EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación y Desarrollo Tecnológico

Tesis de Doctorado

Metodología para complementar la evaluación de UX interpretando la carga de trabajo a partir de datos fisiológicos de los usuarios

presentada por

MC. Edgar Omar Bañuelos Lozoya

como requisito para la obtención del grado de

Doctor en Ciencias de la Computación

Director de tesis

Dr. Juan Gabriel González Serna

Codirector de tesis

Dr. Nimrod González Franco

Cuernavaca, Morelos, México. Septiembre de 2023

cenidet
Centro Nacional de Investigación y Desarrollo Tecnológico

EDUCACIÓN
22 AGO 2023
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SERVICIOS ESCOLARES
RECIBIDO
1230

ESC\FORDOC09

Cuernavaca, Morelos, 15/agosto/2023

ASUNTO: ACEPTACIÓN DEL TRABAJO DE TESIS DOCTORAL

MARÍA YASMÍN HERNÁNDEZ PÉREZ
JEFA DEL DEPARTAMENTO DE CIENCIAS COMPUTACIONALES
PRESENTE

Los abajo firmantes, miembros del Comité Tutorial de la Tesis Doctoral del alumno **EDGAR OMAR BAÑUELOS LOZOYA** manifiestan que después de haber revisado su trabajo de tesis doctoral titulado **"METODOLOGÍA PARA COMPLEMENTAR LA EVALUACIÓN DE UX INTERPRETANDO LA CARGA DE TRABAJO A PARTIR DE DATOS FISIOLÓGICOS DE LOS USUARIOS"**, realizado bajo la dirección de Juan Gabriel González Serna y la codirección de Nimrod González Franco, el trabajo se ACEPTA para proceder a su impresión.

ATENTAMENTE
"Excelencia en Educación Tecnológica®
"Educación Tecnológica al Servicio de México"

JUAN GABRIEL GONZÁLEZ SERNA
CENIDET

NIMROD GONZÁLEZ FRANCO
CENIDET

ANDREA MAGADÁN SALAZAR
CENIDET

NOÉ ALEJANDRO CASTRO SÁNCHEZ

DANTE MÚJICA VARGAS
CENIDET

JOSÉ ALEJANDRO REYES ORTIZ
UNIVERSIDAD AUTÓNOMA METROPOLITANA

C.c.p.: María Elena Gómez Torres / Jefa del Depto. de Servicios Escolares
Dr. Carlos Manuel Astorga Zaragoza / Subdirector Académico
Expediente

SEP TecNM CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
RECIBIDO
22 AGO 2023
SUBDIRECCIÓN ACADÉMICA

Cuernavaca, Mor., 22/agosto/2023
No. De Oficio: SAC/139/2023
Asunto: Autorización de impresión de tesis

EDGAR OMAR BAÑUELOS LOZOYA
CANDIDATO AL GRADO DE DOCTOR EN CIENCIAS DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"METODOLOGÍA PARA COMPLEMENTAR LA EVALUACIÓN DE UX INTERPRETANDO LA CARGA DE TRABAJO A PARTIR DE DATOS FISIOLÓGICOS DE LOS USUARIOS"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

A T E N T A M E N T E
Excelencia en Educación Tecnológica®
"Conocimiento y tecnología al servicio de México"

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SUBDIRECCIÓN ACADÉMICA

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO

EDUCACIÓN | TECNOLÓGICO NACIONAL DE MÉXICO
23 AGO 2023
CENTRO NACIONAL DE INVESTIGACIÓN Y DESARROLLO TECNOLÓGICO
SERVICIOS ESCOLARES
RECIBIDO

C. c. p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/lmz

A mi abuelo Manuel.

Agradecimientos

Al Conahcyt por el apoyo económico otorgado.

A los miembros de mi comité revisor, por los consejos, crítica y recomendaciones que acertadamente me brindaron y que contribuyeron a fortalecer este trabajo.

A mi Director y Codirector de tesis, por la guía y el apoyo que me otorgaron durante toda la investigación, agradezco la confianza y espero haber cumplido con sus expectativas.

Doctores Andrea, Noé, Alex, Dante, Nimrod y Gabriel, a todos ustedes les reitero mi respeto, confío en que sigamos colaborando y podamos mantener una relación profesional próspera.

A mis compañeros Eddy, Jorge y Derick, les deseo éxito profesional y enhorabuena por la nueva etapa de vida que cada uno está por emprender.

Al personal del centro y a todas las personas que facilitaron mi estancia en la institución, en particular a Don Manuel, Doña Magos y al resto del personal de la cafetería.

Al TECNM y a su campus Instituto Tecnológico de Parral por permitirme esta aventura, me comprometo a responder desde mi trinchera académica y mantener el ímpetu en pro de nuestros estudiantes.

A Don Héctor, Mariano, Paty y sus familias, la convivencia con ustedes hizo más ameno el periodo incierto de pandemia que vivimos y valoro la amistad que nos brindaron a mi esposa y a mi.

A mis padres, hermanos, sobrinos, abuela, suegros, cuñados y al resto de mi familia, gracias por su apoyo incondicional y alentarme a seguir adelante a pesar de la distancia.

Gracias a mi esposa, Maleny, por darme el empuje para iniciar este reto y el ánimo para mantenerme enfocado cuando el estrés se presentaba, nos esperan nuevas oportunidades por tomar y dificultades por vencer juntos, deseo que sigamos esforzándonos por explotar nuestro potencial individual, profesional y como pareja, te amo.

Omar B.

Resumen

La evaluación de experiencia de usuario tiene una naturaleza subjetiva y aunque diversos enfoques han buscado complementar las técnicas tradicionales a partir de datos inherentes al usuario, persisten algunas problemáticas y la mayoría se habían centrado en la interpretación de emociones.

Por esta razón, en este documento se describe una investigación doctoral que resulta en una metodología para complementar la evaluación de experiencia de usuario con base en el reconocimiento del estado cognitivo de carga de trabajo a partir de datos fisiológicos del usuario.

La metodología consta de siete componentes que definen el constructo, las características de los estímulos y el experimento, las herramientas para la adquisición de datos, las tareas de preprocesamiento de datos y extracción de características, el entrenamiento de un modelo de aprendizaje y la predicción de la carga de trabajo en cada segmento evaluado, por participante o de manera general.

La metodología fue validada en un experimento de evaluación de experiencia de usuario de un sotfware de edición de documentos, obteniendo, entre otros aspectos, una exactitud del 87 % en la clasificación de carga de trabajo evaluando el modelo con un enfoque LOSO y correspondencia entre la carga de trabajo predicha y la esperada para cada segmento.

Palabras clave: UX, datos fisiológicos, estados cognitivos, carga de trabajo.

Abstract

User experience evaluation has a subjective nature, and although several approaches have sought to complement traditional techniques based on user-inherent data, some problems persist, and most focus on the interpretation of emotions.

For this reason, this thesis describes doctoral research that results in a methodology to complement user experience assessment based on the recognition of the cognitive state of "workload" from the physiological data of the users.

The methodology consists of seven components that define the construct, the stimulus and experiment, the data acquisition tools, the data preprocessing and feature extraction tasks, the training of a learning model, and the workload prediction in each evaluated segment, per participant or overall.

It was validated in an experiment to evaluate the user experience of a document edition software, obtaining, among other aspects, a model accuracy of 87 % in the workload classification using a LOSO validation and coincidence between the predicted and expected workload for each segment.

Keywords: UX, physiological data, cognitive states, workload.

Contenido

1 | Introducción

La experiencia de usuario o *User Experience* (UX), es tradicionalmente evaluada con técnicas de naturaleza subjetiva que dependen de lo reportado por los usuarios y del análisis del evaluador, influenciados por su percepción y criterio, entre otros factores [1]-[3]. Aunque varios enfoques complementan la evaluación tradicional con la inferencia de estados mentales a partir de datos inherentes al usuario (p. ej., [4] y [5]), estos enfoques se han centrado en el análisis de estados emocionales del modelo de afecto presentado por Posner, Russell y Peterson [6].

La carga de trabajo es un estado cognitivo que puede expresarse como una experiencia subjetiva y manifestarse de manera fisiológica y con variaciones en el rendimiento de la tarea [7]. Este estado ha sido detectado a partir de tecnologías como electroencefalografía (EEG) y seguimiento ocular (ET, del inglés *Eye Tracking*), debido a su relación con la actividad eléctrica cerebral y con la variación en la dilatación de la pupila (p. ej., [8] y [9]), y se ha estudiado en contextos como: juegos virtuales, conducción, control de tráfico aéreo, entre otros (p. ej., [10]-[12]).

Enmarcada en el contexto de UX, en este documento se describe una investigación doctoral en la que se desarrolló una metodología para complementar la evaluación con base en el reconocimiento de carga de trabajo a partir de datos fisiológicos.

La metodología propuesta consta de siete componentes y define, entre otros, un experimento inicial para capturar el comportamiento fisiológico de los usuarios en tareas cognitivas específicas, un segundo experimento que corresponde a la sesión de interacción con el producto digital a evaluar y un esquema para obtener la carga de trabajo en cada segmento, ya sea por participante o de manera general a través de un mecanismo de votación.

El documento se organiza de la siguiente manera: el primer capítulo introduce la problemática y demás aspectos que definieron la investigación, el capítulo 2 presenta el marco teórico y el análisis de los trabajos relacionados del estado del arte, el capítulo 3 muestra las pruebas realizadas con el *dataset* CogLoad y la construcción y experimentación con conjuntos de datos propios, el capítulo 4 describe la metodología propuesta y la evaluación de su desempeño, por último, el capítulo 5 expone las conclusiones, aportaciones y trabajos futuros a partir de esta investigación.

1.1. Problemática

Los métodos de evaluación de UX pueden clasificarse utilizando cuadrantes divididos por los ejes objetivo-subjetivo y cualitativo-cuantitativo, por ejemplo, los mostrados en la Figura 1.1. De forma general, los métodos objetivos se basan en las respuestas de los usuarios durante la interacción y los subjetivos en sus expresiones posterior a ella, los métodos cuantitativos se basan en el análisis de los datos recogidos y los cualitativos en las interpretaciones por parte de los evaluadores [2].

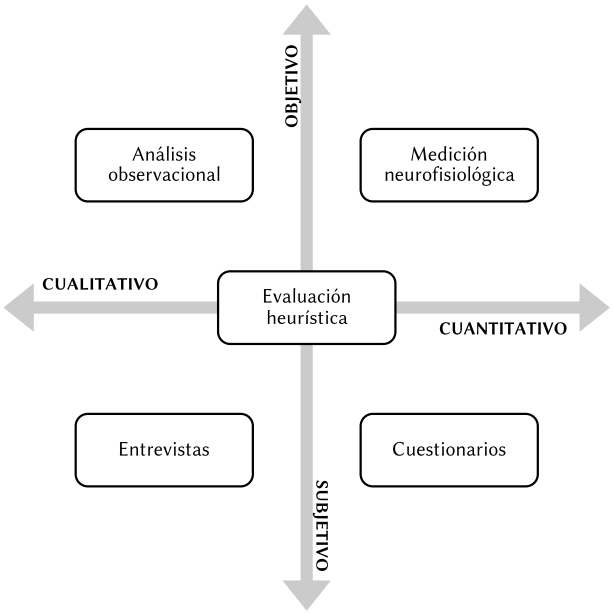


Figura 1.1: Clasificación de métodos de evaluación de UX
Fuente: Laar, Gürkök, Bos *et al.* [2]

Técnicas tradicionales como cuestionarios, entrevistas y pensar en voz alta (*think-aloud*) se ubican en el lado subjetivo del plano, con problemas inherentes relacionados con la capacidad de lenguaje, expresión y restricciones de memoria del usuario [13], así como la experiencia y capacidad de interpretación del propio evaluador.

Las mediciones a partir de señales fisiológicas del usuario son métodos cuantitativos-objetivos que proporcionan datos de manera continua, sin embargo, problemas identificados previamente [2] persisten: la correlación que tienen estas mediciones con la UX o sus componentes aún no está bien definida y algunos sensores todavía producen incomodidad al usuario, restringiendo sus movimientos e influyendo en la propia experiencia.

Adicionalmente, los datos fisiológicos son difíciles de contextualizar, interpretar y generalizar [14]-[17]; por ejemplo, señales como las de electroencefalografía son sujetas a artefactos y dependientes de la persona, seguimiento ocular a variaciones en la iluminación y electromiografía a los movimientos de la persona; con variados tiempos de respuesta después de un estímulo, como en el caso del retraso mayor de la respuesta galvánica en piel (GSR, del inglés *Galvanic Skin Response*) en comparación con otras señales [14][18][19].

Por otro lado, las demandas de la industria exigen que las soluciones que se desarrollen sean certeras, entreguen resultados en un tiempo adecuado y ayuden a generar recomendaciones significativas; cumpliendo con los tiempos de los procesos de desarrollo ágil, utilizando métodos efectivos para visualizar los resultados, con la mayor automatización y la menor intervención humana [20].

Adicionalmente, hay un consenso de que enfoques multimodales son necesarios para entender contextos relacionados con UX, sin embargo, resalta la escasez de más y mejores conjuntos de datos y de mecanismos para hacerlos compatibles e integrarlos [21]; limitando la reutilización de conjuntos de datos validados para experimentar el reconocimiento de carga cognitiva y otros estados mentales y obtener resultados comparativos.

La hipótesis de investigación es la siguiente:

Hipótesis: Es posible complementar la evaluación de UX detectando la carga de trabajo de los usuarios utilizando modelos de aprendizaje automático entrenados con datos fisiológicos.

1.2. Objetivos

El objetivo general es desarrollar una metodología para complementar la evaluación de UX con base en el reconocimiento de carga de trabajo, modelos de aprendizaje automático y datos fisiológicos de los usuarios.

Se definieron los siguientes objetivos específicos:

1. Identificar las características de las señales fisiológicas que se emplean en el reconocimiento de carga de trabajo y analizar la relación entre el estado mental y la UX.

2. Construir un conjunto de datos con estímulos relacionados con UX, seleccionar un conjunto similar del estado del arte y evaluar modelos tradicionales de aprendizaje automático en el reconocimiento de carga de trabajo.
3. Definir y evaluar la metodología para complementar la evaluación de UX con base en el reconocimiento de carga de trabajo.

1.3. Justificación

Tomando en consideración los atributos descritos por Charlton [22], se determinó factible medir el estado mental de carga de trabajo debido a que:

- Es posible detectar cambios en los estados cognitivos de carga de trabajo y otros, en contextos similares a UX (p. ej., [9][23][24]).
- Algunos de los dispositivos para recopilar datos fisiológicos, en particular los de electroencefalografía, pueden resultar molestos y requieren de un tiempo considerable para su colocación y calibración, sin embargo, tienen una alta aplicabilidad en entornos controlados [19] y no interfieren en el desarrollo de tareas que solo requieren el uso de ratón y teclado (p. ej., [25]).
- Se cuenta con equipo disponible para sensar datos EEG, GSR, fotopletismografía (PPG, del inglés *Photoplethysmography*) y ET, además de la herramienta *UXLab* y otras de acceso libre para gestionar la captura de las señales fisiológicas; contemplando a estudiantes del TecNM/Cenidet u otras instituciones como participantes en las pruebas.
- La evaluación de UX es un tópico en constante investigación y contribuye finalmente al desarrollo de mejores aplicaciones de software. Además, el reconocimiento de estados cognitivos puede extrapolarse a otros contextos.

El valor adicional de las mediciones fisiológicas en la evaluación de UX es proporcionar una mayor objetividad, superando las limitaciones del usuario para expresar y recordar adecuadamente su percepción. Hoy en día, todavía es necesario que estas mediciones complementen las técnicas tradicionales, sin embargo, esta investigación contribuye al desarrollo de enfoques para obtener resultados con una menor dependencia a datos subjetivos que auxilien a los evaluadores en la interpretación de las pruebas y a generar recomendaciones significativas en un tiempo adecuado.

2 | Marco teórico

Este capítulo presenta el marco conceptual, los trabajos que fungen como antecedente a esta investigación y el análisis del estado del arte acorde a varios tópicos relacionados.

2.1. Marco conceptual

2.1.1. Experiencia de usuario

Definida en el estándar ISO 9241-210, la experiencia de usuario o *User Experience* (UX) se refiere a las "percepciones y respuestas de un usuario que resultan del uso o uso anticipado de un sistema, producto o servicio" [26]; involucra varias facetas (instrumental, emocional-afectiva y experiencia) y es consecuencia del estado interno del usuario, las características del sistema o aplicación y el contexto donde la interacción ocurre [27].

La Tabla 2.1 muestra un resumen de aspectos importantes que caracterizan la experiencia de usuario como tópico teórico-práctico [28], desde su impulso y bases teóricas, hasta el enfoque de investigación y la perspectiva de negocio que la motiva.

Tabla 2.1: Resumen de características de UX

Aspecto	UX
Impulso	Principalmente humano
Bases teóricas	Fuertes y diversas, dado su enfoque inicial multidisciplinario
Enfoque principal	Evaluar y entender la experiencia de usuario/el proceso de experimentar, recopilar información para diseñar y crear productos y servicios que aporten más valor, experiencias placenteras y que permitan el cumplimiento de las metas
Diseños de investigación	Ambos (cuantitativo y cualitativo), con un fuerte énfasis en investigación cualitativa
Objetivos de investigación	Entender, modelar
Enfoque de investigación	Enfoque holístico
Perspectiva de negocio	Poca atención directa a la dimensión monetaria

Fuente: Adaptada de Wechsung y De Moor [28]

La evaluación tradicional de UX comprende la aplicación de pruebas en donde se encarga a los usuarios interactuar con un producto digital, con uno o más evaluadores obteniendo información –con técnicas como: observación, entrevista, *think-aloud*, recorrido cognitivo (*cognitive walkthrough*), cuestionarios estandarizados, entre otras (ver Figura 2.1)– y analizándola posteriormente para identificar problemáticas y generar recomendaciones de mejora.

La experiencia de usuario está relacionada con otros conceptos como calidad de la experiencia

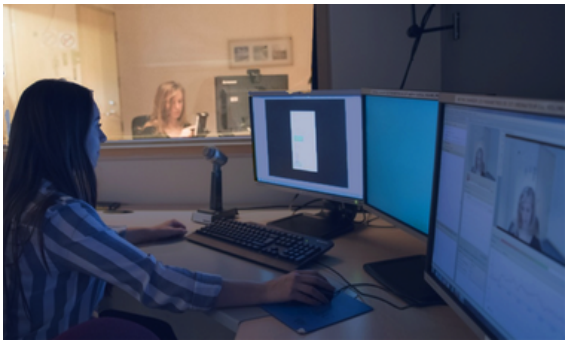


Figura 2.1: Entorno de evaluación tradicional de UX
Fuente: Portada del curso *UX Evaluation: User Testing*

(QoE, del inglés *Quality of Experience*) y calidad de la experiencia de usuario (QUX, del inglés *Quality of User Experience*).

QoE busca obtener el grado de satisfacción o molestia del usuario y reflejarlo como una medida de calidad [29], teniendo similitudes con UX[1] pero con diferencias en su definición y aplicación que parten desde su origen: UX de grupos de investigación de Interacción Humano-Computadora y usabilidad, QoE de grupos relacionados con telecomunicaciones y del concepto de calidad en el servicio [28].

