Universidad Politécnica
de Madrid

**Escuela Técnica Superior de
Ingenieros Informáticos**

Grado en Ciencia de Datos e Inteligencia Artificial

Trabajo Fin de Grado

Análisis Exploratorio de la Confianza Humana basado en Señales EEG mediante Técnicas de Aprendizaje Automático e Interpretabilidad

Autor: Lucía Rebolledo Romillo
Tutora: Laura Melgar García
Cotutor: Javier Bajo Pérez

Madrid, Enero 2026

Este Trabajo Fin de Grado se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Grado
Grado en Ciencia de Datos e Inteligencia Artificial
Título: Análisis exploratorio de la confianza humana basado en señales EEG mediante técnicas de aprendizaje automático e interpretabilidad.
Enero, 2026

Autor: Lucía Rebolledo Romillo

Tutor: Laura Melgar García
Cotutor: Javier Bajo Pérez
Departamento de Inteligencia Artificial
ETSI Informáticos
Universidad Politécnica de Madrid

Agradecimientos

Tras 4 años de carrera que parecía que iban a ser muchos más, esto llega a su fin.

Lo primero de todo, quiero agradecer a Laura y Javier por darme la oportunidad de hacer este trabajo con ellos. Gracias por estar pendientes a cualquier hora para resolver dudas, por enseñarme tanto a nivel profesional y por confiar en mí desde el principio.

Quiero agradecer a mis padres por siempre creer en mí, y sobre todo cuando yo no lo hacía. Por darme ánimo en cada examen, saliera bien o saliera mal, y por confiar en mí para sacar esta carrera adelante incluso cuando alguna asignatura se resistía.

A mi hermana, por ser mi referente cada día, por animarme a poner en mis estudios y en mi futuro laboral la misma pasión que ella tiene, y por empujarme a no abandonar y a luchar siempre por lo que quiero.

Agradezco a mis mejores amigas, las BLIMIS, por ser mi segunda familia, por cuidarme siempre. Gracias por darme fuerzas durante toda la carrera y por escucharme hablar de estas tecnologías tan nuevas sin entender nada.

A mis amigos del colegio, por no fallar nunca en un día de biblioteca (FU), por animarnos en los peores momentos y por las risas en épocas de exámenes, cuando todo parecía tan negro.

Quiero agradecer al Erasmus por todo lo que me hizo crecer y a los amigos que hice por enseñarme a disfrutar cada momento.

Agradezco también a mis amigos de la carrera, por apoyarnos mutuamente durante estos años, por los viajes, las risas, los agobios a las mil de la noche y por hacer este camino mucho más llevadero.

Quiero agradecerle a Dios, por todas las personas que ha puesto en mi camino durante estos cuatro años, por la familia que me ha dado y por cuidarme y sostenerme.

Y, por último, quiero agradecerme a mí misma: por el esfuerzo, la dedicación, las horas de trabajo, las lágrimas y también la felicidad tras cada entrega o cada aprobado; por no rendirme nunca, incluso cuando parecía imposible, y por sacar esta carrera adelante.

Resumen

El estudio de la confianza (trust) a partir de señales de electroencefalograma (EEG) ha despertado interés en los últimos años debido a su posible aplicación en ámbitos como la interacción humano-máquina, la neurociencia cognitiva y los sistemas adaptativos. Sin embargo, el análisis de la confianza mediante EEG presenta dificultades relevantes, ya que las señales son ruidosas y varían de forma considerable entre sujetos, y la evaluación del trust se basa en la percepción subjetiva de cada usuario, lo que complica su modelado mediante técnicas de aprendizaje automático.

El objetivo principal de este Trabajo Fin de Grado es explorar la relación entre las señales EEG y los estados de confianza, analizando la viabilidad de identificar patrones asociados al trust mediante técnicas de análisis de datos y aprendizaje automático. El enfoque del trabajo es exploratorio y se centra en estudiar el comportamiento de distintas representaciones de la variable trust y su relación con las características extraídas del EEG, sin buscar como objetivo el desarrollo de un sistema predictivo final.

El estudio se basa en una base de datos EEG adquirida por el grupo NeuroTechAI de la Universidad Politécnica de Madrid, que incluye tanto señales en crudo como señales procesadas, registradas en dos contextos experimentales: un experimento controlado de carácter neutral y un experimento de tipo ecológico. Las señales en crudo se emplean para extraer características en los dominios frecuencial y temporal. Posteriormente, los canales EEG se agrupan para construir regiones cerebrales —frontal, central, parietal, temporal y occipital— que se utilizan en los análisis posteriores con el objetivo de estructurar los datos y facilitar su interpretación.

Para ello, se emplea una base de datos EEG en la que las etiquetas de referencia se obtienen a partir de la autoevaluación del usuario sobre su nivel de confianza. Dichas etiquetas se analizan bajo tres codificaciones diferentes: una escala completa (5 niveles), una codificación binaria (baja y alta confianza) y una codificación terciaria (baja, media y alta confianza), con el fin de estudiar el impacto de la representación de la variable trust en el rendimiento de los modelos.

En una primera fase se realizó un análisis exploratorio de los datos, que incluyó análisis estadístico, temporal y topográfico de las características EEG. Este análisis permitió estudiar la organización de los datos en relación con los distintos niveles de confianza y evaluar la correspondencia entre las características extraídas y las etiquetas de trust. Los resultados muestran una

variabilidad inter-sujeto y una correspondencia limitada entre los niveles de confianza.

Posteriormente, se aborda el aprendizaje automático supervisado, evaluando distintos clasificadores entrenados a partir de las características EEG y las etiquetas de trust. El entrenamiento de los modelos se realiza mediante ajuste de hiperparámetros, y la evaluación se lleva a cabo utilizando métricas como accuracy, balanced accuracy y F1- score. Los resultados alcanzados muestran un rendimiento moderado, lo que confirma la dificultad de clasificar estados de confianza a partir de señales EEG.

Con el fin de mejorar la interpretación de los modelos supervisados, se incorporan técnicas de explicabilidad basadas en SHAP (SHapley Additive exPlanations). Este análisis permite identificar las características que más contribuyen a las decisiones de los modelos y analizar la relevancia de determinadas bandas de frecuencia y regiones cerebrales en la estimación del trust. Estas técnicas aportan una visión complementaria al análisis de rendimiento y facilitan la comprensión del comportamiento de los modelos.

Como complemento al análisis principal, se realiza una experimentación exploratoria con datos EEG propios, cuyo objetivo es validar el procedimiento de adquisición y registro de señales. Esta experimentación no se integra en el análisis principal del trabajo y se emplea únicamente para comprobar la viabilidad del protocolo experimental.

En conjunto, este trabajo muestra las posibilidades y limitaciones del uso de señales EEG para el análisis de la confianza mediante aprendizaje automático. Los resultados indican que la estimación del trust a partir de EEG constituye una tarea compleja, condicionada por la elevada variabilidad de las señales y la limitada separabilidad entre clases, sin observarse una diferenciación clara entre los distintos niveles de confianza. No obstante, los análisis realizados sugieren una mayor contribución de la banda low gamma y de la actividad en regiones frontales, temporales y occipitales, aunque estas tendencias no permiten definir patrones robustos y generalizables. Estos resultados ponen de manifiesto la necesidad de continuar investigando con bases de datos más amplias y enfoques que tengan en cuenta la variabilidad individual.

Abstract

The study of trust based on electroencephalogram (EEG) signals has attracted interest in recent years due to its potential application in areas such as human-machine interaction, cognitive neuroscience, and adaptive systems. However, analysing trust using EEG presents significant difficulties, as the signals are noisy and vary considerably between subjects, and the assessment of trust is based on the subjective perception of each user, which complicates its modelling using machine learning techniques.

The main objective of this Final Degree Project (Trabajo Fin de Grado) is to explore the relationship between EEG signals and states of trust, analysing the feasibility of identifying patterns associated with trust using data analysis and machine learning techniques. The approach of the project is exploratory and focuses on studying the behaviour of different representations of the trust variable and its relationship with the characteristics extracted from the EEG, without seeking to develop a final predictive system.

The study is based on an EEG database acquired by the NeuroTechAI group at the Polytechnic University of Madrid, which includes both raw and processed signals recorded in two experimental contexts: a neutral controlled experiment and an ecological experiment. The raw signals are used to extract features in the frequency and time domains. Subsequently, the EEG channels are grouped to construct brain regions—frontal, central, parietal, temporal, and occipital—which are used in subsequent analyses with the aim of structuring the data and facilitating its interpretation.

To do this, an EEG database is used in which reference labels are obtained from the user's self-assessment of their confidence level. These labels are analysed under three different encodings: a full scale, a binary encoding (low and high confidence) and a tertiary encoding (low, medium and high confidence), in order to study the impact of the representation of the trust variable on model performance.

In the first phase, an exploratory analysis of the data was carried out, including statistical, temporal and topographical analysis of the EEG characteristics. This analysis allowed us to study the organisation of the data in relation to the different levels of trust and to evaluate the correspondence between the extracted characteristics and the trust labels. The results show inter-subject variability and limited correspondence between trust levels.

Subsequently, supervised machine learning is addressed, evaluating different classifiers trained based on EEG characteristics and trust labels. The models are trained by adjusting hyperparameters, and evaluation is carried out using

metrics such as accuracy, balanced accuracy, and F1 score. The results achieved show moderate performance, confirming the difficulty of classifying trust states based on EEG signals.

In order to improve the interpretation of supervised models, explanation techniques based on SHAP (SHapley Additive exPlanations) are incorporated. This analysis allows us to identify the characteristics that contribute most to the models' decisions and analyse the relevance of certain frequency bands and brain regions in estimating trust. These techniques provide a complementary view to performance analysis and facilitate understanding of model behaviour.

As a complement to the main analysis, exploratory experimentation is carried out with our own EEG data, with the aim of validating the signal acquisition and recording procedure. This experimentation is not integrated into the main quantitative analysis of the work and is used solely to verify the viability of the experimental protocol.

Overall, this study shows the possibilities and limitations of using EEG signals for trust analysis through machine learning. The results indicate that estimating trust based on EEG is a complex task, conditioned by the high variability of the signals and the limited separability between classes, with no clear differentiation between different levels of trust. However, the analyses suggest a greater contribution from the low gamma band and activity in the frontal, temporal, and occipital regions, although these trends do not allow for the definition of robust and generalisable patterns. These results highlight the need for further research using larger databases and approaches that take individual variability into account.

Tabla de contenidos

1 Introducción

1.1 Contexto del estudio

La confianza entre humanos y sistemas inteligentes se ha convertido en un elemento principal en el desarrollo de tecnologías automatizadas. Se encuentra en sectores como la conducción asistida, la robótica colaborativa o la inteligencia artificial aplicada a la toma de decisiones. Destaca la capacidad de un usuario para confiar adecuadamente en el sistema que determina tanto su aceptación como su eficacia. Se debe ajustar la confianza para así evitar la sobredependencia o al rechazo de su uso, afectando así al rendimiento y la seguridad de la interacción [1], [2].

Tradicionalmente, la confianza se ha estudiado desde la psicología social, donde se define como la expectativa de fiabilidad hacia los demás[3]. Con la aparición de la automatización, este concepto se trasladó al ámbito tecnológico como un mecanismo adaptativo que regula la relación entre el control humano y la autonomía del sistema [1] [4]. Sin embargo, la confianza en los sistemas inteligentes está evolucionando a lo largo del tiempo según la experiencia, los fallos o el grado de transparencia del sistema.

En los últimos años, esta nueva visión ha impulsado el desarrollo de modelos computacionales capaces de estimar y predecir la confianza en tiempo real. En particular, se han aplicado técnicas de aprendizaje supervisado, que permiten clasificar los estados de confianza a partir de señales fisiológicas como el EEG o la respuesta galvánica de la piel [5].

Asimismo, la incorporación de herramientas de explicabilidad de modelos (Explainable Artificial Intelligence, XAI) como SHAP o LIME ha permitido comprender qué variables o regiones cerebrales contribuyen más a las predicciones, favoreciendo la transparencia y la interpretación neurofisiológica de los resultados [6].

1.2 Objetivos

El presente trabajo aborda de forma exploratoria el análisis de la actividad cerebral asociada a distintos niveles de confianza, realizando un análisis estadístico, temporal y topográfico de las señales EEG, y empleando enfoques de aprendizaje automático supervisado y no supervisado, junto con técnicas de explicabilidad.

El objetivo general es explorar la existencia de patrones neuronales relevantes que permitan determinar la confianza, con el fin de aportar una visión más profunda sobre cómo esta se refleja en la actividad cerebral durante la interacción humano–sistema.

De forma más específica, se plantean los siguientes objetivos.

1. **Analizar las señales EEG** y extraer sus características más relevantes en las diferentes sesiones (neutral y ecológica).
2. **Explorar la variable Trust** en sus versiones completa, binaria y terciaria, en ambos conjuntos de datos.
3. **Realizar análisis exploratorios: estadísticos, temporales y topográficos**, comparando resultados entre regiones cerebrales (frontal, temporal, central, parietal y occipital).
4. **Aplicar técnicas de aprendizaje no supervisado** (K-Means, DBSCAN, Agglomerative Clustering, GMM …) para detectar patrones cerebrales asociados a distintos niveles de confianza.
5. **Aplicar técnicas de aprendizaje supervisado** (KNN, SVM, Naive Bayes, Random Forest, XGBoost) para crear y evaluar modelos de clasificación de la confianza (Trust).
6. **Evaluar la interpretabilidad de los resultados** mediante técnicas de explicabilidad (SHAP) para identificar las variables más influyentes en la predicción de confianza.
7. **Documentar el proceso completo**, los resultados y las conclusiones, contribuyendo al desarrollo de modelos de Human-System Trust basados en inteligencia artificial.

2 Estado del Arte

2.1 HST (Human System Trust)

El estudio de la confianza en sistemas automatizados surge de la necesidad de comprender cómo las personas interactúan con tecnologías cada vez más autónomas. En contextos donde el ser humano delega parte del control en un sistema como la aviación, la conducción automatizada o la robótica colaborativa, la confianza determina el grado de aceptación, supervisión y dependencia hacia la máquina [4].

Desde la psicología, Rotter (1967) definió la confianza como una expectativa generalizada sobre la fiabilidad de los demás [3], mientras que Mayer, Davis y Schoorman (1995) la describieron como la voluntad de una parte de ser vulnerable ante las acciones de otra, basada en tres dimensiones fundamentales: habilidad, benevolencia e integridad [7]. Estos componentes se trasladan directamente al ámbito tecnológico: los usuarios evalúan la competencia del sistema (habilidad), su alineación con los objetivos humanos (benevolencia) y su coherencia o previsibilidad (integridad).

Con la expansión de la automatización, Lee y See (2004) propusieron el concepto de trust in automation, entendiendo la confianza como un mecanismo adaptativo que regula el equilibrio entre la dependencia y el control del usuario como se observa en la Figura 2.1 [1]. Desde esta perspectiva, la confianza no debe maximizarse sin límite, sino calibrarse: es decir, ajustarse a las capacidades reales del sistema. Una confianza mal calibrada puede derivar en sobreconfianza (uso indebido o complacencia) o subconfianza (rechazo o falta de uso), con efectos negativos en el rendimiento y la seguridad [4].

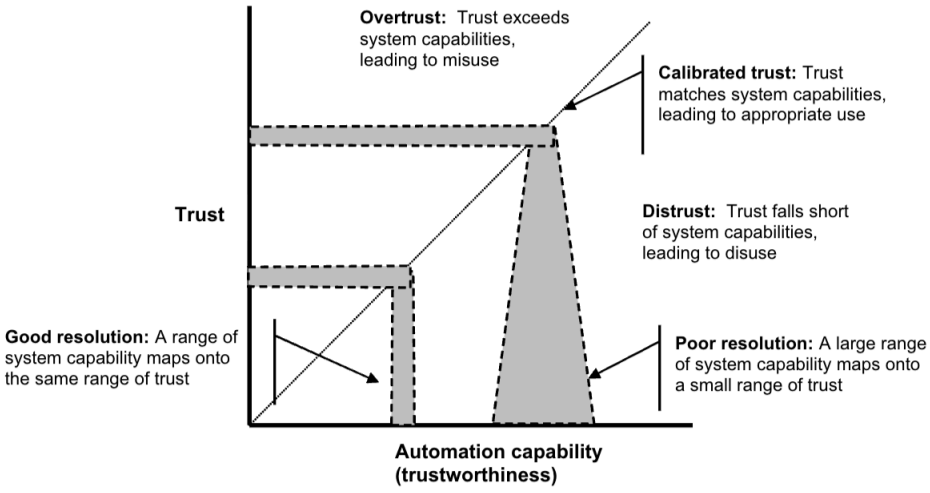

Figura 2.1. Relación conceptual entre la confianza del usuario (Trust) y la capacidad del sistema (Automation capability), ilustrando los estados de confianza calibrada, sobreconfianza (overtrust) y desconfianza (distrust), así como el efecto de la resolución del sistema. Adaptado de Lee y See [1].

Diversos estudios han demostrado que la confianza se construye a partir de la fiabilidad percibida del sistema, pero también está influida por factores humanos y por situaciones. En su metaanálisis, Schaefer et al. (2016) identificaron la fiabilidad, la transparencia, la carga de trabajo y la experiencia previa como los principales determinantes del trust [2]. A medida que los sistemas se vuelven más complejos, la confianza no depende sólo de la precisión técnica, sino de la comprensión que el usuario tiene del funcionamiento interno del sistema.

Otro aspecto importante es la transparencia, entendida como la capacidad del sistema para comunicar de forma clara sus estados, intenciones y limitaciones. La transparencia mejora la calibración de la confianza al permitir que el usuario anticipe el comportamiento del sistema y comprenda sus decisiones. Sin embargo, una transparencia excesiva o mal diseñada puede generar sobrecarga cognitiva o confusión [1].

La confianza en la automatización también es dinámica y evoluciona con la experiencia. Los primeros fallos pueden deteriorarla rápidamente, mientras que la recuperación suele requerir múltiples interacciones exitosas. Este carácter asimétrico, descrito en los estudios de Lee y Moray (1992), muestra que la confianza es más fácil de perder que de recuperar [8]. Además, su desarrollo es individual y contextual: diferentes usuarios presentan ritmos y umbrales distintos de ajuste, lo que ha llevado a explorar mecanismos de trust adaptation en sistemas cognitivos capaces de responder al estado de confianza estimado del usuario [1], [4].

Finalmente, autores recientes como Lyons et al. (2021) destacan que en los entornos de human autonomy teaming, la confianza debe entenderse como un proceso bidireccional. No sólo importa cuánto confía el humano en la máquina, sino también cómo el sistema evalúa la fiabilidad y consistencia del operador. Este concepto de mutual trust resulta especialmente relevante en aplicaciones como la conducción automatizada o la robótica colaborativa, donde la toma de decisiones se comparte entre ambos agentes [9].

Por tanto, el estudio del Human System Trust ofrece el marco conceptual para comprender cómo se forma, calibra y mantiene la confianza entre personas y sistemas inteligentes. Este marco sirve de base para los modelos y las herramientas de medición presentadas en los apartados siguientes.

2.2 Modelos

La confianza no es un estado estático, sino un proceso que se construye y evoluciona con el tiempo. A lo largo de las últimas décadas, se han desarrollado distintos modelos de confianza, desde enfoques teóricos de la psicología social hasta representaciones dinámicas y computacionales aplicadas a la interacción humano máquina.

En las primeras aproximaciones psicológicas, Rotter (1967) la definió como una expectativa generalizada sobre la fiabilidad y honestidad de los demás, situando la base conceptual para su estudio en distintos contextos sociales [3]. A partir de esta definición, Mayer, Davis y Schoorman (1995) propusieron un modelo, mostrado en la Figura 2.2, basado en el ámbito de la organización, en el que la confianza se entiende como una función de tres factores principales mencionados anteriormente: la habilidad, la benevolencia y la integridad del agente en quien se deposita la confianza [7]. Este enfoque aportó una estructura conceptual que más tarde se trasladó al estudio de la interacción entre humanos y sistemas automatizados.

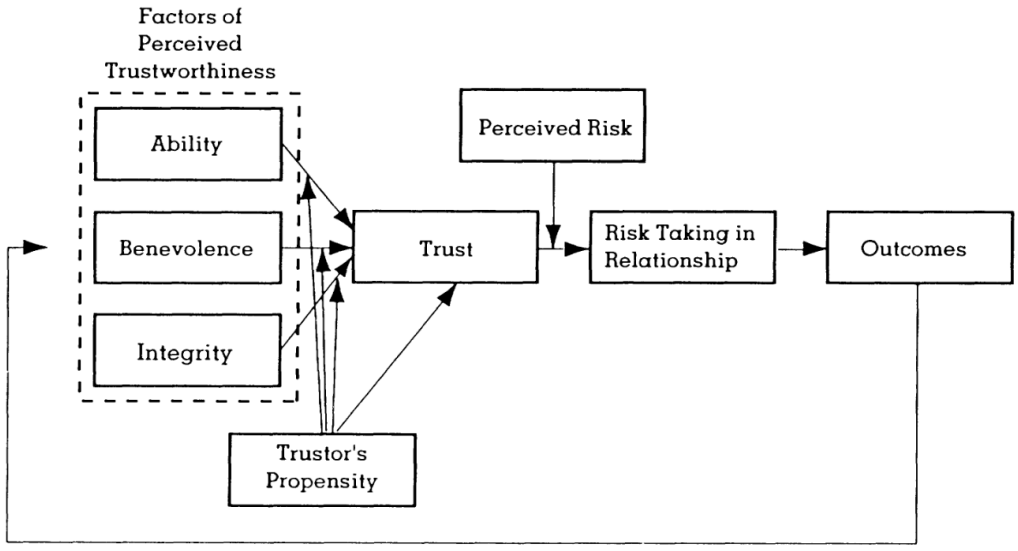

Figura 2.2. Modelo de confianza propuesto por Mayer, Davis y Schoorman (1995) [7].

Con el avance de la automatización, Parasuraman y Riley (1997) señalaron que el comportamiento humano frente a las máquinas podía oscilar entre el uso adecuado, el mal uso o el desuso de la tecnología, dependiendo del grado de confianza que el usuario depositara en el sistema [4]. Esta idea se amplió en el trabajo de Lee y See (2004), quienes introdujeron el concepto de calibración de la confianza, entendido como el equilibrio entre la confianza del operador y la fiabilidad real del sistema [1]. Según estos autores, una confianza mal calibrada

puede provocar sobredependencia cuando el usuario delega en exceso en la máquina o, por el contrario, subutilización cuando se rechaza su asistencia incluso siendo fiable. De este modo, la calibración de la confianza se convierte en un proceso dinámico que debe ajustarse continuamente en función de la experiencia del usuario y del comportamiento del sistema.

Investigaciones posterior han demostrado que la confianza en sistemas inteligentes se actualiza constantemente en función de la experiencia, la transparencia y el rendimiento del sistema [2], [8]. En el contexto del Human Autonomy Teaming, Lyons et al. (2021) destacan que la confianza debe considerarse como un proceso adaptativo y bidireccional, donde tanto el ser humano como el sistema ajustan su comportamiento en función de la interacciones previas [9]. Esta visión resulta clave para diseñar sistemas que regulen su nivel de autonomía o de explicabilidad en función de la confianza percibida.

Los modelos computacionales han permitido formalizar esta evolución temporal. Akash et al. (2018) propusieron un modelo de aprendizaje supervisado basado en señales fisiológicas, a través de los electroencefalogramas (EEG) y la respuesta galvánica de la piel (GSR), con el objetivo de estimar el nivel de confianza del usuario en tiempo real y de manera más específica [5]. Este enfoque permite vincular la confianza con factores neurofisiológicos. Dando lugar a una aproximación más objetiva que los cuestionarios o medidas subjetivas.

A su vez, los grandes avances de la inteligencia artificial explicable (Explainable Artificial Intelligence, XAI) han permitido mejorar la interpretación de estos modelos. Herramientas como SHAP o LIME facilitan la comprensión de variables o regiones cerebrales contribuyendo a las predicciones de confianza, y aumentando en gran medida a la transparencia del proceso de decisión algorítmica [6]. En este sentido, la explicabilidad no solo aporta interpretabilidad técnica, sino que también favorece una calibración más adecuada de la confianza entre el usuario y el sistema automatizado.

En conjunto, los modelos teóricos y dinámicos de la confianza trazan una evolución clara: parten de los primeros conceptos más psicológicos, basados en expectativas sociales, hasta los modelos computacionales actuales, que permiten estimar y explicar la confianza en tiempo real. Este recorrido marca el tránsito de una perspectiva estática a una adaptativa, en la que la confianza se entiende como una variable medible, en constante cambio y esencial para el desempeño eficiente de los sistemas inteligentes colaborativos.

En este contexto, los modelos basados en aprendizaje automático propuestos por Akash et al. (2018) constituyen un punto de partida fundamental para el estudio empírico de la confianza. No obstante, los enfoques más recientes tienden a incorporar modelos dinámicos y técnicas no supervisadas, como los planteados por Xu y Dudek (2015) o Guo y Yang (2020), que permiten capturar la evolución temporal y las fluctuaciones de la confianza en interacción continua [10], [11]. Estas referencias resultan especialmente relevantes para justificar los análisis de clustering y modelado dinámico que se abordarán en fases posteriores del presente trabajo.

2.3 Herramientas y métodos de medición de la confianza

La medición de la confianza en la interacción humano sistema ha pasado de aproximaciones psicológicas basadas en cuestionarios a enfoques neurofisiológicos y computacionales que permiten estimarla en tiempo real. En este contexto, el uso de electroencefalogramas (EEG) se ha convertido en una herramienta clave para analizar cómo las variaciones en la actividad cerebral reflejan diferentes niveles de confianza durante la interacción de sistemas automatizados [5]. Este cambio responde a la necesidad de obtener indicadores objetivos del Human System Trust (HST) sin depender exclusivamente de evaluaciones subjetivas [2]. Este avance refleja la búsqueda de métricas más precisas y adaptativas que capten las variaciones de confianza a lo largo del tiempo.

Los primeros estudios utilizaron escalas de autoinforme para evaluar la confianza percibida, como la propuesta por Rotter (1967), centrada en las expectativas de fiabilidad interpersonal [3]. Con la introducción de sistemas automatizados, autores como Parasuraman y Riley (1997) relacionaron la confianza en el uso, mal uso y desuso de la automatización [4]. Estas herramientas conceptuales fueron el punto de partida para modelos posteriores que incorporaron mediciones objetivas.

A medida que se desarrollaron técnicas de registro fisiológico, las investigaciones se orientaron hacia medidas continúas basadas en señales EEG, GSR, ECG y parámetros oculares, que permiten inferir el estado de confianza del usuario durante la interacción. Como Akash et al. (2018) emplearon registros EEG y de conductancia de la piel junto con modelos de machine learning para clasificar estados de alta y baja confianza [5]. De forma complementaria, Xu y Dudek (2016) y Guo y Yang (2021) desarrollaron modelos probabilísticos dinámicos basados en inferencia bayesiana, capaces de predecir la evolución temporal de la confianza en entornos colaborativos [10], [11].

Las herramientas utilizadas en los distintos trabajos mencionados fueron los entornos de programación científica como Python o MATLAB. Por otro lado, librerías de aprendizaje automático ampliamente utilizadas scikit-learn, XGBoost, TensorFlow además de paquetes especializados en neurociencia como MNE-Python o EEGLAB. Estas herramientas facilitan el preprocesamiento de señales, la extracción de características y la clasificación automática de estados fisiológicos.

Las nuevas tecnologías combinan estas técnicas con métodos de explicabilidad de modelos (Explainable AI, XAI). Recursos como SHAP (SHapley Additive exPlanations) permiten interpretar qué variables influyen más en las predicciones, mejorando la transparencia de los modelos y su valor neurofisiológico [6].

2.3.1 Herramientas empleadas en el presente trabajo

En el presente trabajo se ha utilizado Python 3.11 como lenguaje principal de desarrollo, empleando librerías del ecosistema científico de Python para el análisis de señales EEG y la aplicación de técnicas de aprendizaje automático. Las herramientas empleadas son las siguientes:

- **NumPy y pandas**, para la gestión y manipulación de los datos.
- **MNE,** para el análisis de las señales EEG y la topografía
- **scikit-learn,** para la implementación, entrenamiento y evaluación de modelos de aprendizaje automático.
- **SHAP,** para el análisis de la explicabilidad e interpretación de los modelos.
- **seaborn y matplotlib**, utilizadas como apoyo en la visualización de resultados y en el análisis exploratorio.

El uso conjunto de estas herramientas permite abordar de forma integrada el tratamiento de señales EEG, el modelado automático y la interpretación de los resultados obtenidos.

3 Desarrollo

En este capítulo se describe la metodología empleada a lo largo del trabajo, incluyendo la descripción de los conjuntos de datos utilizados, su tratamiento y el análisis exploratorio realizado. Asimismo, se presentan las condiciones experimentales consideradas y las técnicas de análisis aplicadas, que sirven de base para los capítulos posteriores.

La Figura 3.1 muestra las etapas principales seguidas durante el desarrollo del Trabajo Fin de Grado, junto con el capítulo en el que se presentan. Como se observa en la figura, el trabajo parte de la base de datos proporcionada por el grupo de investigación NeuroTechAI. A partir de esta base de datos, en el Capítulo 3 se realiza la descripción de los datos, el agrupamiento por regiones cerebrales y un análisis exploratorio inicial. Posteriormente, en el Capítulo 4, se aborda el aprendizaje supervisado, incluyendo la preparación de los datos y el entrenamiento de los distintos modelos de clasificación. En el Capítulo 5 se desarrolla el análisis de interpretabilidad de los modelos mediante técnicas SHAP. Finalmente, los resultados obtenidos se presentan en el Capítulo 6, y el Capítulo 7 recoge una experimentación propia de carácter exploratorio destinada a validar el procedimiento de adquisición y registro de señales EEG

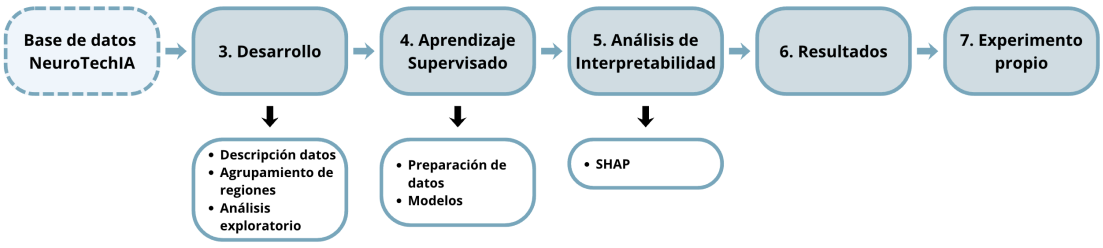

Figura 3.1. Flujo de trabajo y organización por capítulos del Trabajo Fin de Grado

3.1 Descripción del conjunto de datos

Los conjuntos de datos empleados en este Trabajo Fin de Grado proceden de un proyecto de investigación desarrollado por el grupo de investigación NeuroTechAI de la Universidad Politécnica de Madrid (UPM), cuyo objetivo es el estudio de la confianza humano–sistema en tareas cognitivas mediante el registro de señales psicofisiológicas. En particular, la finalidad del proyecto es medir la confianza de pilotos de aeronaves en las indicaciones que le proporciona el sistema. En dicho proyecto se diseñaron distintos escenarios experimentales con el fin de analizar el comportamiento del usuario y su

respuesta ante sistemas automatizados en contextos controlados y progresivamente más complejos.

