft, "Coherence in children with attention-deficit/hyperactivity

disorder and excess beta activity in their eeg," *Clinical Neurophysiology*, vol. 118, no. 7, pp. 1472–1479, 2007. [Online]. Available: https://doi.org/10.1016/j.clinph.2007.04.006

Anexo A. Etiquetas de canales

Número	Canal
1	Fp1
2	Fp2
3	F3
4	F4
5	C3
6	C4
7	P3
8	P4
9	O1
10	O2
11	F7
12	F8
13	T7
14	T8
15	P7
16	P8
17	Fz
18	Cz
19	Pz

Anexo B. Coeficientes de características más relevantes

Característica	Coeficiente
katz_Fz	0,2147
area_positiva_F4	0,1914
katz_P7	0,1587
alfa_power_F4	0,1043
area_positiva_P8	0,0985
IQR_Cz	0,0972
petrosian_O2	0,0965
variance_P4	0,0921
theta_power_Cz	0,0914
higuchi_Fp1	0,0847
complexity_F8	0,0790
mean_P7	0,0779
petrosian_P8	0,0710
mean_F3	0,0700
amplitud_absoluta_Pz	0,0695
alfa_power_Fp1	0,0674
IQR_P7	0,0667
mean_P4	0,0663
mean_P8	0,0651
katz_P3	0,0573
petrosian_F4	0,0542
amplitud_absoluta_F7	0,0529
activity_P4	0,0517
mobility_Fz	0,0506
higuchi_T8	0,0502
median_Fp1	0,0497
kurtosis_Cz	0,0490
alfa_power_Cz	0,0475
petrosian_P3	0,0472
IQR_T7	0,0469
kurtosis_Fp2	0,0463
higuchi_P7	0,0455
katz_C3	0,0454
skewness_T8	0,0452
kurtosis_T8	0,0390

Característica	Coeficiente
skewness_Fp2	0,0371
median_Pz	0,0345
kurtosis_F8	0,0341
katz_Pz	0,0340
higuchi_F3	0,0308
median_P4	0,0285
alfa_power_P8	0,0284
alfa_power_Fp2	0,0279
median_C4	0,0261
IQR_Pz	0,0247
beta_power_Fp2	0,0226
median_Cz	0,0217
median_F3	0,0194
beta_power_O1	0,0190
amplitud_absoluta_C3	0,0183
amplitud_absoluta_Fp2	0,0176
petrosian_F8	0,0175
kurtosis_P7	0,0149
beta_power_C4	0,0148
higuchi_Cz	0,0135
beta_power_F8	0,0123
skewness_O2	0,0115
skewness_F7	0,0109
beta_power_O2	0,0087
area_positiva_C4	0,0087
theta_power_F3	0,0080
skewness_F4	0,0075
complexity_T8	0,0046
theta_power_O1	0,0045
skewness_Fz	0,0042
alfa_power_O1	0,0036
mobility_P8	0,0027
amplitud_absoluta_T7	0,0020
mean_Cz	8,10E-05
mean_Pz	7,26E-05

Anexo C. Hiperparámetros clasificadores

Clasificador	**Hiperparámetro**
SVM	`kernel: lineal`
Random Forest	`n_trees=500`, `class_weight=balanced`
Regresión Logística	`max_iter=1000`, `class_weight=balanced`
Naive Bayes	`none`

UNIVERSIDAD DE CONCEPCION – FACULTAD DE INGENIERIA
RESUMEN DE MEMORIA DE TITULO

Departamento:	Departamento de Ingeniería Eléctrica
Carrera:	Ingeniería Civil Biomédica
Nombre del memorista:	Aura Toledo Araneda
Título de la memoria:	Detección del trastorno de déficit de atención e hiperactividad utilizando algoritmos de clasificación de machine learning basado en señales de electroencefalografía
Fecha de la presentación oral:	28 de enero de 2025
Profesor(es) Guía:	Esteban Pino
Profesor(es) Revisor(es):	Pamela Guevara y Geronimo Bolado
Concepto:	
Calificación:	

Resumen

El TDAH es uno de los trastornos del neurodesarrollo más prevalentes a nivel global, afectando a millones de niños y adolescentes. Este proyecto desarrolló una metodología basada en aprendizaje automático para clasificar señales de EEG de participantes con TDAH, utilizando una base de datos con 121 sujetos (61 con TDAH y 60 de control). El proceso incluyó el filtrado y la eliminación de artefactos de las señales, seguido de la extracción de características en el dominio del tiempo y en el dominio de la frecuencia. Luego, se aplicó LASSO para la selección de características más relevantes, lo que redujo la dimensionalidad de los datos y mejoró el rendimiento de los modelos.

El análisis estadístico mostró que las características más relevantes se concentraban en las regiones frontal y parietal del hemisferio derecho, con predominio de ritmos alfa y beta. Se evaluaron cuatro clasificadores: SVM, Regresión Logística, Random Forest y Naive Bayes, mostrando sus resultados en términos de accuracy, precisión, recall y F1-score. Regresión Logística obtuvo el mejor desempeño, alcanzando un 85.83 % de precisión en validación cruzada con 10 folds. La metodología demostró ser eficaz para clasificar señales de EEG y tiene potencial para ser aplicada en futuros estudios sobre el TDAH.