Por otro lado, QUX es presentado como un constructo para extender los modelos para QoE/UX agregando la dimensión eudaimónica (bienestar, significado y propósito de uso), sin embargo, solo se han identificado implicaciones y retos para guiar la agenda de investigación de manera integral [30].

2.1.2. Estados mentales

Se define un estado mental como "una disposición a la acción, es decir, cada aspecto del estado interno de un organismo que podría contribuir a su comportamiento u otras respuestas" [31]; puede comprender un gran número de variables como todos los pensamientos, sentimientos, creencias, intenciones, recuerdos activos, percepciones, etc., que están presentes en un momento dado.

Salzman y Fusi [31] interpretan los ejes de valencia e intensidad que caracterizan una emoción (ver Figura 2.2) como dos variables que componen el estado mental actual, de tal forma que tienen

[1] Por ejemplo en su evaluación, en QoE se utiliza el cuestionario *Mean Opinion Score* (MOS) para obtener una medida subjetiva de calidad, con semejanzas en su aplicación con los que se emplean en UX.

el mismo estatus ontológico que las variables que describen procesos cognitivos; proponiendo que las interacciones entre emoción y cognición pueden entenderse en el contexto de los estados mentales que pueden ser cambiados o controlados por eventos internos o externos.

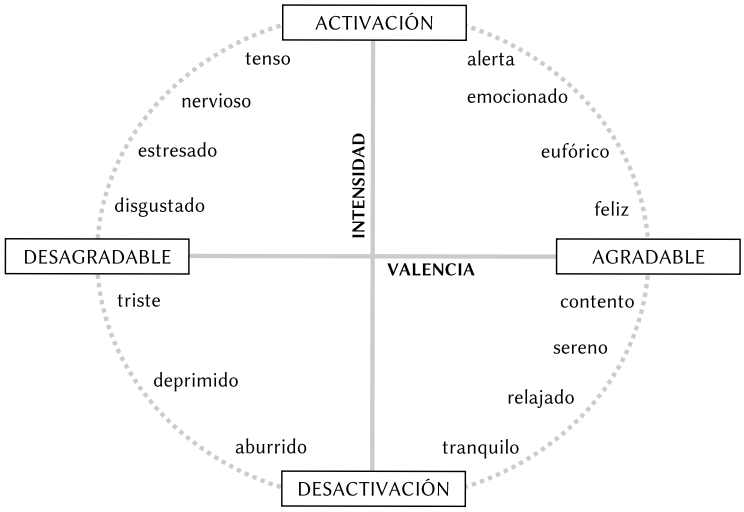


Figura 2.2: Representación gráfica del modelo de afecto
Fuente: Posner, Russell y Peterson [6]

La relación entre cognición y emoción ha sido tratada por otros autores; Robinson, Watkins y Harmon-Jones [32] hacen una revisión general de enfoques propuestos en varias subdisciplinas concluyendo que la cognición y la emoción interactúan de formas complejas que necesitan apreciarse en términos matizados, requiriendo el análisis detallado del contexto.

En el contexto de evaluación de UX varios enfoques tratan con el reconocimiento de estados emocionales (p. ej., [4], [15], [17], [25]), sin embargo, no consideran estados cognitivos que se han estudiado en otros dominios (p. ej., [10]-[12]).

La cognición se refiere a procesos como la memoria, la atención, el lenguaje, la resolución de problemas y la planificación [33]. A partir de estos procesos se identifican estados como carga de trabajo, estrés mental, atención, entre otros. Los estados cognitivos pueden manifestarse de diferentes formas, por ejemplo, los usuarios pueden exhibir carga de trabajo con variaciones en el rendimiento en la tarea y con manifestaciones fisiológicas [7], o el estrés mental puede generar respuestas físicas como agitación, ansiedad y sudoración [23].

2.1.3. Carga de trabajo

Vidulich y Tsang [7] presentan la carga de trabajo o carga cognitiva[1] como una función entre la oferta y demanda de recursos de atención o procesamiento; identificando dos determinantes principales: el suministro interno de los recursos (percepción, actualización de memoria, toma de decisiones, etc.) y las demandas externas de la tarea (dificultad, prioridad, etc.; representadas por el "mundo", ver la Figura 2.3).

La conciencia de la situación o *situation awareness* representa el contenido de los recursos cognitivos o de atención en un momento determinado y se caracteriza en tres fases: la percepción de los elementos en el ambiente, la comprensión de la situación actual y la proyección del estado futuro de eventos y elementos en el ambiente [22].

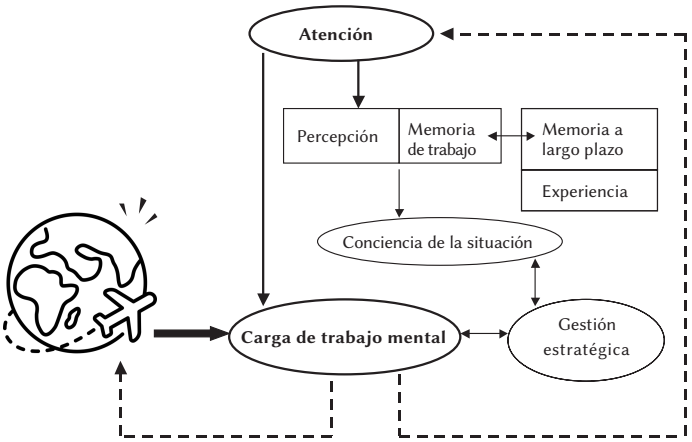


Figura 2.3: Relación entre carga de trabajo mental y la conciencia de la situación
Fuente: Vidulich y Tsang [7]

Según Zhou, Yu, Chen *et al.* [38], la carga de trabajo puede incrementarse por varios factores como: dificultad de la tarea, los materiales o herramientas utilizados, el contexto situacional, el contexto social, la experiencia de la persona en el campo, entre otros.

Adicionalmente, los mismos autores identifican cuatro tipos de técnicas de medición para determinar la carga de trabajo: subjetivas, de rendimiento, fisiológicas y de comportamiento; teniendo al cuestionario NASA-TLX (del inglés *NASA Task Load Index*) como el principal instrumento subjetivo y describiendo relaciones entre el estado cognitivo y las señales fisiológicas de actividad eléctrica del corazón y del cerebro, conductancia de la piel y actividad ocular.

[1] Diversos autores identifican el mismo constructo como carga de trabajo, carga de trabajo mental, carga cognitiva o incluso carga de trabajo cognitiva (p. ej., [34]-[37]).

2.1.4. Tecnologías para captura de datos

En la estimación de estados emocionales y cognitivos se utilizan diversas tecnologías para capturar datos inherentes al usuario. Estas tecnologías pueden agruparse en tres categorías [39]:

- Basadas en percepción, capturan elementos de expresión humana y comportamiento, tales como: expresiones faciales, entonación y modulación de voz, movimientos corporales, información contextual (ejemplo: análisis de uso de dispositivos de entrada), etc.
- Fisiológicas, centradas en las respuestas subconscientes del cuerpo humano, como: latidos cardíacos, presión sanguínea, actividad cerebral, etc.; relacionadas con los sistemas neuroendocrino y nerviosos central y autónomo.
- Subjetivas, las conforman los instrumentos para obtener los autoreportes de los individuos acerca de cómo perciben su estado, son menos dependientes a dispositivos electrónicos que las dos anteriores.

A continuación se describen algunas de las tecnologías más utilizadas para datos fisiológicos.

La mayoría de los datos de seguimiento ocular (ET) son de comportamiento porque representan hacia donde mira el usuario en un determinado momento, por cuanto tiempo y la ruta que sus ojos siguen. Los movimientos son capturados a través de cámaras y métodos que iluminan el ojo, identifican la reflexión en la cornea y en la pupila y establecen el punto de mirada relacionado [40] (ver Figura 2.4); obteniendo características como fijaciones –pausas breves del movimiento del ojo en un área específica– y sacadas –movimientos rápidos del ojo que suceden entre una fijación y otra–.

Las mediciones de pupilometría pueden ser capturadas con dispositivos de ET generalmente de alto costo. Los datos del diámetro de las pupilas se consideran fisiológicos por estar directamente relacionados con el sistema nervioso autónomo [41].

La electroencefalografía (EEG) es una tecnología utilizada para capturar señales relacionadas con la actividad eléctrica cerebral, se registra mediante electrodos fijados en el cuero cabelludo distribuidos comúnmente bajo el estándar 10-20 [42]. Las ondas cerebrales se manifiestan como voltajes eléctricos oscilantes de millonésimas de voltio [43]. La Tabla 2.2 muestra las características de los principales cinco ritmos u ondas cerebrales reconocidos –con sus rangos de frecuencia

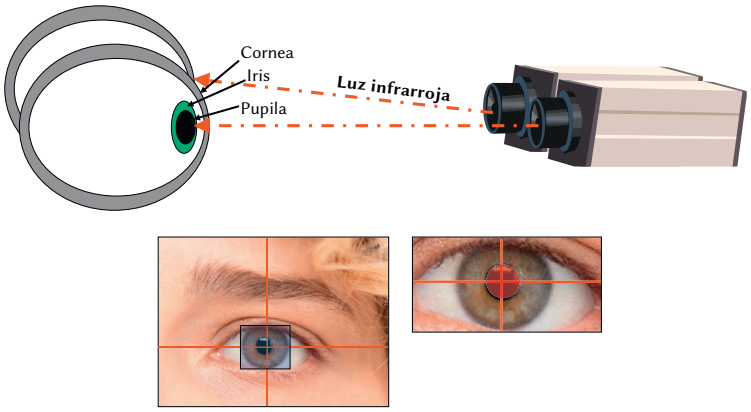


Figura 2.4: Figura conceptual del funcionamiento de la tecnología ET
Fuente: Schall y Romano Bergstrom [40]

específicos y estados relacionados– y la Figura 2.5 muestra ejemplos de las formas de onda de cada uno de ellos.

Tabla 2.2: Características de las cinco ondas cerebrales básicas

Banda de frecuencias	Frecuencia	Estados cerebrales
Gamma (γ)	>30 Hz	Concentración
Beta (β)	12-35 Hz	Ansiedad dominante, activo, atención externa, relajación
Alfa (α)	8-12 Hz	Mucha relajación, atención pasiva
Theta (θ)	4-8 Hz	Relajación profunda, centrado en sí mismo
Delta (δ)	0.5-4 Hz	Sueño

Fuente: Abhang, Gawali y Mehrotra [43]

Un electrocardiograma (ECG) permite capturar la señal eléctrica generada por la actividad muscular del corazón, se graba colocando un conjunto de electrodos en el tórax y ocasionalmente en las extremidades, dependiendo de la aplicación [42]. Un latido tiene cinco diferentes ondas (P, Q, R, S y T) que permiten determinar el ritmo y frecuencia cardíacas. La Figura 2.6 muestra el desglose de un fragmento de señal ECG identificando las ondas y los principales intervalos, puntos y segmentos.

La frecuencia cardíaca o tasa de latidos (HR, del inglés *Heart Rate*) es el número de veces que el corazón late en un minuto; el ritmo cardíaco es el patrón que siguen los latidos y puede ser descrito como regular/irregular o rápido/lento [45]. HR se mide calculando la distancia del punto R al punto R (pico a pico, ver Figura 2.7); la secuencia de intervalos de tiempo entre latidos determina la variabilidad de la frecuencia cardíaca (HRV, del inglés *Heart Rate Variability*) [46].

Tanto HR como HRV, pueden obtenerse a través de dispositivos de fotopletismografía (PPG).

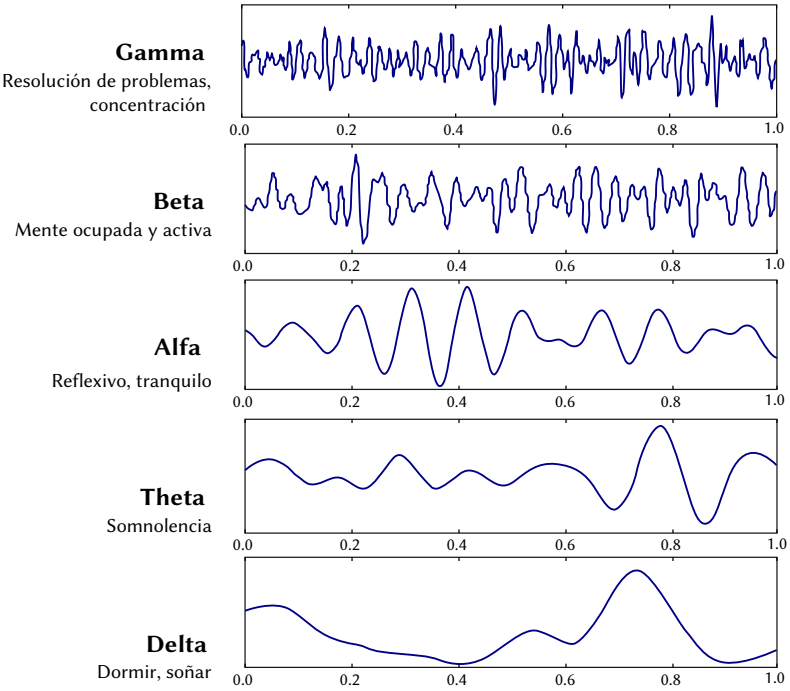


Figura 2.5: Ejemplos de las formas de ondas de cada ritmo cerebral y estados relacionados
Fuente: Abhang, Gawali y Mehrotra [43]

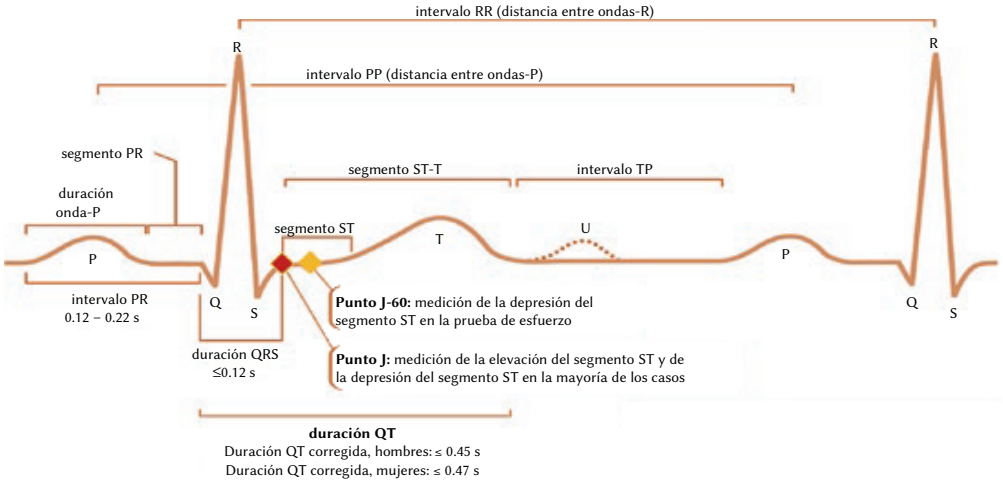


Figura 2.6: Ondas ECG, intervalos básicos, puntos y segmentos
Fuente: Okutucu y Oto [44]

La tecnología PPG mide los cambios del volumen sanguíneo en la vasculatura dérmica utilizando una fuente de luz para iluminar el tejido y un fotodetector para detectar las variaciones mínimas en la intensidad de la luz reflejada o transmitida asociadas a los cambios en el volumen de sangre [47]. Para la medición PPG el sensor puede ser colocado en dedos, muñecas o en los lóbulos de las orejas, por lo que resulta menos invasivo que ECG (ver Figura 2.8).

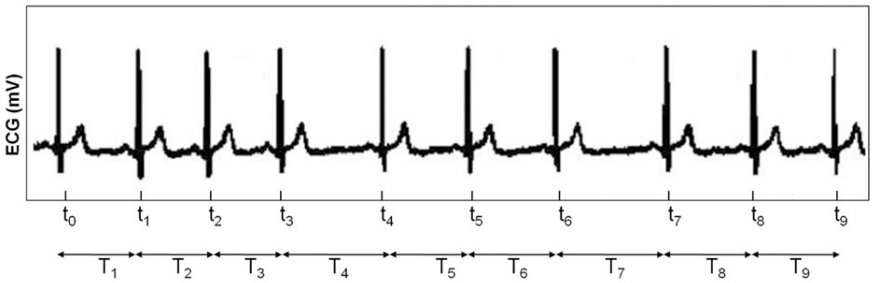


Figura 2.7: Ejemplo de una serie temporal con intervalos RR
Fuente: Baig y Kavakli [46]

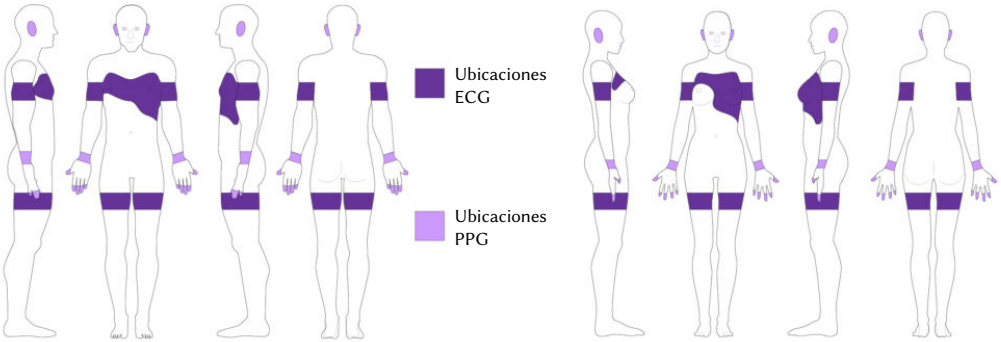


Figura 2.8: Ubicaciones para sensores ECG y PPG
Fuente: Zeagler [48]

La respuesta galvánica de la piel (GSR), también conocida como actividad electrodermal (EDA, del inglés *Electrodermal Activity*), proporciona una medición de la resistencia eléctrica de la piel al colocar dos electrodos en las falanges distales de los dedos medio e índice, la cual puede aumentar o disminuir acorde a la variación de la sudoración del cuerpo humano [49].

Relacionado con la captura de datos fisiológicos pero no propiamente con un tipo de dato en específico, *Lab Streaming Layer* (LSL)[1] es un sistema para la adquisición y sincronización de datos multimodales que consiste en una librería base y una serie de herramientas que incluyen el programa de grabación *LabRecorder*, importadores de archivo y aplicaciones propietarias u *open source* para trabajar con datos diversos (EEG, PPG, ET, GSR, etc.) y de varios fabricantes (Emotiv, BrainVision, SMI, Tobii, iMotions, etc.).

Junto con LSL, se diseñó la especificación de archivo XDF (del inglés *Extensible Data Format*), este formato es utilizado para almacenar en forma binaria los datos capturados y estructurar los metadatos en un esquema XML.