Durante el desarrollo del proyecto de investigación se recopilaron distintas señales psicofisiológicas, entre las que se incluyen electroencefalogramas (EEG), electrocardiogramas (ECG), señales de eye-tracking y actividad electrodérmica (EDA), registrada mediante el dispositivo Bitbrain Ring. Sin embargo, en el presente trabajo se han descartado estas señales adicionales, centrándose exclusivamente en el análisis de las señales EEG y en las variables de confianza asociadas. Esta decisión se ha tomado con el objetivo de delimitar el alcance del trabajo y profundizar en el análisis de una única señal.

En cuanto a la muestra, inicialmente participaron 22 sujetos, que completaron las dos sesiones experimentales consideradas (condición neutral y condición ecológica). No obstante, tras el análisis de calidad de los datos y la aplicación de los criterios de exclusión establecidos, el número final de participantes se redujo a 15 sujetos para la condición neutral y 11 sujetos para la condición ecológica, siendo únicamente estos conjuntos los empleados en los análisis posteriores.

A lo largo del proyecto de investigación se plantearon distintas condiciones experimentales, diseñadas para introducir progresivamente un mayor nivel de complejidad. En primer lugar, se diseñó un experimento neutral, en el que los participantes realizaban exclusivamente la tarea Stroop (tarea cognitiva) en un entorno controlado. En segundo lugar, se planteó un experimento ecológico, en el que se introducía una tarea adicional representativa del pilotaje de una aeronave, incrementando la carga cognitiva del participante. Adicionalmente, se desarrollaron ejercicios de validación posteriores con tareas de mayor complejidad, que no se consideran en el presente trabajo.

En el marco de este estudio, el análisis se centra exclusivamente en las condiciones experimentales neutral y ecológica. A continuación, se describen ambos experimentos con mayor detalle, incluyendo la estructura de las tareas, las fases experimentales y el procedimiento seguido durante la adquisición de datos.

3.1.1 Experimento Neutral

La condición neutral consiste en un experimento en el que el participante realiza la tarea Stroop en un entorno controlado y sin interferencias externas. El objetivo es registrar la actividad cerebral del participante a través del EEG en

una situación de interacción sencilla con el sistema, que sirve como referencia para el análisis posterior.

La tarea Stroop consiste en una interacción en la que en la pantalla aparecen escritos los colores yellow (amarillo), red (rojo), green (verde) y blue (azul). Estas palabras aparecen pintadas en un color que puede coincidir o no con el significado de la palabra, y el participante debe seleccionar en la pantalla la inicial del color en el que está pintado el texto. La Figura 3.2 muestra un ejemplo de la tarea Stroop, en el que aparece la palabra blue coloreada en rojo; en este caso, el participante debe seleccionar la letra “R”, correspondiente al color rojo en inglés.

Figura 3.2. Ejemplo del funcionamiento de la tarea Stroop sin automatización.

El experimento está estructurado en distintas fases. En una fase inicial de preparación se configuran y colocan los sensores, se comprueba la calidad de las señales y se informa al participante del procedimiento experimental. A continuación, el participante realiza una fase de práctica de la tarea Stroop sin asistencia del sistema, con el objetivo de familiarizarse con la dinámica del experimento.

Posteriormente, se desarrolla la fase principal del experimento, en la que se introduce un sistema de automatización que actúa como apoyo a la decisión, sugiriendo una de las posibles respuestas en cada ensayo del Stroop. Durante esta fase, la fiabilidad de la automatización varía por bloques, alternando periodos de alta y baja fiabilidad. Esta variación permite analizar los cambios en la confianza del participante en el sistema en función de su comportamiento. Tras cada interacción, el participante debe indicar cuánto ha confiado en la automatización en una escala del 1 al 5, y al final de cada bloque debe seleccionar la confianza percibida a lo largo de dicho bloque.

La Figura 3.3 muestra un ejemplo del funcionamiento de la tarea Stroop con automatización, en la que el sistema proporciona una sugerencia de respuesta al participante. En este caso, la automatización sugiere incorrectamente la letra

“B” (color azul), cuando la respuesta correcta es la letra “R”, correspondiente al color rojo en el que está pintada la palabra.

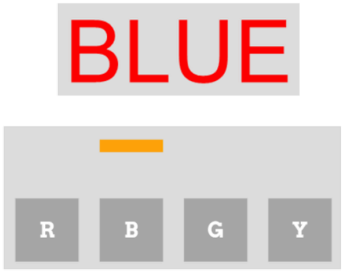

Figura 3.3. Ejemplo del funcionamiento de la tarea Stroop con automatización.

Durante toda la condición neutral se registra de forma continua la señal EEG del participante, así como las respuestas de confianza obtenidas durante la tarea.

3.1.2 Experimento Ecológico

Para añadir una mayor complejidad, se plantea la condición ecológica, en la que se simula un entorno más realista y con una mayor carga cognitiva. En este caso, además del desarrollo de la prueba Stroop, el participante debe prestar atención a una tarea representativa del pilotaje de una aeronave. Para recrear esta situación se emplea la herramienta OpenMATB, utilizada para simular tareas de monitorización en entornos operativos y que requiere atención sostenida sobre distintos indicadores, incrementando así la carga cognitiva del participante.

En el marco del proyecto de investigación, se ha seleccionado la tarea que se muestra en la Figura 3.4 con el objetivo de que los participantes (pilotos en la aplicación futura de este proyecto) monitoricen distintos tanques de combustible, simulando una tarea de supervisión relevante en el contexto aeronáutico.

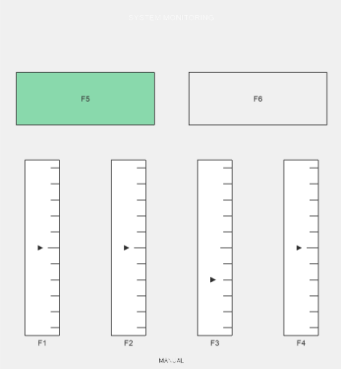

Figura 3.4. Ejemplo de la tarea de monitorización empleada en la condición ecológica mediante la herramienta OpenMATB.

El experimento se divide en tres fases. En primer lugar, se desarrolla una fase inicial en la que se explica al participante el funcionamiento de la tarea de monitorización OpenMATB y se comprueba que la adquisición de datos por parte del sensor se realiza correctamente.

A continuación, se incluye una fase breve de práctica, cuyo objetivo es permitir al participante adaptarse a la realización simultánea de ambas tareas. Esta fase se reduce con respecto a la condición neutral con el fin de evitar una duración excesiva del experimento y minimizar la aparición de fatiga, teniendo en cuenta que ambas condiciones se realizan dentro de la misma sesión experimental.

Por último, se lleva a cabo la fase de automatización. En este caso, se consideran únicamente dos bloques: un primer bloque con una fiabilidad de la automatización del 50 % y un segundo bloque con una fiabilidad del 100 %. Se realiza una reducción de la fase con el objetivo de no prolongar excesivamente el experimento.

La estructura de los bloques y los cuestionarios administrados durante y al final de cada uno de ellos es equivalente a la utilizada en el experimento neutral, lo que permite una comparación directa entre ambas condiciones.

3.1.3 Datos en crudo y preprocesados

Los datos utilizados en este trabajo proceden del registro continuo de señales EEG durante la realización de los experimentos descritos en el apartado anterior. Estas señales constituyen los datos en crudo, es decir, registros

directos de la actividad cerebral adquiridos a través de los electrodos del casco EEG, sin aplicar ningún tipo de procesamiento previo.

La configuración elegida para los experimentos se muestra en la Figura 3.5. Esta configuración consta de 19 sensores distribuidos uniformemente alrededor de la cabeza, con el objetivo de capturar una cantidad significativa de todos los tipos posibles de ondas cerebrales, junto con un sensor adicional colocado en la oreja que sirve como conexión a tierra.

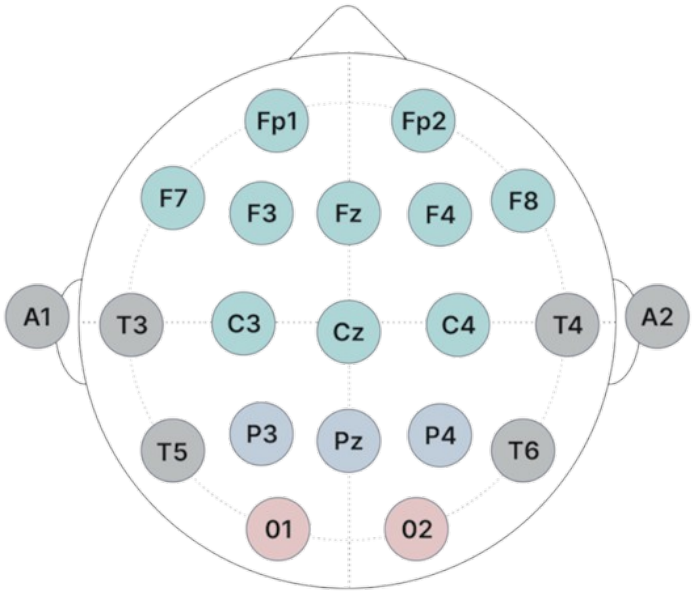

Figura 3.5. Distribución y posiciones de los electrodos EEG correspondientes a un sistema de 19 sensores.

La correspondencia entre la nomenclatura de cada orificio del gorro EEG (celdas sombreadas en azul) y los números de los sensores (celdas sombreadas en blanco) se muestra en la Tabla 3.1.

Esta numeración se emplea de forma consistente en las distintas etapas del trabajo, incluyendo el preprocesado y el análisis de las señales EEG, con el objetivo de facilitar la gestión y referencia de los canales.

Tabla 3.1. Correspondencia entre posiciones y números de sensores.

1	Fp1	11	C4
2	Fp2	12	T8 (T4)
3	F7	13	P7 (T5)
4	F3	14	P3
5	Fz	15	Pz
6	F4	16	P4
7	F8	17	P8 (T6)
8	T7 (T3)	18	01
9	C3	19	02
10	Cz	Reference	A1

A partir de estas señales en crudo, se llevó a cabo un proceso de preprocesado y extracción de características con el objetivo de obtener representaciones más compactas y manejables de la información contenida en el EEG. En primer lugar, las señales continuas se segmentaron en épocas (epochs) en función de los eventos definidos durante el experimento, de modo que cada época representa un intervalo temporal de interés asociado a una fase concreta de la tarea experimental.

Dado que las épocas presentan duraciones variables, se aplicó posteriormente un proceso de segmentación en ventanas temporales deslizantes dentro de cada época, con el fin de obtener segmentos de longitud fija que permitieran una extracción homogénea de características. Cada ventana representa así un intervalo temporal concreto de actividad cerebral dentro de una época.

Para cada ventana temporal se extrajeron características de todos los canales EEG registrados y de distintas bandas de frecuencia. En concreto, se calcularon once tipos de características para cada uno de los 19 canales y para seis bandas de frecuencia, lo que dio lugar a un total de 1254 características por ventana temporal. Las bandas consideradas fueron delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), sigma (13–16 Hz), beta (16–24 Hz) y low gamma (24–60 Hz).

Sobre cada una de estas ventanas se calcularon distintas características extraídas tanto en el dominio frecuencial como en el dominio temporal. En el dominio frecuencial, se obtuvieron medidas basadas en la densidad espectral de potencia (PSD), la potencia relativa y la entropía diferencial, calculadas por canal y por banda de frecuencia.

Adicionalmente, sobre las señales EEG previamente filtradas en cada banda se calcularon estadísticas descriptivas y parámetros de complejidad temporal,

incluyendo la media, la desviación estándar, el rango pico a pico, la asimetría, la curtosis y los parámetros de Hjorth (actividad, movilidad y complejidad).

Las características extraídas se almacenaron de forma estructurada junto con información adicional, como el identificador de la ventana temporal y las etiquetas de confianza asociadas. En función del análisis realizado en cada apartado, se trabajó tanto con las señales en crudo como con las características preprocesadas, lo que permitió estudiar el comportamiento de los datos en distintos niveles de representación.

3.2 Agrupación de canales EEG por regiones cerebrales

En estudios basados en señales EEG, es habitual agrupar los electrodos en regiones cerebrales amplias como frontal, temporal, central, parietal y occipital, con el objetivo de facilitar la interpretación de los resultados y reducir la complejidad inherente en los datos. Esta decisión se apoya en consideraciones de tipo neurofisiológico, interpretativo y metodológico, que se desarrollan a continuación.

Desde el punto de vista **neurofisiológico**, la señal EEG no puede interpretarse como la actividad de una región cerebral aislada, sino como el resultado de la actividad conjunta de amplias poblaciones neuronales. Además, la propagación del campo eléctrico a través del cráneo y el cuero cabelludo limita la resolución espacial del EEG, de modo que electrodos cercanos tienden a registrar señales similares y correlacionadas. Por este motivo, las medidas obtenidas no representan activaciones puntuales, sino la actividad integrada de regiones cerebrales más amplias [12].

Desde una perspectiva **interpretativa**, el análisis a nivel regional permite relacionar los resultados con funciones cognitivas, evitando interpretaciones excesivamente locales que no son apropiadas para EEG. De este modo, resulta más sencillo interpretar los patrones observados, por ejemplo, asociando la actividad frontal con procesos de control cognitivo o la actividad occipital con el procesamiento visual, tal y como se describe de forma recurrente en la literatura [13].

Por último, desde el punto de vista de la **reducción de la complejidad** de los datos, el análisis regional contribuye a disminuir la dimensionalidad del problema, reduciendo el número de variables y atenuando la variabilidad intercanal e inter-sujeto. Este enfoque resulta útil en fases iniciales del análisis,

donde un número elevado de características altamente correlacionadas puede dificultar la interpretación de los datos y el análisis de su estructura interna.

En conjunto, la agrupación de electrodos por regiones cerebrales constituye una estrategia equilibrada que permite conservar información funcional relevante y mejorar la interpretabilidad de los resultados. Además, al reducir la dimensionalidad y la variabilidad entre canales, este enfoque resulta especialmente útil en las fases de análisis exploratorio y en el estudio de patrones globales, sirviendo como marco interpretativo para los resultados obtenidos posteriormente. La Figura 3.6 ilustra la agrupación de los electrodos en las distintas regiones cerebrales empleadas en este estudio.

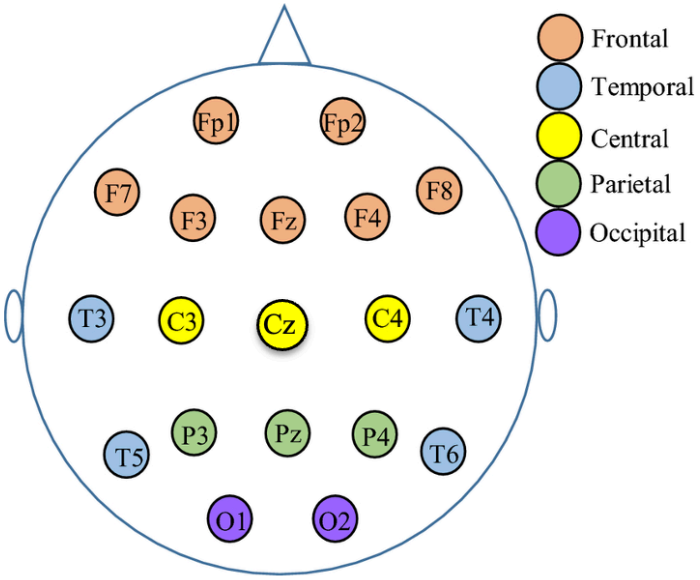

Figura 3.6. Distribución de los electrodos EEG agrupados por regiones cerebrales (frontal, temporal, central, parietal y occipital) para un sistema de 19 sensores. Adaptada de [14].

3.2.1 Descripción funcional de las regiones cerebrales

Región frontal

La región frontal se asocia de manera general con funciones cognitivas de alto nivel, como el control ejecutivo, la planificación, la atención sostenida y la toma de decisiones. En estudios EEG, la actividad frontal se ha relacionado

frecuentemente con procesos de control cognitivo, monitorización del rendimiento y regulación de la conducta, especialmente en tareas que requieren mantener objetivos, inhibir respuestas automáticas o evaluar la información disponible [15].

En el contexto de tareas cognitivas y estudios de confianza humano–sistema, las regiones frontales resultan particularmente relevantes, ya que intervienen en la evaluación de la situación, la gestión de la incertidumbre y la adaptación del comportamiento ante distintos niveles de fiabilidad del sistema. Por este motivo, la actividad frontal constituye un foco de interés habitual en estudios que analizan procesos de toma de decisiones y control cognitivo.

Región temporal

Las regiones temporales están implicadas en el procesamiento sensorial, la memoria y diversos aspectos emocionales [16]. En estudios basados en electroencefalografía (EEG), la actividad registrada en regiones temporales se ha relacionado con la integración de información auditiva y visual, así como con procesos de memoria y reconocimiento [12].

En tareas cognitivas complejas, la región temporal participa en la integración de estímulos y en la interpretación del contexto de la tarea. En este trabajo, se incluye esta región con el objetivo de analizar su posible papel en el procesamiento de la información durante la interacción con el sistema.

Región central

La región central se asocia principalmente con áreas motoras y sensorimotoras, implicadas en la planificación, ejecución y monitorización de respuestas motoras. En estudios EEG, esta región es especialmente relevante en tareas que requieren respuestas manuales o decisiones que se traducen en acciones físicas.

En el contexto de tareas cognitivas experimentales, la actividad registrada en la región central puede reflejar procesos relacionados con la preparación y ejecución de la respuesta y, en este trabajo, su análisis se incluye principalmente con fines exploratorios y de control.

Región parietal

Las regiones parietales desempeñan un papel fundamental en los procesos atencionales, la integración multisensorial y la gestión de la carga cognitiva. La actividad parietal se ha relacionado con la atención sostenida, el procesamiento

espacial y la integración de información procedente de distintas modalidades sensoriales [17].

En tareas cognitivas como el Stroop u otras tareas de control atencional, la región parietal se ha relacionado con procesos de atención y resolución de conflictos. En el presente estudio, su análisis permite evaluar posibles efectos asociados a la carga cognitiva.

Región occipital

La región occipital está principalmente asociada al procesamiento visual. En EEG, esta región muestra una actividad característica en respuesta a estímulos visuales y es especialmente sensible a cambios en la presentación, complejidad o relevancia de dichos estímulos [13].

Dado que las tareas experimentales se basan en la presentación de estímulos visuales, la región occipital resulta especialmente relevante para el análisis de la respuesta sensorial asociada a su procesamiento, particularmente en la condición ecológica.

3.2.2 Aplicación del enfoque regional al presente trabajo

En este Trabajo Fin de Grado, los análisis de las señales EEG se realizan principalmente a nivel regional, en lugar de centrarse exclusivamente en electrodos individuales. Esta decisión responde a varios motivos metodológicos y prácticos.

En primer lugar, el análisis regional permite aumentar la robustez de los resultados, al reducir la influencia de variaciones locales, ruido o irregularidades presentes en canales concretos. En segundo lugar, mejora la interpretabilidad, ya que los resultados pueden relacionarse de forma más directa con funciones cognitivas generales, evitando interpretaciones excesivamente locales que no son apropiadas para EEG.

Además, el enfoque regional facilita la comparabilidad entre sujetos, al mitigar diferencias individuales en la colocación del gorro, la conductividad del cuero cabelludo o la morfología craneal. En este trabajo, la información regional se emplea principalmente como herramienta de análisis exploratorio y como marco interpretativo para los resultados obtenidos en las fases posteriores de aprendizaje automático, en particular en el análisis de interpretabilidad mediante SHAP.

Aunque este enfoque implica una pérdida de resolución espacial fina, dicha limitación se considera asumible en el contexto de un análisis exploratorio orientado a comprender patrones globales y a facilitar la interpretación de los resultados finales del estudio.

3.3 Análisis exploratorio

Antes de abordar el modelado mediante técnicas de aprendizaje automático, se realizó un análisis exploratorio de las señales EEG con el objetivo de comprender su estructura, variabilidad y su relación con los niveles de confianza (Trust). Dada la complejidad de este tipo de señales, caracterizadas por ser no estacionarias, multicanal y dependientes tanto del individuo como del contexto experimental, se optó por una estrategia de análisis progresiva que combina enfoques estadísticos, temporales y espaciales.

En primer lugar, se analizaron las relaciones globales entre la potencia espectral y los niveles de confianza. A continuación, se estudió la evolución temporal de la señal EEG cruda a lo largo de las sesiones experimentales. Finalmente, se examinó la distribución espacial de la actividad registrada mediante representaciones topográficas.

Este enfoque permite obtener una visión conjunta del comportamiento de las señales EEG y sirve de base para justificar las decisiones adoptadas en las fases posteriores de modelado.

3.3.1 Análisis estadístico con comparaciones experimentales

Con el fin de facilitar la interpretación de un conjunto de datos EEG de alta dimensionalidad, se optó por analizar la señal en el dominio frecuencial mediante la densidad espectral de potencia (PSD). Esta medida permite describir cómo se distribuye la energía de la señal en distintas bandas de frecuencia, y es ampliamente utilizada en estudios EEG por su capacidad para resumir de forma robusta información relevante de señales no estacionarias [18] [13].

Los canales EEG se agruparon en cinco regiones cerebrales —frontal, temporal, central, parietal y occipital— y, para cada región, se calcularon valores medios de PSD en las bandas de frecuencia delta, theta, alpha, beta y low gamma. Esta organización regional y espectral proporciona una interpretación más

estructurada de los datos y contribuye a reducir parcialmente el ruido, facilitando la comparación entre participantes y condiciones experimentales.

Con el fin de analizar la variable Trust desde distintos enfoques se emplearon tres codificaciones complementarias:

- **Trust completo**, manteniendo la escala original.
- **Trust binario**, diferenciando entre estados de alta y baja confianza.
- **Trust terciario**, introduciendo un nivel intermedio.

En primer lugar, la variable Trust se mantuvo en su escala original para realizar análisis descriptivos y exploratorios, con el objetivo de estudiar la distribución general de las puntuaciones y su comportamiento entre participantes y condiciones experimentales.

Posteriormente, dichas puntuaciones se transformaron en variables categóricas con el fin de poder emplearlas en modelos de clasificación y de mitigar el desequilibrio entre clases. Asimismo, esta transformación permitió tener en cuenta la variabilidad inter-sujeto en el uso de la escala de confianza, ya que distintos participantes pueden interpretar y utilizar los valores de la escala de manera diferente.

Para la codificación binaria, se definió un umbral específico para cada participante, correspondiente a la media de sus puntuaciones de Trust. Las puntuaciones inferiores a dicho umbral se asignaron a la clase de confianza baja, mientras que las superiores se asignaron a la clase de confianza alta. Este enfoque personalizado permite capturar diferencias relativas en la percepción de la confianza, evitando el uso de umbrales absolutos comunes a todos los participantes.

Para la codificación terciaria, se aplicó una discretización basada en percentiles a la distribución de puntuaciones de cada participante. En concreto, se calcularon los percentiles 33 y 67, que se utilizaron como puntos de corte para definir tres niveles categóricos: confianza baja, media y alta. Este procedimiento, guiado por la distribución de los datos, permite introducir un nivel intermedio manteniendo un reparto más equilibrado de las clases.

Cada una de estas representaciones aporta una perspectiva distinta del estudio. Mientras que la representación binaria facilita contrastes claros entre estados extremos, la representación terciaria permite capturar transiciones graduales, y la escala completa preserva toda la información disponible sobre la confianza.

Los resultados se visualizaron mediante heatmaps de PSD región–banda, generados de forma independiente para cada participante y condición experimental (neutral y ecológica). Este tipo de representación facilita la detección visual de patrones de activación y permite comparar de forma directa cómo varía la potencia espectral en función del nivel de confianza.

Condición Neutral

Como ejemplo representativo del comportamiento individual, en las Figuras 3.7, 3.8 y 3.9 se muestran los mapas de calor de la PSD media por región y banda correspondientes al participante 4 en la condición neutral, empleando las tres representaciones de la variable Trust (completa, binaria y terciaria).

En este participante se observa una mayor actividad en la región frontal, especialmente en la banda low gamma. En concreto, los valores de potencia son más elevados para niveles bajos de confianza y disminuyen progresivamente a medida que aumenta el nivel de Trust. Este patrón se aprecia de forma consistente en las distintas representaciones de la variable Trust, aunque con diferentes grados de detalle.

Además, se observa un incremento de la actividad en la región temporal en la banda low gamma para niveles intermedios de confianza. Este patrón se manifiesta principalmente en esta banda y no se observa de forma consistente en todos los participantes.

La comparación entre las tres representaciones de Trust permite extraer información complementaria. La escala completa facilita una visualización detallada de cada nivel de confianza, mientras que la representación terciaria resulta especialmente útil para destacar estados intermedios. Por su parte, la representación binaria reduce notablemente el ruido y permite diferenciar de forma más clara entre estados de alta y baja confianza, a costa de perder información sobre las transiciones entre niveles.

Al considerar de manera conjunta los resultados obtenidos para todos los participantes en la condición neutral, se observa una elevada variabilidad inter-sujeto. Este comportamiento es esperable en señales EEG y pone de manifiesto la fuerte dependencia individual de la actividad cerebral registrada.

No obstante, a pesar de esta variabilidad, se identifican patrones recurrentes que aparecen de forma consistente en varios participantes. En particular, la región frontal destaca como una de las zonas más sensibles a los cambios en el nivel de confianza, especialmente en la banda low gamma, que es la que

presenta las variaciones más pronunciadas a lo largo de los distintos niveles de Trust. En muchos participantes, esta banda muestra valores más elevados asociados a niveles bajos de confianza, con una disminución progresiva a medida que aumenta el nivel de Trust.

En menor medida, también se observan variaciones en la región temporal, principalmente en la banda low gamma, aunque este comportamiento resulta más dependiente del individuo y menos consistente que el observado en la región frontal. La región occipital presenta asimismo variaciones en algunos participantes, lo que puede relacionarse con el carácter visual de la tarea experimental, si bien su contribución no es homogénea en el conjunto de sujetos. De forma global, el análisis conjunto de las distintas regiones cerebrales indica que la banda low gamma concentra las variaciones más destacadas en relación con el nivel de confianza.

Desde el punto de vista de las distintas representaciones de la variable Trust, el análisis global confirma las tendencias observadas a nivel individual. En particular, la representación binaria muestra patrones más estables y menos ruidosos entre participantes, mientras que la escala completa y la representación terciaria presentan una mayor dispersión.

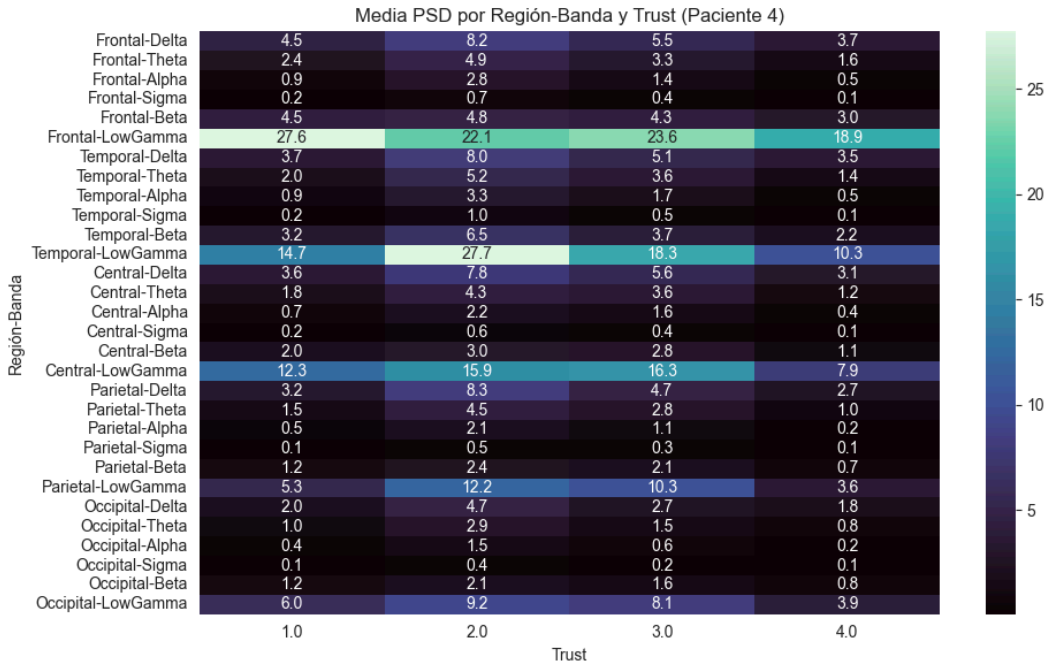

Figura 3.7. Heat map de la media de la PSD por región cerebral y banda de frecuencia para todos los niveles de Trust en la condición Neutral (participante 4).