[1] Documentación disponible en `https://labstreaminglayer.readthedocs.io/index.html`

2.1.5. NASA Task Load Index

NASA-TLX [50] es un cuestionario estandarizado que brinda una medición de carga de trabajo considerando las ponderaciones de cada usuario a las subescalas de: demanda mental, demanda física, demanda temporal, rendimiento, esfuerzo y frustración. Su aplicación se realiza después de que el usuario termina cada tarea e incluye los siguientes pasos:

1. Valoración de subescalas: el usuario valora cada subescala marcando alguna de las 21 líneas de gradación para cada una de ellas, las cuales representan desde el valor "muy bajo" hasta el valor "muy alto" (salvo la subescala de rendimiento que va de "perfecto" a "fracaso"), correspondientes a valores numéricos desde el 0 hasta el 100. La Figura 2.9 muestra la versión moderna en inglés para la aplicación en papel de esta actividad.
2. Elección de origen de carga de trabajo: el usuario elige entre parejas de subescalas para indicar la que considera es más preponderante para la carga de trabajo experimentada. Está actividad puede realizarse para cada tarea o utilizar los mismos pesos para toda la sesión, dependiendo de los estímulos utilizados y el criterio del investigador.
3. Cálculo de pesos de subescalas: el investigador utiliza las tarjetas de comparación de subescalas y determina el peso de cada una acorde a su frecuencia de selección.
4. Cálculo de calificación ponderada de carga de trabajo: el investigador utiliza los pesos y la valoración de las subescalas para obtener la calificación ajustada de cada subescala; la calificación ponderada se obtiene de sumar las calificaciones ajustadas y dividir el resultado entre quince, obteniendo un valor entre 0 y 100.

NASA-TLX maneja el término de "carga de trabajo" como un constructo hipotético relacionado con el costo en el que incurre un humano para alcanzar un nivel específico de rendimiento en una tarea, con diferente relevancia para diferentes individuos y resultado de una combinación implícita de factores; utilizando las subescalas como punto de partida para luego obtener una medida integrada, tomando en cuenta la dificultad de los participantes para cuantificar, recordar y verbalizar sus impresiones a través de una sola escala general.

A pesar de que conceptualmente la "carga de trabajo mental" pudiera estar más relacionada con la subescala de "demanda mental", esta variable es difícil de cuantificar directamente y se encontró que está altamente correlacionada con la carga de trabajo general y con otras variables (p.

ej., frustración), por lo que su aporte de información se ve reducido por su falta de independencia [50].

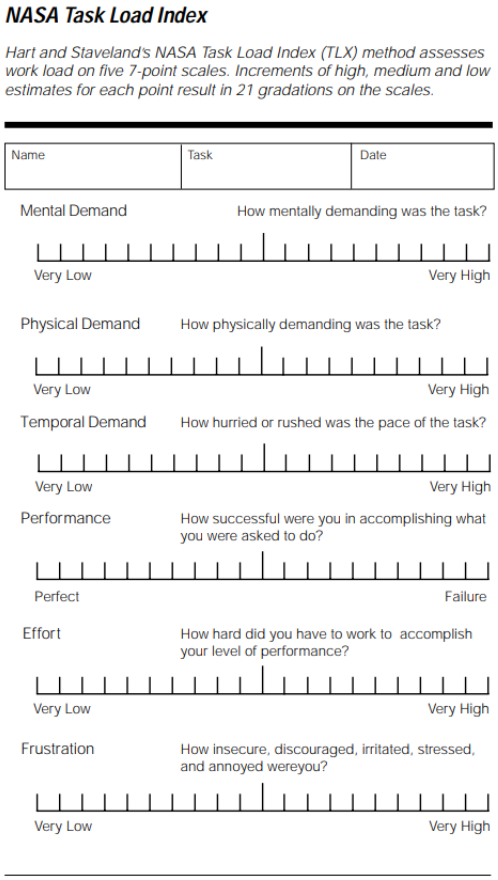


Figura 2.9: Ponderación de subescalas, NASA-TLX
Fuente: *NASA TLX Paper and Pencil Version*

2.1.6. Aprendizaje automático

El aprendizaje automático o *machine learning* se identifica como un área de investigación en ciencias de la computación, específicamente en inteligencia artificial, y se relaciona con algoritmos mediante los cuales una computadora puede aprender a través de un proceso de entrenamiento con un conjunto de ejemplos de entrada. Los ejemplos son relevantes para una tarea, por lo que el algoritmo transforma los datos de entrada en representaciones útiles que se aproximen a la salida esperada, llegando eventualmente a descubrir reglas que permitan automatizarla [51].

Aunque es común tratar de manera indiscriminada los términos algoritmo y modelo de aprendizaje automático, un modelo representa la salida o producto del entrenamiento de un determinado algoritmo con datos e hiperparámetros específicos [52].

Goodfellow, Bengio y Courville [53] describen el término "aprendizaje" retomando la definición de Mitchell [54]: "se dice que un programa de cómputo aprende de la experiencia E con respecto a alguna clase de tareas T y medida de rendimiento P, si su rendimiento en tareas T, medido por P, mejora con la experiencia E". La instanciación de las variables anteriores en el contexto de esta tesis es:

- Tarea T: Clasificación binaria de carga de trabajo.
- Medida de rendimiento P: Exactitud.
- Experiencia E: Conjunto de datos fisiológicos (aprendizaje supervisado).

La exactitud o *accuracy* (acc) es una métrica que representa la proporción de ejemplos correctamente clasificados entre el número total de ejemplos examinados [55], se determina por la expresión

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

donde las variables corresponden a los valores en la matriz de confusión de: verdaderos positivos TP, falsos positivos FP, verdaderos negativos TN y falsos negativos FN.

2.2. Antecedentes

En el grupo de Sistemas Híbridos Inteligentes del TecNM/CENIDET se han desarrollado tesis de doctorado y de maestría que se consideran como antecedente directo a esta investigación.

En la tesis doctoral de Alejandres Sánchez [56] se describen un par de experimentos utilizando sistemas de recomendación sensibles al contexto, con y sin interfaces de realidad aumentada y con diferentes esquemas de explicación, realizados con la finalidad de obtener una valoración cuantitativa de aspectos subjetivos de la UX (efectividad, confianza y satisfacción) empleando cuestionarios como instrumentos de evaluación.

González Franco [57] presentó en su tesis doctoral la metodología de evaluación de UX denominada *UXEeg*, aplicable a tecnologías de asistencia o rehabilitación para personas con discapacidad. Definió tres tipos de evaluación y seis escenarios para la selección de actividades con la finalidad de desarrollar experimentos para identificar correlaciones entre estados mentales

y componentes de la UX utilizando datos EEG y otros instrumentos de evaluación.

En su investigación de maestría, Fouilloux Quiroz [58] implementó un método para capturar e integrar datos EEG, multimedia y de otros sensores en pruebas de evaluación de UX. Utilizando los lenguajes de programación Python y C#, desarrolló una aplicación de grabación y otra de reproducción, respectivamente, superando problemas de sincronización de datos de diversos dispositivos y presentando la información a través de cuadros de video y gráficas con manipulación a través de una línea de tiempo.

En sus estudios de maestría, García Pinzón [59] continuó los desarrollos previos, depurando código, agregando módulos y nombrando formalmente la plataforma como *UXLab*, presentando su primera versión. Además, realizó pruebas de clasificación emocional con los algoritmos de Naive Bayes, máquinas de soporte vectorial (SVM, del inglés *Support Vector Machine*) y K-vecinos más cercanos (kNN, del inglés *k-Nearest Neighbor*); con datos capturados utilizando la metodología de Soriano Terrazas [60] para inducción emocional a través de realidad virtual inmersiva.

Finalmente, Lagunes Ramírez [61] y Morales Morante [62] complementaron la herramienta *UXLab* con los resultados de sus investigaciones de maestría; el primero, permitiéndole estructurar datos de ET, generar videos y analizar métricas oculares (mapas de calor, mapas de rutas, áreas de interés, entre otras); el segundo, agregando código para el procesamiento de datos fisiológicos, además de modelos para la predicción de emociones con base en las propiedades de valencia y activación.

2.3. Estado del arte

La búsqueda y análisis del estado del arte se realizó bajo un protocolo de Revisión Sistemática de Literatura (SLR, del inglés *Systematic Literature Review*)[1], el objetivo fue identificar y analizar investigaciones que contemplaran el reconocimiento de estados cognitivos en el contexto de UX, en particular aquellas donde utilizaran algoritmos de aprendizaje automático y datos inherentes al usuario.

A continuación se describen los artículos más relevantes y los principales hallazgos encontrados, agrupados acorde a varios tópicos generales.

[1] Para definir el protocolo se consideraron las recomendaciones de Kitchenham y Charters [63].

2.3.1. Clasificación de estados cognitivos

En esta sección se describen los trabajos que tratan directamente con la clasificación de estados cognitivos con modelos de aprendizaje automático.

Dos trabajos abordan la predicción de confusión con modelos RF (*Random Forest*). Lallé, Conati y Carenini [64] identificaron las características relacionadas con el diámetro de las pupilas como las más importantes en la predicción de este estado cognitivo. Salminen, Nagpal, Kwak *et al.* [65] trabajaron con datos de ET (sin pupilometría), edad y género; empleando dos técnicas para aumento de datos y encontrando a SMOTE [66] como la de mejor rendimiento y a la edad como característica más influyente.

La investigación de Mathur, Lane y Kawsar [67] es la única del estado del arte que describe un estudio de larga duración, tres meses, donde su objetivo fue el reconocimiento de compromiso con las bitácoras de uso del teléfono celular y el contexto del usuario. Por otro lado, el trabajo de Huang, Li, Ngai *et al.* [23] destaca porque utilizaron datos de una cámara web convencional y del ratón, proponiendo un enfoque no intrusivo para inferir el nivel de estrés mental con base en el patrón clic-mirada.

En cuanto a estudios relacionados con carga de trabajo. En [10] descubrieron que el rendimiento en una tarea era mejor cuando se utilizaba el teclado en comparación con una interfaz *touch* y que además los usuarios informaban de un menor índice de carga de trabajo. Jimenez-Molina, Retamal y Lira [9] probaron varios modelos para reconocer carga de trabajo mental al navegar en un sitio web ficticio, agrupando y etiquetando los datos fisiológicos considerando la relación entre diámetro de la pupila y la carga mental.

Gjoreski, Kolenik, Knez *et al.* [68] introdujeron el conjunto de datos CogLoad en un contexto de clasificación entre tareas de descanso y tareas cognitivas que inducen carga de trabajo; realizaron pruebas considerando enfoques de transformación de características por participante y resaltaron la importancia de utilizar un esquema de validación LOSO (*Leave One Subject Out*) para evaluar los modelos de aprendizaje. También presentaron Snake, un conjunto de datos construido usando de estímulo el tradicional juego de víbora para dispositivos móviles.

La Tabla 2.3 resume las características de los estudios descritos previamente en esta sección,

incluyendo el mejor modelo de clasificación acorde a la métrica reportada.

Tabla 2.3: Resumen de artículos con clasificación de estados cognitivos

Ref.	Estados cognitivos	Modelos con mejor rendimiento	Participantes (Femenino/-Masculino)	Estímulo	Datos
[64]	Confusión	RF, sensibilidad 0.61, especificidad 0.926	136 (75F/61M)	Software de visualización de datos	Autoreportes, ET (con pupilometría), clicks
[65]	Confusión	RF, rango de exactitud 72.6-99.1 %	29 (14F/15M)	Hojas de datos personales	ET, edad, género
[23]	Estrés	RF, nivel-clic usuario-dependiente f1-score 0.66; clasificador logístico, nivel-sesión usuario-independiente f1-score 0.79	20 (7F/13M)	Software de preguntas aritméticas	ET (de video), clicks
[67]	Compromiso	SVM, f1-score 0.82	10 (3F/7M), 10 (3F/7M), 130 (34F/96M)	Uso de teléfono celular	Estudios 1 y 2: EEG y bitácoras de uso; estudio 3: bitácoras, contexto y datos demográficos
[10]	Carga de trabajo, atención	LDA, exactitud: 92 % carga de trabajo y 86 % atención	12 (3F/9M)	Juego de laberinto virtual	Autoreportes, EEG, comportamiento de teclado y *touch*
[9]	Carga de trabajo	MLP, exactitud 93.7 %	61 (19F/42M)	Navegación web	EDA, PPG, temperatura, ECG, EEG, ET (con pupilometría)
[68]	Carga de trabajo (CogLoad)	Tipo *bagging* utilizando árboles de decisión, exactitud 68.2 %	23 (4F/19M)	Tareas cognitivas y n-back	Autoreportes, GSR, PPG, temperatura, PPG y acelerómetro
[68]	Carga de trabajo (Snake)	XGB, exactitud 82.3 %	23 (7F/16M)	Juego Snake	Autoreportes, GSR, PPG, temperatura, PPG y acelerómetro

Ninguno de los enfoques previos utilizó modelos de aprendizaje profundo en alguna parte del proceso. En otros contextos se han observado buenos resultados con arquitecturas de tipo autoencoder (p. ej., [69], [70]) y de tipo convolucional (p. ej., [71], [72]); sin embargo, su uso es ineficiente si el número de participantes en los experimentos es reducido y se capturan pocos datos, ya que los modelos de aprendizaje profundo requieren una cantidad significativa de datos para aprovechar su potencial [73].

Las investigaciones [65] y [64] consideraron técnicas como SMOTE o ADASYN [74] para el aumento de datos y el equilibrio de clases; no se encontró el uso de otras técnicas o modelos para generar datos sintéticos, como los basados en redes tipo GAN [75], que están siendo estudiados y evaluados en otros contextos (p. ej., [76], [77]).

Se analizaron trabajos relacionados con carga de trabajo en otros dominios para conocer las características preponderantes que se utilizan de cada señal fisiológica, no se describen a detalle por no estar vinculados con UX, sin embargo se consideran para mencionar algunos hallazgos.

Es común el uso de datos EEG para reconocer carga de trabajo, fundamentado en correlaciones reportadas con ese estado (p. ej., [14]). De entre varias investigaciones ([8]-[10], [12], [71],

[78]-[83]), en la mayoría identificaron los ritmos cerebrales calculando la densidad espectral de poder (PSD, del inglés *Power Spectral Density*) utilizando FFT, agregando características estadísticas como media, desviación estándar, etc. No se encontraron trabajos recientes que consideraran potenciales relacionados con el evento, como se reportaba en revisiones previas (p. ej., [84]).

Considerando varios estudios ([9], [24], [68], [78], [82], [85]) se observa que para el reconocimiento de carga de trabajo es común utilizar datos ECG/PPG en conjunto con los de otras señales; extrayendo características del dominio de la frecuencia (LF, HF, etc.), del dominio del tiempo –basadas en intervalos RR– y otras como las basadas en la gráfica Poincaré, la cual permite observar la variabilidad de la tasa de latidos contrarrestando los intervalos RR con los subsecuentes.

Tomando en cuenta trabajos que capturan datos GSR ([9], [68], [78], [82]), prevalecen las características estadísticas a partir de los valores obtenidos de conductancia, además, se observa la combinación sugerida de combinar estas con las obtenidas a partir de datos ECG/PPG [86].

Analizando investigaciones que consideran datos ET ([9], [11], [87], [88]), se confirma que las características de pupila son preponderantes para el reconocimiento de carga de trabajo, como se concluía en revisiones previas (p. ej., [21], [39]). Sin embargo, las mediciones de pupilometría se obtienen con precisión generalmente con dispositivos ET de alto costo, lo que limita su utilización.

En general, las investigaciones no informan del tiempo de preparación dedicado a cada participante, lo cual puede estar limitado por el tipo y el número de dispositivos de medición que deben configurarse. En los dispositivos de EEG no invasivos un mayor número de electrodos puede implicar más tiempo de calibración. En los dispositivos de ET, el tiempo de calibración es menor pero deben tomarse en cuenta las condiciones de iluminación del entorno.

En el caso del monitoreo de la actividad cardíaca, se obtiene una mayor información y precisión con ECG, con el inconveniente de que los sensores son más intrusivos y su instalación requiere de un protocolo más estricto en comparación con los basados en PPG. En el caso de GSR, los sensores suelen colocarse en los dedos, limitando el movimiento pero dedicando poco tiempo a su preparación.

Para seleccionar adecuadamente el tipo y cantidad de dispositivos de medición utilizados

en las evaluaciones de UX, se pueden tener en cuenta las recomendaciones de Zeagler [48] para dispositivos vestibles o *wearables* y las de Erins, Minejeva, Kivlenieks *et al.* [89] en el contexto de la detección de fatiga, ya que la intrusividad e interferencia con la tarea debe ser mínima y para ello es necesario considerar aspectos como la percepción del peso, el movimiento del usuario, entre otros.

Antes de determinar los sensores a utilizar, es necesario evaluar la conveniencia de medir el conjunto de estados cognitivos propuestos en una determinada aplicación, para ello pueden considerarse los atributos aportados por Charlton [22] relativos a sensibilidad, intrusión, diagnóstico, conveniencia de la medición, relevancia, transferibilidad y aceptación.

Es común que los experimentos informen la edad y el sexo de los participantes pero no presenten conclusiones diferenciadas. Se ha observado que las diferencias individuales dadas por diversos factores pueden influir en las señales fisiológicas [19]; sin embargo, pocos estudios tienen en cuenta estos factores (como por ejemplo lo hacen en [90]). Además, se reafirma lo encontrado en [46] en relación a que no se realizan experimentos estandarizados y la falta de uniformidad dificulta establecer comparaciones entre los resultados.

En relación a los métodos para evaluar el rendimiento de los modelos de aprendizaje, LOSO permite una mejor generalización porque se realiza el entrenamiento dejando los datos no conocidos de un participante para su evaluación, lo que justifica su uso a pesar de que tiende a generar métricas con valores más bajos en comparación con enfoques de validación cruzada de *n-folds* [78][91]-[93].

La generalización de los modelos es un reto en UX y otros contextos, donde las diferencias entre los estudios académicos y las pruebas en la vida real prevalecen. Por ejemplo, en los experimentos académicos que han tratado con señales fisiológicas como complemento comúnmente han reclutado tantos participantes como fue posible –una revisión encontró una media de 24 participantes en una muestra de 33 investigaciones [94]– y las pruebas en la vida real usan principalmente mediciones cualitativas y con pocos participantes.

Incluso, según el análisis de algunos líderes de la industria [95], cinco participantes podrían ser suficientes para evaluar la UX, pudiendo variar dependiendo de los recursos disponibles y otras condiciones.

2.3.2. Arquitecturas de evaluación

En esta sección se describen investigaciones que contemplan estados cognitivos pero que hacen énfasis en arquitecturas de evaluación.

La plataforma Lean UX es presentada por Hussain, Khan, Hur *et al.* [17], su objetivo es apoyar a los evaluadores en la interpretación de medidas observacionales, fisiológicas y tradicionales. Su arquitectura se compone de varias capas e incluye módulos para el reconocimiento de emociones y estrés mediante el análisis de datos de EEG y ET, así como para el reconocimiento de emociones mediante el análisis de expresiones faciales, lenguaje corporal y voz a partir de vídeos y sonidos capturados con una cámara web y un micrófono.