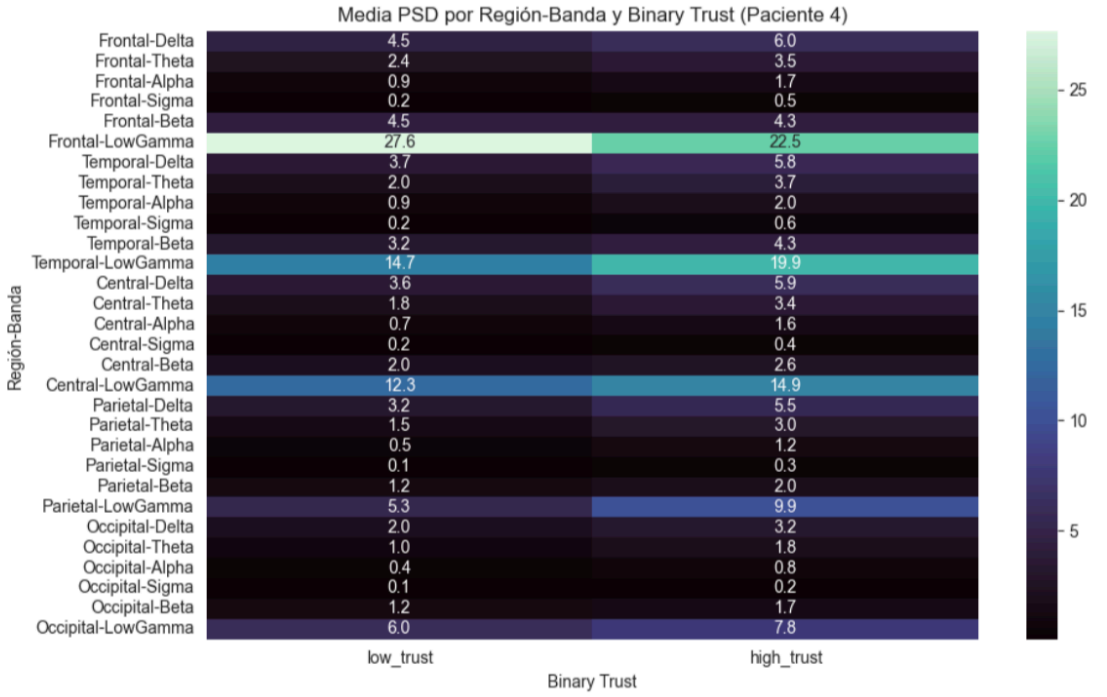

Figura 3.8. Heat map de la media de la PSD por región cerebral y banda de frecuencia para Trust binario en la condición Neutral (participante 4).

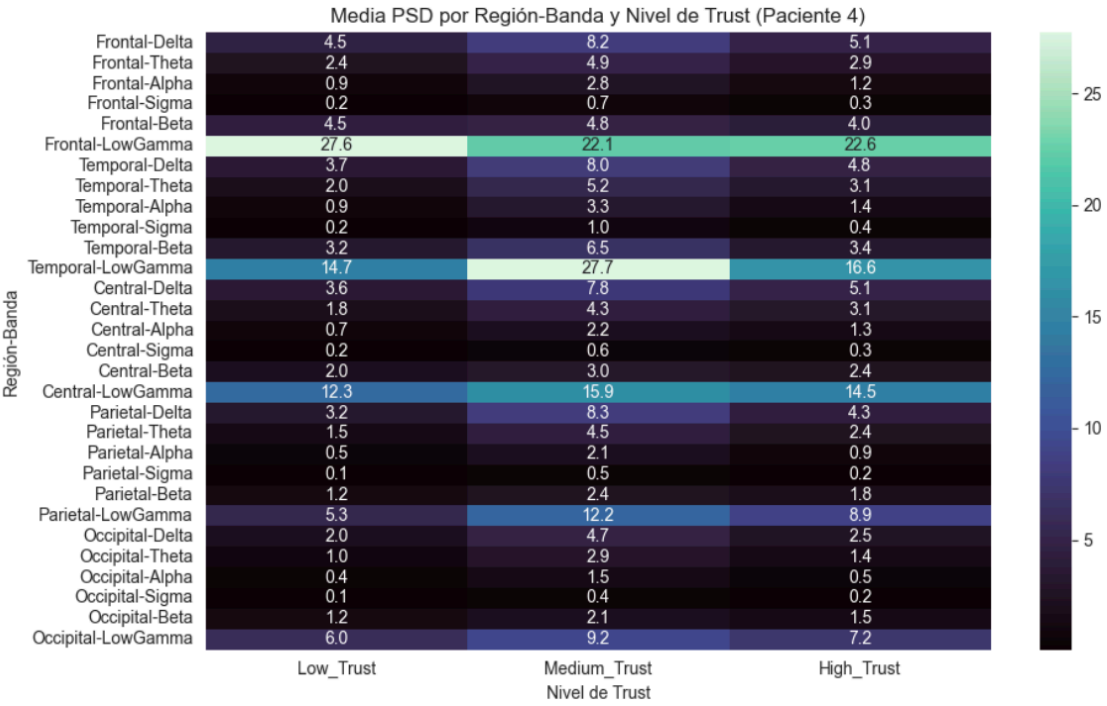

Figura 3.9. Heat map de la media de la PSD por región cerebral y banda de frecuencia para Trust terciario en la condición Neutral (participante 4).

De forma global, los resultados obtenidos en la condición neutral sugieren que la relación entre la actividad EEG y el nivel de confianza no se expresa de forma uniforme entre participantes, pero presenta regularidades claras, especialmente en la región frontal y en la banda low gamma. Estas observaciones refuerzan la importancia de considerar enfoques exploratorios y de analizar diferentes representaciones de la variable Trust antes de abordar el modelado mediante técnicas de aprendizaje automático.

Condición Ecológica

Como ejemplo representativo del comportamiento individual en la condición ecológica, en las Figuras 3.10, 3.11 y 3.12 se muestran los mapas de calor de la PSD media por región y banda correspondientes al participante 12, empleando las tres representaciones de la variable Trust.

En este participante se observa nuevamente una destacada contribución de la región frontal, especialmente en la banda low gamma, donde los valores de potencia son más elevados para niveles bajos de confianza y disminuyen de forma progresiva a medida que aumenta el nivel de Trust. Este patrón se mantiene de manera consistente en las distintas representaciones de la variable Trust, lo que refuerza la relevancia de esta región y banda en el contexto de la condición ecológica.

Adicionalmente, en esta condición se aprecia una mayor implicación de la región occipital, con valores de potencia elevados en varias bandas de frecuencia, lo que resulta coherente con el carácter más visual del entorno experimental. Esta activación occipital es más marcada que la observada en la condición neutral para este participante, sugiriendo una mayor carga asociada al procesamiento visual de los estímulos. La comparación entre las representaciones de Trust muestra un comportamiento similar al observado en la condición neutral.

En conjunto, los resultados individuales en la condición ecológica muestran una mayor complejidad y variabilidad de la señal EEG respecto a la condición neutral, con una participación más marcada de regiones relacionadas con el procesamiento visual, manteniéndose al mismo tiempo la relevancia de la región frontal en relación con los cambios en el nivel de confianza.

Al analizar de forma conjunta los resultados correspondientes a todos los participantes en la condición ecológica, se observa nuevamente una elevada variabilidad inter-sujeto, incluso más pronunciada que en la condición neutral.

Los patrones de activación difieren notablemente entre participantes, lo que refleja la mayor complejidad y heterogeneidad de la señal EEG en este entorno experimental.

A pesar de esta variabilidad, se identifican tendencias comunes que aparecen de manera recurrente en varios sujetos. En particular, la región frontal vuelve a destacar como una de las zonas más sensibles a los cambios en el nivel de confianza, especialmente en la banda low gamma, donde se observan las modulaciones más consistentes. En numerosos participantes, esta banda presenta valores de potencia más elevados asociados a niveles bajos de confianza, con una disminución progresiva conforme aumenta el nivel de Trust.

En comparación con la condición neutral, en el entorno ecológico se aprecia una mayor implicación de la región occipital, con incrementos de actividad más frecuentes y pronunciados en distintos participantes. Este comportamiento resulta coherente con el carácter más visual y dinámico del experimento ecológico, y sugiere una mayor carga asociada al procesamiento de estímulos visuales. De forma adicional, la región temporal también muestra variaciones relevantes en algunos sujetos, aunque de manera menos consistente.

Desde el punto de vista de las distintas representaciones de la variable Trust, el análisis global en la condición ecológica muestra un comportamiento similar al observado en la condición neutral, siendo la representación binaria la que ofrece patrones más estables entre participantes.

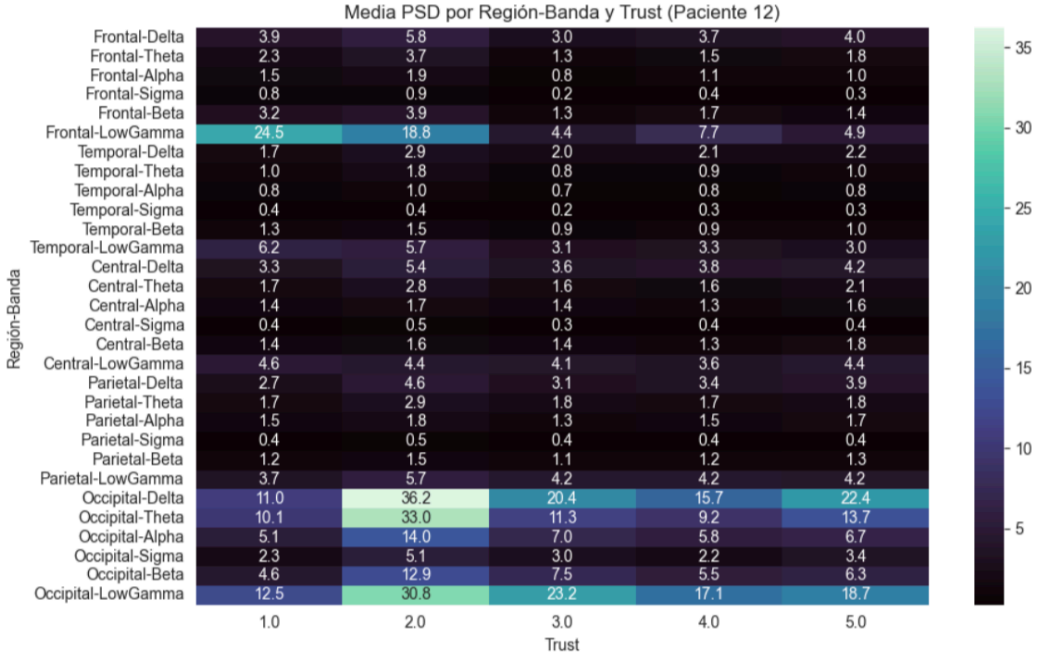

Figura 3.10. Heat map de la media de la PSD por región cerebral y banda de frecuencia para todos los niveles de Trust en la condición Ecológica (participante 12).

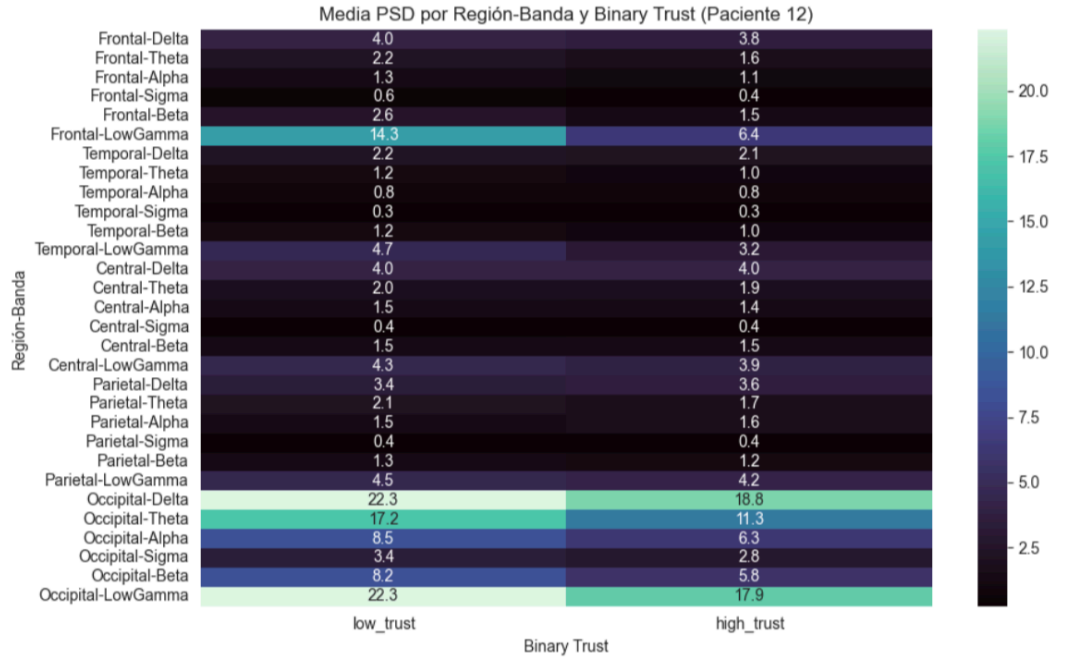

Figura 3.11. Heat map de la media de la PSD por región cerebral y banda de frecuencia para Trust binario en la condición Ecológica (participante 12).

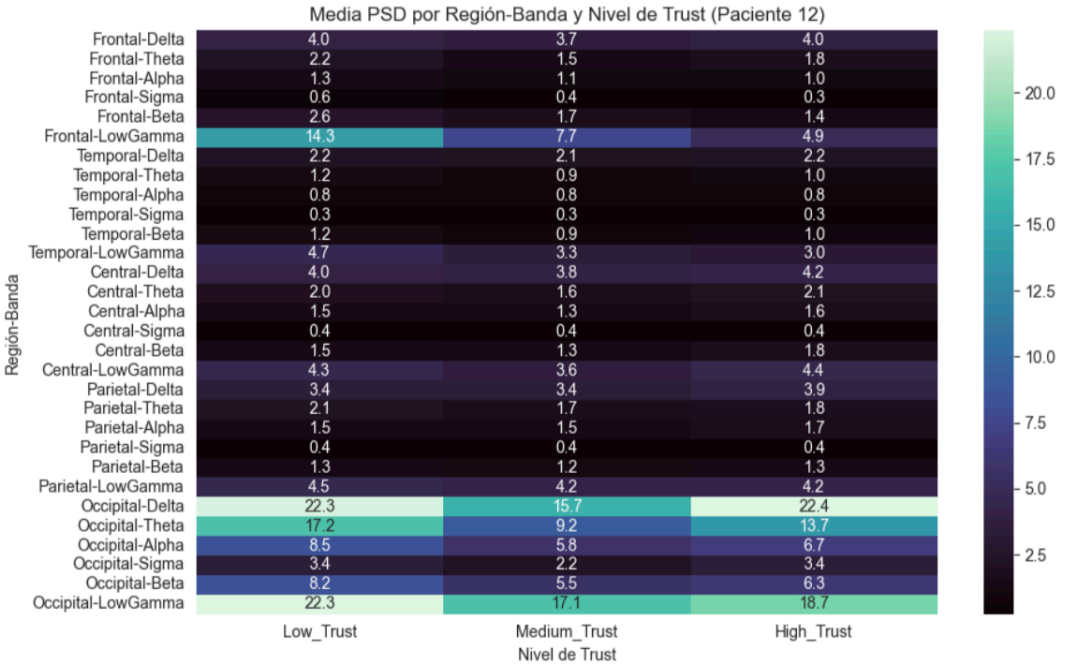

Figura 3.12. Heat map de la media de la PSD por región cerebral y banda de frecuencia para Trust terciario en la condición Ecológica (participante 12).

En conjunto, los resultados globales en la condición ecológica indican que la relación entre la actividad EEG y el nivel de confianza muestra mayores diferencias entre participantes que en la condición neutral. No obstante, se mantienen patrones claros, especialmente en la región frontal y occipital además de la importancia de la banda low gamma, lo que refuerza su relevancia en el análisis exploratorio de la confianza humano–sistema.

3.3.2 Análisis temporal de las señales EEG

Los resultados obtenidos mediante el análisis estadístico en el dominio frecuencial permiten identificar patrones globales y diferencias generales asociadas a los niveles de confianza. Sin embargo, este enfoque no proporciona información sobre la evolución temporal de la señal EEG ni sobre la estabilidad de dichos patrones durante la ejecución de las tareas experimentales.

Por este motivo, se llevó a cabo un análisis temporal de las señales EEG, de carácter fundamentalmente exploratorio, cuyo objetivo es estudiar la dinámica y variabilidad de la señal cruda a lo largo de las sesiones experimentales y evaluar si los cambios en el nivel de confianza se reflejan de forma directa en el dominio temporal.

El análisis temporal se realizó a partir de la señal EEG multicanal continua, representando de forma conjunta la evolución temporal de los distintos electrodos a lo largo de cada sesión experimental. Estas representaciones permiten observar el comportamiento general de la señal y contextualizar su evolución en relación con los cambios en el nivel de Trust asociados a cada intervalo experimental, identificando bloques temporales homogéneos y las transiciones entre estados.

En las representaciones temporales se muestra la señal EEG registrada por cada canal a lo largo del tiempo, junto con una señalización visual de los cambios en el nivel de Trust. Esta indicación facilita la identificación de los intervalos temporales asociados a cada nivel de confianza, así como de las transiciones entre ellos, permitiendo contextualizar visualmente la evolución de la señal durante la tarea.

De manera complementaria, se realizó un análisis descriptivo basado en el cálculo de la amplitud media de la señal EEG para cada canal y nivel de Trust. Este enfoque, aplicado directamente sobre la señal cruda, permite comparar de forma global el comportamiento medio de los distintos electrodos entre niveles de confianza, manteniendo la resolución espacial por canal y sin pretender establecer relaciones causales directas.

Condición neutral

Como ejemplo representativo del comportamiento individual, en la Figura 3.13 se muestran la señal EEG temporal y la amplitud media por canal correspondientes al participante 6 en la condición neutral.

A partir de las representaciones temporales, se observa que la señal EEG presenta variaciones continuas a lo largo del tiempo, pero no muestra patrones temporales claros ni cambios abruptos sincronizados de forma consistente con las transiciones entre niveles de Trust. Esto dificulta la identificación de relaciones directas entre la señal cruda temporal y los cambios en la confianza, lo que justifica el uso de características calculadas sobre épocas y ventanas temporales.

El análisis de la amplitud media por canal, mostrado en la Figura 3.14, revela diferencias dependientes del participante, con variaciones más visibles en canales frontales, especialmente en electrodos como F7 y F8 en varios sujetos. En el caso del participante 6, estas regiones muestran variaciones más pronunciadas entre niveles de Trust, en concordancia con los resultados obtenidos previamente.

A nivel global, los resultados de la condición neutral ponen de manifiesto una elevada variabilidad inter-sujeto. Mientras que en algunos participantes se observan variaciones claras en regiones frontales, en otros la señal presenta un comportamiento más homogéneo entre niveles de confianza. Esta heterogeneidad refuerza la naturaleza exploratoria del análisis y confirma que los efectos asociados al Trust no se manifiestan de forma uniforme en el dominio temporal.

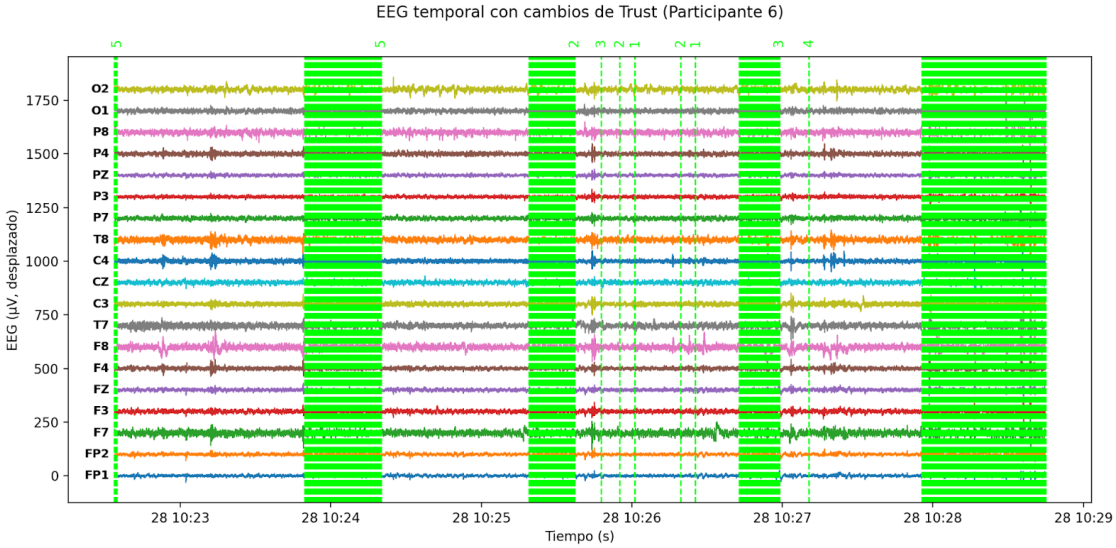

Figura 3.13. Representación temporal de las señales EEG por canal en la condición Neutral, incluyendo los cambios en el nivel de Trust (participante 6).

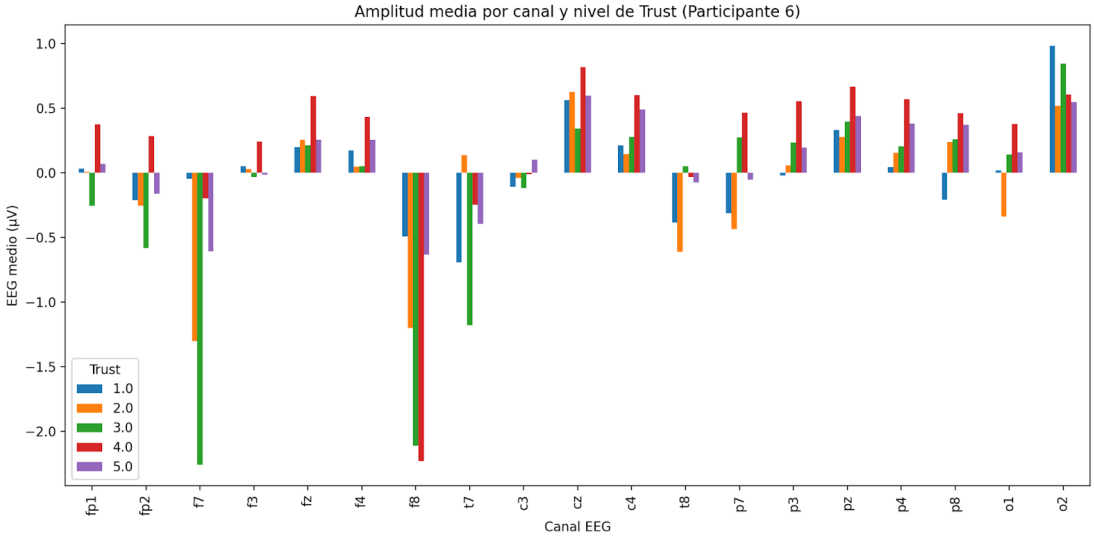

Figura 3.14. Amplitud media de la señal EEG por canal para los distintos niveles de Trust en la condición Neutral (participante 6).

Condición ecológica

En la condición ecológica, se analizaron de forma análoga las representaciones temporales y la amplitud media por canal. Como ejemplo representativo, la Figura 3.15 muestra la evolución temporal de las señales EEG por canal para

el participante 10, mientras que la Figura 3.16 presenta la amplitud media por canal para los distintos niveles de Trust en este mismo participante.

Las representaciones temporales en la condición ecológica mostradas en la Figura 3.15 evidencian una señal más variable en comparación con la condición neutral, con fluctuaciones de amplitud más pronunciadas en determinados intervalos temporales. Este comportamiento refleja una mayor irregularidad de la señal EEG a lo largo del tiempo en este entorno experimental, posiblemente asociada a la mayor complejidad y carga cognitiva del entorno ecológico.

Por su parte, el análisis de la amplitud media por canal, representado en la Figura 3.16, indica que, aunque las diferencias absolutas entre niveles de Trust son reducidas, se mantienen variaciones relevantes en regiones frontales, especialmente en los electrodos F7 y F8, en concordancia con lo observado en el análisis previo. Además, en algunos participantes se aprecia una mayor implicación de la región occipital, asociada al carácter más visual del experimento ecológico.

A nivel global, la condición ecológica presenta una mayor variabilidad inter-sujeto y temporal que la condición neutral. Los patrones de activación difieren notablemente entre participantes, y las variaciones asociadas al Trust aparecen de forma más sutil y dependiente del contexto experimental.

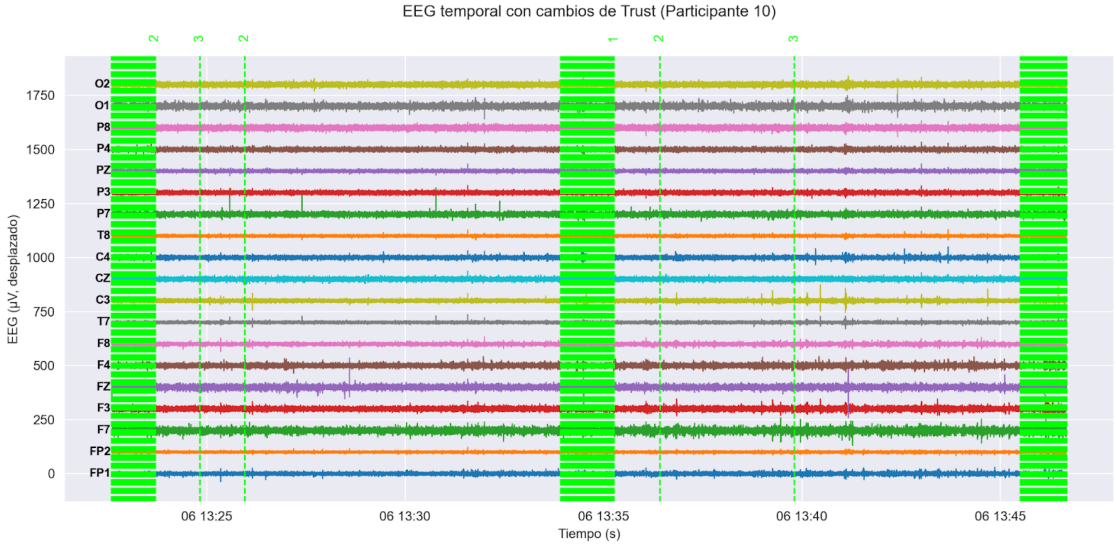

Figura 3.15. Representación temporal de las señales EEG por canal en la condición Ecológica, incluyendo los cambios en el nivel de Trust (participante 10).

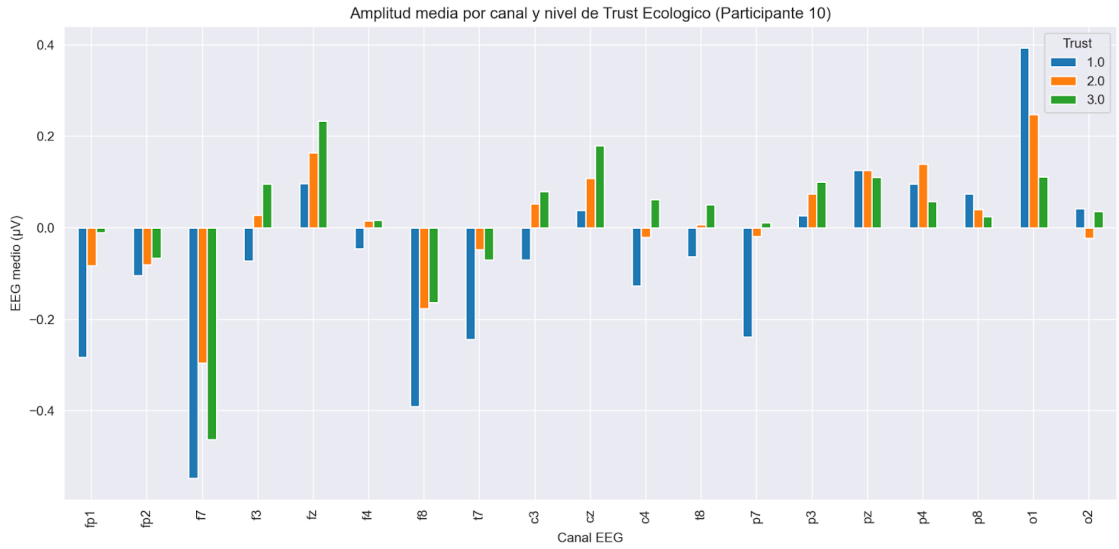

Figura 3.16. Amplitud media de la señal EEG por canal para los distintos niveles de Trust en la condición Ecológica (participante 10)

En conjunto, el análisis temporal confirma que las diferencias asociadas a los niveles de Trust no se expresan de forma clara y sistemática en la señal EEG cruda, sino que aparecen como pequeñas variaciones, altamente dependientes del individuo y del contexto experimental. Estos resultados refuerzan las conclusiones obtenidas en el análisis estadístico y justifican el uso de estrategias basadas en la extracción de características más informativas, especialmente en el dominio frecuencial y espacial, en lugar de trabajar directamente con la señal temporal cruda.

3.3.3 Análisis topográfico

El análisis topográfico de señales EEG consiste en representar sobre la silueta del cuero cabelludo la distribución espacial de medidas extraídas de la señal como se muestra en la Figura 3.17, permitiendo explorar patrones espaciales globales de la actividad cerebral registrada por los distintos electrodos. En este trabajo, el análisis topográfico se emplea con un carácter exploratorio, con el objetivo de evaluar si los distintos niveles de confianza y las condiciones experimentales (Neutral y Ecológica) se asocian a distribuciones espaciales diferenciables de la actividad EEG.

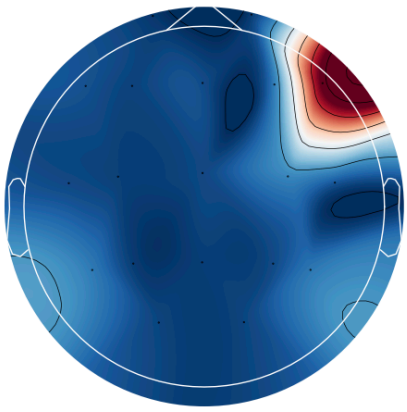

Figura 3.17 Silueta del cuero cabelludo.

Este análisis complementa los análisis estadístico y temporal previamente presentados, aportando una visión espacial que facilita la identificación de regiones que muestran una contribución recurrente o variaciones asociadas a Trust.

El análisis topográfico se realizó utilizando la densidad espectral de potencia (PSD) en la banda Low Gamma (30–50 Hz), seleccionada por su relevancia en estudios previos y por los resultados obtenidos en los análisis exploratorios iniciales.

Para cada participante y para cada nivel de confianza (Trust 1–5), se calculó el valor medio de la PSD por canal. A partir de estos valores, se generaron mapas topográficos mediante interpolación espacial basada en la disposición de los electrodos sobre el cuero cabelludo.

El análisis se realizó de manera individual para las dos condiciones experimentales y para cada una se obtuvieron dos tipos de representaciones:

- **Mapas individuales**, que permiten analizar la variabilidad intra-sujeto y observar la distribución espacial media de la actividad EEG para cada participante y para cada nivel de Trust.

- **Mapas grupales**, obtenidos a partir del promedio de los valores de PSD de todos los participantes para cada nivel de confianza, con el objetivo de evaluar la consistencia de los patrones espaciales a nivel global.

Condición Neutral

La Figura 3.18 muestra los mapas topográficos de la densidad espectral de potencia (PSD) en la banda low gamma para el participante 6 en la condición Neutral, correspondientes a los niveles de Trust 1 a Trust 5. En los distintos mapas se observa una mayor actividad localizada en regiones fronto-laterales, mientras que el resto de las regiones presenta valores inferiores.

A nivel individual, los mapas topográficos de la condición Neutral presentan una distribución espacial con valores más elevados en regiones frontales y laterales. Las regiones posteriores muestran valores inferiores, mientras que la región central presenta una distribución más homogénea.

Al comparar los distintos niveles de confianza dentro de esta condición, no se observan cambios espaciales entre Trust 1 y Trust 5. Las diferencias se manifiestan como variaciones graduales de intensidad en regiones similares, lo que indica una estabilidad espacial intra-sujeto elevada.

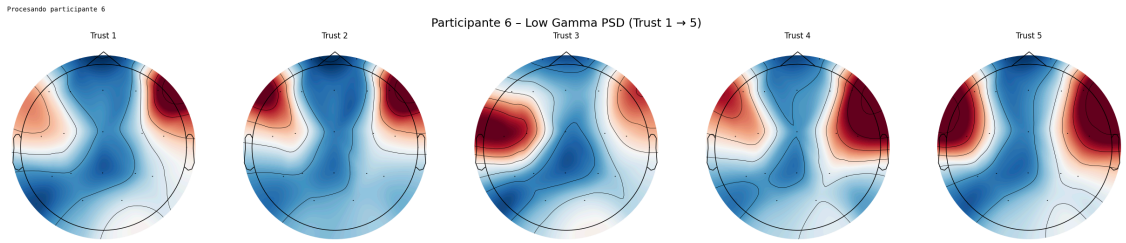

Figura 3.18. Mapas topográficos de la densidad espectral de potencia (PSD) en la banda Low Gamma para los distintos niveles de Trust (1–5) en la condición Neutral (participante 6).

La Figura 3.19 presenta los mapas topográficos grupales de la densidad espectral de potencia (PSD) en la banda low gamma para la condición Neutral, correspondientes a los niveles de Trust 1 a Trust 5.

Como se observa en la figura, a nivel grupal la distribución espacial de la potencia es similar entre los distintos niveles de confianza. Los mapas correspondientes a Trust 2, Trust 3 y Trust 5 muestran patrones espaciales similares, con una mayor actividad en regiones occipitales y temporales. En cambio, en los niveles Trust 1 y Trust 4 se observa un aumento localizado de la actividad en regiones frontales y temporales. Las regiones centrales y parietales presentan valores más bajos en todos los niveles de confianza.

La ausencia de desplazamientos de las zonas de mayor potencia entre Trust 1 y Trust 5 indica que, en la condición Neutral, las diferencias entre niveles de confianza se reflejan como cambios de intensidad en regiones similares, sin cambios en la distribución espacial global.

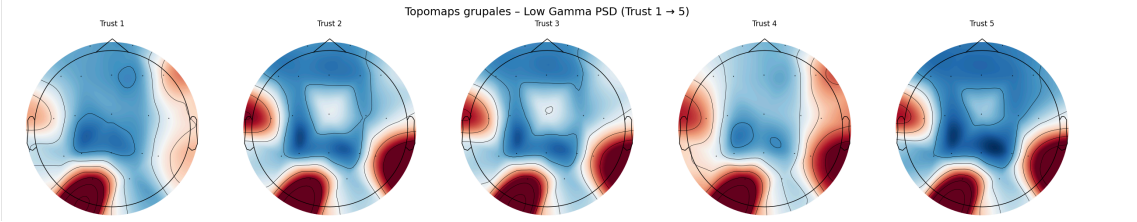

Figura 3.19. Muestra los mapas topográficos grupales para los distintos niveles de Trust en la condición Neutral.

Condición Ecológica

A continuación, se presentan los mapas topográficos correspondientes a la condición Ecológica, en la que los participantes interactúan con un entorno más dinámico y próximo a situaciones reales. En comparación con la condición Neutral, este contexto experimental introduce una mayor carga perceptiva y cognitiva, lo que puede reflejarse en una distribución espacial de la actividad EEG más compleja y variable. El análisis se centra, de nuevo, en la banda Low Gamma y en la comparación de los distintos niveles de confianza, con el objetivo de evaluar si el aumento de la complejidad del entorno se traduce en diferencias en los patrones espaciales de la actividad cerebral, tanto a nivel individual como grupal.

Como se observa en la Figura 3.20, los mapas topográficos individuales de la condición Ecológica muestran una mayor variabilidad entre participantes que en la condición Neutral. La actividad en la banda low gamma aparece distribuida entre regiones frontales, temporales y occipitales. La intensidad y la localización de esta actividad varían entre sujetos. En algunos casos, niveles bajos de Trust, como Trust 1, presentan una menor intensidad en regiones occipitales en comparación con otros niveles. Este comportamiento no aparece en todos los participantes.

En la condición Ecológica, al comparar los distintos niveles de Trust, se observan cambios de intensidad en la región occipital. Los niveles de Trust más bajos presentan una menor intensidad en dicha región, mientras que los niveles

de Trust más altos muestran valores mayores. Este comportamiento se observa también en otros participantes, aunque no de forma uniforme en todos los casos.

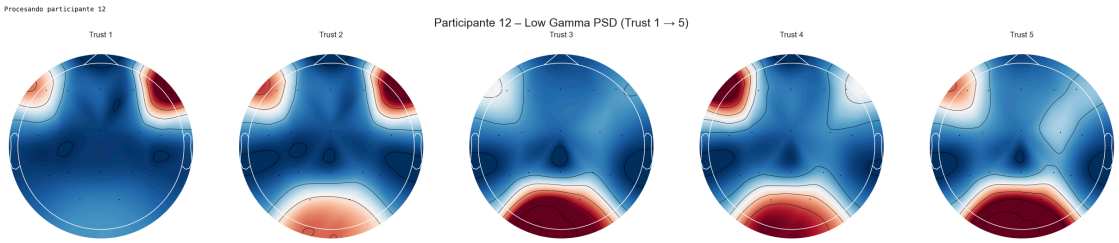

Figura 3.20. Mapas topográficos de la densidad espectral de potencia (PSD) en la banda Low Gamma para los distintos niveles de Trust (1–5) en la condición Ecológica (participante 12).

En el análisis grupal, representado en la Figura 3.21, la distribución espacial de la potencia en la condición Ecológica varía con el nivel de Trust. Para Trust 1 se observa una menor intensidad en regiones occipitales y temporales. En niveles intermedios de Trust, la intensidad aumenta principalmente en la región occipital y, en menor medida, en regiones temporales. A medida que aumenta el nivel de Trust, se aprecia un incremento progresivo de la intensidad en estas regiones, y en Trust 5 se observa además actividad en regiones fronto-laterales. Las regiones centrales y parietales presentan valores bajos en todos los niveles de confianza.

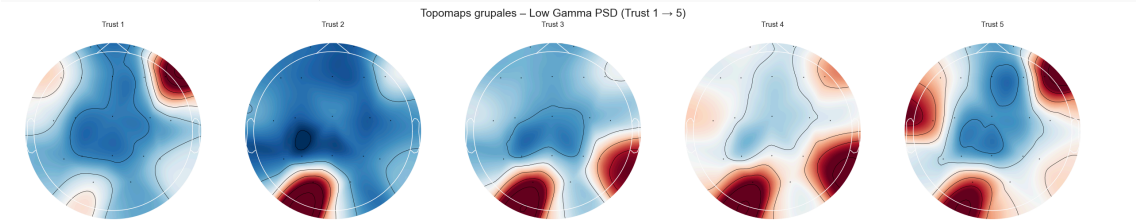

Figura 3.21. Muestra los mapas topográficos grupales para los distintos niveles de Trust en la condición Ecológica.

La comparación entre las condiciones Neutral y Ecológica indica que el contexto experimental influye en la complejidad y variabilidad espacial de la actividad EEG, siendo esta mayor en la condición Ecológica. Sin embargo, en ambas condiciones, los niveles de confianza se asocian principalmente a cambios de

intensidad sobre patrones espaciales similares, predominando las áreas occipital, frontal y temporal.

Estos resultados son coherentes con los análisis estadístico y temporal previos, en los que no se identificaron diferencias categóricas claras entre niveles de confianza de manera constante. En conjunto, el análisis topográfico refuerza la idea de que la relación entre la actividad EEG y la confianza se manifiesta de forma sutil y distribuida, sin reorganizaciones espaciales marcadas en la banda Low Gamma.

3.3.4 Aprendizaje No supervisado

Las técnicas de aprendizaje automático pueden clasificarse, de forma general, en dos grandes paradigmas: el aprendizaje supervisado y el aprendizaje no supervisado. En el primero, los modelos se entrenan utilizando tanto las características de entrada como las etiquetas asociadas a cada observación, mientras que en el aprendizaje no supervisado no se dispone de información previa sobre las clases o salidas, y el objetivo principal es identificar patrones, estructuras o agrupaciones internas en los datos.

Con el objetivo de explorar la posible existencia de patrones cerebrales relacionados con la confianza sin emplear etiquetas predefinidas, se aplicaron distintas técnicas de aprendizaje no supervisado sobre las señales EEG. En este contexto, se buscó analizar si los propios registros eran capaces de organizarse de forma natural en grupos diferenciados que pudieran corresponderse con distintos niveles de confianza, sin utilizar la variable Trust ni su versión binaria ni terciaria como variable objetivo. El flujo metodológico seguido en este análisis se resume de forma esquemática en la Figura 3.22.

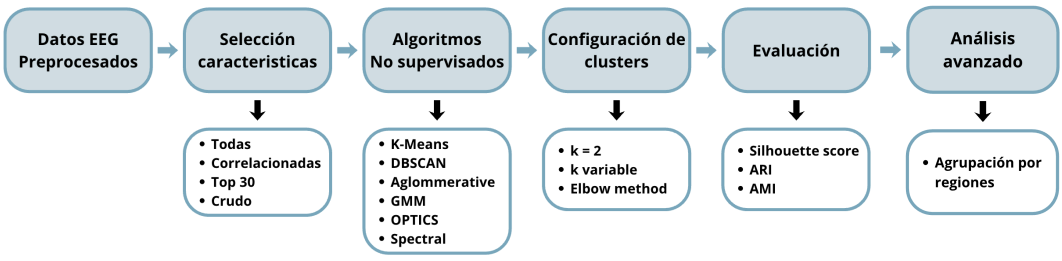

Figura 3.22. Esquema del flujo metodológico seguido en el análisis no supervisado de las señales EEG.

En este trabajo se emplearon técnicas de aprendizaje no supervisado. En particular, se aplicaron distintos algoritmos de cluster, entre ellos KMeans, DBSCAN, Agglomerative Clustering u OPTICS, entre otros [19]. Para la evaluación de los clústeres obtenidos se empleó silhouette score como métrica interna para evaluar la cohesión y separación de los clústeres obtenidos, independientemente de las etiquetas proporcionadas por los participantes [19]. Adicionalmente, se utilizaron las métricas externas ARI (Adjusted Rand Index) y AMI (Adjusted Mutual Information) para comparar los resultados obtenidos con los valores reales de la variable Trust [20].

En una primera fase se realizaron experimentos con el algoritmo K-Means, inicialmente fijando el número de clústeres en dos, en concordancia con la clasificación binaria de la confianza empleada posteriormente. Para estos experimentos se utilizaron distintos conjuntos de variables: el conjunto completo de características disponibles, las diez más correlacionadas con la variable Trust y las características más relevantes identificadas en el informe de los datos obtenidos. Los resultados mostraron cierta diferenciación visual entre los grupos; sin embargo, las métricas de evaluación indicaron una baja correspondencia entre las agrupaciones obtenidas y los niveles de confianza.

Con el fin de facilitar la interpretación de estos resultados, se aplicó una reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA), que permitió representar los datos en un espacio de menor dimensión y visualizar de forma más clara la distribución de los clústeres. Aunque esta representación, mostrada en la Figura 3.23, facilita el análisis visual de la estructura interna de los datos, no se observó una mejora significativa en la correspondencia con la variable Trust.

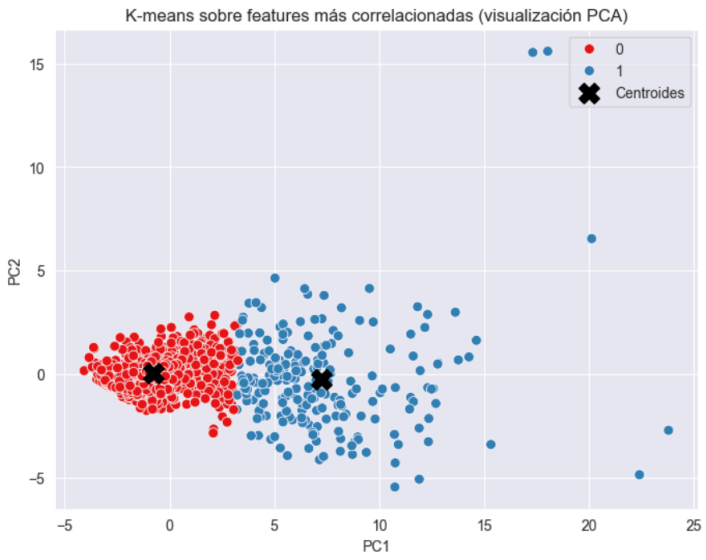

Figura 3.23. Visualización en el espacio PCA de los clústeres obtenidos mediante K-means a partir de las variables seleccionadas, incluyendo la posición de los centroides.

Tras estos resultados iniciales, se amplió la exploración aplicando otros algoritmos de aprendizaje no supervisado con el objetivo de analizar diferentes hipótesis sobre la organización de los datos. Entre los métodos evaluados se incluyeron DBSCAN, clustering aglomerativo, Gaussian Mixture Models (GMM), Optics y Spectral. Asimismo, se repitieron los experimentos con K-Means variando el número de clústeres y utilizando el método del codo (Elbow Method) para estimar un número adecuado de grupos [21]. En todos los casos, las métricas de evaluación continuaron mostrando valores bajos, sin evidenciar una separación clara y consistente entre los clústeres obtenidos y las etiquetas del trust asociadas, como se observa en las Figuras 3.24 y 3.25, donde se representan los valores de Adjusted Rand Index (ARI) y Adjusted Mutual Information (AMI) obtenidos para los distintos modelos de aprendizaje no supervisado.

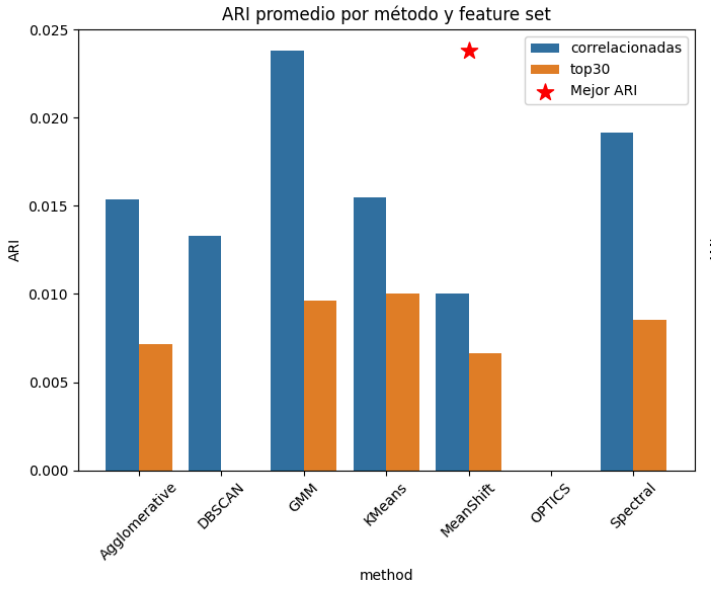

Figura 3.24. Comparación del ARI promedio entre distintos métodos de clustering para EEG.

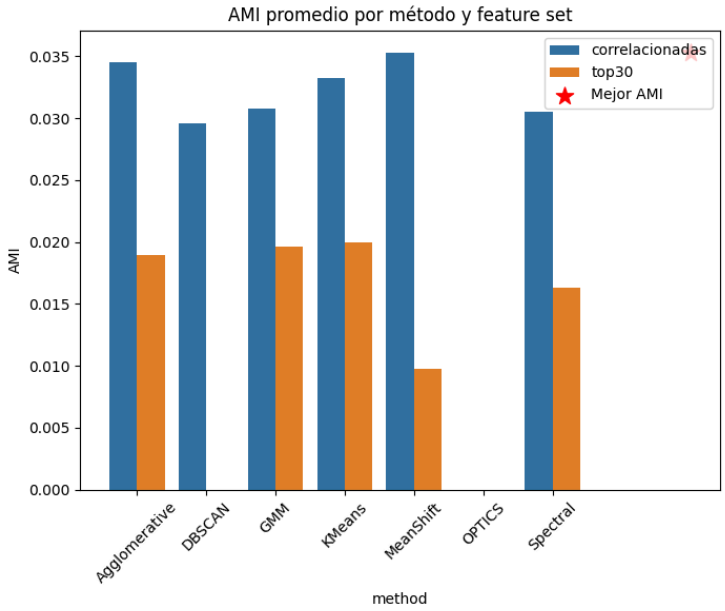

Figura 3.25. Comparación del AMI promedio entre distintos métodos de clustering para EEG.

También se realizaron pruebas utilizando las señales en crudo, sin que se observaran diferencias relevantes respecto a los resultados anteriores.

Para mejorar la interpretabilidad de los resultados, se optó por agrupar los canales EEG por regiones cerebrales (frontal, central, parietal, temporal y occipital). Esta estrategia permitió reducir el ruido asociado a canales individuales y aportar una estructura más coherente a los datos. Se observó que la densidad espectral de potencia (PSD), y en particular la banda Low Gamma, presentaban una mayor relevancia en este análisis, en línea con lo observado anteriormente.

Aunque los valores de ARI y AMI se mantuvieron en niveles bajos, el modelo GMM mostró una mejor separación interna de los datos en comparación con el resto de algoritmos evaluados. Además, se observó una notable variabilidad inter-sujeto, lo que sugiere que los patrones cerebrales asociados a la confianza pueden diferir significativamente entre participantes, dificultando la obtención de una estructura global común mediante técnicas no supervisadas.

En este contexto, la Figura 3.26 muestra la proyección en el espacio PCA de los clústeres obtenidos mediante el algoritmo K-Means para el conjunto de datos analizado, mientras que la Figura 3.27 presenta la misma proyección coloreada según los valores reales de Trust. La comparación visual entre ambas representaciones refuerza los resultados obtenidos, ya que no se observa una

correspondencia clara entre los niveles de confianza y la estructura de los clústeres obtenidos.

Figura 3.26. Proyección en el espacio PCA de los clústeres obtenidos mediante K-Means (k = 4).

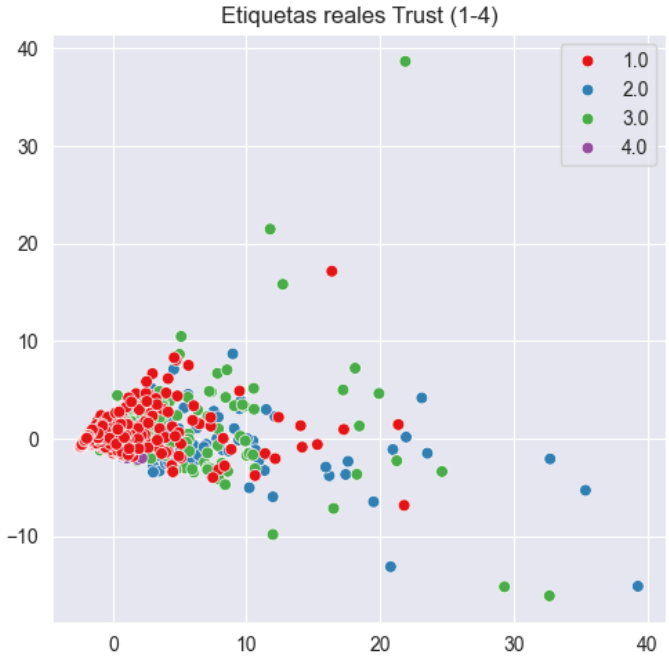

Figura 3.27. Proyección en el espacio de PCA de las muestras EEG, coloreadas según los valores reales de Trust

En conclusión, los resultados presentados corresponden a un análisis exploratorio de los enfoques de agrupación aplicados a las características del EEG en relación con los niveles de confianza. Los análisis realizados no permiten extraer conclusiones definitivas, ya que la separación entre los grupos obtenidos y las etiquetas de confianza subjetivas sigue siendo limitada.

En este sentido, sería necesario profundizar en este tipo de análisis mediante nuevos experimentos específicamente diseñados para estudiar la relación entre los patrones del EEG y la confianza. Dichos experimentos podrían proporcionar datos más estructurados y condiciones más controladas, lo que permitiría que los métodos de agrupación y otros enfoques no supervisados ofrecieran resultados más consistentes e interpretables.

4 Aprendizaje Supervisado

Este capítulo presenta los modelos de aprendizaje automático supervisado aplicados para clasificar la confianza entre el ser humano y el sistema. El capítulo se divide en cuatro secciones principales: preparación de datos, modelos de clasificación, optimización de hiperparámetros y, por último, evaluación de modelos.

El objetivo principal de este estudio es analizar cómo influyen las distintas variables en la clasificación de la confianza, de manera que los resultados obtenidos puedan interpretarse y relacionarse con el comportamiento observado durante los experimentos. Para ello, se emplean posteriormente técnicas de explicabilidad que permiten comprender mejor el funcionamiento de los modelos.

Inicialmente, se planteó el problema de clasificación de la confianza tanto en un esquema binario como en un esquema terciario. No obstante, tras una evaluación preliminar del rendimiento de los modelos, el enfoque terciario mostró resultados significativamente inferiores y una menor estabilidad entre participantes. Por este motivo, y con el objetivo de garantizar una comparación más robusta y consistente, el análisis supervisado se centra finalmente en un esquema de clasificación binaria.

Con el fin de facilitar la comprensión del procedimiento seguido, la Figura 4.1 muestra de forma esquemática el flujo metodológico empleado en el aprendizaje supervisado de las señales EEG. En dicho esquema se representan las distintas etapas del proceso, desde la preparación de los datos hasta la evaluación de los modelos. Asimismo, se incluye de forma diferenciada la fase de explicabilidad, representada en otro color, con el objetivo de indicar que este análisis se desarrolla en un capítulo posterior y no forma parte del procedimiento descrito en el presente capítulo.

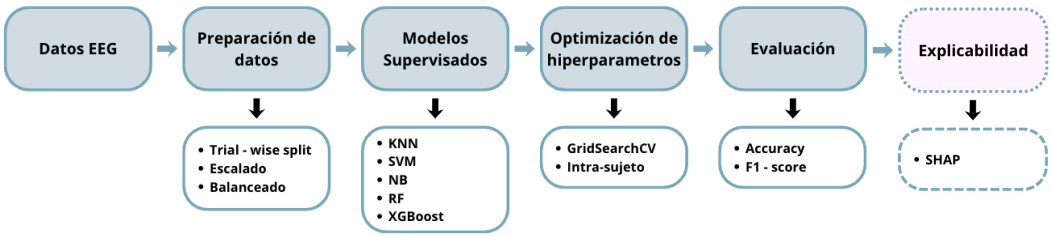

Figura 4.1. Esquema del flujo metodológico seguido en el aprendizaje supervisado de las señales EEG.

4.1 Preparación de los datos

La preparación de los conjuntos de datos es una etapa fundamental en el desarrollo de modelos de aprendizaje automático. En el caso de los modelos de clasificación, propios del aprendizaje supervisado, es necesario disponer de etiquetas de clase conocidas que sirvan como referencia durante el proceso de entrenamiento. En este trabajo, dichas etiquetas corresponden a las respuestas de los participantes a las preguntas de evaluación de la confianza tras cada prueba Stroop, que se utilizan como variable objetivo para entrenar y evaluar los modelos.

Para preparar los datos para la aplicación de técnicas de aprendizaje automático, los registros experimentales deben someterse a varios pasos de preprocesamiento: división en conjuntos de entrenamiento y prueba, normalización y corrección del desequilibrio de clases. Estos tres aspectos se detallan en las siguientes subsecciones de este documento.

4.1.1 Trial wise

Dividir los datos registrados en conjuntos de entrenamiento y prueba es un componente clave de la preparación de datos. Por lo general, esta división se realiza de forma aleatoria, asignando entre el 70 % y el 80 % de los datos al conjunto de entrenamiento y el resto al conjunto de prueba. Esta división aleatoria ayuda a minimizar el sesgo de muestreo y a mejorar la capacidad de generalización de los modelos [22].

Sin embargo, en este proyecto, la división de los datos no es totalmente aleatoria. Dado que el conjunto de datos ha sido preprocesado y segmentado en épocas, y posteriormente subdividido en ventanas, es esencial evitar que los datos de la misma época aparezcan tanto en el conjunto de entrenamiento como en el de prueba. En otras palabras, todas las instancias que pertenecen a una época determinada deben asignarse exclusivamente al conjunto de entrenamiento o al de prueba. La selección de las épocas que pertenecen a cada conjunto se lleva a cabo de forma aleatoria, siguiendo las prácticas estándar de aprendizaje automático.

Este enfoque se conoce como estrategia de división de entrenamiento-prueba por ensayo [22][23]. Su objetivo principal es evitar la fuga de datos, ya que las

instancias que se originan en la misma época están altamente correlacionadas, lo que se ajusta a escenarios de aplicación realistas.

Por tanto, se realiza una prueba 70%, 30% a través de este método. La figura 4.2 ilustra una implementación correcta de la estrategia trial-wise.

La partición trial-wise se mantiene fija a lo largo de todas las ejecuciones con el fin de garantizar la reproducibilidad de los resultados y evitar variaciones debidas a divisiones aleatorias.

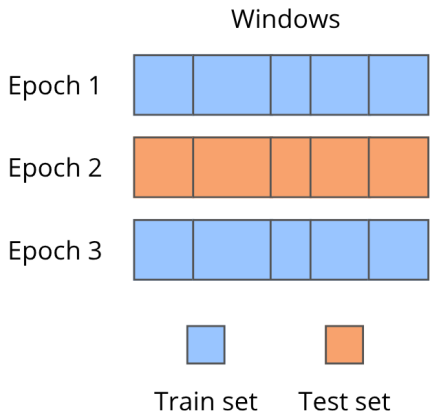

Figura 4.2. Esquema del procedimiento de partición train–test basado en ensayos (trial-wise) utilizado en este estudio.

4.1.2 Escalado

Después de dividir los datos, se puede llevar a cabo el escalado de características. Aunque este paso no es estrictamente necesario en todos los flujos de trabajo de aprendizaje automático, se considera una buena práctica cuando se utilizan determinados algoritmos de clasificación.

El escalado de características tiene como objetivo unificar los rangos de las variables. Entre las técnicas más comunes y ampliamente adoptadas se encuentran la normalización y la estandarización, presentadas en [22].

Este paso será únicamente realizado para los modelos SVM y Naive Bayes los cuales se mencionan posteriormente.

4.1.3 Balanceo de clases

Las etiquetas utilizadas como clases en los modelos de aprendizaje supervisado se derivan de la variable Trust, previamente descrita en la Sección 3.3.1. En el análisis supervisado se emplea exclusivamente la codificación binaria de la confianza, diferenciando entre estados de confianza baja y alta.

Esta transformación se aplica con el objetivo de mitigar el desequilibrio de clases y de adaptar la variable objetivo al uso de modelos de clasificación. Dado que los participantes pueden utilizar la escala de confianza de forma distinta, se definió un umbral específico para cada sujeto, calculado como la media de sus puntuaciones de Trust. Las observaciones con valores inferiores a dicho umbral se asignaron a la clase de confianza baja, mientras que las superiores se asignaron a la clase de confianza alta.

Este enfoque permite capturar diferencias relativas en la percepción de la confianza y reduce el impacto de la variabilidad inter-sujeto en el entrenamiento de los modelos, favoreciendo una distribución de clases más equilibrada y comparable entre participantes.

4.2 Modelos

En este trabajo se evalúan cinco modelos de clasificación utilizados en problemas de aprendizaje supervisado: K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest y XGBoost. Estos algoritmos representan distintos enfoques de clasificación y permiten analizar su comportamiento bajo un mismo conjunto de datos [22].

4.2.1 KNN

El algoritmo K-Nearest Neighbours (KNN) es un método de aprendizaje supervisado no paramétrico y basado en instancias. Para clasificar una nueva observación, el algoritmo identifica las K muestras más cercanas del conjunto de entrenamiento utilizando una métrica de distancia, como la euclídea, Manhattan o Minkowski [22]. La clase asignada corresponde a la más frecuente entre dichos vecinos.

El rendimiento de KNN depende principalmente de la elección del número de vecinos K y de la métrica de distancia empleada. Al no disponer de una fase de entrenamiento explícita, el coste computacional se concentra en la fase de predicción, lo que puede resultar ineficiente para conjuntos de datos de gran

tamaño. Asimismo, el algoritmo es sensible a la escala de las características, por lo que suele requerir un preprocesamiento previo de los datos.

4.2.2 SVM

Support Vector Machine (SVM) es un algoritmo de clasificación supervisada cuyo objetivo es encontrar un hiperplano de decisión óptimo que separe las diferentes clases maximizando el margen entre ellas [24]. En su forma básica, SVM construye un clasificador lineal, aunque puede incorporar un margen suave mediante un parámetro de regularización que permite cierto grado de error con el fin de mejorar la capacidad de generalización del modelo.

Cuando los datos no son linealmente separables, SVM puede extenderse mediante el uso de funciones kernel, dando lugar a la denominada SVM kernelizada. Este enfoque permite proyectar implícitamente los datos a espacios de mayor dimensión, donde la separación lineal resulta posible. Entre los kernels más utilizados se encuentran el lineal, el polinómico y el de base radial. El rendimiento del modelo depende en gran medida de la elección del kernel y de la correcta selección de sus hiperparámetros, así como del parámetro de regularización.

4.2.3 Naive Bayes

Naive Bayes es un clasificador probabilístico basado en el teorema de Bayes que estima la probabilidad de pertenencia de una observación a cada clase a partir de sus características [25]. Para simplificar el cálculo, el modelo asume que cada característica contribuye de forma independiente a la predicción de la clase, dado su valor.

A pesar de esta suposición simplificadora, Naive Bayes suele ofrecer un rendimiento competitivo en la práctica, especialmente en problemas con un gran número de características. Además, destaca por su bajo coste computacional y su rapidez de entrenamiento. En este trabajo se emplea la variante Gaussian Naive Bayes, adecuada cuando las características continuas pueden aproximarse mediante una distribución normal dentro de cada clase [26].