Un conjunto de trabajos relacionados describen un enfoque para evaluar la UX teniendo como herramienta a los mapas de calor fisiológicos [25], los cuales amplían los mapas tradicionales de calor de la mirada para representar el estado mental del usuario al interactuar con la interfaz. Estos mapas fueron validados en un experimento con páginas web [96] y aunque se relacionaron con la complejidad visual, se determinó que para maximizar su utilidad se deben integrar al análisis tradicional (cuestionarios, entrevistas, etc.).

Complementando su investigación, Georges, Courtemanche, Sénécal *et al.* [16] evaluaron con participantes expertos la aceptación y utilidad de informes de UX parcialmente completados con imágenes de mapas de calor fisiológicos, encontrando que su uso es factible en la práctica, recibiendo comentarios positivos y sugerencias de mejora.

Además, en [20] determinaron los requisitos que debe cumplir una herramienta de evaluación de UX que considere datos fisiológicos y autoreportes, destacando la necesidad de automatizar el procesamiento de datos y entregar resultados útiles en tiempo y forma para los equipos de desarrollo de software que siguen metodologías ágiles.

Las arquitecturas de evaluación analizadas consideran varios tipos de sensores y la detección de diversos estados mentales: Hussain *et al.* [17] hacen hincapié en las características y el rendimiento independiente de los modelos utilizados en cada módulo de detección; Courtemanche *et al.* [16][20] [25][96] destacan la importancia de las herramientas para representar los estados mentales de los usuarios y su utilidad con respecto a los evaluadores que las interpretan, considerando los

requisitos que exige la industria.

En general, las arquitecturas definen módulos o capas para la captura de datos y su procesamiento, para el análisis y cálculo de métricas y para la generación y presentación de resultados, iniciando el proceso con el usuario realizando una tarea y terminando con un evaluador experto interpretando los resultados y generando o complementando un informe final con los hallazgos detectados en las pruebas.

2.3.3. Correlaciones con métricas UX o estados cognitivos

En esta sección se describen artículos donde se encontraron correlaciones entre las diferentes señales fisiológicas con métricas de UX y/o estados cognitivos.

Chai, Ge, Liu *et al.* [97] investigaron la relación de la asimetría frontal alfa de EEG con la experiencia y dificultad en la tarea al interactuar con un conjunto de aplicaciones móviles, no encontrando correlaciones significativas.

Yao, Liu, Li *et al.* [98] analizaron la relación de las características de GSR con métricas de rendimiento; identificando que las tareas con una menor tasa de completez tienen una tendencia no significativa a causar valores GSR más altos y que en el caso de atracción, eficiencia, fiabilidad y la novedad la correlación varia en rangos desde 0.46 hasta 0.58.

En [99], se encontraron aumentos significativos en la ondas cerebrales beta y gamma de EEG durante eventos relevantes en un juego de plataforma en comparación con eventos normales del juego y con otra tarea cognitiva. McMahan, Parberry y Parsons [100] también evaluaron el compromiso con la tarea y la activación utilizando índices calculados a partir de las bandas de potencia de EEG y establecieron umbrales y un conjunto de reglas para definir un modelo de flujo o inmersión en el juego.

En una investigación se planteó la hipótesis de que el usuario experimenta menos carga cognitiva cuando el método de clasificación de productos está en consonancia con el objetivo de búsqueda; interpretando la carga cognitiva a partir de datos EEG [101].

En [102] se buscó la correlación entre datos de ET con autoeficacia, percepción de riesgo, facilidad de uso percibida y utilidad percibida en tareas con un asistente de software; encontrando

la correlación más fuerte entre la facilidad de uso percibida y el número de fijaciones que se convierten en clics.

Juanéda, Sénécal y Léger [103] utilizaron fijaciones para medir la atención sobre un producto focal y sobre distractores similares o disímiles en posiciones cercanas o lejanas. Entre otros hallazgos, encontraron que los individuos prestan menos atención al producto focal cuando los distractores están cerca, siendo más acentuado cuando los distractores no son similares.

Desrochers, Léger, Fredette *et al.* [104] evaluaron la actitud de los consumidores hacia un sitio en-línea de compra considerando dos tipos de productos y tareas de diferente complejidad aritmética. Obtuvieron la atención visual y la carga cognitiva a través del análisis de las fijaciones y del diámetro pupilar, respectivamente, encontrando que la atención hacia las imágenes de los productos influía de forma diferente en la actitud hacia el sitio en función de las características de la tarea y de la carga cognitiva relacionada.

Por último, Federici, Mele, Bracalenti *et al.* [13] evaluaron la usabilidad de una aplicación web buscando correlaciones entre cuestionarios subjetivos, EEG y emociones a través de expresiones faciales; concluyeron que las mediciones EEG son necesarias ya que observaron que la disminución en la motivación de uno de los grupos de prueba no se reflejó en los autoreportes pero si en el incremento de actividad cerebral.

La Tabla 2.4 resume las características de las investigaciones descritas en esta sección.

Persiste la utilidad de los cuestionarios de autoreporte, habiendo un consenso de que enfoques multimodales son necesarios para entender totalmente contextos relacionados con UX. Adicionalmente, se mantiene tanto la escasez de más y mejores conjuntos de datos y de mecanismos para hacerlos compatibles e integrarlos, como la necesidad de estandarizar las metodologías para capturar e interpretar las mediciones fisiológicas [21].

2.3.4. Conjuntos de datos

Se analizaron conjuntos de datos que pudieran ser útiles para realizar experimentos de clasificación de estados cognitivos (ver Tabla 2.5 para más detalle):

- HciLab [105], incluye datos ECG, GSR y temperatura durante actividades de conducción

Tabla 2.4: Resumen de artículos con correlaciones con métricas UX o estados cognitivos

Ref.	Objetivo	Participantes (Femenino/-Masculino)	Estímulo	Datos
[97]	Correlaciones entre asimetría frontal alfa, experiencia y dificultad en la tarea	20 (10F/10M)	Tareas en aplicaciones móviles	Autoreportes, EEG
[98]	Correlaciones entre GSR y métricas de rendimiento en la tarea	20 (10F/10M)	Tareas en aplicaciones móviles	Autoreportes, GSR, PPG, HR y respiración
[99]	Análisis de poder EEG durante tareas con diferencias cognitivas	30 (20F/10M)	Tarea cognitiva Dos-Imágenes y videojuego	EEG y videos de pantalla y frontales
[100]	Análisis de estado de flujo con base en compromiso e índices de activación	30 (20F/10M)	Videojuego	EEG y videos de pantalla y frontales
[101]	Análisis de carga de trabajo, ordenamiento de productos y metas de los usuarios	21 (10F/11M)	Tareas de compra en-línea	EEG
[102]	Correlaciones entre ET, aceptación y percepción	10 (7F/3M)	Asistente de creación de bases de datos	Autoreportes, ET (con pupilometría), clicks y video de pantalla
[103]	Atención visual y análisis de rendimiento en la tarea	38 (no indicado)	Tareas de compra en-línea	ET
[104]	Análisis de la actitud ante un sitio web considerando atención visual, carga cognitiva, tipo de producto y complejidad aritmética	38 (17F/21M)	Tareas de compra en-línea	Autoreportes, ET (con pupilometría)
[13]	Evaluación de usabilidad	30 (15F/15M)	Tareas en sitio web	Autoreportes, videos de pantalla y frontales, bitácoras de uso de ratón y teclado, EEG

real con el objetivo de detectar carga de trabajo.

- SWELL-KW [106], presentado en el contexto de investigación en estrés, incluye datos ECG, GSR, bitácoras de computadora, expresiones faciales y posturas corporales de participantes realizando actividades en un escenario realista de oficina y con interrupciones a través de correos electrónicos; capturando autoreportes como NASA-TLX, escala Likert de estrés, entre otros.
- Conjunto de datos de INRIA [107], contiene grabaciones EEG, ECG, EMG, GSR, entre otras; capturadas de participantes realizando actividades para inducir estrés a través de una entrevista de trabajo falsa y para inducir carga cognitiva con tareas *n-back*; utiliza varios cuestionarios estandarizados para el autoreporte subjetivo.
- STEW, construido por Lim, Sourina y Wang [81] con datos EEG de participantes masculinos, autoreporte con una escala de carga de trabajo mental y tareas del test psicológico SIMKAP.
- AffectiveROAD [108], incluye datos GSR, PPG, temperatura y respiración en actividades de conducción real y con etiquetado subjetivo de estrés.
- EEGMat [109], contiene grabaciones EEG de participantes durante la realización de restas mentales, agregando una métrica de rendimiento con base en la cantidad de restas correctas.
- Cogload [68], contiene datos GSR, PPG, temperatura y acelerómetro de participantes

durante actividades cognitivas para inducir carga de trabajo mental.

- Snake [68], contiene datos GSR, PPG, temperatura y acelerómetro de participantes jugando en el tradicional juego *Snake* para dispositivos móviles e inducir carga de trabajo mental.

Tabla 2.5: Conjuntos de datos relacionados con estados cognitivos

Conjunto de datos	Estado cognitivo	Participantes (Femenino/Masculino), edad promedio en años	Datos	Estímulos	Autoreportes
HciLab	Carga de trabajo	10 (3F/7M), 35.6	ECG, GSR y temperatura	Conducción	Ninguno
SWELL-KW	Estrés	25 (8F/17M), 25	ECG, GSR, bitácoras, expresiones faciales y posturas corporales	Actividades en escenario realista de oficina	NASA-TLX, escala de estrés y otros
INRIA	Estrés y carga de trabajo	24 (12F/12M), 24.7	EEG, ECG, EMG, GSR y otros	Entrevista falsa de trabajo y tareas *n-back*	STAI, SAM y RSME
STEW	Carga de trabajo	48 (0F/48M), no indicado	EEG	Test psicológico SIMKAP	Escala de carga de trabajo
AffectiveROAD	Estrés	10 (5F/5M), 29.9	GSR, PPG, temperatura y respiración	Conducción	Escala de estrés
EEGMat	Rendimiento en la tarea	36 (27F/9H), 18.25	EEG	Restas mentales	Entrevista final
Snake	Carga de trabajo	23 (7F/16M), 24.91	GSR, PPG, temperatura, HR, intervalos RR y acelerómetro	Juego *Snake*	NASA-TLX
CogLoad	Carga de trabajo	23 (4F/19M), 29.51	GSR, PPG, temperatura, HR, intervalos RR y acelerómetro	Tareas cognitivas y *n-back*	Hexaco y NASA-TLX

No se encontró un conjunto de datos que incluyera datos EEG, GSR, ECG y ET con estímulos para inducir carga de trabajo en el contexto de UX. Sin embargo, CogLoad destaca por inducir carga de trabajo, contar con autoreportes NASA-TLX y utilizar un estímulo de tareas cognitivas de diferente dificultad.

A continuación se describen las investigaciones relacionadas con CogLoad.

Inicialmente, Gjoreski, Luštrek y Pejović [110] abordaron la predicción de tres clases, modelos específicos a cada tarea, tres enfoques de etiquetado y validación LOSO. Trabajaron con características estadísticas y otras relacionadas con el análisis SCR y HRV extraídas de segmentos de datos de 90 segundos. Sus mejores exactitudes medias fueron del 47 % con un modelo RF y etiquetado basado en NASA-TLX, de 51 % con Naive Bayes y etiquetado por nivel de dificultad y del 46 % con un clasificador GC y etiquetas de opacidad.

En [68] introdujeron CogLoad en un contexto de clasificación de descanso frente a tarea, extrayendo características similares a las del trabajo anterior pero con datos de los últimos 30 segundos de cada tarea. El artículo detalla experimentos con características brutas, normalizadas por participante y estandarizadas por participante; con y sin selección de características utilizando

un método basado en información mutua. La exactitud más alta fue del 68.2 % utilizando un modelo de tipo *bagging* en combinación con características seleccionadas y normalizadas.

En [111], los autores presentaron los resultados de un reto de aprendizaje automático organizado con CogLoad como conjunto de referencia. El enfoque de clasificación fue de "no carga" frente a "carga cognitiva", similar al estudio anterior. De los trece métodos presentados, las exactitudes más altas fueron del 69.4 % con un modelo tipo *ensemble* y validación cruzada de *5-folds*, detallado en [112], y del 67.9 % con un modelo SVM en validación LOSO.

Jaiswal, Chatterjee, Gavas *et al.* [36] utilizaron ventanas de 30 segundos de datos normalizados por participante y extrajeron características con un método de tres niveles, extrayendo 1568 características totales y seleccionando 50 utilizando dos algoritmos del estado del arte. Probaron la clasificación de reposo frente a tareas a partir de CogLoad original y una versión aumentada utilizando SMOTE. En una validación LOSO, su mayor exactitud fue del 69.5 % con los datos originales, el conjunto de características recomendadas y un modelo basado en RF.

Tervonen, Pettersson y Mäntyjärvi [113] analizaron diferentes longitudes de ventana para clasificar la carga cognitiva y el descanso. Comparan los resultados obtenidos con modelos XGB –extrayendo características similares a [110] y [68]– pero en longitudes de ventana de 5s a 30s, con pasos de 5s. Alcanzan una exactitud del 67.6 % con una ventana de 25s, concluyendo que las ventanas más largas tienden a mejorar el rendimiento del modelo.

Finalmente, Salfinger [114] evaluó arquitecturas de aprendizaje profundo, normalización global y por participante y métodos de aumento de datos por desplazamiento temporal y sobremuestreo. El mejor rendimiento fue una exactitud media del 64 %, obtenida a partir de modelos ResNet entrenados con datos normalizados globalmente y sin aumento, considerando ventanas de 30 segundos y conjuntos de entrenamiento-validación-prueba de 12-3-3 sujetos aleatorios.

3 | Trabajo inicial

En este capítulo se describen las pruebas realizadas con un conjunto de datos del estado del arte y dos experimentos de evaluación de experiencia de usuario a sitios web de índole educativo, desde su formulación hasta la entrega de resultados.

Los hallazgos encontrados en esta sección permitieron estructurar la versión final de la metodología que representa el principal producto de esta investigación.

3.1. Pruebas con conjunto de datos del estado del arte

Teniendo en cuenta la escasa disponibilidad de conjuntos de datos en el contexto de UX (ver sección 2.3.4), se consideró CogLoad [68] para la experimentación inicial de clasificación binaria de carga de trabajo. Se determinó utilizar CogLoad principalmente porque induce carga de trabajo utilizando un estímulo que considera percepción visual y factores espaciales, aspectos que también son experimentados por un usuario cuando interactúa con un producto digital [115]-[117].

El estímulo, implementado inicialmente por Haapalainen, Kim, Forlizzi *et al.* [118], consiste de las siguientes seis tareas cognitivas (ver Figura 3.1):

1. Identificando la figura (GC: *Gestalt completion*), observar trazos incompletos de un dibujo y tratar de identificarlo.
2. Encontrando el patrón oculto (HP: *Hidden pattern*), identificar el patrón de trazos de un dibujo dentro de imágenes de comparación.
3. Encontrando las A's (FA: *Finding A's*), encontrar palabras que contienen la letra A.
4. Comparando números (NC: *Number comparison*), observar parejas de números y decidir si son iguales o no.
5. Buscando los caminos (PT: *Pursuit test*), rastrear visualmente líneas que empiezan en un lado e identificar su final en el otro.
6. Localizando las X's dispersas (SX: *Scattered X's*), encontrar letras X distribuidas en la pantalla junto con otras letras aleatorias.

El conjunto contiene datos de temperatura, GSR, HR, RR y acelerómetro de 23 participantes,

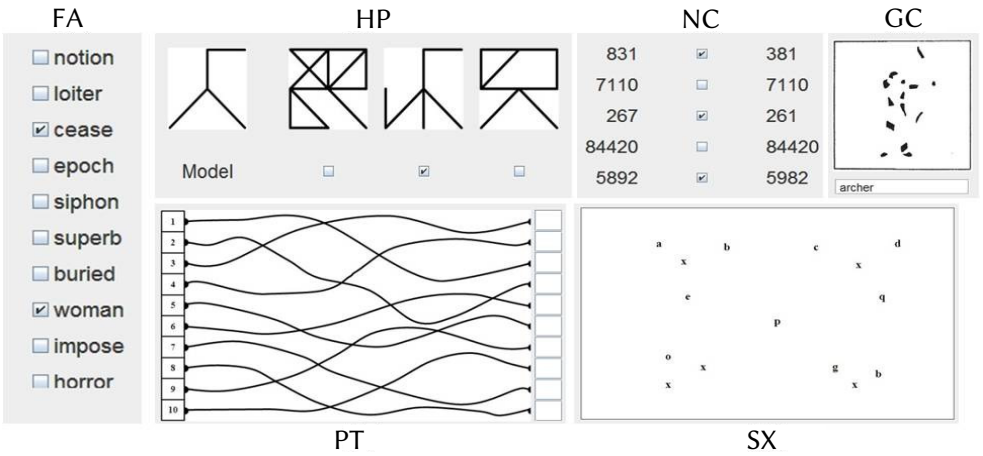


Figura 3.1: Implementación original de las seis tareas cognitivas
Fuente: Haapalainen, Kim, Forlizzi *et al.* [118]

grabados con una pulsera Microsoft Band y remuestreados a 1 Hz. El experimento incluyó dos tareas *n-back* (tareas secundarias), las tareas cognitivas en tres niveles de dificultad (tareas primarias) y una tarea intermedia para liberar los recursos cognitivos de los participantes entre cada tarea primaria. Además, los participantes respondieron un cuestionario de personalidad Hexaco una vez y un cuestionario NASA-TLX para evaluar la carga de trabajo después de cada tarea cognitiva.

Las columnas de las subescalas NASA-TLX contienen valores entre 0 y 5, excepto cuando los usuarios rellenaban cuestionarios o estaban descansando, y la columna TLX_mean es su suma, lo que indica un uso simplificado del instrumento. Además, cada participante sólo tuvo una sesión y cada agrupamiento de columnas de usuario, tarea y nivel representa un segmento de tarea individual con los mismos valores de autoreporte.

La Figura 3.2 presenta las correlaciones Spearman entre los valores de autoreporte y el nivel de dificultad de la tarea, considerando sólo los segmentos de datos de dificultades bajas y altas. Como era de esperar, muestra fuertes correlaciones entre la mayoría de las columnas de las subescalas y correlaciones moderadas, un coeficiente superior a 0.4, entre el nivel y el esfuerzo, el nivel y la demanda mental, y el nivel y el valor ponderado de la carga de trabajo (TLX_mean).