4.2.4 Random Forest

Random Forest es un método de aprendizaje conjunto que combina múltiples árboles de decisión entrenados sobre diferentes subconjuntos de los datos obtenidos mediante muestreo Bootstrap [22]. Además, en cada división de los árboles se selecciona aleatoriamente un subconjunto de características, lo que introduce diversidad entre los modelos individuales y reduce el riesgo de sobreajuste.

Las predicciones finales se obtienen mediante votación mayoritaria entre los árboles del conjunto. Random Forest es capaz de modelar relaciones no lineales complejas, manejar conjuntos de datos de alta dimensión y mantener un buen rendimiento en presencia de ruido. Asimismo, el algoritmo permite estimar la importancia de las características a partir de su contribución a la reducción del error de clasificación.

4.2.5 XGBoost

XGBoost (Extreme Gradient Boosting) es un algoritmo de aprendizaje conjunto basado en árboles de decisión entrenados de forma secuencial mediante gradient boosting [27]. Cada nuevo árbol se ajusta para corregir los errores residuales cometidos por el conjunto de modelos anteriores, mejorando progresivamente la precisión del clasificador.

Este algoritmo incorpora diversas optimizaciones orientadas a mejorar tanto el rendimiento como la eficiencia computacional, entre las que destacan la regularización explícita para evitar el sobreajuste, la paralelización del proceso de entrenamiento y el uso de información de segundo orden para una optimización más precisa. Además, XGBoost gestiona de forma eficiente los valores perdidos y está diseñado para trabajar con conjuntos de datos de gran tamaño.

4.3 Selección de hiperparámetros

Dos pasos importantes en el desarrollo de modelos de aprendizaje automático son la optimización de hiperparámetros y la validación de modelos. La Tabla 4.1 resume la configuración de hiperparámetros evaluados en busca de la combinación óptima.

Tienen un gran impacto en la capacidad del modelo para generalizar a datos nuevos y desconocidos. La validación, por otro lado, evalúa el rendimiento del modelo más allá del conjunto de entrenamiento, asegurando que sus predicciones sigan siendo precisas y fiables cuando se aplican a datos independientes.

Estos dos pasos están estrechamente relacionados: la optimización de los hiperparámetros sin una validación adecuada puede dar lugar a un sobreajuste o un subajuste, mientras que la validación sin ajuste puede dar lugar a un rendimiento subóptimo del modelo.

En este estudio, ambos procesos se implementan utilizando la clase GridSearchCV de la biblioteca scikit-learn de Python [28], que realiza una búsqueda sistemática sobre una cuadrícula de combinaciones predefinidas y permite estimar el rendimiento del modelo mediante validación cruzada.

Para cada clasificador, se realizó una búsqueda sistemática en cuadrícula para identificar la combinación de hiperparámetros que ofrecía el mejor rendimiento en términos de F1-score ponderado (weighted). Este proceso se llevó a cabo exclusivamente sobre el conjunto de entrenamiento, empleando validación cruzada StratifiedGroupKFold, que permite preservar la distribución de las clases y garantizar la independencia entre grupos (epochs). La búsqueda se realizó de forma independiente para cada participante, seleccionando la configuración óptima de cada modelo a nivel individual.

Tabla 4.1. Hiperparámetros explorados con GridSearchCV en la clasificación de Trust binario.

Modelos	**Clasificación**
KNN	- n_neighbors: [3, 5, 7] - weights: ["uniform", "distance"] - metric: ["euclidean", "manhattan"]
SVM	- kernel: ["rbf", "linear"] - C: [0.001, 0.01, 0.1, 1] - gamma: ["scale", "auto"]

Naïve Bayes	- var_smoothing: [1e-12, 1e-10, 1e-8, 1e-6, 1e-5]
Random Forest	- n_estimators: [500, 800, 1000] - max_depth: [12, 14, 16] - min_samples_leaf: [2, 4] - max_features: ["sqrt"]
XGBoost	- n_estimators: [100, 200] - max_depth: [3, 5] - learning_rate: [0.01, 0.1] - subsample: [0.8, 1.0]

4.3.1 Modelización intra-sujeto

Este procedimiento prioriza un enfoque de modelización intra-sujeto, permitiendo capturar las particularidades de los datos de cada participante. En consecuencia, las estrategias descritas en esta sección se aplican de forma independiente a cada individuo, optimizando y validando los hiperparámetros de cada modelo de manera específica para cada participante, con el fin de adaptar la configuración del modelo a los patrones característicos de sus señales.

Una vez evaluadas todas las combinaciones de hiperparámetros para cada participante, se definió adicionalmente una estrategia de selección global basada en el rendimiento medio inter-sujeto. En concreto, para cada combinación de hiperparámetros se calcularon los valores medios del F1-score ponderado obtenidos en validación cruzada a través de todos los participantes, seleccionándose como configuración global aquella que maximizó dicho promedio.

Finalmente, los modelos individuales se reentrenaron utilizando esta configuración global de hiperparámetros, estableciendo una parametrización común que se empleó posteriormente en los análisis de explicabilidad del modelo.

4.4 Métricas de evaluación

Tras el proceso de entrenamiento y optimización de los modelos de aprendizaje automático, se evaluó su rendimiento utilizando las etiquetas de referencia en su codificación binaria. Dado que el problema abordado corresponde a una tarea de clasificación, se emplearon métricas estándar ampliamente utilizadas en la literatura para evaluar la calidad de los modelos.

Entre las métricas consideradas se incluyen la accuracy, el F1-score y la balanced accuracy, que permiten obtener una visión complementaria del rendimiento del clasificador, especialmente en contextos donde puede existir desequilibrio entre clases [22].

Accuracy mide la proporción de predicciones correctas con respecto al número total de muestras y se define como:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Donde TP y TN representan los verdaderos positivos y verdaderos negativos, respectivamente, y FP y FN los falsos positivos y falsos negativos.

El F1-score combina las métricas de precision y recall en una única medida, proporcionando una evaluación equilibrada del rendimiento del modelo cuando existe un compromiso entre ambas:

$$Precision = \frac{TP}{TP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 - Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Esta métrica resulta especialmente útil cuando el conjunto de datos presenta desequilibrio entre clases, ya que penaliza tanto los falsos positivos como los falsos negativos.

Adicionalmente, se empleó la balanced accuracy, que tiene en cuenta el rendimiento del modelo en cada clase de forma independiente, asignando el

mismo peso a todas ellas. Esta métrica se define como la media de las tasas de verdaderos positivos de cada clase:

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

El uso conjunto de estas métricas permite evaluar el rendimiento de los modelos desde distintas perspectivas, proporcionando una visión más completa de su comportamiento. Las definiciones y formulaciones empleadas siguen las utilizadas de forma estándar en la literatura de aprendizaje automático [22].

5 SHAP análisis de interpretabilidad

Este capítulo presenta el marco teórico y metodológico del análisis de interpretabilidad basado en SHAP (SHapley Additive exPlanations) aplicado en este trabajo. En esta sección se describe el uso del método SHAP, su funcionamiento conceptual y su aplicación en modelos basados en árboles, así como los distintos enfoques de interpretación local y global empleados. Los resultados obtenidos y las representaciones gráficas asociadas se presentan en el apartado de Resultados.

Con el fin de interpretar el comportamiento de los modelos Random Forest y XGBoost, se aplicó el método de explicabilidad SHAP (SHapley Additive exPlanations), que permite analizar la contribución de las distintas variables en la predicción de la confianza. Esta técnica permite desglosar las predicciones del modelo y analizar la importancia de cada variable en el proceso de toma de decisiones [29]. En este estudio, se utilizó SHAP para estudiar la relevancia del EEG y sus características extraídas para la clasificación binaria.

El análisis de interpretabilidad se centró en los modelos Random Forest y XGBoost debido a su naturaleza basada en árboles, lo que permite una aplicación directa y eficiente del método SHAP mediante el uso de TreeExplainer.

5.1 Uso del SHAP

En el campo de la inteligencia artificial explicable (XAI), se han propuesto diversos métodos para comprender el funcionamiento de modelos de aprendizaje automático cuyo proceso interno de toma de decisiones no resulta directamente interpretable, y que por ello suelen denominarse modelos de “caja negra” [30]. En este trabajo se ha elegido SHAP (SHapley Additive exPlanations) para interpretar las predicciones de los clasificadores empleados, en particular Random Forest y XGBoost.

SHAP asigna a cada característica una medida de importancia basada en los valores de Shapley de la teoría de juegos cooperativos [31], lo que proporciona una base teórica sólida y consistente para la interpretabilidad. A diferencia de otras técnicas como la importancia de características clásica o métodos locales como LIME [32], SHAP proporciona explicaciones más coherentes con el comportamiento real del modelo, ya que asigna a cada característica una contribución proporcional a su impacto en las predicciones del clasificador. Además, las explicaciones obtenidas mediante SHAP son estables, al no

depender de procesos aleatorios, y permiten analizar el efecto conjunto de varias variables sobre la predicción.

5.2 Funcionamiento conceptual de SHAP

Desde un punto de vista conceptual, SHAP permite explicar una predicción individual del modelo descomponiéndola en un valor base y en las contribuciones de cada característica. El valor base representa la predicción media del clasificador, es decir, el nivel de confianza esperado antes de considerar la información específica de una muestra concreta [31]. A partir de este valor, cada característica extraída del EEG aporta una contribución positiva o negativa que ajusta la predicción final del modelo.

A partir del valor base, determinadas características del EEG pueden contribuir a aumentar o disminuir la probabilidad de predecir un nivel de confianza alto. Por ejemplo, una mayor potencia en la banda gamma en regiones frontales puede contribuir positivamente a aumentar la probabilidad de confianza alta, mientras que una menor potencia en la banda alfa en regiones parietales puede contribuir a reducir dicha probabilidad, favoreciendo la predicción de un nivel de confianza más bajo. La combinación de estas contribuciones permite reconstruir la predicción final del modelo para una instancia concreta.

En este contexto, el signo del valor SHAP indica la dirección de la influencia de cada característica sobre la predicción del modelo: valores positivos empujan la salida del clasificador hacia niveles de confianza más altos, mientras que valores negativos la desplazan hacia niveles de confianza más bajos. Por su parte, la magnitud del valor SHAP refleja la importancia de la característica, de modo que valores cercanos a cero indican una influencia reducida en la decisión del modelo. Esta formulación aditiva facilita la interpretación del comportamiento del clasificador, al permitir identificar de forma clara qué bandas de frecuencia y regiones cerebrales influyen más en la estimación de la confianza del usuario [31][33].

5.3 SHAP en modelos basados en árboles

Otra razón para emplear SHAP en este estudio es su facilidad de integración dentro del flujo de trabajo de aprendizaje automático implementado en Python. En este trabajo, los modelos se optimizaron previamente mediante validación cruzada y búsqueda en cuadrícula, y posteriormente se entrenaron modelos

finales con los hiperparámetros seleccionados, que fueron los utilizados para el análisis de interpretabilidad.

SHAP resulta especialmente adecuado para modelos basados en árboles de decisión, como Random Forest y XGBoost, ya que puede aprovechar su estructura interna para calcular de forma eficiente la contribución de cada característica a la predicción del modelo. En particular, el método TreeExplainer estima los valores SHAP considerando cómo las variables influyen en las decisiones del clasificador a lo largo de los distintos árboles que lo componen [34].

En este estudio, el análisis SHAP se aplicó sobre los modelos finales entrenados, lo que permite una interpretación directa de la contribución de cada característica del EEG. De este modo, SHAP no solo proporciona una medida de importancia global de las variables, sino que permite analizar su impacto en predicciones individuales, ofreciendo una interpretación coherente del comportamiento de los modelos empleados.

5.4 Interpretación global y local mediante SHAP

En comparación con las medidas clásicas de importancia de características, los valores SHAP proporcionan una interpretación más completa del comportamiento del modelo. Las técnicas tradicionales, como la reducción de impureza en árboles de decisión o la importancia por permutación, ofrecen únicamente un ranking global de variables relevantes, sin informar sobre cómo ni en qué sentido cada característica influye en la predicción [35].

Por el contrario, SHAP permite analizar el impacto real de cada característica sobre la salida del modelo, indicando tanto la magnitud como la dirección de su influencia. Además, este enfoque tiene en cuenta el efecto conjunto de las variables, distribuyendo de forma coherente su contribución incluso cuando existen interacciones o correlaciones entre ellas [31]. A través de representaciones como el summary plot, es posible observar cómo una misma característica puede contribuir a aumentar o reducir la predicción en diferentes instancias, en función de su valor y del contexto del resto de variables.

De este modo, SHAP combina la interpretación local de predicciones individuales con una visión global del modelo, permitiendo obtener explicaciones consistentes a distintos niveles de análisis [30]. Esta capacidad facilita una comprensión más fiable del funcionamiento del clasificador y reduce el riesgo de interpretaciones erróneas basadas únicamente en medidas globales de importancia.

Considerando las ventajas descritas, en este estudio se empleó SHAP para analizar la contribución de las señales EEG y de sus características derivadas en la predicción del nivel de confianza del usuario. El análisis de interpretabilidad se centró en los modelos Random Forest y XGBoost, ambos basados en árboles de decisión, y se llevó a cabo de forma independiente para cada contexto experimental: la sesión neutral y la sesión ecológica. Este enfoque permitió evaluar posibles diferencias en el comportamiento del modelo en función del entorno de la tarea.

Asimismo, los resultados obtenidos mediante SHAP se analizaron tanto a nivel individual, proporcionando explicaciones específicas para cada participante, como a nivel global, mediante la agregación de las contribuciones de todos los sujetos. En los apartados siguientes se presentan los principales resultados de interpretabilidad, comenzando por el análisis individual y, posteriormente, por las tendencias globales del modelo.

5.5 Análisis SHAP individual por participante

Se utilizaron dos tipos de representaciones gráficas para el análisis individual: el gráfico beeswarm SHAP (gráfico de enjambre de abejas) y el gráfico de barras SHAP. Estas visualizaciones permiten identificar, para cada participante, las variables que tienen mayor influencia en la predicción del modelo, así como la dirección de dicha influencia.

El gráfico resumen SHAP permite visualizar cómo cada característica influye en las predicciones del modelo. Las variables se ordenan según su importancia global, y cada punto corresponde a una muestra del conjunto de pruebas. La posición a lo largo del eje horizontal refleja la dirección y la magnitud del impacto de la característica en el resultado del modelo, mientras que el color indica el valor relativo de la característica. Esta representación facilita la interpretación de la relación entre los valores de las características y los cambios en la predicción del modelo.

Por su parte, el gráfico de barras ofrece una representación más compacta al resumir la importancia de cada característica a través del valor SHAP absoluto medio. Esto permite una comparación directa de las variables más influyentes entre participantes y entre condiciones experimentales.

5.6 Análisis SHAP global

Con el fin de obtener una perspectiva general del comportamiento de los modelos, se agregaron las contribuciones SHAP de todos los participantes para cada condición experimental. Para ello, se calculó en primer lugar, para cada participante, la importancia media de cada característica como el valor SHAP absoluto medio, promediado sobre todas las instancias del conjunto de prueba. Posteriormente, estas importancias se promediaron entre participantes, asignando el mismo peso a cada sujeto.

Este procedimiento permitió obtener una medida de importancia global de las características, independiente de las variaciones individuales. De este modo, fue posible identificar qué variables del EEG resultaron más relevantes, en promedio, para la predicción de la confianza y analizar si determinadas bandas de frecuencia o tipos de características aportan información de forma consistente al modelo.

Adicionalmente, con el objetivo de facilitar la interpretación de los resultados, se realizó un análisis agrupando las características del EEG por regiones cerebrales. En concreto, las variables se categorizaron según la región cortical del electrodo asociado (frontal, temporal, central, parietal u occipital), y la importancia SHAP por región se calculó como la suma de las importancias medias de las características asociadas a cada una. Esta agrupación permitió trasladar los resultados del modelo a un nivel de análisis regional, facilitando su interpretación.

El análisis se realizó de forma independiente para los modelos Random Forest y XGBoost. Aunque es posible comparar entre modelos las tendencias generales observadas, como las bandas de frecuencia o regiones cerebrales más relevantes, no se comparan directamente los valores numéricos de las importancias SHAP, ya que SHAP asigna las contribuciones en función de la estructura y el funcionamiento interno de cada modelo.

6 Resultados

En este capítulo se presentan los principales resultados obtenidos en el Trabajo Fin de Grado, así como las conclusiones derivadas del análisis realizado. En particular, se analizan los resultados de los modelos de aprendizaje automático supervisado empleados para la estimación del nivel de confianza del usuario a partir de señales EEG, y se evalúa la contribución de las distintas características mediante la explicabilidad a través de SHAP.

6.1 Resultados del Aprendizaje Supervisado

Este apartado presenta los resultados obtenidos a partir de los modelos de aprendizaje automático supervisado desarrollados para estimar los niveles de confianza del usuario. En concreto, se aborda un esquema de clasificación binaria, con el objetivo de analizar la capacidad de los modelos para discriminar entre niveles de confianza bajos y altos en función del contexto experimental.

Los experimentos se realizaron de forma independiente para las condiciones neutral y ecológica, permitiendo evaluar el impacto del aumento de la carga cognitiva y de la complejidad del entorno sobre el rendimiento de los clasificadores.

6.1.1 Mejores hiperparámetros

Para cada uno de los modelos considerados se llevó a cabo un proceso de optimización de hiperparámetros mediante GridSearchCV, aplicado de forma independiente a cada participante.

A partir de los resultados obtenidos, los valores de F1-score ponderado asociados a una misma combinación de hiperparámetros se promediaron entre participantes, seleccionándose como configuración global aquella que presentó el mayor rendimiento medio inter-sujeto.

Los hiperparámetros globales óptimos seleccionados para cada uno de los clasificadores se recogen en la Tabla 6.1, donde se muestran de forma conjunta las configuraciones correspondientes a las fases neutral y ecológica. En dicha tabla se incluyen los cinco clasificadores evaluados: K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Naïve Bayes, Random Forest y XGBoost.

Tabla 6.1. Mejores hiperparámetros óptimos seleccionados para los modelos de clasificación en las condiciones Neutral y Ecológica.

Modelo	Clasificación Neutral	Clasificación Ecologica
KNN	- n_neighbours: 3 - weights: "uniform" - metric: "manhattan"	- n_neighbours: 3 - weights: "uniform" - metric: "manhattan"
SVM	- kernel: "linear" - C: 0.01 - gamma: "scale"	- kernel: "linear" - C: 1 - gamma: "scale"
Naïve Bayes	- var_smoothing: 1e-12	- var_smoothing: 1e-12
Random Forest	- n_estimators: 1000 - max_depth: 12 - min_samples_leaf: 4 - max_features: "sqrt"	- n_estimators: 500 - max_depth: 14 - min_samples_leaf: 4 - max_features: "sqrt"
XGBoost	- n_estimators: 200 - max_depth: 3 - learning_rate: 0.1 - subsample: 0.8	- n_estimators: 100 - max_depth: 3 - learning_rate: 0.1 - subsample: 1.0

De forma general, se observa que los modelos basados en árboles (Random Forest y XGBoost) tienden a requerir una mayor capacidad de modelado, reflejada en un número elevado de estimadores y profundidades moderadas, mientras que los modelos más simples (KNN, SVM y Naïve Bayes) presentan configuraciones más conservadoras.

6.1.2 Métricas: Accuracy, F1-score, ROC-AUC

Con el objetivo de evaluar el rendimiento de los clasificadores, se planteó inicialmente un problema de clasificación binaria, diferenciando entre niveles de confianza bajos y altos. Tal y como se describió en el Capítulo 4, se evaluaron cinco modelos de aprendizaje automático supervisado: KNN, SVM, Naïve Bayes, Random Forest y XGBoost.

Los valores de accuracy (Acc) y F1-score (F1) presentados en las Tablas 6.2 y 6.3 corresponden a la media del rendimiento en el conjunto de test, calculada sobre todos los participantes, empleando en cada caso el conjunto de hiperparámetros globales óptimos previamente seleccionado para cada modelo y condición experimental.

Condición Neutral

Tabla 6.2. Rendimiento de los clasificadores binarios de confianza en la fase neutral.

Classifier	Acc	F1-score
KNN	0.609	0.473
SVM	0.619	0.555
Naïve Bayes	0.512	0.405
Random Forest	0.680	0.664
XGBoost	0.672	0.516

Con el fin de ilustrar la distribución del rendimiento a nivel individual y contextualizar los valores medios reportados, las Figuras 6.1, 6.2, 6.3, 6.4 y 6.5 muestran los resultados obtenidos por cada participante para los distintos clasificadores en la condición neutral. En dichas figuras, las barras representan el rendimiento individual, mientras que las líneas horizontales indican las medias globales inter-sujeto.

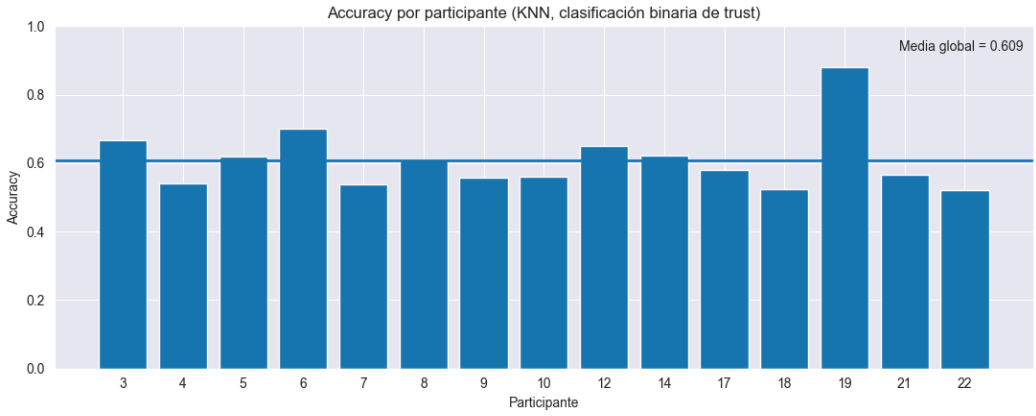

Figura 6.1. Accuracy por participante del clasificador XGBoost en la clasificación binaria de la confianza en la condición neutral. La línea horizontal indica la media global inter-sujeto.

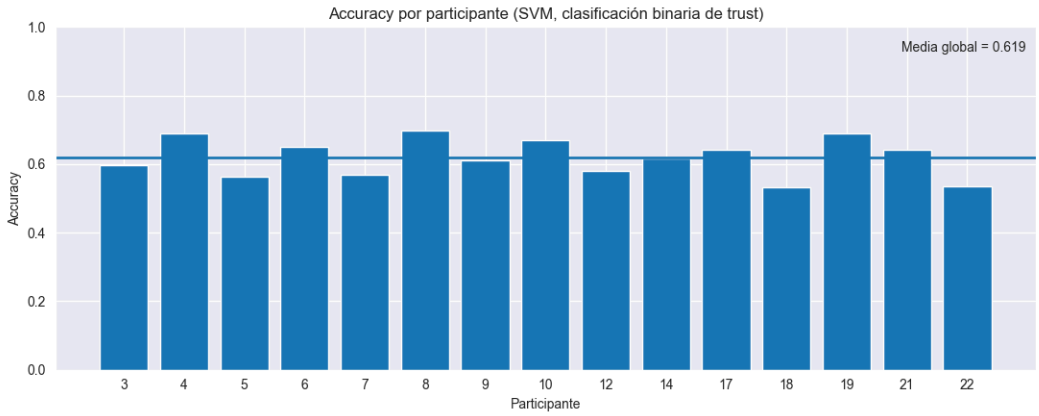

Figura 6.2. Accuracy por participante del clasificador SVM en la clasificación binaria de la confianza en la condición neutral. La línea horizontal indica la media global inter-sujeto.

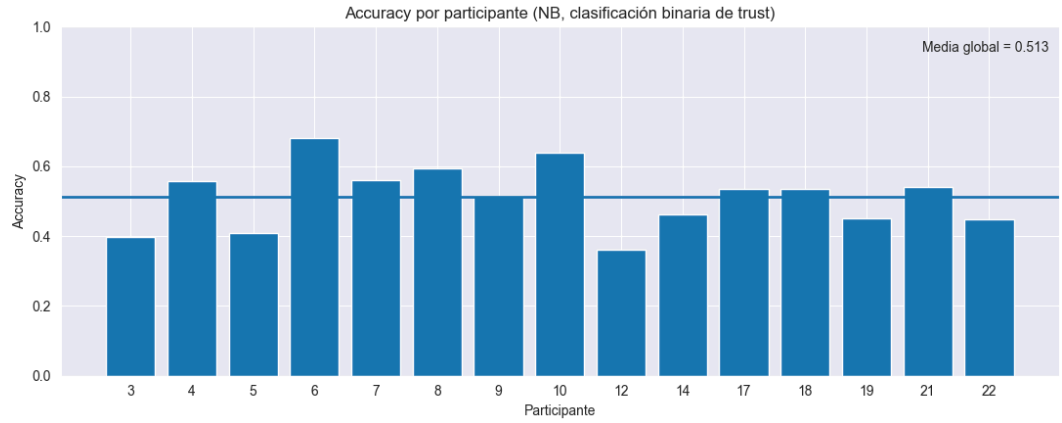

Figura 6.3. Accuracy por participante del clasificador NB en la clasificación binaria de la confianza en la condición neutral. La línea horizontal indica la media global inter-sujeto.

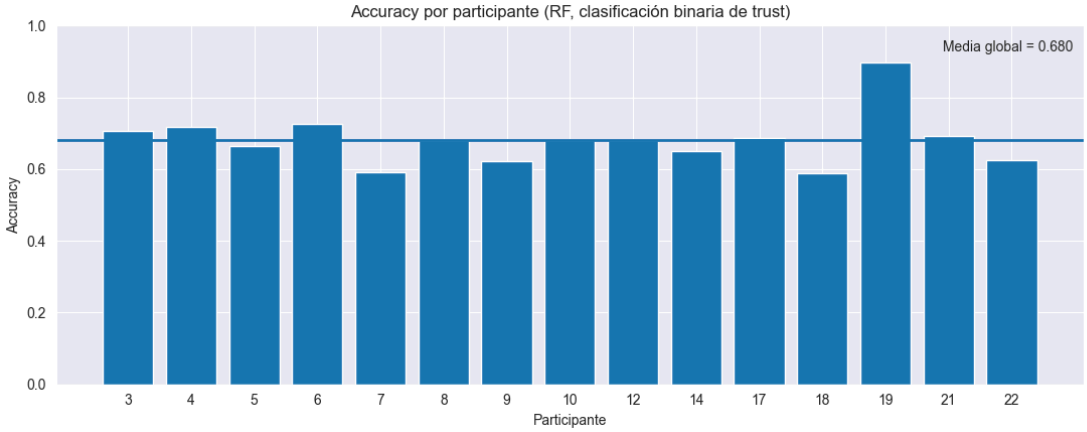

Figura 6.4. Accuracy por participante del clasificador RF en la clasificación binaria de la confianza en la condición neutral. La línea horizontal indica la media global inter-sujeto.

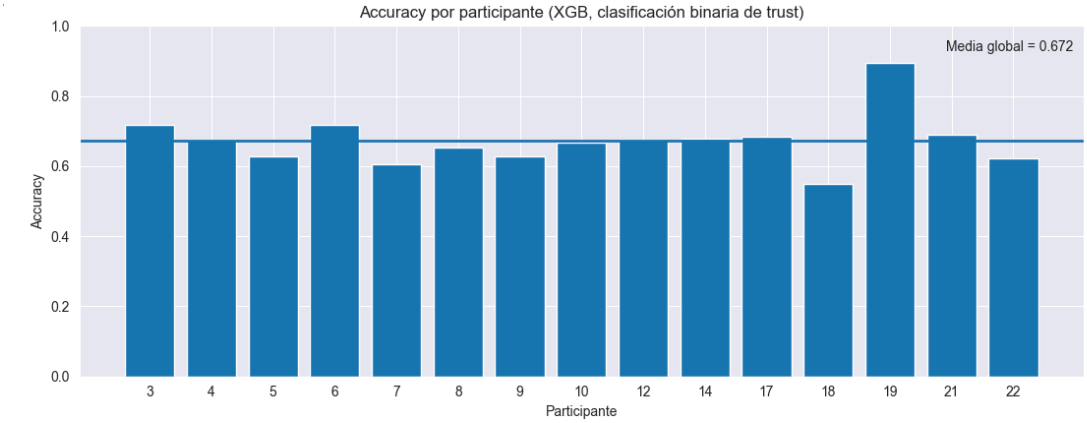

Figura 6.5. Accuracy por participante del clasificador XGB en la clasificación binaria de la confianza en la condición neutral. La línea horizontal indica la media global inter-sujeto.

En la fase neutral, los mejores resultados se obtienen con los modelos basados en árboles. En particular, Random Forest alcanza la mayor precisión (Acc = 0.680) y el F1-score (F1 = 0.664), seguido de XGBoost, con un rendimiento ligeramente inferior. Los modelos KNN y SVM presentan resultados intermedios, mientras que Naïve Bayes muestra el rendimiento más bajo, lo que sugiere que la hipótesis de independencia entre características no se ajusta adecuadamente a la naturaleza de los datos EEG.

Condición Ecológica

Tabla 6.3. Rendimiento de los clasificadores binarios de confianza en la fase ecológica.

Classifier	Acc	F1-score
KNN	0.590	0.519
SVM	0.571	0.513
Naïve Bayes	0.531	0.497
Random Forest	0.631	0.613
XGBoost	0.633	0.547

De forma análoga, las Figuras 6.6, 6.7, 6.8, 6.9 y 6.10 presentan la distribución del rendimiento individual por participante para la condición ecológica. Estas visualizaciones permiten analizar el impacto del entorno experimental más complejo sobre el comportamiento de los clasificadores y la variabilidad inter-sujeto.

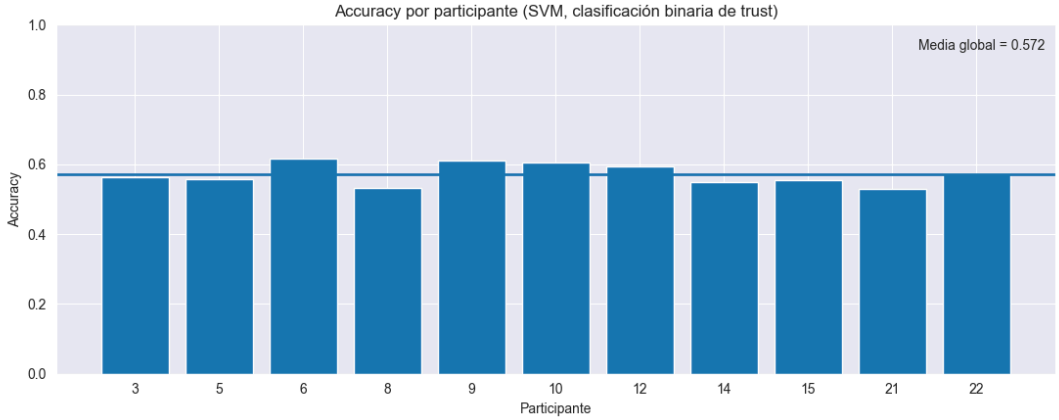

Figura 6.6. Accuracy por participante del clasificador SVM en la clasificación binaria de la confianza en la condición ecológica. La línea horizontal indica la media global inter-sujeto.

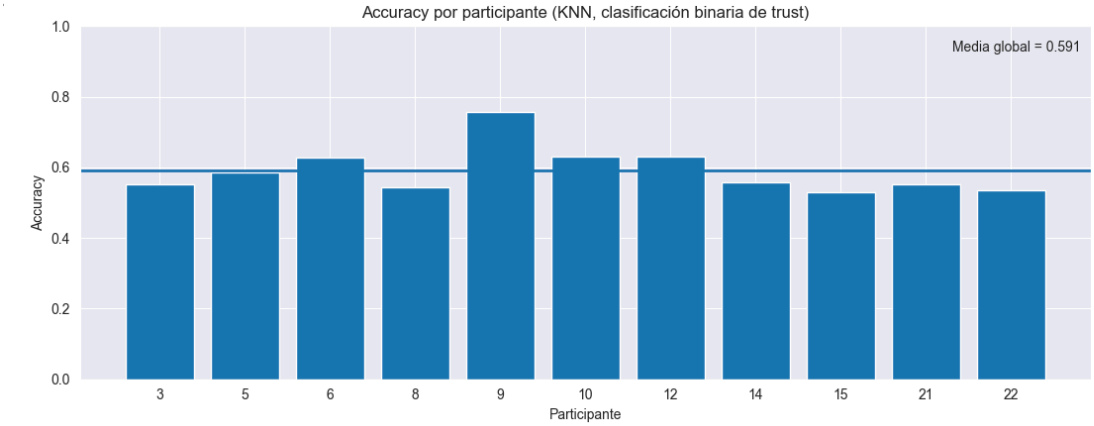

Figura 6.7. Accuracy por participante del clasificador KNN en la clasificación binaria de la confianza en la condición ecológica. La línea horizontal indica la media global inter-sujeto.

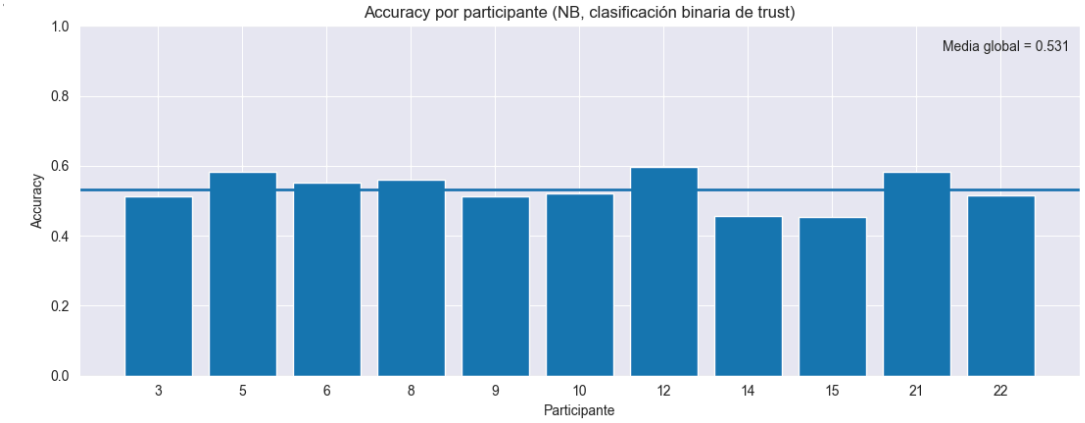

Figura 6.8. Accuracy por participante del clasificador NB en la clasificación binaria de la confianza en la condición ecológica. La línea horizontal indica la media global inter-sujeto.

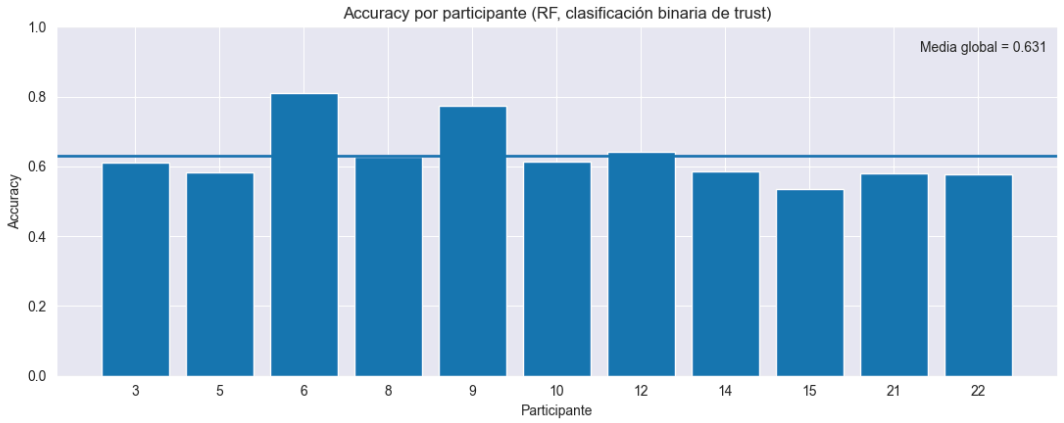

Figura 6.9. Accuracy por participante del clasificador RF en la clasificación binaria de la confianza en la condición ecológica. La línea horizontal indica la media global inter-sujeto.

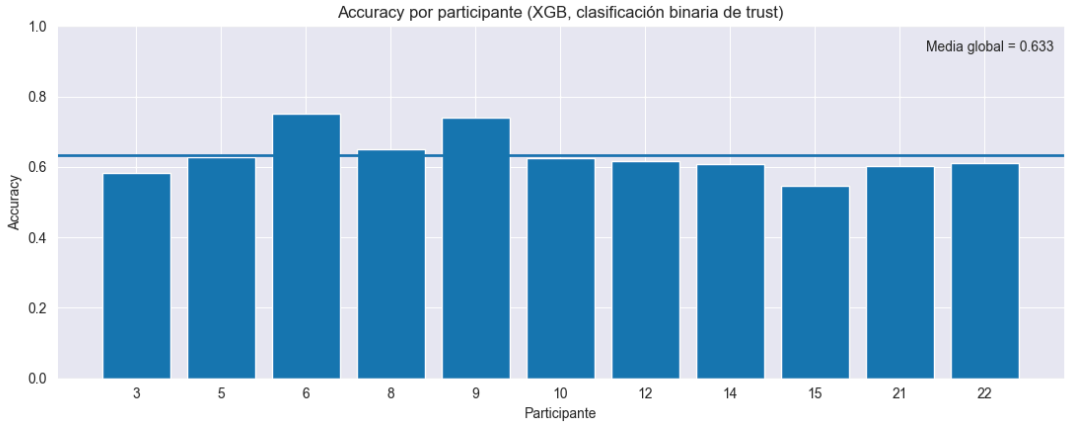

Figura 6.10. Accuracy por participante del clasificador XGB en la clasificación binaria de la confianza en la condición ecológica. La línea horizontal indica la media global inter-sujeto.

En la fase ecológica, el rendimiento global de los modelos disminuye ligeramente, lo cual es coherente con el aumento de la complejidad del entorno experimental y de la carga cognitiva del participante. No obstante, los modelos Random Forest y XGBoost continúan mostrando un comportamiento más robusto que el resto, alcanzando valores de accuracy cercanos a 0.63. De nuevo, Naïve Bayes obtiene los peores resultados relativos.

De forma general, los resultados indican que, aunque los modelos supervisados son capaces de capturar cierta relación entre las características extraídas del EEG y los niveles de confianza del usuario, el rendimiento alcanzado es moderado, lo que pone de manifiesto la dificultad inherente a la estimación de estados cognitivos complejos a partir de EEG, especialmente en entornos más realistas.

6.2 Interpretación de resultados mediante SHAP

En esta sección se presentan los resultados del análisis de interpretabilidad realizado mediante SHAP (SHapley Additive exPlanations), aplicado a los modelos supervisados seleccionados. El objetivo de este análisis es comprender cómo las distintas características extraídas del EEG contribuyen a las

decisiones de los clasificadores, proporcionando una interpretación complementaria a las métricas de rendimiento obtenidas.

Es importante tener en cuenta que los valores de rendimiento alcanzados en el aprendizaje supervisado son moderados, especialmente en la condición ecológica. No obstante, este nivel de desempeño es habitual en problemas de estimación de estados cognitivos complejos a partir de señales EEG. En este contexto, el análisis SHAP se utiliza con un enfoque exploratorio, no para evaluar el rendimiento del modelo, sino para analizar qué características del EEG contribuyen de forma más consistente a sus predicciones.

En los apartados siguientes se presentan, en primer lugar, los resultados del análisis SHAP a nivel individual, que permiten explorar la variabilidad entre participantes, y posteriormente los resultados a nivel global, obtenidos al combinar la información de todos los participantes para identificar tendencias comunes en el comportamiento de los modelos.

6.2.1 Resultados SHAP individual

En este apartado se presentan los resultados del análisis SHAP a nivel individual, con el objetivo de analizar cómo contribuyen las distintas características del EEG a la predicción de la confianza en participantes concretos. Para ello, se muestran ejemplos representativos que permiten ilustrar el comportamiento de los modelos y facilitar la interpretación de los valores SHAP.

Condición Neutral

Las Figuras 6.11 y 6.12 muestran los resultados del análisis SHAP correspondientes al participante 22 en la condición Neutral, utilizando los modelos Random Forest y XGBoost, respectivamente. En la Figura 6.11 se presenta el summary plot (beeswarm) del modelo Random Forest, mientras que la Figura 6.12 muestra el bar plot de importancia media de las características obtenido con XGBoost. Este participante se seleccionó como ejemplo ilustrativo al presentar un patrón de contribuciones estable y fácilmente interpretable.

En ambos modelos, las características con mayor impacto se concentran de forma predominante en la banda low gamma. La relevancia no se limita a la densidad espectral de potencia, sino que incluye métricas que describen la variabilidad temporal de la señal, como la amplitud peak to peak, la desviación estándar y los parámetros de Hjorth. Este patrón indica que el modelo tiene en

cuenta tanto la intensidad media de la actividad EEG como su dinámica temporal.

Además, el summary plot Figura 6.11, permite analizar la dirección de las contribuciones de las características más relevantes. En este caso, se observa un patrón consistente en el que valores elevados de las métricas dominantes, principalmente en la banda low gamma, tienden a asociarse con valores SHAP positivos, mientras que valores bajos contribuyen de forma negativa a la salida del modelo. Este comportamiento sugiere que el clasificador aprende relaciones estables entre la magnitud de estas características y su influencia en la predicción

Desde el punto de vista espacial, las características más relevantes se asocian principalmente a canales parietales y temporales, destacando ch16 (P4), ch14 (P3) y ch17 (T6), junto con contribuciones adicionales de canales frontales como ch7 (F8) y ch4 (F3). Esta distribución sugiere que, en la condición Neutral, la información relacionada con la confianza se apoya en regiones laterales y parietales, con una participación frontal más moderada

La coincidencia entre el summary plot y el bar plot muestra que las características aparecen entre las más relevantes en ambos modelos. Asimismo, Random Forest y XGBoost asignan importancia a bandas, métricas y canales similares en especial los más relevantes, reflejando un comportamiento comparable en este participante. Este ejemplo permite ilustrar de forma clara el análisis SHAP individual en la condición Neutral y sirve como referencia para el análisis global posterior.

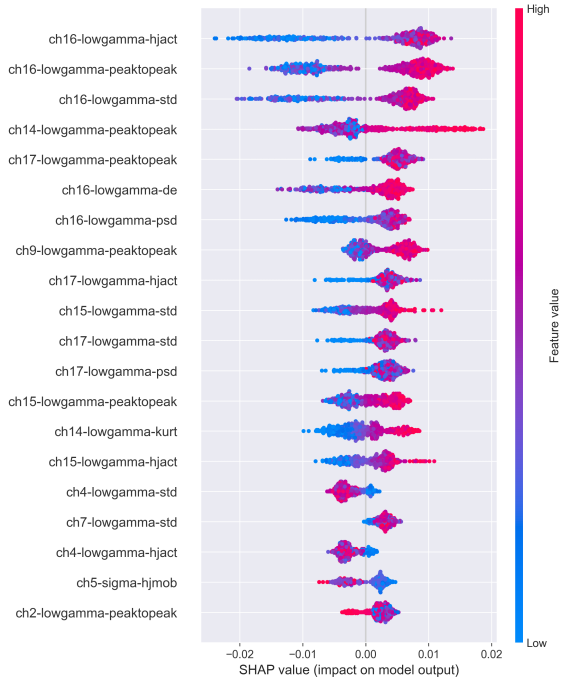

Figura 6.11. SHAP summary plot para el modelo Random Forest en la condición Neutral (participante 22).

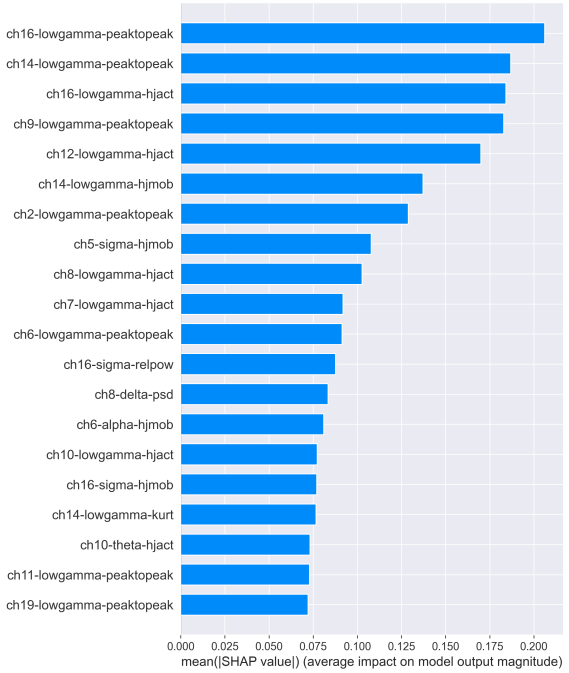

Figura 6.12. SHAP bar plot de importancia de características para el modelo XGBoost en la condición Neutral (participante 22).

Condición Ecológica

Para ilustrar el análisis de interpretabilidad mediante SHAP en la condición ecológica, se seleccionó el participante 10 como ejemplo representativo. Al igual que en la condición neutral, se evaluaron todos los participantes de forma individual para ambos modelos (Random Forest y XGBoost). Sin embargo, en la condición ecológica se observó una mayor heterogeneidad inter-sujeto, con patrones de contribución más diversos y, en muchos casos, menos consistentes entre participantes. Por este motivo, se escogió el participante 10, ya que presenta un patrón de importancias claro y relativamente estable, que permite ilustrar de forma comprensible el funcionamiento de los modelos en este contexto más complejo.

Las Figuras 6.13 y 6.14 muestran, respectivamente, el summary plot de SHAP para el modelo Random Forest y el bar plot de importancias medias de SHAP para el modelo XGBoost correspondientes al participante 10. En ambos modelos se observa una contribución dominante de características asociadas a la banda low gamma, en coherencia con los resultados obtenidos en la condición neutral y con el análisis global del conjunto de participantes.

En el modelo Random Forest, las características más influyentes se concentran principalmente en la banda low gamma y están asociadas, sobre todo, a métricas de variabilidad y complejidad de la señal, como la desviación estándar (std), la actividad de Hjorth (hjact), la entropía diferencial (de) y la densidad espectral de potencia (psd). Estas variables aparecen de forma recurrente en canales frontales y centrales, especialmente en ch6 (F4) y ch4 (F3), lo que indica un peso relevante de la actividad frontal en la predicción de la confianza para este participante. De manera complementaria, también se observan contribuciones de características en otras bandas, como alpha (peak-to-peak, relpow) y sigma (relpow), aunque con una importancia menor en comparación con low gamma.

El summary plot en la Figura 6.13, muestra además un patrón interpretable en la dirección de las contribuciones, aunque más heterogéneo que en la condición neutral. En general, valores elevados de las características más influyentes en la banda low gamma tienden a asociarse con valores SHAP negativo, mientras que valores bajos contribuyen de forma ositiva a la predicción del modelo. No obstante, este comportamiento no es uniforme para todas las métricas ni canales, lo que refleja una relación más compleja entre las características del EEG y la salida del clasificador en la condicion ecológica.

Por su parte, el modelo XGBoost presenta un patrón de contribuciones más diverso, tanto en términos de métricas como de bandas de frecuencia. Junto a características de la banda low gamma, aparecen con un peso relevante variables de la banda alpha, como peak-to-peak, potencia relativa y skewness, así como métricas de complejidad asociadas a las bandas beta y sigma. Este comportamiento sugiere que, en la condición ecológica, el modelo integra información procedente de distintos descriptores del EEG para capturar la variabilidad asociada a la confianza.

Desde el punto de vista espacial, las características más influyentes no se concentran exclusivamente en regiones frontales. Además de canales frontales como ch6 (F4) y ch4 (F3), se observan contribuciones relevantes en otras regiones, ch17 (T6) y ch18 (O1). Esta distribución espacial más amplia refuerza la idea de que, bajo condiciones ecológicas, la actividad cerebral relacionada con la confianza presenta una organización más distribuida en la parte posterior, en comparación con la condición neutral.

En conjunto, el participante 10 constituye un ejemplo adecuado para ilustrar el análisis SHAP en la condición ecológica, ya que refleja tanto la persistencia de la banda low gamma como principal fuente de información, como el aumento de la diversidad en las métricas, bandas y regiones implicadas. Este análisis pone de manifiesto la utilidad de SHAP para interpretar el comportamiento de los modelos a nivel individual y para evidenciar las diferencias en los patrones de contribución entre condiciones experimentales con distintos grados de complejidad.

Figura 6.13. SHAP summary plot para el modelo Random Forest en la condición Ecológica (participante 10).

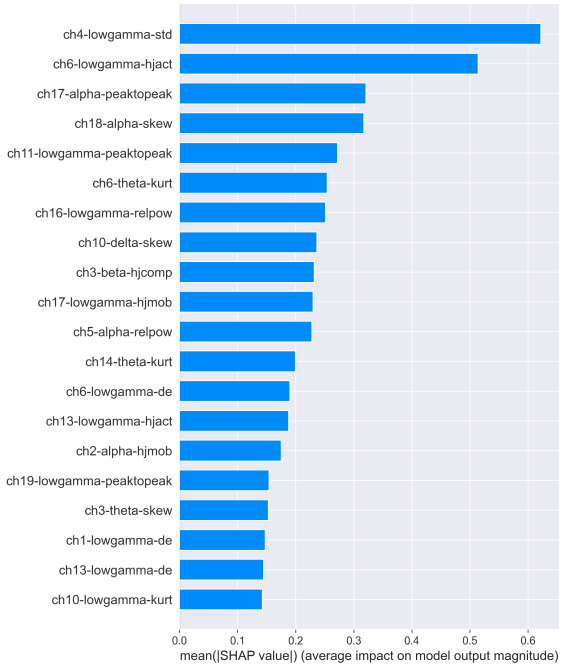

Figura 6.14. SHAP bar plot de importancia de características para el modelo XGBoost en la condición Ecológica (participante 10).

6.2.2 Resultados SHAP global

A partir del procedimiento descrito en la sección metodológica, se realizó un análisis global de interpretabilidad agregando los valores SHAP de todos los participantes para cada condición experimental (Neutral y Ecológica), de forma independiente para los modelos Random Forest y XGBoost.

Para cada participante, se calculó la importancia media de cada característica como el valor SHAP absoluto medio sobre el conjunto de prueba, y posteriormente estas importancias se promediaron entre participantes, otorgando el mismo peso a cada sujeto.

Este enfoque permitió identificar de manera robusta qué características del EEG aportaron información de forma consistente a la predicción de la confianza, independientemente de la variabilidad individual. Es importante señalar que, dado que Random Forest y XGBoost presentan escalas internas distintas en los valores SHAP, las magnitudes absolutas no son directamente comparables entre modelos; por tanto, el análisis se centra en los patrones relativos de importancia y en la recurrencia de determinadas características, más que en la comparación directa de valores numéricos.

6.2.2.1 Importancia global de características

Condición Neutral

Las Figuras 6.15 y 6.16 muestran la importancia global de las 20 características más relevantes según los valores SHAP para los modelos Random Forest y XGBoost, respectivamente, en la condición Neutral. En ambas figuras se observa un patrón común en cuanto a las bandas de frecuencia y tipos de características que aportan más información al modelo.

En los dos modelos, las características más relevantes se concentran mayoritariamente en la banda de frecuencia low gamma. Destacan especialmente métricas relacionadas con la variabilidad y la dinámica temporal de la señal, como la actividad de Hjorth (hjact), la desviación estándar (std) y, en menor medida, la amplitud pico a pico (peak-to-peak).

En el modelo Random Forest, las características con mayor importancia global se asocian principalmente a canales como ch8 (T3), ch12 (T4), ch13 (T5) y ch18 (O1), que corresponden a regiones temporales y occipitales. Asimismo, aparecen de forma recurrente características asociadas a ch4 (F3), indicando una contribución adicional de regiones frontales. Todas estas variables pertenecen a la banda low gamma y están relacionadas con medidas de actividad y dispersión de la señal.

Por su parte, el modelo XGBoost refuerza este patrón general, situando de forma consistente variables de la banda low gamma entre las más influyentes. En particular, destacan las métricas de actividad de Hjorth (hjact) y desviación estándar (std), concentrándose los mayores valores de importancia en los canales ch8 (T3), ch6 (F4), ch13 (T5), ch4 (F3) y ch18 (O1). Junto a estas, aparecen de manera complementaria otros descriptores de la señal, como la movilidad de Hjorth (hjmob), la curtosis (kurtosis) y el rango pico a pico (peak-to-peak), lo que sugiere que el modelo incorpora información procedente de distintos tipos de características.

En conjunto, estos resultados indican que, en la condición Neutral, la información asociada a la variabilidad y a la dinámica rápida de la señal EEG en la banda low gamma aporta una contribución relevante a la predicción de la confianza. Otras bandas de frecuencia aparecen de forma menos recurrente y con menor peso relativo en el ranking global de ambos modelos.

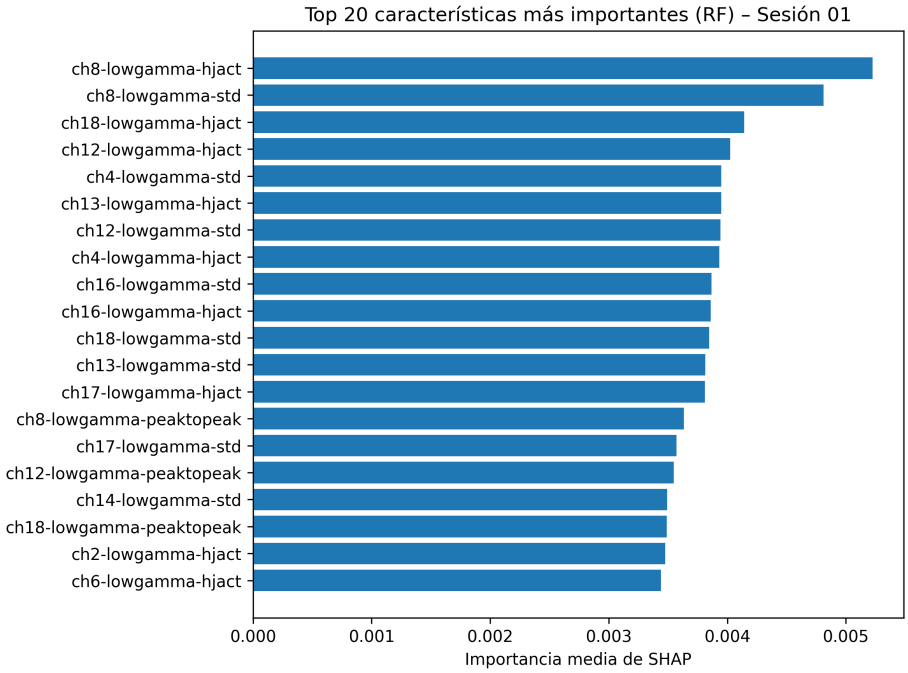

Figura 6.15. Importancia global de las 20 características más relevantes según los valores SHAP para el modelo Random Forest en la condición Neutral.

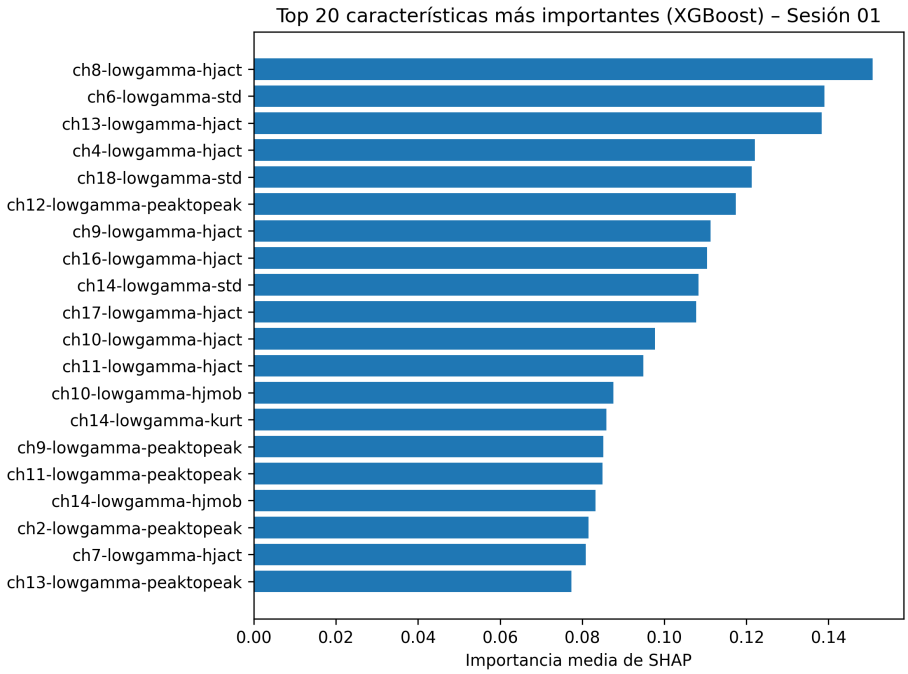

Figura 6.16. Importancia global de las 20 características más relevantes según los valores SHAP para el modelo XGBoost en la condición Neutral.

Condición Ecológica

Las Figuras 6.17 y 6.18 muestran la importancia global de las 20 características más relevantes según los valores SHAP para los modelos Random Forest y XGBoost, respectivamente, en la condición Ecológica. Al igual que en la condición Neutral, en ambos modelos se observa un patrón dominante asociado a la banda de frecuencia low gamma.

En los dos modelos, las características más relevantes pertenecen mayoritariamente a dicha banda y corresponden a métricas relacionadas con la variabilidad, la dispersión y la complejidad de la señal EEG. Entre las variables más influyentes aparecen de forma recurrente la desviación estándar (std), la actividad de Hjorth (hjact) y la amplitud pico a pico (peak-to-peak), así como, en esta condición, otras métricas como la entropía diferencial (de), la curtosis (kurt) y la movilidad de Hjorth (hjmob).

En el caso del modelo Random Forest, las características con mayor importancia se distribuyen entre canales asociados a distintas regiones corticales, todas

ellas vinculadas a la banda low gamma. En particular, el canal ch19 (O2), correspondiente a la región occipital, ocupa los primeros puestos del ranking, seguido de canales como ch9 (C3) en la región central, ch4 (F3) y ch3 (F7) en la región frontal, y ch13 (T5) en la región temporal. Asimismo, se observa la contribución de canales parietales, como ch16 (P4), entre las características más relevantes. Esta distribución sugiere que el modelo Random Forest extrae información relevante de múltiples regiones corticales, sin concentrarse en una única zona del cuero cabelludo.

Por su parte, el modelo XGBoost presenta un patrón parcialmente diferente en la distribución de la importancia de las características, aunque mantiene la relevancia de canales frontales y temporales. Entre las variables más influyentes destacan ch3 (F7), ch7 (F8) y ch6 (F4) en la región frontal, junto con la contribución de canales occipitales como ch19 (O2) y de canales temporales y centrales, como ch13 (T5) y ch9 (C3). Estos resultados indican que, si bien el peso relativo de cada región varía respecto a Random Forest, el modelo XGBoost integra información procedente de múltiples regiones corticales, con una contribución destacada de áreas frontales y temporales en la predicción del estado de confianza.

A pesar de las diferencias observadas en la jerarquía de importancia de las características entre ambos modelos, se identifica un conjunto de regiones que aparece de forma recurrente entre las más relevantes en la sesión ecológica. En concreto, las regiones frontal, temporal y occipital muestran una presencia consistente tanto en Random Forest como en XGBoost, lo que sugiere que la información más discriminativa para la estimación de la confianza se distribuye principalmente entre estas áreas.

En conjunto, la comparación entre la condición Neutral y la condición Ecológica indica que la banda low gamma constituye la principal fuente de información en ambos contextos experimentales. Asimismo, en ambas condiciones se observa la participación recurrente de regiones frontales, temporales y occipitales, sin evidenciarse una concentración clara de la información relevante en una única región del cuero cabelludo.