Los experimentos descritos en esta sección tienen un objetivo de clasificación de niveles de carga de trabajo y difieren de la predicción de descanso vs. tarea de la mayoría de los trabajos citados en la sección 2.3.4, sin embargo, se tomó la métrica de exactitud (*accuracy*) con valor del

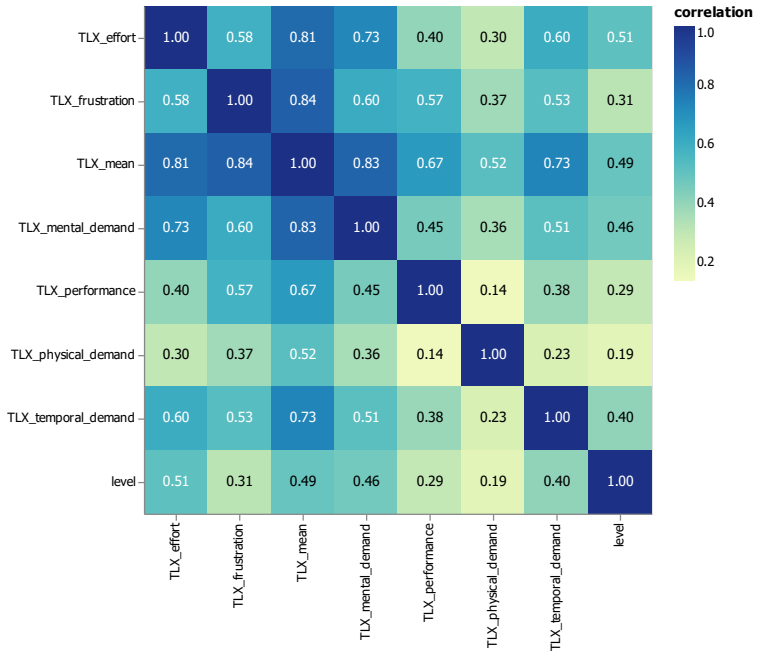


Figura 3.2: Correlaciones Spearman entre las subescalas NASA-TLX y el nivel de dificultad de la tarea

69.5 % como comparativo inicial ya que es el mejor resultado reportado con CogLoad.

Aunque distinguir niveles de carga de trabajo utilizando CogLoad es un reto [111], se buscó un modelo y condiciones de experimentación que permitieran distinguir al menos los niveles de carga de trabajo bajo y alto según los autoreportes de NASA-TLX.

3.1.1. Preprocesamiento de datos y extracción de características

Se codificó un flujo de trabajo de aprendizaje automático utilizando el lenguaje Python. Este flujo descarta inicialmente las columnas del acelerómetro, la temperatura y las subescalas de NASA-TLX, dejando sólo los valores fisiológicos de GSR, HR y RR, teniendo en cuenta el equipo de laboratorio con el que se podrían capturar datos similares en futuros experimentos y realizar comparativas.

Como siguiente paso, se eliminaron los datos de tareas no primarias y se etiquetaron los segmentos de tareas primarias utilizando el valor medio TLX_mean del usuario o de la escala. Para el primer enfoque, se asumió que si las tareas de baja y alta dificultad tienen una relación con la carga cognitiva elicitada [118], sería posible tomar la media de los valores TLX_mean de cada usuario como punto de corte para etiquetar sus datos, los valores superiores como clase alta

y los valores inferiores como clase baja. En el segundo enfoque, se consideró el intervalo 0-30 de los valores TLX_mean y se utilizó el valor central, quince, como punto de corte en general. Adicionalmente, se detectaron y removieron tareas con valores fisiológicos repetidos generados probablemente por un problema en el proceso de captura.

De forma similar a los trabajos relacionados, se extrajeron características del dominio temporal: media, desviación estándar, asimetría, curtosis, primer y tercer cuartil, desviación entre cuartiles, coeficiente de variación, media de la primera derivada, media de la segunda derivada y diferencia entre los valores mínimo y máximo; teniendo un total de 36 características, tres señales fisiológicas en bruto y 33 de tipo estadístico. Para efectos de experimentación, la extracción de características se realizó en ventanas de longitud fija y ventanas variables en función de la duración de cada segmento.

3.1.2. Modelos y experimentación

Al igual que en el artículo original [68], se generaron diferentes versiones del conjunto de datos variando el escalado de las características: sin escalado, con normalización min-max específica para cada participante y con estandarización específica para cada participante. Además, se experimentó con la eliminación de los primeros 20 segundos de cada segmento, dejándolos con una duración mínima de 30 segundos, teniendo en cuenta el posible efecto del inicio de tarea diferente en la carga de trabajo de los participantes.

Los modelos de aprendizaje automático utilizados fueron RF, XGB y SVM. Se comenzó con parámetros predeterminados y luego se tomaron los resultados de búsquedas de hiperparámetros aleatorias y de cuadrícula; usando principalmente los algoritmos implementados en la librería scikit-learn [119]; en el caso de XGB se utilizó la implementación de Chen y Guestrin [120].

Para conocer su influencia en el rendimiento de los modelos, también se experimentó eliminando las tareas de dificultad media y datos de participantes con valores de autoreporte NASA-TLX inesperados según dificultades bajas y altas, es decir, con segmentos de baja dificultad etiquetados con alta carga de trabajo o viceversa.

De igual forma, se realizaron pruebas con un modelo autoencoder convolucional para realizar una reducción de dimensionalidad y evaluar el rendimiento de los modelos; así como pruebas de

clasificación de tres niveles de carga de trabajo.

El enfoque de validación fue LOSO ya que se pretendía tratar el problema con un acercamiento que solventara las diferencias individuales, entrenando con los datos de todos los participantes menos uno y los datos del participante restante como conjunto de prueba. Se consideró la exactitud media entre participantes, se eligió esta métrica para fines comparativos, por ser la más reportada en los trabajos relacionados.

3.1.3. Resultados

La Figura 3.3 muestra el número de filas etiquetadas por usuario en el conjunto de datos filtrado tras eliminar las tareas no primarias y primarias con valores repetidos, indicando las clases altas y bajas en barras para cada enfoque de etiquetado. En el enfoque por media del usuario en TLX_mean todo el conjunto de datos presenta un desequilibrio del 7.3 %, con un desequilibrio medio del 16.4 % entre los subconjuntos de usuarios. Considerando el enfoque del valor central de la escala, el desequilibrio del conjunto de datos completo es del 8.5 %; sin embargo, el desequilibrio medio entre usuarios es del 49.5 %.

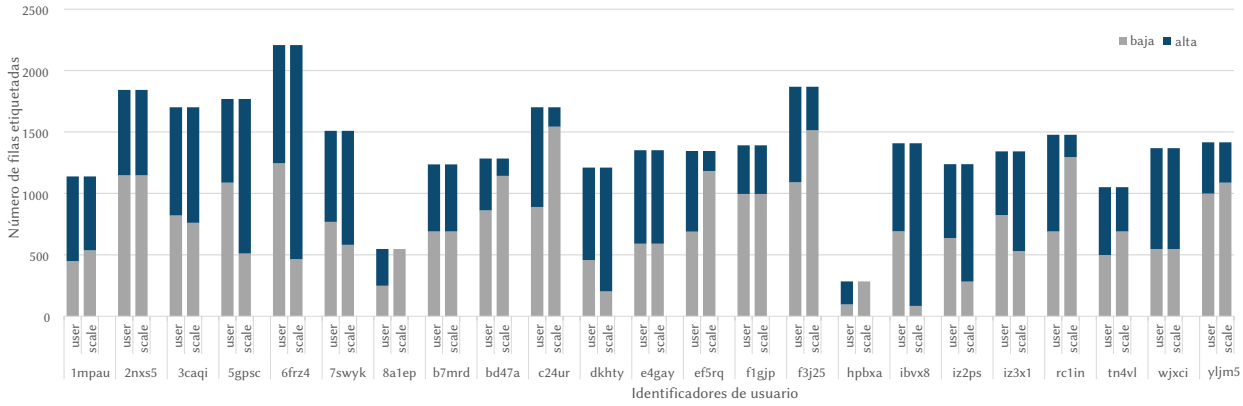


Figura 3.3: Datos etiquetados por user_id usando los dos enfoques propuestos, por la media del usuario en TLX_mean (user) y por el valor central de la escala (scale), indicando las clases baja y alta

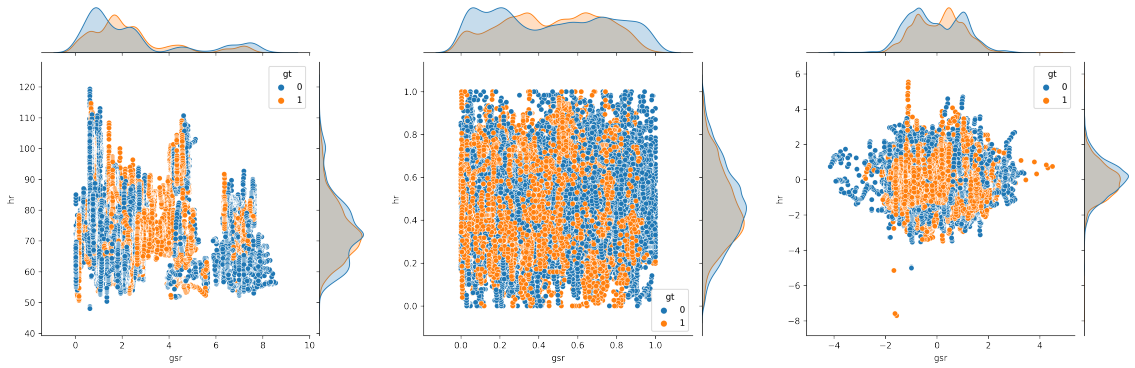
Se utilizaron las expresiones 3.1 y 3.2 [121] para normalizar y estandarizar los datos por participante.

$$xscaled = \frac{x - min(x)}{max(x) - min(x)} \tag{3.1}$$

$$z = \frac{x - mean(x)}{std(x)} \tag{3.2}$$

Con la primera expresión cada característica (x) fue escalada a valores entre 0 y 1 sustrayéndole el valor mínimo (min) y dividiéndola entre la diferencia del valor máximo (max) y mínimo; con la segunda, a cada característica se le restó la media (mean) y se dividió entre la desviación estándar (std).

La Figura 3.4 muestra gráficos de dispersión entre las variables fisiológicas GSR y HR considerando los esquemas de escalado de características utilizando el conjunto de datos filtrado y con etiquetado por el valor central de la escala.



(a) Base (b) Normalizado (c) Estandarizado

Figura 3.4: Dispersión de datos entre las variables GSR y HR considerando los enfoques de escalado de características, los valores 0 y 1 en gt representan carga baja y alta, respectivamente.

La Tabla 3.1 muestra las combinaciones de modelo y condiciones de experimentación con las que se obtuvieron los mejores resultados.

La mejor exactitud promedio del 73 % se alcanzó con un modelo SVM[1], utilizando el conjunto etiquetado acorde a la media de TLX_mean por usuario, considerando solo los participantes con etiquetado esperado –descartando a dos participantes–, omitiendo los segmentos de datos de dificultad media, con estandarización por participante y extracción de características con ventanas de tiempo variable acorde a la duración de cada segmento pero omitiendo los primeros 20s considerando una longitud mínima de 30s.

[1]Principales parámetros: C=1, gamma=0.001 y kernel=rbf

Tabla 3.1: Modelo y condiciones de experimentación con los mejores rendimientos

	Modelo	Etiquetado	Participantes	Con dificultad media	Exactitud promedio (acc)
1	SVM	Media por usuario	Solo con etiquetado esperado (21)	No	73.0 %
2	SVM	Centro de la escala	Solo con etiquetado esperado (11)	No	72.6 %
3	SVM	Media por usuario	Solo con etiquetado esperado (21)	No	71.7 %
4	SVM	Media por usuario	Todos los participantes (23)	No	71.3 %
5	SVM	Centro de la escala	Solo con etiquetado esperado (11)	No	71.2 %
6	SVM	Centro de la escala	Solo con etiquetado esperado (11)	No	70.8 %
7	SVM	Media por usuario	Solo con etiquetado esperado (21)	No	70.4 %
8	SVM	Media por usuario	Todos los participantes (23)	Si	68.6 %
9	XGB[1]	Media por usuario	Todos los participantes (23)	Si	68.1 %
10	XGB[2]	Centro de la escala	Todos los participantes (23)	Si	68.1 %

[1] Extracción de características con ventanas variables de tiempo acorde a la longitud completa del segmento; todos los demás con ventanas variables pero eliminando los primeros 20s considerando un mínimo de 30s
[2] Sin escalado de características, todos los demás con estandarización

La combinación con rendimiento más alto se repite en las posiciones 1, 3 y 7; con diferencias solamente en los parámetros del modelo. De los diez mejores resultados el 80 % corresponde a modelos SVM y el resto a modelos XGB, el mejor modelo RF se localizó en la posición número #22 con una exactitud del 66.4 %. De igual manera, no se obtuvieron resultados satisfactorios, todos menores al 58 % acc, utilizando ventanas de tiempo de 10s, 18s y 25s para la extracción de características.

En la Tabla 3.1 se resalta también que la mayoría de las mejores combinaciones consideran solo a los participantes con etiquetado esperado, observando que al etiquetar bajo el enfoque de media por usuario se descarta al 9 % de los participantes (2 personas) y con etiquetado por el valor central de la escala se descarta al 52 % de los participantes (12 personas).

La experimentación de clasificación de tres niveles de trabajo se llevo a cabo con *ground-truth* con base en los valores de TLX_mean, estableciendo un valor 0 (carga de trabajo baja) cuando fuera menor a 13, de 1 (carga media) cuando estuviera entre 12 y 18 y de 2 (alta) cuando fuera mayor a 17; sin embargo el mejor rendimiento alcanzado fue de 55.34 % acc con un modelo SVM.

En cuanto a pruebas de clasificación considerando la reducción previa de dimensionalidad o número de características, se tomó la implementación en Python de un autoencoder convolucional descrito en [122], el cual forma parte del *toolkit* pyEDA. Un aspecto relevante del modelo es que la primera capa lineal reduce el tamaño de la entrada a la potencia de dos más cercana, haciéndolo más flexible y escalable.

Aunque la dimensionalidad no es alta, 36 características, se configuró el modelo para reducirlas

a ocho, utilizando el resto de los parámetros de forma predeterminada (100 épocas, tamaño 10 del batch). La tasa de aprendizaje predeterminada fue 0.001, con función MSE de pérdida y optimizador Adam. Una vez entrenado el modelo de codificación, se probaron los mejores modelos de clasificación encontrados previamente, sin embargo el mejor resultado fue del 53.37 % acc.

3.2. Estudio piloto

El objetivo principal del estudio piloto fue realizar un experimento para construir un conjunto de datos propio y realizar pruebas iniciales para reconocer el estado cognitivo de carga de trabajo en un contexto de evaluación de experiencia de usuario.

Lo anterior implicó definir y evaluar varios aspectos, como por ejemplo: el equipo disponible para la captura de datos fisiológicos así como el protocolo para su instalación, calibración y mantenimiento; el software para captura y sincronización de datos así como sus requerimientos de instalación y funcionamiento; el protocolo de experimentación, incluyendo el tipo y orden de las tareas, duración aproximada de las pruebas, condiciones de las instalaciones físicas y de seguridad e higiene; entre otros.

Las siguientes secciones describen las características del experimento y los resultados obtenidos.

3.2.1. Participantes

Se capturaron datos de 22 participantes, 9 mujeres y 13 hombres, con una media de edad de 23.7 años. El 55 % fueron reclutados de la carrera de Ingeniería en Sistemas Computacionales del TecNM Parral y el resto de círculos sociales y familiares. Del total de los participantes, 15 eran estudiantes y el resto egresados de diversas carreras.

Todos los participantes llenaron previamente un formulario en-línea con sus datos personales y escolares, condiciones de salud preexistentes, entre otros, que permitieron determinar su idoneidad para el experimento y definir la agenda; una vez en el lugar de la prueba se les solicitó firmar un consentimiento escrito.

La mayor parte del experimento fue realizado en la sala de juntas del Depto. de Sistemas y Computación, ubicada dentro del edificio del Laboratorio de Cómputo del TecNM Parral, con

previa autorización de la Dirección y atendiendo las recomendaciones del Comité de Higiene y Seguridad de la institución. La Figura 3.5 muestra a dos de los participantes realizando una de las tareas.

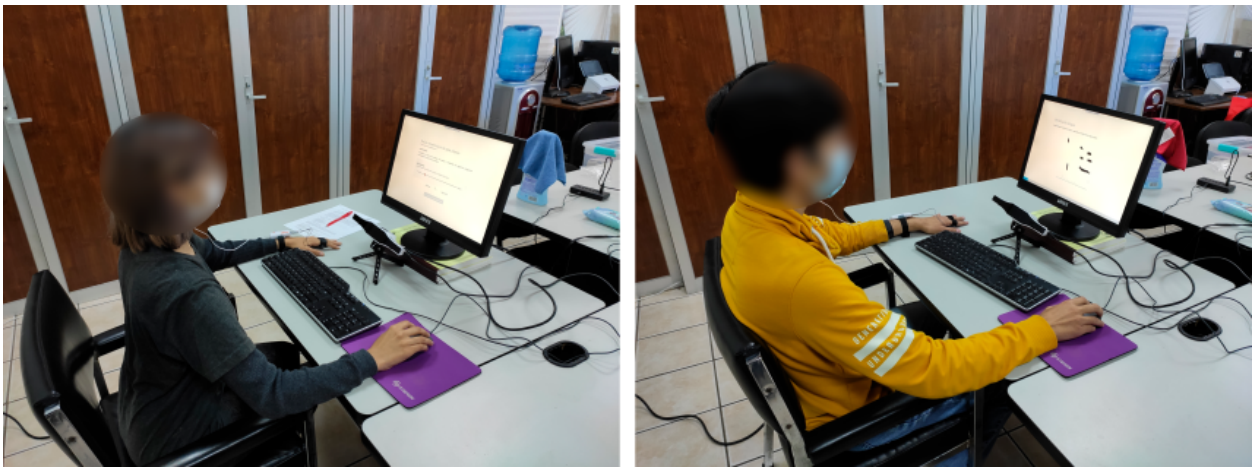


Figura 3.5: Participantes realizando una tarea en el experimento del estudio piloto.

3.2.2. Equipo y datos recolectados

Todos los datos fisiológicos fueron recolectados utilizando un mismo equipo de cómputo. Para la captura se utilizaron tres diferentes dispositivos (ver Tabla 3.2), aunque finalmente EPOC Flex fue descartado por problemas en la instalación del gorro y calibración con los primeros participantes.

Tabla 3.2: Dispositivos para captura de datos fisiológicos

	Shimmer3 GSR+	Gazepoint GP3	Emotiv EPOC Flex
Datos	GSR y tasa de latidos a través de PPG	Seguimiento ocular con pupilometría	EEG
Tasa de muestreo	102 Hz	60 Hz	128 Hz
Ubicación de sensores	Dos dedos de una mano y clip en oreja	Frente al usuario, entre el teclado y el monitor	Gorro de 32 canales colocado en la cabeza
Conexión	Bluetooth	USB	Bluetooth
Software que lo utiliza	Consensys (configuración) y ShimmerCapture (captura y transmisión LSL)	Gazepoint Control (calibración, captura y transmisión TCP/IP) y Gazepoint GP3-LSL (transmisión LSL)	EmotivPro (calibración, captura y transmisión LSL)

3.2.3. Software

El proceso de captura de datos fisiológicos se fundamentó en el sistema LSL, para lo que se utiliza software intermediario o *middleware* que interactúa con el equipo de sensado de señales fisiológicas y emplea librerías LSL para propagar los flujos de datos por la red a través del

protocolo TCP/IP. Una vez disponibles, el sistema administra la recolección, sincronización y almacenamiento de los flujos de datos locales o remotos, lo cual puede realizarse con el programa de grabación incluido, generando archivos de datos con un formato XDF.

A continuación se describen las herramientas de software y sus funciones principales.

PruebasCognitivas fue desarrollada para los fines de esta investigación; considera las tareas cognitivas utilizadas en CogLoad para inducir carga de trabajo, añadiendo una barra de tiempo límite como factor de estrés, como sugieren otros autores [23]. Adicionalmente, permite definir tareas genéricas para interactuar con un sitio web o aplicación de escritorio, ya que despliega las instrucciones en una ventana sobrepuesta para que el usuario pueda leer y dar seguimiento a cada uno de los pasos.

Las funciones de PruebasCognitivas son: registrar los datos de los participantes y de las plantillas de pruebas; aplicar las pruebas y almacenar resultados, marcadores de segmentos y respuestas de autoreportes; enviar flujo de datos de marcadores hacia la herramienta de grabación LSL; calcular valores de carga de trabajo acorde a NASA-TLX y exportar archivos de resultados de las pruebas. La Figura 3.6 muestra algunas capturas de pantalla del software, los datos de los participantes se desenfocaron para mantener su privacidad.

PruebasCognitivas permite calcular y exportar los resultados de carga de trabajo acorde a la metodología de NASA-TLX, sin embargo, su configuración se basa en plantillas, por lo que permite adicionalmente aplicar cualquier cuestionario estandarizado cuyas respuestas tengan un formato tipo Likert[1].

De igual forma, se desarrolló GazepointGP3-LSL como intermediario con Gazepoint Control para recibir los datos del dispositivo de seguimiento ocular y enviarlos hacía la herramienta de grabación. El software es una versión con interfaz gráfica de usuario del código publicado por Moein Razavi[2] que consume la Open Gaze API. La Figura 3.7 muestra una captura de pantalla del software.

Para la codificación de PruebasCognitivas y GazepointGP3-LSL se uso la plataforma .NET y el lenguaje de programación C#, con tecnologías UWP (del inglés *Universal Windows Platform*) y

[1]Con escala que contengan un valor menor, un valor mayor y valores con distribución equitativa entre ellos.
[2]Disponible en `https://github.com/moeinrazavi/Gazepoint-Eyetracking-GSR-HeartRate--LSL`

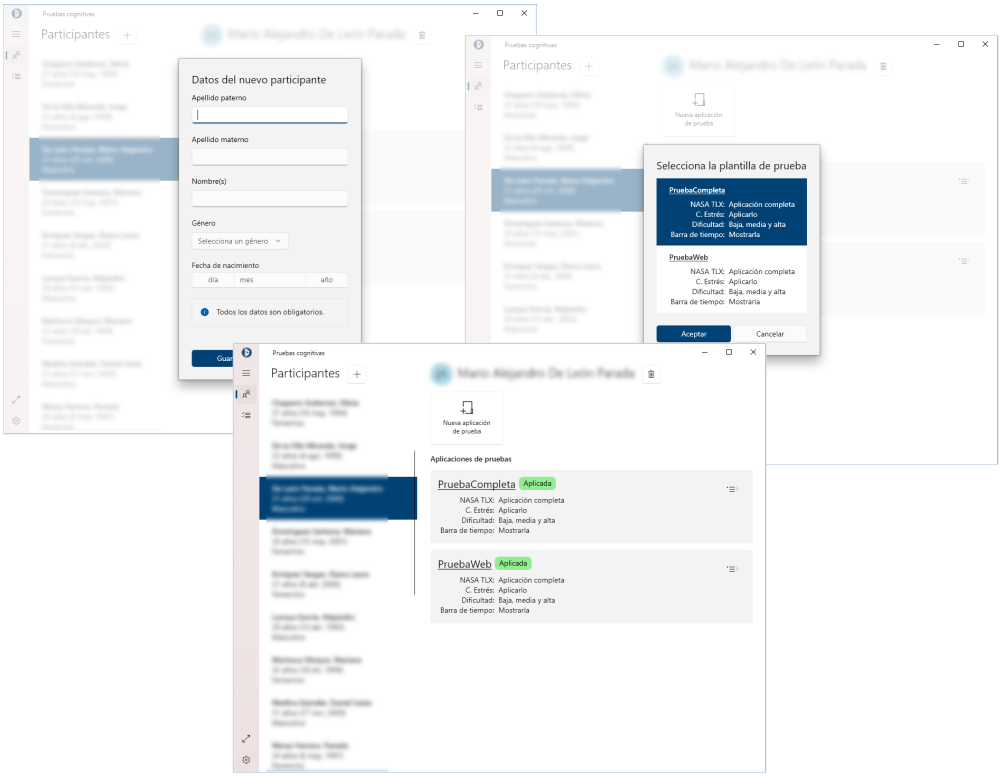


Figura 3.6: Capturas de pantalla de PruebasCognitivas

Windows Forms respectivamente.

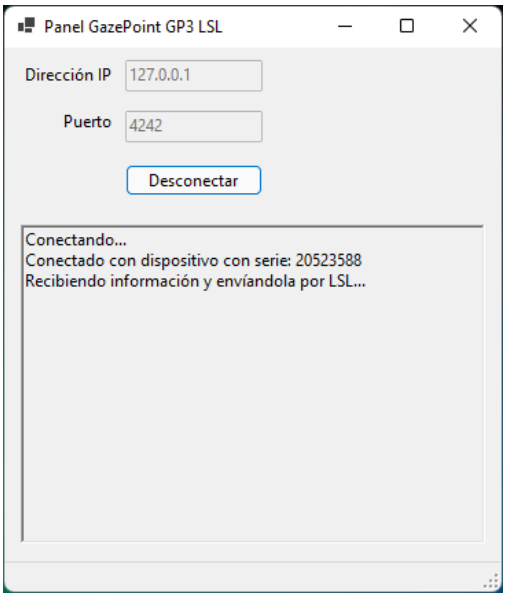


Figura 3.7: Captura de pantalla de GazepointGP3-LSL

Gazepoint Control es la herramienta base para calibrar el dispositivo Gazepoint GP3, capturar datos y publicarlos mediante conexiones de red bajo un formato propietario. La Figura 3.8 muestra

el software en funcionamiento.

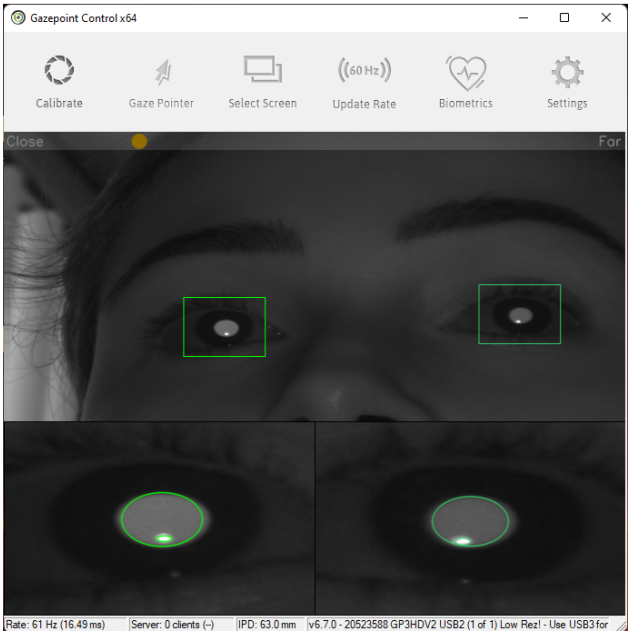


Figura 3.8: Captura de pantalla de Gazepoint Control

ShimmerCapture es una herramienta publicada por el fabricante Shimmer[1] con permiso para redistribución y uso, por lo que se modificó para agregar la funcionalidad de retransmitir los datos capturados utilizando el sistema LSL. Sus funciones son calibrar dispositivo Shimmer GSR+, capturar datos y enviarlos hacia la herramienta de grabación (ver Figura 3.9). También de Shimmer, el software ConsensysBASIC se utiliza para la configuración inicial del dispositivo, acción que se realiza una sola vez previo a llevar a cabo el experimento.

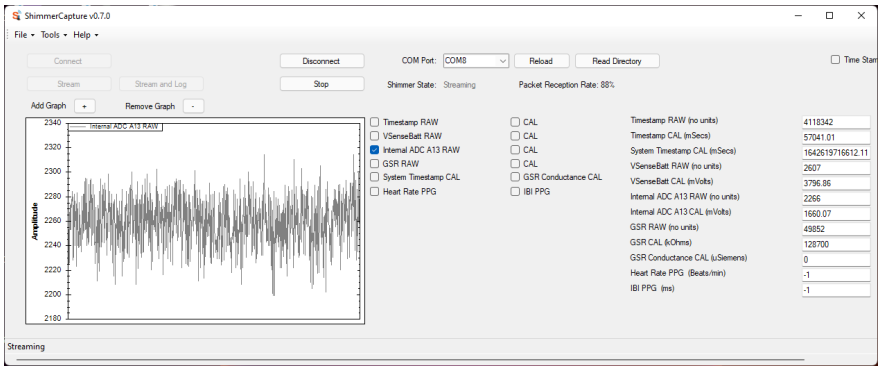


Figura 3.9: Captura de pantalla de ShimmerCapture

El software EmotivPRO permite la calibración y captura de datos de electroencefalograma. Además, incluye la opción para retransmitir los datos bajo el sistema LSL (ver Figura 3.10).

[1]Disponible en `https://github.com/ShimmerEngineering/Shimmer-C-API`

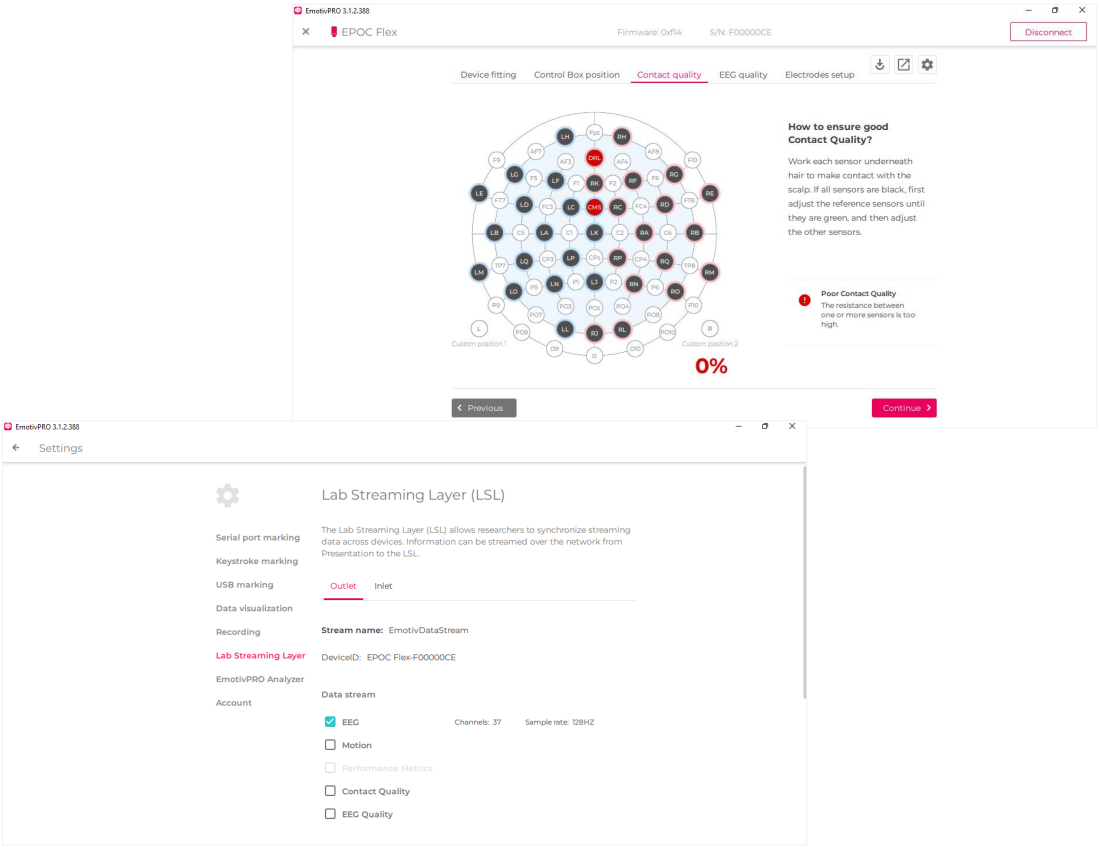


Figura 3.10: Capturas de pantalla de EmotivPRO

Finalmente, LabRecorder es la herramienta de grabación de flujos de datos mediante el sistema LSL, siendo la parte medular dentro del proceso de captura. La Figura 3.11 muestra un proceso de grabación activo, con los flujos de datos del dispositivo Shimmer, Gazepoint y de PruebasCognitivas –aunque no representa un dispositivo físico, PruebasCognitivas transmite los datos de inicio y fin de cada segmento con una tasa de muestreo irregular–.

ConsensysBASIC, Gazepoint Control, EmotivPRO y LabRecorder se utilizaron “tal cual” son proporcionadas por sus desarrolladores.

3.2.4. Procedimiento y estímulos

El experimento consistió de dos sesiones de pruebas con tareas presentadas al usuario a través de la herramienta de software PruebasCognitivas. El participante utilizaba un monitor externo al cual se le compartía el escritorio en modo extendido y el experimentador supervisaba algunos parámetros a través del monitor de la computadora portátil.

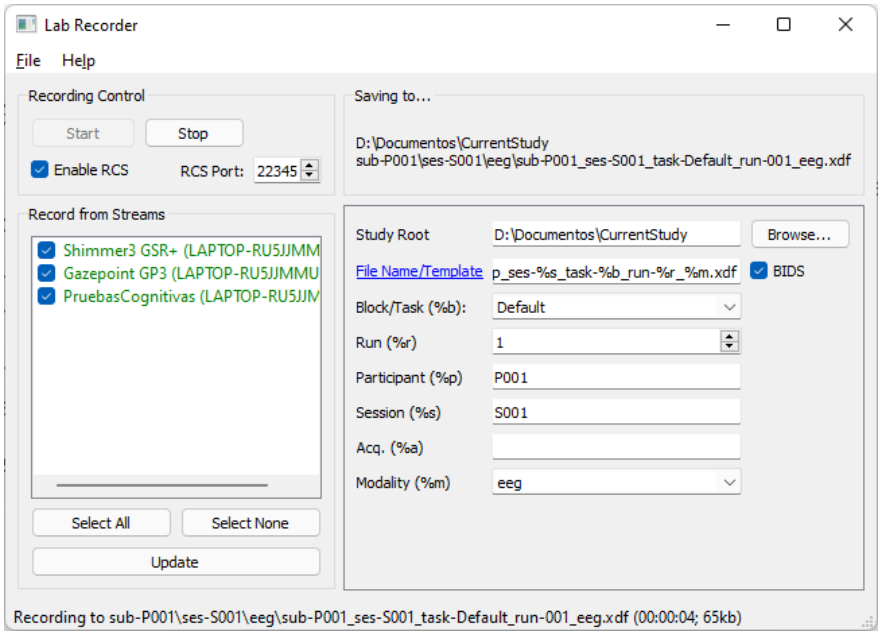


Figura 3.11: Captura de pantalla de LabRecorder

En la primera sesión se despliegan a pantalla completa tres bloques con seis tareas cognitivas cada uno y con una tarea de relajación entre cada bloque (ejercicio de respiración guiada, ver Figura 3.12). Las actividades de cada bloque tienen una misma dificultad, intercalando los bloques en dificultad baja, alta y media. Antes de cada tarea se muestran las instrucciones para llevarla a cabo y después de cada tarea se solicita contestar el cuestionario de autoreporte NASA-TLX/Estrés (ver Figura 3.13a). Previo a finalizar la prueba el participante responde el cuestionario de comparaciones entre subescalas NASA-TLX para identificar las fuentes de carga de trabajo (ver Figura 3.13b).

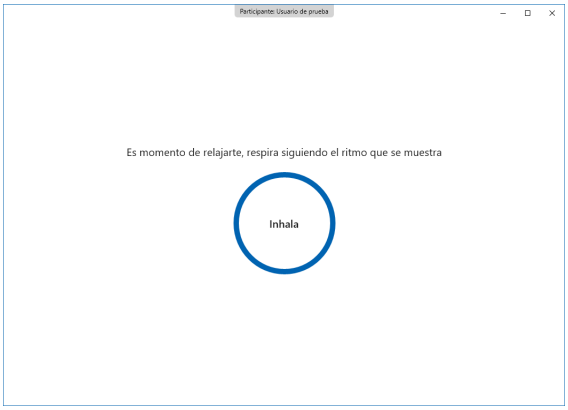
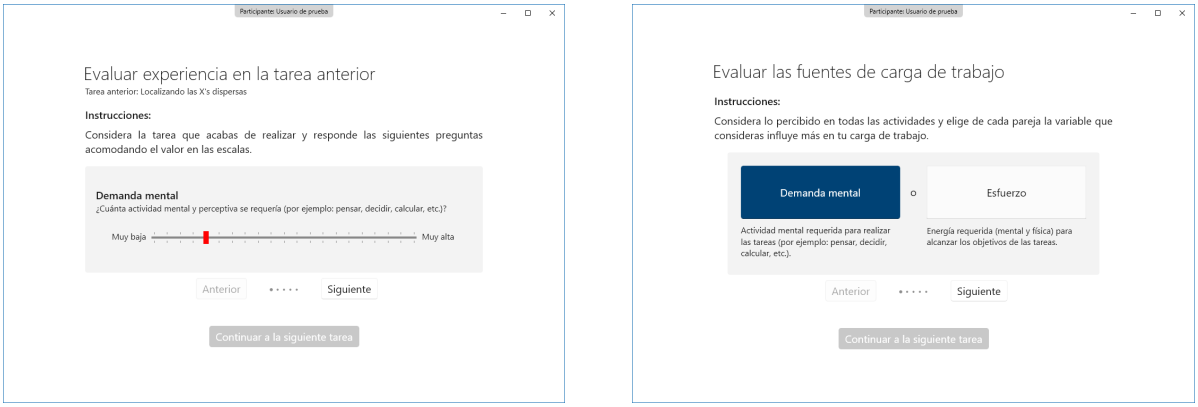


Figura 3.12: Tarea de relajación

Las tareas cognitivas se adaptaron de Haapalainen, Kim, Forlizzi *et al.* [118] (descritas inicial-



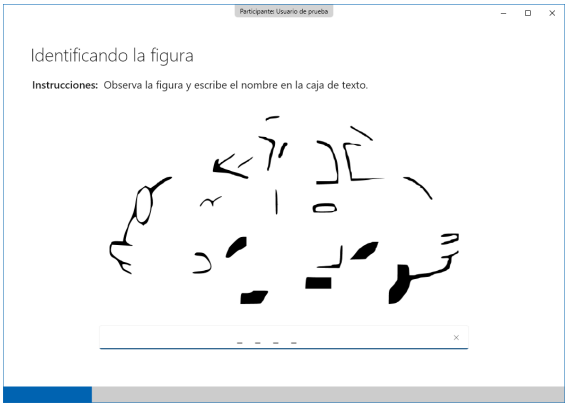
(a) Autoreporte NASA-TLX/Estrés (b) Comparaciones NASA-TLX

Figura 3.13: Cuestionarios de autoreporte

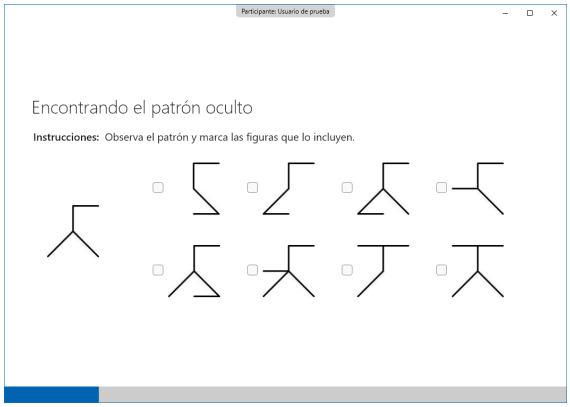
mente en la sección 3.1) agregando una barra de tiempo límite; a continuación se describen las tareas con mayor detalle:

1. Identificando la figura (ver Figura 3.14a): observar trazos incompletos de un dibujo y tratar de identificarlo, niveles de dificultad variando la complejidad de las imágenes.
2. Encontrando el patrón oculto (ver Figura 3.14b): identificar los trazos de una imagen dentro de otra, se utilizan ocho imágenes para comparación aumentando la complejidad en cada nivel de dificultad.
3. Encontrando las A's (ver Figura 3.14c): encontrar las palabras que incluyan la letra "a", variando la dificultad aumentando la longitud de las palabras.
4. Comparando números (ver Figura 3.14d): comparar dos números y decidir si son o no el mismo, la dificultad es manipulada incrementando el número de dígitos y el número de dígitos que tienen que comparar para identificar la primer diferencia.
5. Buscando los caminos (ver Figura 3.14e): seguir visualmente líneas e identificar su destino en el lado opuesto –se oculta el cursor sobre la imagen–, dificultad añadida al agregar puntos de cruce y longitud a las lineas.
6. Localizando las X's dispersas (ver Figura 3.14f): encontrar "x" distribuidas junto con otras letras en la pantalla, el criterio de dificultad considera el número de letras, la proximidad o distribución de las mismas y la existencia de letras de forma similar o rotadas.