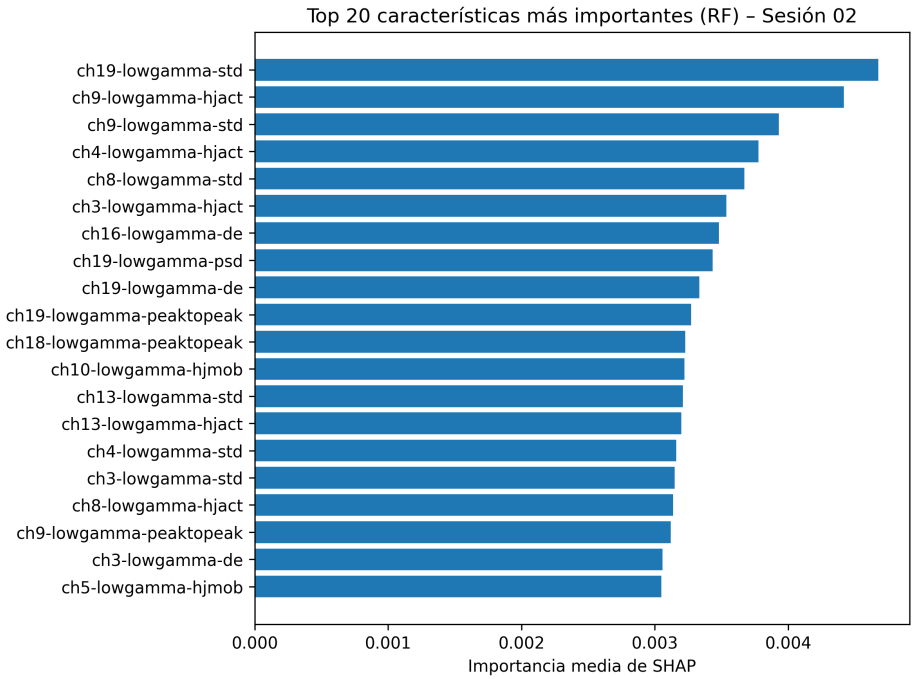

Figura 6.17. Importancia global de las 20 características más relevantes según los valores SHAP para el modelo Random Forest en la condición Ecológica.

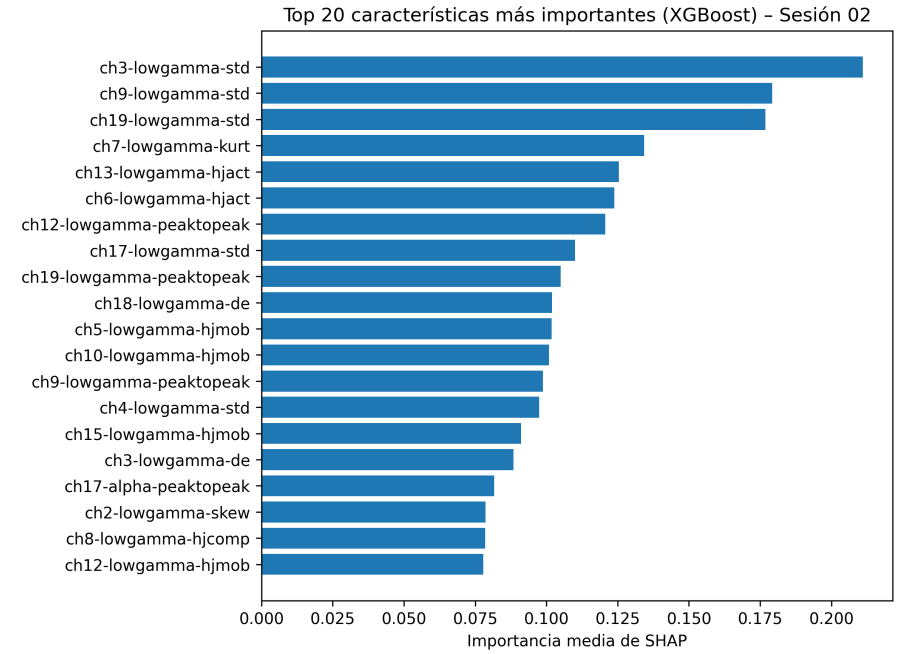

Figura 6.18. Importancia global de las 20 características más relevantes según los valores SHAP para el modelo XGBoost en la condición Ecológica.

6.2.2.2 Resultados del análisis global regional

Para el análisis por regiones cerebrales, la importancia regional se calculó como la suma de las importancias SHAP medias de todas las características asociadas a los electrodos de cada región. Esta medida permite reflejar la contribución total de cada región al modelo.

Este análisis se realizó de forma independiente para los modelos Random Forest y XGBoost y para ambas condiciones experimentales. Dado que los valores SHAP dependen de la estructura interna de cada modelo, la interpretación se centra en la distribución relativa de la importancia entre regiones, y no en la comparación directa de los valores numéricos absolutos entre modelos.

Condición Neutral

Las Figuras 6.19 y 6.20 muestran la importancia total de SHAP por región cerebral en la condición Neutral, correspondiente a los modelos Random Forest y XGBoost, respectivamente.

En ambos modelos se observa un patrón similar en la distribución regional de la importancia. La región frontal presenta la mayor contribución global, seguida de la región temporal. Las regiones central, parietal y occipital muestran valores de importancia inferiores en comparación con las anteriores.

Este resultado sugiere que, en la condición Neutral, las características extraídas de electrodos frontales aportan la información más relevante para la clasificación de la confianza. La contribución intermedia de la región temporal indica una participación secundaria de procesos relacionados con el procesamiento contextual, mientras que la menor importancia asignada al resto de regiones apunta a un papel limitado dentro de este contexto experimental.

Aunque los valores absolutos de importancia difieren entre Random Forest y XGBoost, la jerarquía relativa entre regiones es comparable en ambos modelos, lo que refuerza la estabilidad de este patrón a nivel global. El análisis conjunto de los valores SHAP permite así identificar tendencias espaciales generales que no son evidentes en los análisis individuales, donde existe una mayor variabilidad entre participantes.

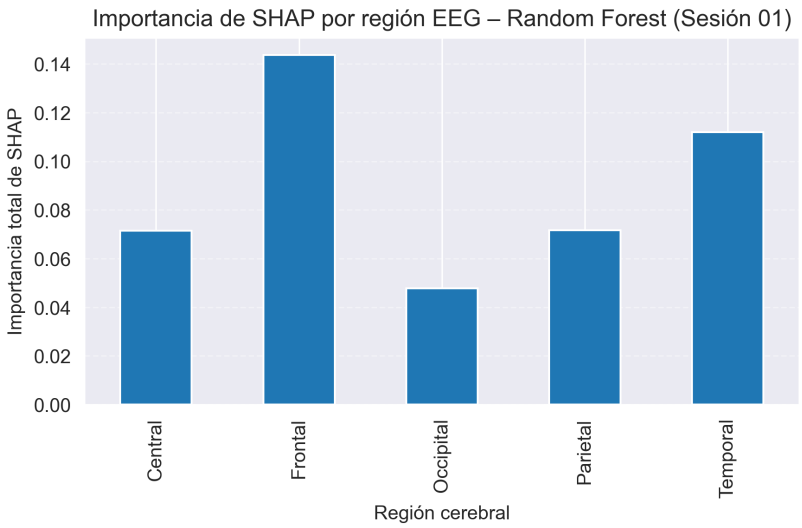

Figura 6.19. Importancia global por región cerebral calculada a partir de los valores SHAP para el modelo Random Forest en la condición Neutral.

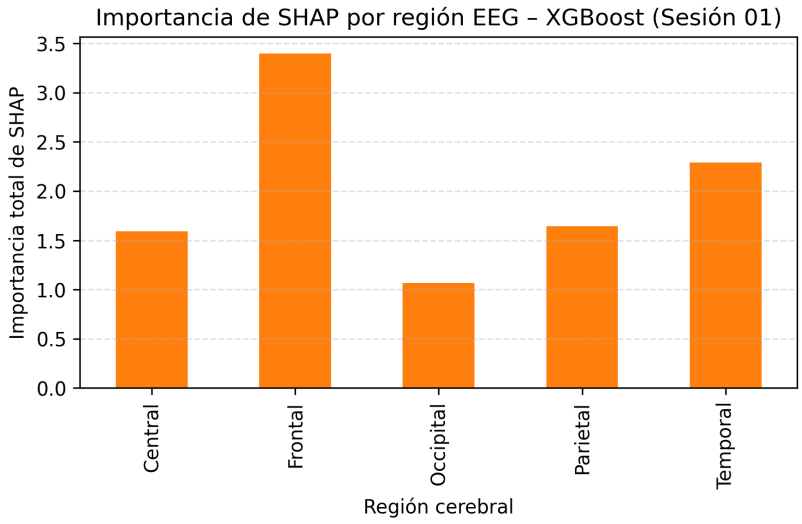

Figura 6.20. Importancia global por región cerebral calculada a partir de los valores SHAP para el modelo XGBoost en la condición Neutral.

Condición Ecológica

Las Figuras 6.21 y 6.22 muestran la importancia por región cerebral obtenida a partir de los valores SHAP para la condición Ecológica, en los modelos Random Forest y XGBoost, respectivamente. La importancia regional se calculó como la suma de las importancias SHAP medias de todas las características asociadas a los electrodos de cada región.

En ambos modelos, la región frontal presenta la mayor contribución global, lo que indica que las características extraídas de electrodos frontales aportan información relevante de forma consistente también en un entorno ecológico. Las región temporal muestra buenas contribuciones, mientras que las regiones parietal, central y occipital presentan valores más bajos de importancia global.

En comparación con la condición Neutral, se mantienen patrones de importancia regional similares, aunque se observa una disminución en la contribución relativa de la región parietal.

La región occipital muestra una importancia global menor en ambos modelos. Este resultado está directamente relacionado con el criterio de agregación empleado, ya que la importancia regional se ha calculado como la suma de las importancias de las características asociadas a cada región. Dado que la región occipital está representada por un número reducido de variables, su contribución total es menor, aunque algunas de sus características aparezcan entre las más relevantes en el análisis global de características individuales.

A pesar de las diferencias en las magnitudes absolutas de los valores SHAP entre Random Forest y XGBoost, la distribución relativa de la importancia entre regiones es similar en ambos modelos. Este patrón indica que, en la condición Ecológica, la información relevante para la predicción de la confianza se distribuye principalmente entre regiones frontales y temporales.

Finalmente, cabe señalar que este análisis describe tendencias globales obtenidas a partir del conjunto de participantes. Las diferencias observadas a nivel individual se atenúan al considerar la suma de las contribuciones por región, lo que permite identificar patrones regionales más estables a nivel global.

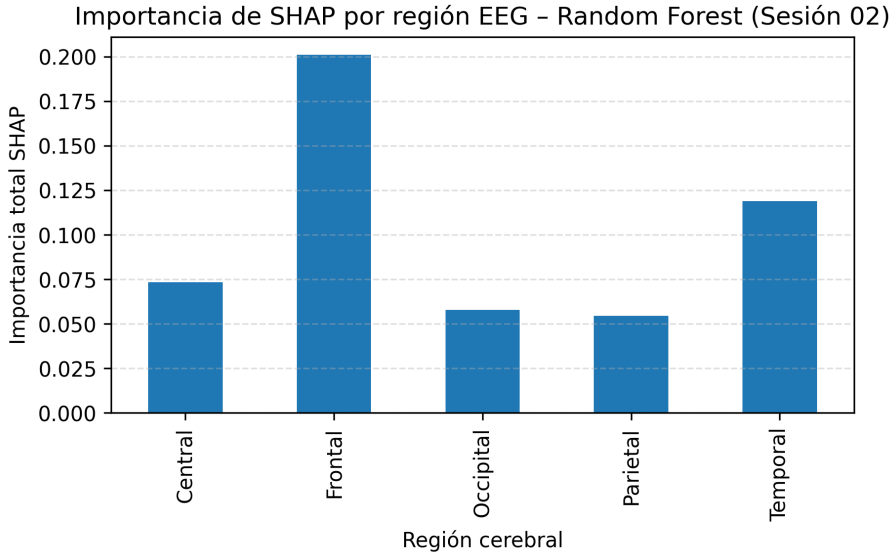

Figura 6.21. Importancia global por región cerebral calculada a partir de los valores SHAP para el modelo Random Forest en la condición Ecológica.

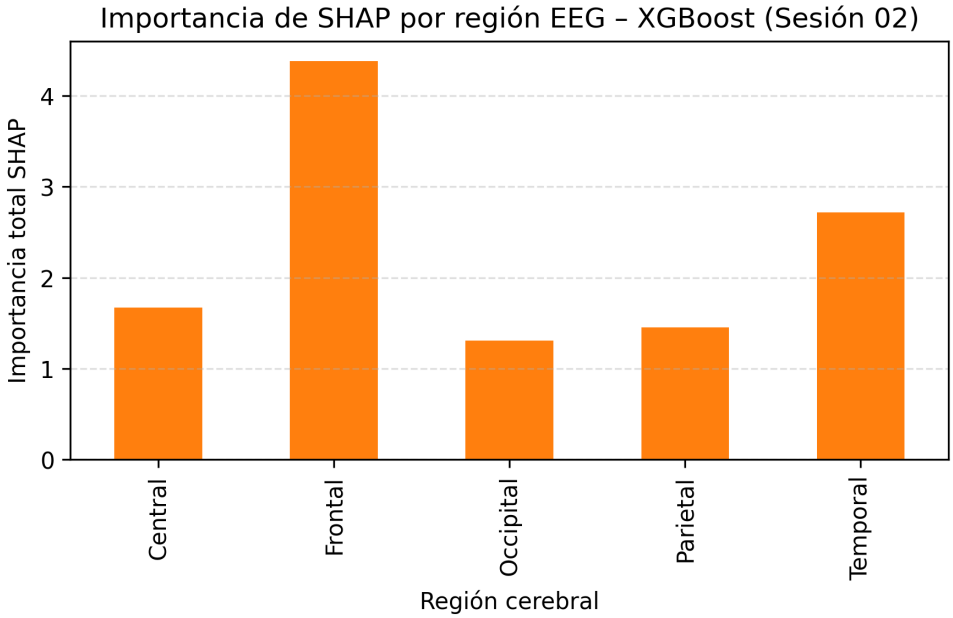

Figura 6.22. Importancia global por región cerebral calculada a partir de los valores SHAP para el modelo XGBoost en la condición Ecológica.

En conjunto, el análisis de interpretabilidad mediante SHAP ha permitido identificar qué características del EEG y qué regiones cerebrales aportan información de forma consistente a los modelos supervisados, tanto a nivel

individual como global. Los patrones observados muestran diferencias entre condiciones experimentales y una elevada variabilidad inter-sujeto, especialmente en el entorno ecológico. Estos resultados se integran y discuten junto con el resto de análisis en el apartado de conclusiones finales.

7 Experimentación propia

En este apartado se decidió recoger datos propios de EEG con el objetivo de validar el procedimiento de adquisición y registro de señales. Esta experimentación tiene un carácter exploratorio y complementario, y no se integra en el análisis principal del estudio, que se basa en la base de datos original empleada en este trabajo.

7.1 Equipamiento EEG

Para la adquisición de las señales EEG se utilizó un dispositivo de la empresa Bitbrain, perteneciente a la serie Versatile EEG 16 [36]. Se trata de un casco EEG inalámbrico de 16 canales, diseñado para facilitar una colocación rápida y cómoda, permitiendo la adquisición de señales en tiempo real.

El dispositivo EEG empleado en este trabajo pertenece a la familia de sistemas water-based EEG comercializados por Bitbrain. El sistema emplea electrodos activos de tipo semi-seco, basados en almohadillas que deben humedecerse previamente con agua para asegurar una correcta conductividad y una baja impedancia en el contacto con el cuero cabelludo. Al tratarse de un sistema water-based, no requiere el uso de gel electrolítico, lo que simplifica el proceso de preparación y mejora la comodidad del participante. Adicionalmente, el dispositivo dispone de un electrodo de tierra, que se coloca mediante una almohadilla específica en el lóbulo de la oreja izquierda.

El sistema registra las señales EEG con una resolución de 24 bits y una frecuencia de muestreo de 256 Hz, valores adecuados para el análisis de la actividad cerebral asociada a tareas cognitivas. El dispositivo se compone de tres elementos principales: el casco EEG, disponible en diferentes tallas para adaptarse a la morfología de la cabeza; los sensores, que se insertan en el casco; y el amplificador EEG, encargado de la adquisición y transmisión de la señal.

Los datos registrados pueden almacenarse directamente en una tarjeta SD integrada en el dispositivo y/o transmitirse de forma inalámbrica mediante conexión Bluetooth, con un alcance superior a los 10 metros. Según las especificaciones técnicas del fabricante, el sistema permite una operación continua de varias horas, más que suficiente para la duración del experimento realizado en este trabajo. Asimismo, el dispositivo está diseñado para ofrecer una monitorización fiable incluso en presencia de cierto ruido electromagnético, lo que contribuye a la calidad de las señales adquiridas.

La Figura 7.1 muestra la interfaz de configuración del dispositivo Versatile EEG 16, utilizada para la selección del equipo y la activación de las señales EEG durante la fase de preparación del experimento.

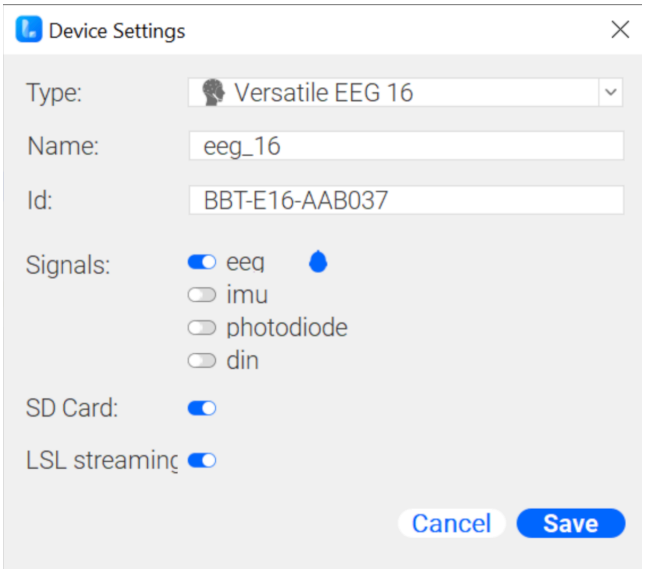

Figura 7.1. Interfaz de configuración del dispositivo Versatile EEG 16

Para este estudio se seleccionaron los 16 canales disponibles en el dispositivo, correspondientes a las siguientes posiciones del sistema internacional 10–20: Fp1, Fp2, F7, F3, Fz, F4, F8, T7 (T3), C3, Cz, C4, T8 (T4), P7 (T5), P3, Pz y P4, siendo AFz el electrodo utilizado como toma de tierra. En la Figura 6.2 se muestra la disposición de los electrodos sobre el casco EEG y su colocación durante la sesión experimental.

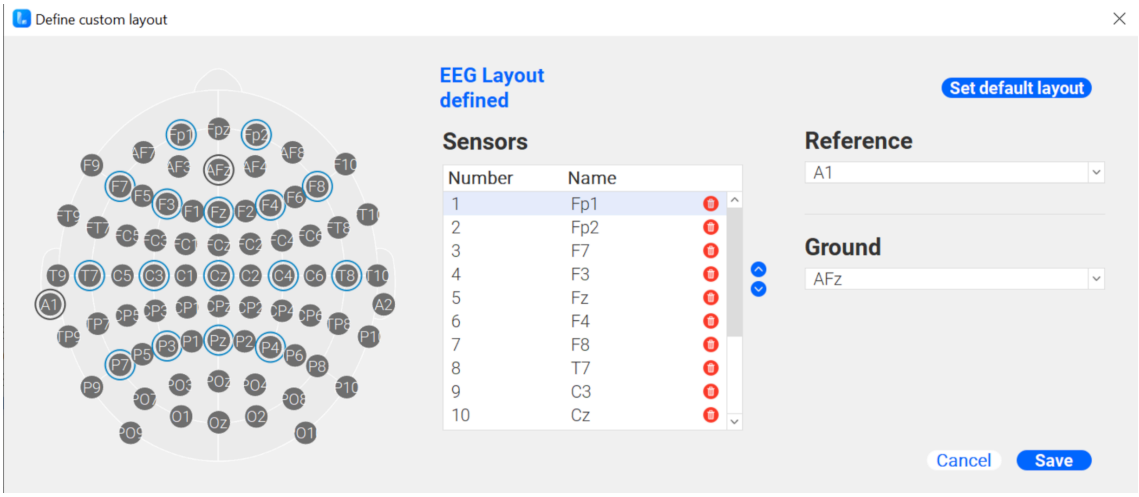

Figura 7.2. Interfaz de definición del layout EEG, mostrando la disposición de los electrodos seleccionados y la configuración de referencia (A1) y tierra (AFz).

Una vez colocado el casco y ajustados los electrodos en sus respectivas posiciones, se humedecieron las almohadillas con agua corriente y se situó el electrodo de tierra en el lóbulo izquierdo de la oreja. Posteriormente, se procedió a comprobar el correcto funcionamiento del sistema mediante la aplicación de adquisición proporcionada por Bitbrain.

En dicha aplicación se verificó la correcta conexión del dispositivo vía Bluetooth, la activación de la tarjeta SD como sistema de respaldo para el almacenamiento de datos, así como la correcta selección de las señales EEG a registrar. A continuación, se accedió al panel de comprobación de sensores, donde se dejó transcurrir un breve periodo de tiempo para permitir la estabilización de la señal. Idealmente, todos los sensores deberían mostrarse en color verde, indicando una buena calidad de señal; no obstante, se consideró aceptable que la mayoría de ellos presentaran una señal adecuada, aunque alguno apareciera en rojo, siempre que la señal registrada fuese interpretable, tal y como se muestra en la Figura 7.3.

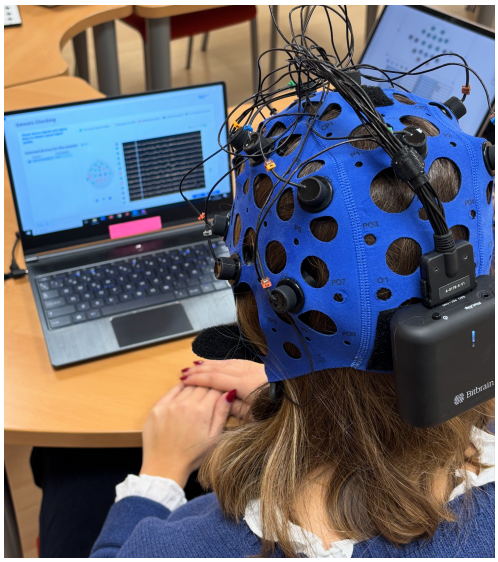

Figura 7.3. Proceso de colocación y ajuste del casco EEG durante la fase de configuración, con el objetivo de verificar el correcto funcionamiento de los electrodos y la adecuada adquisición de la señal.

7.2 Software de adquisición y registro

La adquisición y monitorización de las señales EEG se realizó mediante la aplicación SenssLite proporcionada por Bitbrain. Esta herramienta permitió visualizar en tiempo real la señal EEG de cada canal, comprobar la calidad de los sensores y gestionar el inicio y finalización de las grabaciones.

En la Figura 7.4 se muestra la interfaz de la aplicación SenssLite, en la que se asigna el identificador del participante, se selecciona la ruta de almacenamiento del archivo de datos y se especifica el sensor a utilizar. Asimismo, la aplicación presenta un resumen de la configuración del sensor asociado al participante antes del inicio de la grabación.

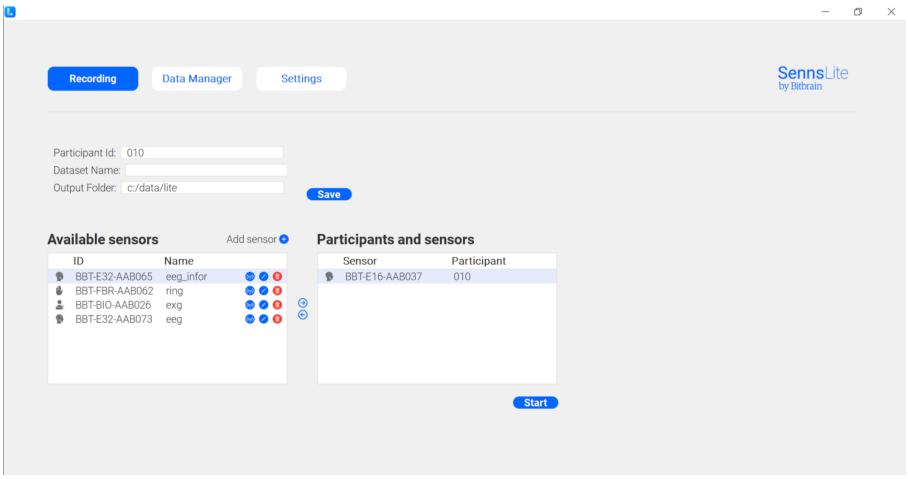

Figura 7.4 Interfaz de la aplicación SenssLite para la selección de sensores.

Los datos fueron almacenados de forma simultánea en el ordenador de adquisición y en la tarjeta SD del dispositivo, garantizando así la integridad de los registros. Durante la sesión, la señal EEG se registró de forma continua.

De forma paralela, se ejecutó la aplicación PsychoPy, donde se encontraba implementado el experimento Stroop estándar descrito en secciones anteriores. Esta aplicación fue la encargada de presentar los estímulos y registrar las respuestas del participante, a partir de las cuales se obtuvieron las medidas de confianza empleadas en este trabajo.

7.3 Procedimiento experimental

Una vez comprobada la calidad de la señal EEG y configurado el sistema de adquisición, se inició la grabación de los datos. Durante el experimento, el participante realizó la tarea Stroop mientras se registraba de manera continua la actividad cerebral.

Al finalizar la tarea, se detuvo la grabación, obteniéndose así los datos EEG en bruto correspondientes a la sesión experimental.

La Figura 7.5 muestra la interfaz de la herramienta de adquisición empleada durante el experimento para la configuración de la grabación, la asignación de sensores al participante y el control del inicio y finalización del registro de señales EEG.

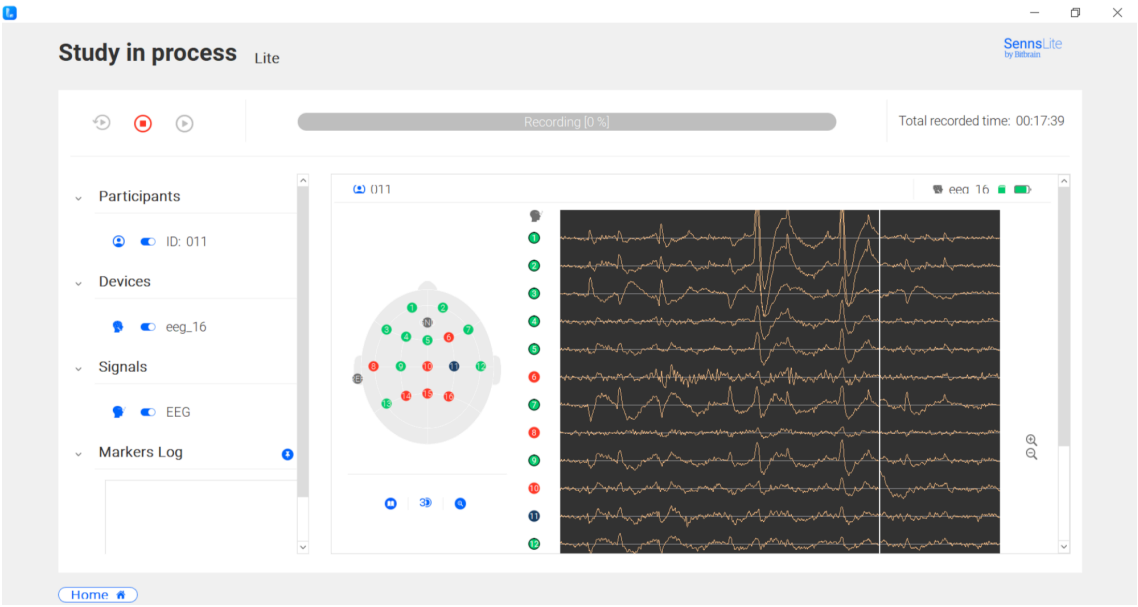

Figura 7.5. Interfaz de la herramienta de adquisición utilizada para la configuración de la grabación y el inicio del registro de señales EEG.

7.4 Participante y consideraciones experimentales

El experimento fue realizado por una única participante, la autora del presente Trabajo Fin de Grado, tal y como se muestra en la Figura 7.6. La participante no presentaba antecedentes neurológicos conocidos y fue informada previamente del procedimiento experimental.

Durante la adquisición de las señales EEG, se procuró mantener una postura estable, minimizando movimientos corporales, parpadeos excesivos y tensiones musculares, con el objetivo de reducir la aparición de artefactos en la señal EEG. Asimismo, se siguieron las recomendaciones básicas para la correcta colocación del casco y la adecuada hidratación de los electrodos, con el fin de asegurar una calidad de señal aceptable durante toda la sesión.

Aunque el número de participantes es limitado, esta experimentación se plantea como un estudio exploratorio y complementario, cuyo objetivo principal es validar el procedimiento de adquisición y registro de las señales EEG, así como evaluar la viabilidad experimental del procedimiento propuesto, sin que los

datos obtenidos se integren en el análisis principal del trabajo, centrado en la base de datos original empleada en este estudio.

Figura 7.6 Desarrollo de la condición Neutral durante la realización del experimento, con el participante equipado con el casco EEG y ejecutando la tarea experimental.

7.5 Datos obtenidos

Como resultado del experimento se obtuvo un registro continuo de señales EEG correspondientes a los 16 canales seleccionados guardados como se muestra en la Figura 7.7. Estos datos constituyen las señales EEG en bruto, obtenidas durante la sesión neutral. Estos datos se emplean como validación del proceso de adquisición, pero no se integran en el análisis cuantitativo desarrollado en los capítulos siguientes, centrado en la base de datos original.

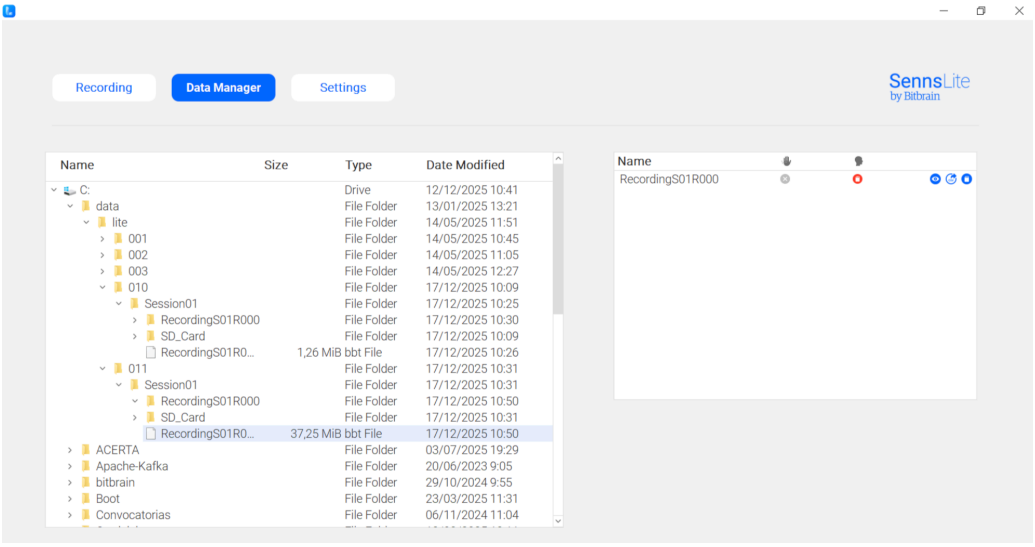

Figura 7.7. Almacenamiento de las señales EEG en bruto correspondientes a los 16 canales registrados durante el experimento.