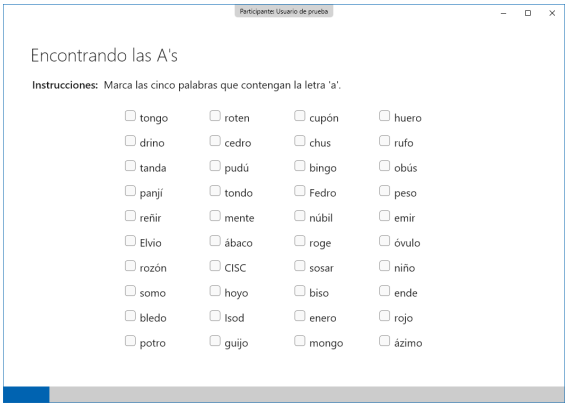
La segunda sesión de pruebas consiste de tres tareas guiadas dentro de sitios web. De manera similar se muestran las instrucciones y el cuestionario de autoevaluación NASA-TLX/Estrés antes



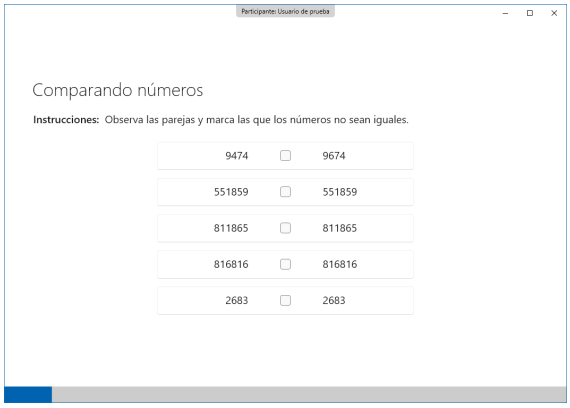
(a) Identificando la figura



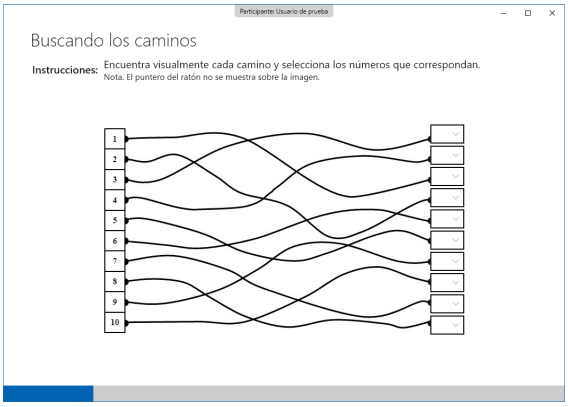
(b) Encontrando el patrón oculto



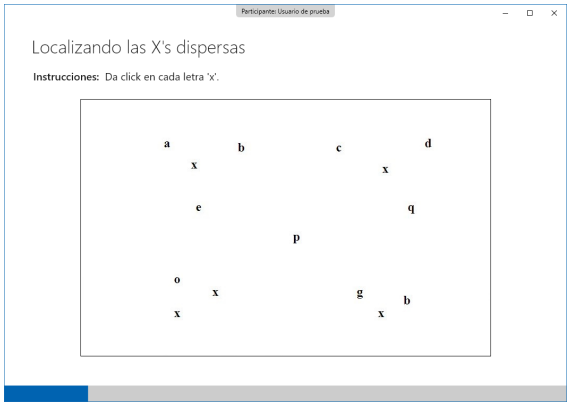
(c) Encontrando las A’s



(d) Comparando números



(e) Buscando los caminos



(f) Localizando las X’s dispersas

Figura 3.14: Tareas cognitivas

y después de cada tarea, con el ejercicio de relajación entre ellas. Las tareas se realizan en un navegador web y la descripción de cada paso se muestra en una ventana –con configuración “siempre al frente”– en la esquina superior derecha. En este caso no se aplica el cuestionario de comparaciones NASA-TLX ya que los resultados se toman de las respuestas de la primera sesión.

El objetivo de cada una de las tres tareas "tipo web" fue el siguiente:

1. Conociendo el sitio (ver Figura 3.15a): identificar la estructura de la página principal del TecNM Cenidet y algunos otros elementos del sitio.
2. Conociendo el Chat bot (ver Figura 3.15b): interactuar con el asistente virtual *chat bot* del sitio web del Sistema de Registro de Aspirantes del TecNM Cenidet y conocer algunos detalles de los programas de estudio y requisitos.
3. Conociendo el Sistema de Registro (ver Figura 3.15c): recorrer las secciones principales del sitio web del Sistema de Registro de Aspirantes del TecNM Cenidet.

Se contemplaba una evaluación UX más completa al Sistema de Registro, interactuando con algunas otras funciones, sin embargo, el experimento tuvo que redefinirse porque coincidió con el periodo oficial de admisión y no fue posible trabajar sobre una versión de prueba del sitio.

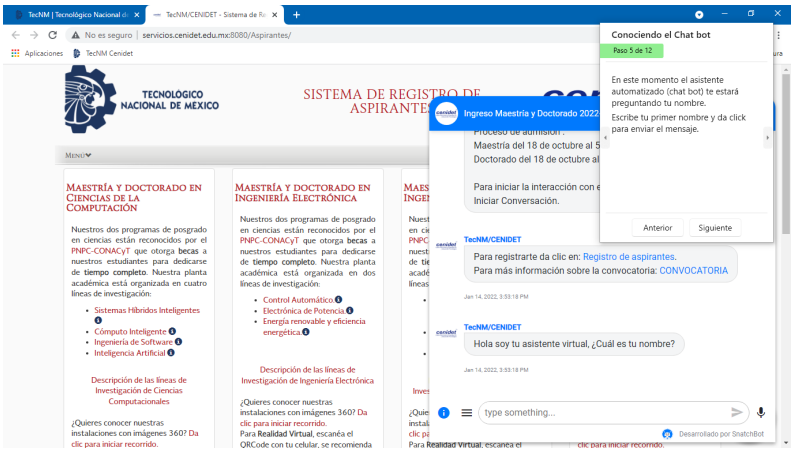
3.2.5. Preprocesamiento de datos y extracción de características

El preprocesamiento se llevó a cabo con un flujo de trabajo codificado en Jupyter *notebooks* y un kernel Python, además de librerías auxiliares como: pandas y numpy para manejo de datos en general; pyxdf para lectura y procesado de archivos XDF y scikit-learn para entrenamiento y evaluación de modelos; entre otras.

Las tareas de preprocesamiento se dividen en dos partes. En la primera parte se agrupan los datos de todos los participantes de las dos sesiones de prueba, tomando como entrada los archivos de datos, generados por la herramienta de grabación, y de autoreporte, generados desde PruebasCognitivas; incluyendo tareas para: leer los flujos de datos y agruparlos, sobremuestrando los datos de ET; añadir marcadores de segmento e identificadores del participante y la sesión; y etiquetar cada segmento acorde a los valores autoreportados y/o calculados de carga de trabajo, estrés, rendimiento y duración.

Aunque cada uno de los flujos de datos contiene estampas de tiempo acorde a la configuración de los dispositivos de captura, la tarea de sincronización se delega a pyxdf porque utiliza la estampa de tiempo de alta resolución generada por el sistema LSL y que tiene como base el reloj local de la computadora, en este caso con mayor precisión debido a que provienen del mismo

(a) Conociendo el sitio



(b) Conociendo el Chat bot

(c) Conociendo el Sistema de Registro

Figura 3.15: Capturas de pantalla de las tareas web

equipo[1].

[1]Mas información en `https://labstreaminglayer.readthedocs.io/info/time_synchronization.html`

En la segunda parte se toma como entrada el conjunto de datos generado previamente y se realizan algunos ajustes. Primero, se eliminan columnas no necesarias entregadas por Shimmer GSR+ y de seguimiento ocular no relacionados con pupilometría. Luego, se validan los datos de pupila para cada ojo considerando la métrica reportada por el dispositivo y se eliminan registros sin autoreporte. El objetivo de los experimentos es clasificar niveles de carga de trabajo, por lo que se descartaron datos de segmentos de relajación o de llenado de cuestionarios que no eran útiles.

Finalmente, se ajustó el valor de verdad a dos niveles de carga de trabajo considerando el enfoque de etiquetado por valor central de la escala descrito en la sección 3.1.1, pero en este caso con un rango de 0 a 100, por lo que los valores entre 0 y 50 fueron remplazados por 0 (carga baja) y los valores mayores de 50 por 1 (carga alta).

Se agruparon los registros que determinan cada segmento –por usuario y tarea– y se extrajeron características estadísticas de los datos fisiológicos de respuesta galvánica de la piel, tasa de latidos y diámetros de pupila izquierda y derecha –GSR, HR, PUPILL y PUPILR, respectivamente–, estas fueron: media, desviación estándar, asimetría, curtosis, primer cuartil, tercer cuartil, desviación de cuartiles, diferencia entre valores máximos y mínimos, coeficiente de variación, valor medio de la primera derivada y valor medio de la segunda derivada.

La extracción de características se llevó a cabo considerando ventanas variables acorde a la duración de cada segmento de datos, descartando previamente los datos no válidos de las columnas de tasa de latidos y de dilatación de pupilas. Posteriormente se eliminaron las columnas relacionadas con el nivel, la tarea y las señales con valores no válidos; quedando 48 características totales.

3.2.6. Modelos y experimentación

El estudio piloto tenía la principal finalidad de desarrollar y probar las herramientas de software, verificar el equipamiento y los protocolos de instalación y calibración con los participantes, así como implementar las tareas relacionadas con el preprocesamiento de los datos fisiológicos.

Por lo anterior, las pruebas de clasificación con modelos de aprendizaje se limitaron a obtener y comparar los resultados entre el enfoque de validación con separación aleatoria 70/30 y el enfoque LOSO, ambos con el conjunto sin escalado de características y un modelo RF con 100

árboles y parámetros predeterminados usando la implementación de la librería scikit-learn.

3.2.7. Resultados

Posterior al preprocesamiento y previo a la extracción de características, se obtuvo un conjunto de datos con un total de 5,384,224 registros y nueve columnas (ver Figura 3.16); las primeras columnas corresponden a los valores fisiológicos, continuando con el marcador de segmento, el identificador del participante y el rendimiento en la tarea, finalizando con el *ground-truth* para carga de trabajo y estrés.

```
Cantidad original: 8569549
Cantidad post eliminar sin autoreporte: 5384224

Descripción básica:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5384224 entries, 0 to 5384223
Data columns (total 9 columns):
 #   Column           Dtype
---  ------           -----
 0   gsr              float64
 1   hr               float64
 2   pupilL           float64
 3   pupilR           float64
 4   marker           object
 5   participante_id  int64
 6   rendimiento      float64
 7   gt_carga         int32
 8   gt_estres        int32
dtypes: float64(5), int32(2), int64(1), object(1)
memory usage: 328.6+ MB
```

Figura 3.16: Descripción del conjunto de datos posterior al preprocesamiento

Se encontró un marcado desequilibro entre clases con una proporción tres a uno –4,057,686 registros para carga baja (75 %) y 1,326,538 registros para clase alta (25 %) –, esto debido a que los segmentos de dificultad media tuvieron una tendencia de ser autoreportados como de carga de trabajo baja, probablemente como consecuencia del orden en que se presentaron los bloques de tareas: de dificultad baja, alta y luego media.

En la Figura 3.17 se presenta el análisis de correlaciones entre las señales fisiológicas, el rendimiento en la tarea y los niveles de carga y estrés; como se esperaba, se observan correlaciones altas entre los diámetros de pupila así como entre los niveles de carga de trabajo y estrés.

En cuanto a los resultados de la clasificación binaria de carga de trabajo. En el enfoque de validación con separación aleatoria 70/30, para los conjuntos de entrenamiento y prueba, se obtuvo una exactitud del 100 %, con características más importantes a la media y al coeficiente de variación de HR (ver Figura 3.18), observando un sobreaprendizaje en el modelo.

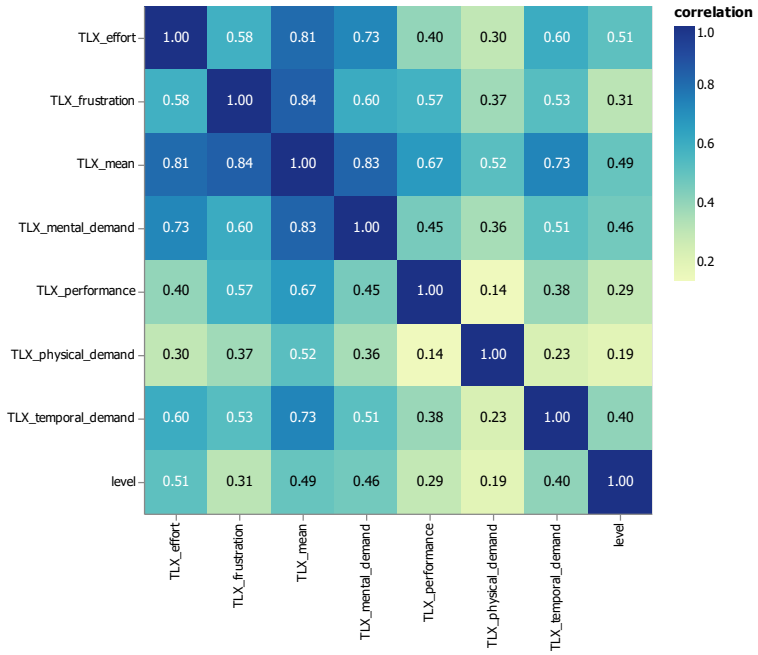


Figura 3.17: Correlaciones entre datos fisiológicos, rendimiento y niveles de carga de trabajo y estrés

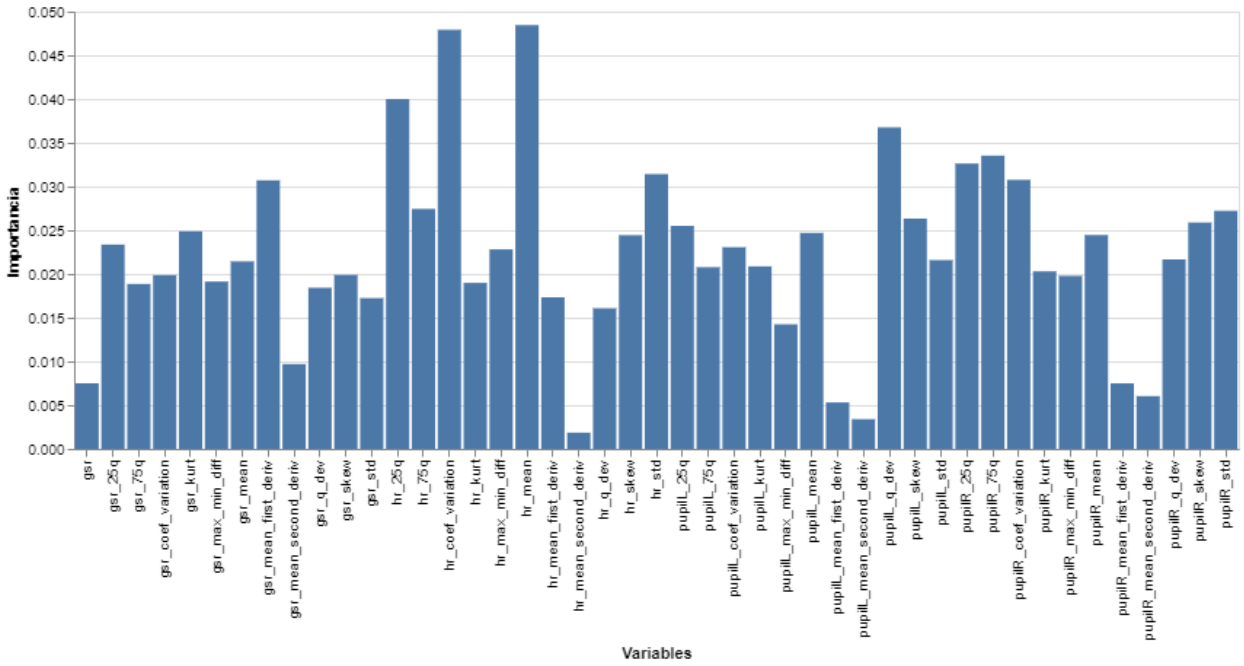


Figura 3.18: Importancias de características, separación 70/30

Bajo el enfoque de validación LOSO se obtuvo una exactitud promedio del 75.57 %. Sin embargo, el modelo es inservible, ya que al observar a detalle los resultados por participante, ver Tabla 3.3, se tiene que hay una nula recuperación para la clase alta de carga de trabajo –todas las métricas con valores cero– como consecuencia del desbalance del etiquetado de clases en el conjunto de

datos, donde incluso en algunos modelos solo se tuvieron datos de prueba etiquetados con clase baja (participantes 3, 7, 11, 15 y 21).