8 Conclusiones y trabajo futuro

El objetivo principal de este Trabajo Fin de Grado ha sido analizar la relación entre la actividad EEG y el nivel de confianza del usuario en un sistema, considerando tanto un entorno controlado (condición Neutral) como un entorno más real (condición Ecológica). Para ello, se ha abordado el problema desde una perspectiva progresiva, combinando análisis exploratorio, técnicas de aprendizaje no supervisado y supervisado, y métodos de interpretabilidad basados en SHAP.

Los **análisis exploratorios** muestran que la confianza no se manifiesta en la señal EEG mediante cambios abruptos ni patrones claramente diferenciados, sino a través de variaciones graduales y altamente dependientes del individuo y del contexto experimental. **El análisis temporal** de la señal en crudo puso de manifiesto una elevada variabilidad inter-sujeto y temporal, sin una asociación directa y consistente con los cambios en el nivel de Trust, lo que justifica el uso de estrategias basadas en la extracción de características. En este sentido, el **análisis estadístico** en el dominio frecuencial resultó más informativo, identificándose la banda low gamma como la más relevante. En particular, la región frontal aparece de forma recurrente como sensible a los cambios en la confianza, observándose en numerosos participantes valores de potencia más elevados asociados a niveles bajos de Trust, con una disminución progresiva a medida que aumenta la confianza. En la condición Ecológica, estas modulaciones se extienden con mayor frecuencia a regiones occipitales, en coherencia con el carácter más visual y complejo del entorno experimental.

El **análisis topográfico** refuerza estas observaciones, mostrando que las diferencias entre niveles de confianza se expresan principalmente como variaciones de intensidad sobre patrones espaciales similares, sin reorganizaciones topográficas claras. Mientras que en la condición Neutral la distribución espacial de la actividad es más estable y con predominio de regiones frontales y temporales, en la condición Ecológica se observa una mayor dispersión espacial y una implicación más marcada de regiones occipitales y temporales. No obstante, en ambos contextos la confianza no se asocia a patrones espaciales exclusivos, sino a variaciones progresivas dependientes del contexto.

Las técnicas de **aprendizaje no supervisado** no revelan una organización natural de los datos en clústeres claramente asociados a los niveles de confianza. Las métricas internas y externas indican que la estructura intrínseca de las características EEG no se corresponde de forma directa con las etiquetas de Trust, lo que sugiere que la confianza no constituye una variable dominante en la organización global del espacio de características. Este resultado es

coherente con la naturaleza continua y subjetiva de la confianza, así como con la elevada variabilidad inter-sujeto propia de las señales EEG.

En el **análisis supervisado**, los modelos basados en árboles (Random Forest y XGBoost) muestran un rendimiento superior al resto de algoritmos evaluados, tanto en la condición Neutral como en la Ecológica, aunque con valores de accuracy y F1 moderados. Este resultado indica que los modelos son capaces de capturar cierta información relevante relacionada con la confianza, pero también pone de manifiesto la dificultad asociada a la estimación de estados cognitivos complejos a partir de EEG, especialmente en contextos más realistas. El descenso de rendimiento observado en la condición Ecológica refuerza la idea de que el aumento de la complejidad experimental introduce una mayor variabilidad en la señal y dificulta la generalización de los modelos.

El **análisis de interpretabilidad mediante SHAP** aporta una perspectiva complementaria y coherente con el resto de resultados obtenidos, permitiendo profundizar en cómo los modelos supervisados utilizan la información contenida en las señales EEG para estimar el nivel de confianza. A nivel individual, se observa que los modelos se apoyan de forma consistente en características asociadas a la banda low gamma, no limitándose a la potencia espectral, sino incorporando también métricas que describen la variabilidad y la dinámica temporal de la señal, como la desviación estándar, la amplitud peak-to-peak y los parámetros de Hjorth. A nivel global, el análisis conjunto de los valores SHAP confirma la relevancia recurrente de esta banda en ambas condiciones experimentales y pone de manifiesto que la información utilizada por los modelos se distribuye entre distintas regiones cerebrales.

El **análisis global y el análisis regional basados en SHAP** resultan complementarios y permiten una interpretación más completa de la contribución de las distintas regiones cerebrales. En el análisis global, las regiones frontal, temporal y occipital aparecen como las más destacadas en términos de importancia, manteniéndose las regiones frontal y temporal como las más relevantes también en el análisis regional. En el caso de la región occipital, su contribución agregada disminuye en el análisis regional, lo que puede atribuirse al reducido número de características asociadas a dicha región; no obstante, algunas de estas características individuales aparecen entre las más influyentes a nivel global. Estos resultados ponen de manifiesto la importancia de considerar ambos niveles de análisis, ya que la agregación regional puede modificar la contribución relativa de cada región sin que ello implique una pérdida de relevancia de determinadas características específicas. En conjunto, el análisis exploratorio sugiere que la información relevante para la estimación de la confianza se distribuye principalmente entre las regiones frontal, temporal y occipital.

En conjunto, los resultados obtenidos indican que la relación entre la actividad EEG y la confianza humano–sistema se manifiesta de forma sutil, distribuida y dependiente del contexto experimental. No se identifican marcadores EEG simples ni patrones categóricos claramente asociados a niveles discretos de confianza; sin embargo, se observan regularidades consistentes en determinadas bandas de frecuencia y tipos de características, especialmente en la banda low gamma. Estas regularidades aparecen de manera recurrente en regiones frontales y temporales, con la participación adicional de regiones occipitales, cuya relevancia se hace más evidente en la condición Ecológica. En este sentido, el presente trabajo pone de manifiesto la utilidad de combinar análisis exploratorio, modelado supervisado e interpretabilidad para abordar el estudio de estados cognitivos complejos a partir de señales EEG, y establece una base metodológica sólida para futuras investigaciones en entornos más realistas.

8.1 Limitaciones del estudio

A pesar de los resultados obtenidos, este trabajo presenta una serie de limitaciones que deben tenerse en cuenta a la hora de interpretar los hallazgos y valorar su alcance.

En primer lugar, la **naturaleza de la variable Trust** condiciona el análisis. La confianza es un constructo subjetivo, continuo y dependiente del contexto, que en este estudio se recoge mediante autoinforme a través de una escala de cinco puntos. Este enfoque permite recoger diferencias individuales en la percepción de la confianza, pero introduce variabilidad inter-sujeto, ya que cada participante puede utilizar la escala siguiendo criterios propios. Como consecuencia, las etiquetas empleadas como variable objetivo reflejan valoraciones relativas más que estados objetivos, lo que dificulta la identificación de patrones EEG consistentes entre niveles de confianza.

En segundo lugar, las señales EEG presentan de forma inherente una **elevada variabilidad inter- e intra-sujeto**, así como una relación señal-ruido limitada. Esta característica se acentúa en la condición ecológica, donde los participantes están expuestos a un mayor número de estímulos y a un entorno experimental más complejo. Aunque se aplicaron procedimientos de preprocesado y extracción de características, parte de la información relevante puede quedar oculta por ruido. Esta limitación se deriva de las propias características del EEG como técnica de registro y no del diseño experimental del estudio.

Otra limitación importante está relacionada con el **rendimiento moderado de los modelos supervisados**. Si bien los modelos basados en árboles superan al

resto de algoritmos evaluados, los valores de accuracy y F1-score indican que la capacidad predictiva es limitada. Esta restricción implica que los resultados del análisis de interpretabilidad mediante SHAP deben interpretarse con cautela, ya que reflejan el comportamiento de modelos que capturan solo parcialmente la relación entre EEG y confianza.

Asimismo, **el tamaño efectivo del conjunto de datos**, condicionado por el número de participantes y por la segmentación temporal de las señales, puede limitar la capacidad de generalización de los modelos entrenados. Aunque se emplearon estrategias para mitigar el desequilibrio entre clases y evaluar el rendimiento de forma adecuada, incluyendo el uso de métricas como la balanced accuracy, un mayor número de sujetos permitiría obtener estimaciones más robustas y reducir la influencia de la variabilidad individual. Una limitación adicional del estudio está relacionada con lo que **dificulta la identificación de patrones EEG** diferenciables y consistentes entre niveles de confianza. A lo largo de los análisis exploratorios, temporales, topográficos y de aprendizaje automático, las diferencias entre niveles de Trust se manifiestan principalmente como variaciones sutiles de la intensidad o de la relevancia de determinadas características, más que como patrones espaciales o temporales claramente diferenciados.

Esta ausencia de cambios marcados y sistemáticos sugiere que la confianza no se refleja en el EEG mediante patrones claramente separables, sino a través de variaciones distribuidas de la actividad cerebral, dependientes del individuo y del contexto experimental. Este comportamiento es coherente con las limitaciones actuales del EEG en términos de resolución espacial y sensibilidad para capturar estados cognitivos complejos, y no debe interpretarse como un fallo del diseño experimental. En consecuencia, la capacidad de los modelos para discriminar niveles de confianza de forma consistente se ve limitada, y los resultados deben entenderse como la identificación de tendencias generales más que como la detección de patrones neuronales definidos.

Por último, este estudio se centra exclusivamente en el análisis de características extraídas de señales EEG. Esta decisión se adoptó con el objetivo de acotar el alcance del trabajo y profundizar en el estudio del EEG de forma específica. No obstante, la ausencia de información multimodal puede limitar la capacidad de los modelos para capturar de manera más completa el estado cognitivo de confianza, aspecto que se plantea como una línea clara de trabajo futuro

8.2 Futuras investigaciones

El presente trabajo se ha planteado con un enfoque exploratorio, centrado en el análisis de señales EEG y su relación con los niveles de confianza del usuario. A partir de los resultados obtenidos, se identifican diversas líneas de investigación que podrían abordarse en trabajos futuros con el objetivo de ampliar y mejorar el alcance del estudio.

Una primera línea de investigación relevante consiste en la incorporación de un enfoque multimodal, combinando el EEG con otras señales psicofisiológicas registradas durante los experimentos, como electrocardiograma (ECG), eye-tracking o actividad electrodérmica. La integración de distintas modalidades podría aportar información complementaria y permitir una caracterización más completa de estados cognitivos complejos como la confianza, mejorando potencialmente la robustez y generalización de los modelos.

Otra posible línea de trabajo se relaciona con la exploración de estrategias alternativas de extracción y selección de características. Aunque en este estudio se han empleado descriptores ampliamente utilizados en la literatura, futuros trabajos podrían evaluar la utilidad de medidas que describan la interacción entre distintas regiones cerebrales, así como de descriptores más complejos de la señal EEG o métodos que permitan reducir el número de variables manteniendo la información más relevante. Estas estrategias podrían contribuir tanto a mejorar el rendimiento de los modelos como a facilitar su interpretación.

Asimismo, la ampliación del conjunto de datos, mediante la inclusión de un mayor número de participantes o sesiones experimentales, permitiría analizar con mayor detalle la estabilidad de los patrones observados y reducir la influencia de la variabilidad inter-sujeto. En este contexto, enfoques de adaptación entre sujetos o técnicas de aprendizaje transferido podrían resultar de interés para mejorar la capacidad de generalización de los modelos.

Por último, una extensión natural de este trabajo consistiría en el análisis de escenarios experimentales de mayor complejidad, incluyendo tareas más próximas a situaciones operativas reales. Aunque estos entornos introducen un mayor nivel de variabilidad en las señales registradas, también permiten evaluar la robustez y aplicabilidad de los modelos en condiciones más cercanas al uso final. Una estrategia progresiva, partiendo de entornos controlados y avanzando hacia contextos más complejos, facilitaría este tipo de análisis.

9 Análisis de Impacto

En este capítulo se analiza el impacto del presente Trabajo Fin de Grado desde distintas perspectivas: personal y académica, científica y tecnológica, social y ética, así como su relación con los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030. Este análisis permite valorar el alcance y las implicaciones del trabajo más allá de los resultados técnicos obtenidos, evaluando su posible contribución y destacando su relevancia en el ámbito de la interacción humano–sistema.

9.1 Impacto personal y académico

La realización de este trabajo ha supuesto un impacto significativo tanto a nivel personal como académico. Desde el punto de vista académico, el desarrollo del proyecto ha permitido aplicar conocimientos adquiridos a lo largo del grado, especialmente en áreas como el procesamiento de señales, el análisis de datos experimentales y el aprendizaje automático. Antes de comenzar el trabajo, no se tenía experiencia previa en el análisis de señales EEG, más allá de haberlas visto en contextos médicos, sin ser consciente de su aplicación en el ámbito tecnológico. Del mismo modo, aunque se había trabajado anteriormente con conjuntos de datos públicos, esta ha sido la primera experiencia con datos reales procedentes de un proyecto de investigación, lo que ha supuesto un cambio relevante en la forma de abordar el análisis y la interpretación de los resultados.

El trabajo con señales EEG reales ha supuesto un reto adicional debido a la complejidad inherente de este tipo de datos, caracterizados por su variabilidad intersujeto, la presencia de ruido y la dependencia del contexto experimental. Uno de los principales retos del proyecto ha sido la comprensión del funcionamiento de la señal EEG y su relación con los procesos cognitivos estudiados. Más allá de la implementación del código, el aspecto más complejo ha sido entender el significado de los datos y contextualizar los resultados obtenidos. Asimismo, la interpretación de los resultados ha supuesto una dificultad adicional, ya que en algunos casos no presentan patrones claros o rendimientos elevados, lo que ha obligado a reflexionar sobre las limitaciones del enfoque y a justificar de manera crítica las decisiones metodológicas adoptadas.

A nivel personal, el proyecto ha contribuido al desarrollo de competencias como la capacidad de aprendizaje autónomo, la planificación del trabajo a medio y largo plazo y la resolución independiente de problemas complejos. Además, este

trabajo ha despertado un interés particular por el análisis de datos y el aprendizaje automático aplicado a contextos centrados en el ser humano, así como por el ámbito de la interacción humano–sistema y la neurotecnología. En este sentido, el proyecto se consolida como una experiencia clave dentro del grado y como un punto de partida para un posible desarrollo académico o profesional en este campo.

9.2 Impacto científico y tecnológico

Desde una perspectiva científica, este trabajo se enmarca en el estudio de la confianza humano–sistema, un factor clave en la interacción entre personas y sistemas automatizados. La confianza del usuario influye directamente en la aceptación, el uso y la eficacia de estos sistemas, especialmente en contextos donde la automatización desempeña un papel relevante en la toma de decisiones. En este sentido, el trabajo adopta un enfoque principalmente exploratorio y metodológico, orientado a comprender cómo actúa el cerebro en situaciones relacionadas con la confianza.

El uso de señales EEG para analizar estados cognitivos asociados a la confianza representa una línea de investigación emergente con un alto potencial de desarrollo. Al inicio del proyecto, la cantidad de información que puede extraerse de una señal EEG resultó especialmente relevante, teniendo en cuenta que se trataba de un ámbito completamente nuevo. Este trabajo contribuye a dicha línea explorando la relación entre la actividad cerebral y distintos niveles de confianza, aportando evidencia adicional sobre la viabilidad de emplear información neurofisiológica como fuente de datos objetiva para caracterizar el estado cognitivo del usuario.

El objetivo principal del proyecto no ha sido el desarrollo de un sistema final aplicable directamente, sino la comprensión de los datos, el análisis de su comportamiento y la evaluación de distintas técnicas para extraer información relevante. En este sentido, el trabajo se sitúa más cerca del análisis y la interpretación de señales que de la implementación de una solución tecnológica cerrada, sentando las bases para futuras investigaciones y posibles mejoras metodológicas.

Desde el punto de vista tecnológico, los resultados obtenidos pueden servir como punto de partida para el desarrollo de sistemas inteligentes adaptativos, capaces de modificar su comportamiento en función del estado cognitivo del usuario. En entornos como la aviación, la conducción automatizada o los sistemas de apoyo a la toma de decisiones, la capacidad de estimar el nivel de

confianza del operador podría permitir ajustar el grado de automatización, ofrecer información adicional o activar mecanismos de apoyo cuando se detecten niveles bajos de confianza.

Este tipo de tecnologías puede contribuir a mejorar la seguridad, la eficiencia y la fiabilidad de los sistemas automatizados, favoreciendo una interacción más fluida y centrada en el ser humano. Aunque el presente trabajo tiene un carácter exploratorio y no plantea una implementación directa en entornos reales, sus resultados se alinean con las tendencias actuales en el desarrollo de sistemas inteligentes y en la integración de factores humanos en el diseño tecnológico.

9.3 Impacto social y ético

El impacto social de este TFG está estrechamente relacionado con la mejora de la interacción entre personas y sistemas tecnológicos desde una perspectiva centrada en el ser humano. En un contexto donde la automatización esta creciendo, resulta fundamental diseñar sistemas que tengan en cuenta el estado cognitivo del usuario, con el fin de evitar una pérdida de control, una confianza excesiva o una desconfianza injustificada en la tecnología. En este sentido, el análisis de la confianza a partir de señales EEG puede contribuir, a largo plazo, al desarrollo de sistemas que actúen como herramientas de apoyo al usuario, en lugar de sustituir su capacidad de decisión.

La posibilidad de medir y analizar la confianza del usuario puede ayudar a reducir errores humanos, aumentar la seguridad y mejorar la experiencia de uso en entornos complejos. Sistemas que se adapten al nivel de confianza del operador pueden facilitar una toma de decisiones más informada y reducir situaciones de estrés o sobrecarga cognitiva, lo que repercute positivamente en el bienestar de las personas y en la aceptación de las tecnologías automatizadas.

Desde el punto de vista ético, el uso de señales neurofisiológicas como el EEG plantea importantes desafíos, ya que este tipo de señales contienen información altamente sensible sobre el estado cognitivo de las personas. Durante el desarrollo del proyecto, y especialmente al haber participado personalmente en la adquisición de algunos datos, se ha tomado conciencia del carácter personal de esta información, reforzando la necesidad de garantizar la privacidad, el anonimato y el consentimiento informado de los participantes. En el contexto de este trabajo, los datos han sido utilizados exclusivamente con fines académicos y de investigación, dentro de un entorno experimental controlado.

Asimismo, resulta fundamental que las tecnologías basadas en la monitorización del estado cognitivo se utilicen de forma responsable y transparente, evitando usos invasivos o discriminatorios. El diseño de estos sistemas debe orientarse a complementar las capacidades humanas y mejorar el desempeño del usuario, manteniendo siempre al ser humano como responsable final de la toma de decisiones. La confianza en este tipo de tecnologías no debe ser ciega, especialmente en etapas tempranas de desarrollo, y su implementación debe estar guiada por principios éticos que favorezcan una integración social responsable y aceptable.

9.4 Impacto medioambiental y vinculación con los ODS

El impacto medioambiental directo de este TFG es limitado, ya que se basa principalmente en el análisis computacional de datos EEG previamente adquiridos y no requiere la realización de experimentos con un consumo elevado de recursos materiales o energéticos. Si bien el desarrollo del trabajo ha implicado el uso de equipamiento electrónico, como un casco EEG reutilizable y un ordenador personal para el procesamiento de los datos, este uso se corresponde con un consumo energético reducido y comparable al de otras actividades académicas habituales. Además, el empleo de herramientas software para el análisis y procesamiento de las señales, junto con el uso puntual de una única sesión de adquisición de datos EEG, permite realizar el estudio sin necesidad de repetir experimentos físicos, contribuyendo así a minimizar la huella ambiental asociada al desarrollo del proyecto. No obstante, de forma indirecta, el avance en el diseño de sistemas automatizados más eficientes y adaptativos puede favorecer una optimización del uso de recursos en distintos ámbitos tecnológicos.

En relación con los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030, este trabajo se alinea con varios de ellos, destacándose los siguientes:

- **ODS 3: Salud y bienestar:** Este trabajo se alinea de manera clara con el ODS 3, al contribuir al estudio de señales neurofisiológicas y a la comprensión de estados cognitivos relacionados con el bienestar, la carga mental y la interacción con sistemas tecnológicos. El análisis de la confianza del usuario a partir de señales EEG puede facilitar el diseño de sistemas que reduzcan el estrés, la fatiga cognitiva y el riesgo de error humano, especialmente en entornos de alta responsabilidad, favoreciendo así la seguridad y el bienestar de las personas.

- **ODS 8: Trabajo decente y crecimiento económico:** Asimismo, el presente TFG se relaciona con el ODS 8, ya que el desarrollo de tecnologías más seguras, eficientes y adaptadas al estado cognitivo del usuario puede contribuir a mejorar las condiciones de trabajo en entornos altamente automatizados. La integración de factores humanos en el diseño de sistemas tecnológicos puede ayudar a reducir riesgos laborales, prevenir errores derivados de la sobrecarga cognitiva y aumentar la eficiencia y la productividad de manera sostenible.

- **ODS 9: Industria, innovación e infraestructura:** El trabajo también se vincula con el ODS 9, al fomentar la investigación y la innovación en sistemas inteligentes centrados en el ser humano. El uso de técnicas de análisis de señales EEG y aprendizaje automático contribuye al desarrollo de soluciones tecnológicas avanzadas que integran información cognitiva en el diseño de sistemas automatizados, promoviendo infraestructuras más seguras, resilientes y adaptativas.

En conjunto, este Trabajo Fin de Grado se enmarca en una línea de desarrollo alineada con los principios de innovación responsable y desarrollo sostenible, contribuyendo de manera indirecta al avance de tecnologías más seguras, eficientes y orientadas al bienestar de las personas.

10 Bibliografía

[1] «Trust in Automation: Designing for Appropriate Reliance», *Hum. Factors*, vol. 46, n.º 1, pp. 50-80, mar. 2004, doi: 10.1518/HFES.46.1.50_30392.

[2] K. E. Schaefer *et al.*, «A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Human-Robot Interaction»:, Defense Technical Information Center, Fort Belvoir, VA, jul. 2014. doi: 10.21236/ADA607926.

[3] J. B. Rotter, «A new scale for the measurement of interpersonal trust1», *J. Pers.*, vol. 35, n.º 4, pp. 651-665, dic. 1967, doi: 10.1111/j.1467-6494.1967.tb01454.x.

[4] R. Parasuraman y V. Riley, «Humans and automation: Use, misuse, disuse, abuse», *Hum. Factors*, vol. 39, n.º 2, pp. 230-253, 1997, doi: 10.1518/001872097778543886.

[5] K. Akash, W.-L. Hu, N. Jain, y T. Reid, «A Classification Model for Sensing Human Trust in Machines Using EEG and GSR», *ACM Trans. Interact. Intell. Syst.*, vol. 8, n.º 4, pp. 1-20, dic. 2018, doi: 10.1145/3132743.

[6] F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, y H. Wallach, «Manipulating and Measuring Model Interpretability», 15 de agosto de 2021, *arXiv*: arXiv:1802.07810. doi: 10.48550/arXiv.1802.07810.

[7] R. C. Mayer, J. H. Davis, y F. D. Schoorman, «An Integrative Model of Organizational Trust», *Acad. Manage. Rev.*, vol. 20, n.º 3, pp. 709-734, 1995, doi: 10.2307/258792.

[8] J. Lee y N. Moray, «Trust, control strategies and allocation of function in human-machine systems», *Ergonomics*, vol. 35, n.º 10, p. 1243, 1992, doi: 10.1080/00140139208967392.

[9] J. B. Lyons, K. Sycara, M. Lewis, y A. Capiola, «Human–Autonomy Teaming: Definitions, Debates, and Directions», *Front. Psychol.*, vol. 12, may 2021, doi: 10.3389/fpsyg.2021.589585.

[10] A. Xu y G. Dudek, «Towards Modeling Real-Time Trust in Asymmetric Human–Robot Collaborations», en *Robotics Research*, vol. 114, M. Inaba y P. Corke, Eds., en Springer Tracts in Advanced Robotics, vol. 114. , Cham: Springer International Publishing, 2016, pp. 113-129. doi: 10.1007/978-3-319-28872-7_7.

[11] Y. Guo y X. J. Yang, «Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian Inference Approach», *Int. J. Soc. Robot.*, vol. 13, n.º 8, pp. 1899-1909, dic. 2021, doi: 10.1007/s12369-020-00703-3.

[12] P. L. Nunez y R. Srinivasan, *Electric Fields of the Brain: The Neurophysics of EEG*. Oxford University Press, 2006.

[13] S. J. Luck, *An Introduction to the Event-Related Potential Technique, second edition*. MIT Press, 2014.

[14] «EEG-electrode-placement-based-on-international-10-20-system-The-18-electrodes-cover-5.png (850×705)». Accedido: 4 de enero de 2026. [En línea]. Disponible en: https://www.researchgate.net/publication/371908856/figure/fig4/AS:1

1431281170873208@1687920982622/EEG-electrode-placement-based-on-international-10-20-system-The-18-electrodes-cover-5.png
[15] E. K. Miller y J. D. Cohen, «An Integrative Theory of Prefrontal Cortex Function», *Annu. Rev. Neurosci.*, vol. 24, n.º Volume 24, 2001, pp. 167-202, mar. 2001, doi: 10.1146/annurev.neuro.24.1.167.
[16] E. R. Kandel, J. Koester, S. Mack, y S. Siegelbaum, Eds., *Principles of neural science*, Sixth edition. New York: McGraw Hill, 2021.
[17] M. Corbetta y G. L. Shulman, «Control of goal-directed and stimulus-driven attention in the brain», *Nat. Rev. Neurosci.*, vol. 3, n.º 3, p. 201, 2002, doi: 10.1038/NRN755.
[18] P. Welch, «The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms», *IEEE Trans. Audio Electroacoustic*s, vol. 15, n.º 2, pp. 70-73, jun. 1967, doi: 10.1109/TAU.1967.1161901.
[19] C. X. Gao *et al.*, «An overview of clustering methods with guidelines for application in mental health research», *Psychiatry Res.*, vol. 327, p. 115265, sep. 2023, doi: 10.1016/j.psychres.2023.115265.
[20] N. X. Vinh, N. X. Vinh, J. Epps, J. Epps, y J. Bailey, «Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance».
[21] R. L. Thorndike, «Who Belongs in the Family?», *Psychometrika*, vol. 18, n.º 4, pp. 267-276, dic. 1953, doi: 10.1007/BF02289263.
[22] A. C. Mu, «Introduction to Machine Learning with Python».
[23] C. Xu, C. Zhang, Y. Zhou, Z. Wang, P. Lu, y B. He, «Trust Recognition in Human-Robot Cooperation Using EEG», 8 de marzo de 2024, *arXiv*: arXiv:2403.05225. doi: 10.48550/arXiv.2403.05225.
[24] C. Cortes y V. Vapnik, «Support-vector networks», *Mach. Learn.*, vol. 20, n.º 3, pp. 273-297, sep. 1995, doi: 10.1007/BF00994018.
[25] K. P. Murphy, *Machine learning: a probabilistic perspective*, 4. print. (fixed many typos). en Adaptive computation and machine learning series. Cambridge, Mass.: MIT Press, 2013.
[26] S. Tibshirani y H. Friedman, «Valerie and Patrick Hastie».
[27] T. Chen y C. Guestrin, «XGBoost: A Scalable Tree Boosting System», en *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, en KDD '16. New York, NY, USA: Association for Computing Machinery, ago. 2016, pp. 785-794. doi: 10.1145/2939672.2939785.
[28] F. Pedregosa *et al.*, «Scikit-learn: Machine Learning in Python», *Mach. Learn. PYTHON*.
[29] T. Gan, S. Wang, G. Mo, S. Li, Y. Lu, y J. Li, «Machine learning prediction and SHAP interpretability analysis of heart failure risk in patients with hyperuricemia», *Front. Cardiovasc. Med.*, vol. 12, dic. 2025, doi: 10.3389/fcvm.2025.1689607.
[30] «Explainable Artificial Intelligence (XAI) : Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI | PDF | Artificial Intelligence | Intelligence (AI) & Semantics», Scribd. Accedido: 9 de enero de 2026. [En línea]. Disponible en: https://www.scribd.com/document/726622217/1

[31] S. M. Lundberg y S.-I. Lee, «A Unified Approach to Interpreting Model Predictions», presentado en Neural Information Processing Systems, may 2017. Accedido: 10 de enero de 2026. [En línea]. Disponible en: https://www.semanticscholar.org/paper/442e10a3c6640ded9408622005e3c2a8906ce4c2
[32] M. T. Ribeiro, S. Singh, y C. Guestrin, «"Why Should I Trust You?": Explaining the Predictions of Any Classifier», en *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, en KDD '16. New York, NY, USA: Association for Computing Machinery, ago. 2016, pp. 1135-1144. doi: 10.1145/2939672.2939778.
[33] L. S. Shapley, «A Value for N-Person Games», mar. 1952. Accedido: 10 de enero de 2026. [En línea]. Disponible en: https://www.rand.org/pubs/papers/P295.html
[34] «(PDF) Explainable AI for Trees: From Local Explanations to Global Understanding», ResearchGate. Accedido: 10 de enero de 2026. [En línea]. Disponible en: https://www.researchgate.net/publication/333077391_Explainable_AI_for_Trees_From_Local_Explanations_to_Global_Understanding
[35] L. Breiman, «Random Forests», *Mach. Learn.*, vol. 45, n.º 1, pp. 5-32, oct. 2001, doi: 10.1023/A:1010933404324.
[36] «Water-based EEG», Bitbrain. Accedido: 10 de enero de 2026. [En línea]. Disponible en: https://www.bitbrain.com/neurotechnology-products/water-based-eeg

11 Anexos

Enlace al repositorio de GitHub dónde se encuentra alojado el proyecto:

https://github.com/luciarebolledo/TFG-analisis-exploratorio-confianza-humana-EEG-ml-interpretabilidad

Este documento esta firmado por

	**Firmante**	CN=tfgm.fi.upm.es, OU=CCFI, O=ETS Ingenieros Informaticos - UPM, C=ES
	Fecha/Hora	Wed Jan 14 18:32:16 CET 2026
	Emisor del Certificado	EMAILADDRESS=camanager@etsiinf.upm.es, CN=CA ETS Ingenieros Informaticos, O=ETS Ingenieros Informaticos - UPM, C=ES
	Numero de Serie	561
	Metodo	urn:adobe.com:Adobe.PPKLite:adbe.pkcs7.sha1 (Adobe Signature)