Tabla 3.3: Resultados de varias métricas por participante y por clase

Participante	Clase	Precisión	Recall	F1-score	Soporte	Exactitud
1	0 1	0.94 0	1 0	0.97 0	236440 15128	0.94
2	0 1	0.92 0	1 0	0.96 0	169013 14238	0.92
3	0	1	1	1	265083	1
4	0 1	0.32 0	1 0	0.49 0	110433 230890	0.32
5	0 1	0.62 0	1 0	0.77 0	162640 98474	0.62
6	0 1	0.86 0	1 0	0.92 0	210494 34806	0.86
7	0	1	1	1	242078	1
8	0 1	0.28 0	1 0	0.44 0	67606 175149	0.28
9	0 1	0.94 0	1 0	0.97 0	212598 12993	0.94
10	0 1	0.86 0	1 0	0.93 0	217201 34296	0.86
11	0	1	0.86	0.92	227596	0.85
12	0 1	0.72 0	1 0	0.83 0	193484 76774	0.72
13	0 1	0.88 0	1 0	0.93 0	187001 26328	0.88
14	0 1	0.66 0	1 0	0.8 0	148223 76366	0.66
15	0	1	0.98	0.99	215158	0.98
16	0 1	0.52 0	1 0	0.69 0	111669 102687	0.52
17	0 1	0.93 0	0.82 0	0.87 0	198156 11876	0.77
18	0 1	0.9 0	1 0	0.95 0	247219 26717	0.9
19	0 1	0.58 0	1 0	0.73 0	165530 122185	0.58
20	0 1	0.61 0	1 0	0.76 0	180471 114895	0.61
21	0	1	0.98	0.99	191629	0.98
22	0 1	0.43 0	1 0	0.6 0	97964 131487	0.43

3.3. Experimento de evaluación UX de sitio web académico

Producto de los hallazgos observados en el estudio piloto se realizaron ajustes en la estructura de las tareas cognitivas, se consideraron nuevas técnicas para el preprocesamiento de datos y se evaluaron nuevos modelos de clasificación, entre otros aspectos. Aspectos que fueron puestos a prueba en un escenario de evaluación UX de un sitio web académico con la finalidad de analizar su idoneidad para la definición formal de la metodología de evaluación de experiencia de usuario.

Las siguientes secciones describen las características del experimento y los resultados obtenidos. Debido a que se utilizaron los mismos dispositivos de captura y se recolectaron los mismos tipos de datos que en el estudio piloto (ver sección 3.2.2), y que se usaron las mismas herramientas de software que auxilian en el proceso de aplicación de las tareas y la captura de los datos fisiológicos (ver sección 3.2.3), no se describen a detalle esos elementos y se simplifican otras descripciones que tienen similitud con el estudio piloto.

3.3.1. Participantes

Se capturaron datos de 19 participantes, 7 mujeres y 12 hombres, con una media de edad de 28.4 años; 84 % estudiantes del TecNM Cenidet y el resto de otras instituciones. El experimento se realizó en un cubículo dentro del edificio de Computación atendiendo las recomendaciones de higiene y seguridad. La Figura 4.7 muestra una participante realizando una tarea.

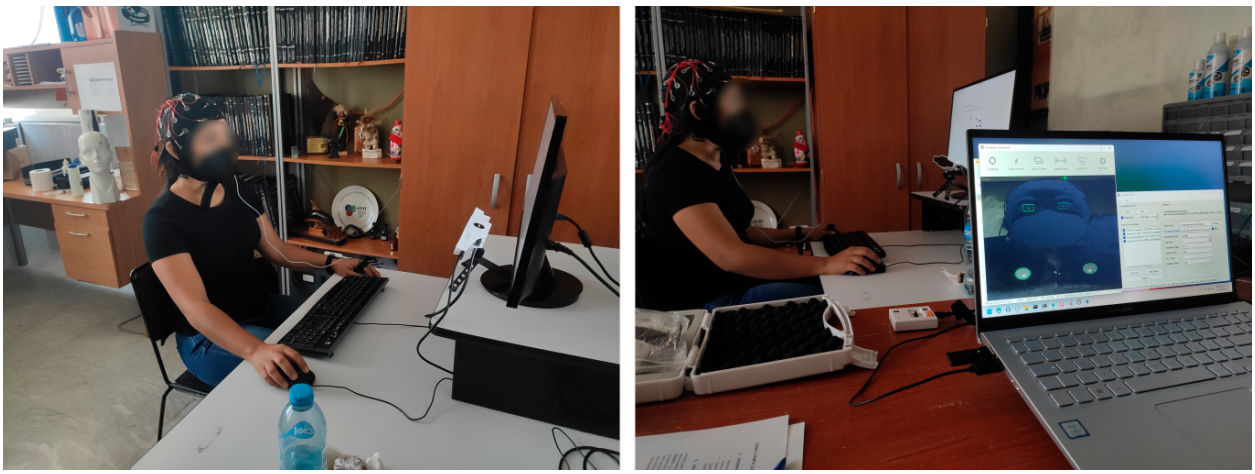


Figura 3.19: Participante realizando una tarea en el experimento de evaluación UX

3.3.2. Procedimiento y estímulos

El experimento consistió de dos sesiones de prueba: sesión de tareas cognitivas y sesión de evaluación UX.

En la primera sesión se aplicaron dos bloques con las seis tareas cognitivas cada uno y con una tarea de relajación previa a cada bloque (tareas descritas en sección 3.2.4). Las actividades de cada bloque con una misma dificultad, iniciando con baja y luego alta. Antes de cada tarea se muestran las instrucciones y después el cuestionario NASA-TLX. Al finalizar se aplica el cuestionario de comparaciones NASA-TLX para identificar las fuentes de carga de trabajo.

En el estudio piloto se esperaba una correlación entre la carga de trabajo y la dificultad de las tareas cognitivas, como se había observado en otros estudios (p. ej., [35], [78], [109]). Sin embargo, analizando la proporción de carga de trabajo autoreportada (ver Tabla 3.4), esto solo se cumplió para el conjunto de tareas de dificultad baja, por lo que se hipotetizó que las tareas de dificultad alta deberían ser más difíciles y que pasar de dificultad alta a media provocaba una percepción de mayor facilidad en estas últimas.

Tabla 3.4: Carga de trabajo autoreportada por dificultad del bloque en el estudio piloto

	Carga de trabajo	
Dificultad	% Baja	% Alta
Baja	82	18
Alta	55	45
Media	77	23

Adicionalmente, en el estudio piloto la duración total de ambas sesiones de prueba –fueron realizadas una tras otra solo con un breve descanso intermedio– fue en promedio de 65 minutos (duración mínima de 50 min. y máxima de 87 min.), sin contar la instalación y calibración inicial de sensores, observando cansancio en los participantes durante la última etapa.

Debido a lo anterior, se realizaron ajustes a las tareas cognitivas, aumentando la dificultad del bloque alto y descartando las tareas de dificultad media.

En la segunda sesión, sesión de evaluación UX, se definieron cinco tareas para realizar en la aplicación web del Departamento de Ciencias Computacionales del TecNM Cenidet (ver Figura 3.20): 1) modificando datos personales, 2) agregando teléfonos, 3) agregando otros datos, 4) agregando una publicación y 5) eliminando datos. Se muestran las instrucciones y el cuestionario

NASA-TLX antes y después de cada tarea, respectivamente, con un segmento de relajación al inicio de la prueba.

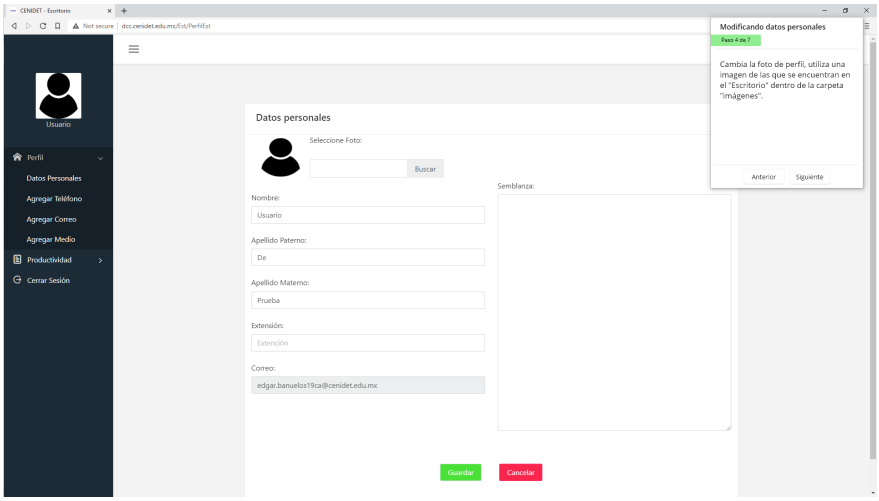


Figura 3.20: Captura de pantalla de tarea en sesión de evaluación UX

3.3.3. Preprocesamiento de datos y extracción de características

En este experimento se incluyeron datos de electroencefalografía, por lo que de entre varios métodos del estado del arte para realizar preprocesamiento EEG (Autoreject [123], ADJUST [124], DETECT [125], entre otros), se eligió realizarlo con PREP [126] por las siguientes razones:

- Parte de la premisa de realizar solo el "suficiente" preprocesamiento para uniformar el comportamiento estadístico de los datos entre diferentes dispositivos y paradigmas experimentales. Esto permite incluso aplicar alguno de los otros métodos posterior a PREP. Además, en esta investigación se utilizó el dispositivo Emotiv Epoc Flex, sin embargo era importante elegir un esquema que tuviera más flexibilidad y uniformidad ante el uso de un dispositivo diferente.
- Considera varios métodos para buscar canales malos: por valores nulos (NaN); por canales *flat*, que tienen valores constantes o muy pequeños por un periodo de tiempo significativo; por desviación estándar, para detectar amplitudes extremas; por ruido de alta frecuencia; por correlación entre canales; por la relación ruido-señal (SNR); y RANSAC, el cual toma una muestra aleatoria de datos de un canal bueno para evaluar el comportamiento de otros canales. Esta variedad de métodos permite identificar las razones por las que se etiqueta un canal como malo y analizar las problemáticas en la calidad de los datos.

- Incluye un mecanismo (heredado del framework EEGLAB para Matlab) para interpolar los canales malos dentro de un proceso iterativo, introduciendo un algoritmo para obtener y remover previamente una señal de referencia calculada entre los sensores.
- Tiene varias implementaciones, principalmente en Python, R y Matlab, que facilitan su utilización.

Además del uso del método PREP, se probó la remoción de artefactos oculares en EEG a través de análisis de componentes independientes (ICA) utilizando la librería mne[1].

Cuando se trabaja en clasificación con modelos de aprendizaje utilizando datos de electroencefalograma es común que las características de densidad espectral de poder relacionadas con los ritmos cerebrales se extraigan utilizando el método de Welch [127] a partir de la transformada rápida de Fourier (p. ej., [8], [46], [57], [71]), sin embargo, en esta investigación se decidió utilizar el método *multitaper* [128] ya que se ha encontrado que como estimador mejora los métodos estándar de análisis espectral y permite generar espectrogramas que incluso pueden ser mejor interpretados visualmente [129][130], esta última característica era relevante para la investigación debido a que originalmente se planteaba realizar experimentos con modelos de redes neuronales convolucionales y se anticipaba entrenarlos con una mejor representación.

Las tareas de procesamiento se implementaron en el método `preprocess` de una clase Python denominada `PprocFile`. Este método se ejecuta para cada uno de los datos de prueba por participante en ambas sesiones –el pseudocódigo 3.1 muestra las tareas que gestiona–. Recibe como parámetros un archivo de datos fisiológicos de un participante en una sesión, el tipo de prueba (cognitiva o web), el archivo de exportación de PruebasCognitivas con la información de etiquetado y el tamaño de la ventana de tiempo para la extracción de características.

El método `preprocess` genera como salida cinco archivos en formato de serialización de Python (tipo *pickle*) conteniendo: 1) datos preprocesados, 2) metadatos, 3) datos *raw* después de la limpieza de artefactos oculares, 4) datos preprocesados normalizados y 5) datos preprocesados estandarizados.

La descripción de cada tarea es la siguiente:

[1] Sitio web de MNE: `https://mne.tools/stable/index.html`

Algoritmo 3.1 Método preprocess

```
datos ← cargarDatos()
datosEtiquetados ← etiquetarDatos(datos)
datosPREP, canalesValidos ← realizarPREP(datosEtiquetados)
if canalesValidos >= 9 then
    if |canalesValidos ∩ [Fp1, Fp2, F7, F8])| > 0 then
        datosFiltrados ← filtrarDatos(datosPREP)
        datosProcesados ← removerArtefactosOculares(datosFiltrados)
        datosProcesados ← extraerCaracteristicasPSD(datosProcesados)
        datosProcesados ← validarPupilas(datosProcesados)
        datosProcesados ← eliminarColumnasInnecesarias(datosProcesados)
        datosProcesados ← eliminarCanalesEEGNoValidos(datosProcesados)
        datosProcesados ← extraerCaracteristicasEstadisticasNoEEG(datosProcesados)
        datosProcesados ← eliminarColumnasDatosNoValidos(datosProcesados)
        datosProcesados ← eliminarSegmentosDescartados(datosProcesados)
        datosProcesados ← eliminarDatosNulos(datosProcesados)
        datosProcesados ← extraerCaracteristicasEstadisticasEEG(datosProcesados)
        datosProcesados ← ajustarGroundTruth(datosProcesados)
        datosProcesados ← normalizar(datosProcesados)
        datosProcesados ← estandarizar(datosProcesados)
        generarSalida()
        return "Preprocesamiento y extracción de características terminados"
    else
        return "Insuficientes canales de referencia para realizar ICA"
    end if
else
    return "Insuficientes canales EEG válidos"
end if
```

1. *Cargar datos*: Obtener los flujos de datos separados y agruparlos sobremuestreando acorde a la tasa más alta), adicionar los marcadores de segmento y agregar los datos de participante y sesión.
2. *Etiquetar datos*: Obtener la información de etiquetado del participante y etiquetar cada segmento.
3. *Realizar análisis PREP*: Gestionar el análisis PREP de los datos EEG.
4. *Filtrar datos*: Aplicar un filtro paso banda de 0.1-40 Hz.
5. *Remover artefactos oculares*: Crear una copia de los datos filtrados y aplicar un filtro paso alto de 1 Hz [131]; realizar el análisis ICA tomando como referencia los canales Fp1, Fp2, F7, F8; remover los componentes oculares en los datos filtrados originales.
6. *Extraer características PSD*: Seccionar los datos acorde a la ventana de tiempo, calcular y agregar la densidad espectral de poder (PSD) absoluta y relativa de cada segmento, en las bandas de las ondas cerebrales.

7. *Validar pupilas*: Actualizar los datos de diámetro de pupila izquierda y derecha acorde a la validez reportada por el sensor.
8. *Eliminar columnas no necesarias*: Eliminar columnas que no se utilizan pero que son entregadas por los sensores.
9. *Eliminar canales EEG no válidos*: Eliminar las columnas de los canales EEG no válidos.
10. *Extraer características estadísticas*: Acorde a la ventana de tiempo, calcular y agregar características estadísticas (media, desviación estándar, valor máximo, valor mínimo, entre otras) de los datos GSR, HR, EEG (por canal válido) y pupilas.
11. *Eliminar columnas con datos no válidos*: Eliminar las columnas de datos fisiológicos que poseen fragmentos de datos no válidos (HR y pupilas).
12. *Eliminar segmentos descartados*: Eliminar segmentos de respiración y cuestionarios de autoevaluación.
13. *Eliminar datos nulos*: Eliminar registros con algún dato nulo (NaN).
14. *Extraer características estadísticas EEG*: De cada registro, agrupar los datos EEG de canales válidos y calcular y agregar características estadísticas; por ejemplo, la desviación estándar representa al valor GFP [132].
15. *Ajustar valores de verdad*: Ajustar el *ground-truth* de carga de trabajo a solo dos niveles (menor a 50: baja; 50 hacia arriba: alta).
16. *Normalizar*: Normalizar los datos con enfoque min-max por participante.
17. *Estandarizar*: Estandarizar los datos con enfoque por participante.
18. *Generar salida*: Generar los archivos *pickle*.

Se definieron otros métodos en la clase `PprocFile` para conjuntar los datos de cada sesión considerando solo los usuarios con suficientes canales EEG válidos, generando los conjuntos para cada enfoque de escalado de características que se utilizaron en la clasificación con modelos de aprendizaje.

Se generó también un conjunto de datos (con sus versiones normalizadas y estandarizadas) sin considerar el análisis de canales válidos EEG, el filtrado, la eliminación de artefactos y la extracción de otras características EEG que se describieron anteriormente. En resumen, este conjunto contiene la información de etiquetado, los datos de los canales EEG en bruto y las

características estadísticas de cada canal EEG, de GSR, de HR y de pupilas.

3.3.4. Modelos y experimentación

Fueron considerados dos enfoques para el entrenamiento y prueba con los modelos de aprendizaje automático.

En el primer enfoque, entrenamiento y prueba con sesión cognitiva, el objetivo fue encontrar modelos y ajustes con un buen desempeño en la clasificación considerando validación LOSO y solo los datos de la sesión cognitiva. Se utilizaron modelos RF y XGB, descartando SVM dado que inicialmente el tiempo de entrenamiento era considerablemente mayor.

Se realizaron pruebas con el conjunto de datos con procesamiento EEG completo y con los siguientes ajustes:

- *statEEG*: Utilizar solo las características estadísticas de los canales válidos EEG.
- *gtMedia*: Ajustar el *ground-truth* considerando la media por usuario en la sesión cognitiva, si está por debajo de la media se etiqueta como carga baja y por arriba como carga alta.
- Combinación de los dos ajustes previos.

Para fines comparativos, se realizaron también varias pruebas de clasificación con el conjunto de datos sin preprocesamiento EEG.

En el segundo enfoque, entrenamiento con sesión cognitiva y pruebas con sesión web o de evaluación UX, se utilizó el conjunto de datos con preprocesamiento EEG y los datos de la sesión cognitiva para el entrenamiento y los datos de la sesión web para la validación con un enfoque LOSO. Agregando los siguientes ajustes:

- *sin20s*: Descartar los primeros 20s de cada segmento considerando una duración mínima de 30s.
- *segP*: Considerar para el entrenamiento solo los tres segmentos altos con mayor carga autoreportada y los tres segmentos bajos con la menor carga autoreportada.
- *sinEEG*: Descartar todos los datos EEG.
- Combinación de ajustes previos.

Para completar el proceso de evaluación de experiencia de usuario, se definió un mecanismo

por votación, en donde se utiliza el modelo de aprendizaje con mejor rendimiento, obtenido en el segundo enfoque, para obtener la carga de trabajo de cada segmento de evaluación UX considerando las predicciones mayoritarias de cada participante; los segmentos con carga de trabajo alta se relacionan con un detrimento en la experiencia del usuario debido a la demanda de recursos requerida.

3.3.5. Resultados

Analizando la proporción de carga de trabajo autoreportada (ver Tabla 3.5) se observó una mejor distribución de clases, acorde a lo esperado por la dificultad de los segmentos, concluyendo que el ajuste en la presentación de los bloques de tareas cognitivas y el aumento de complejidad de las tareas de dificultad alta brindó resultados satisfactorios.

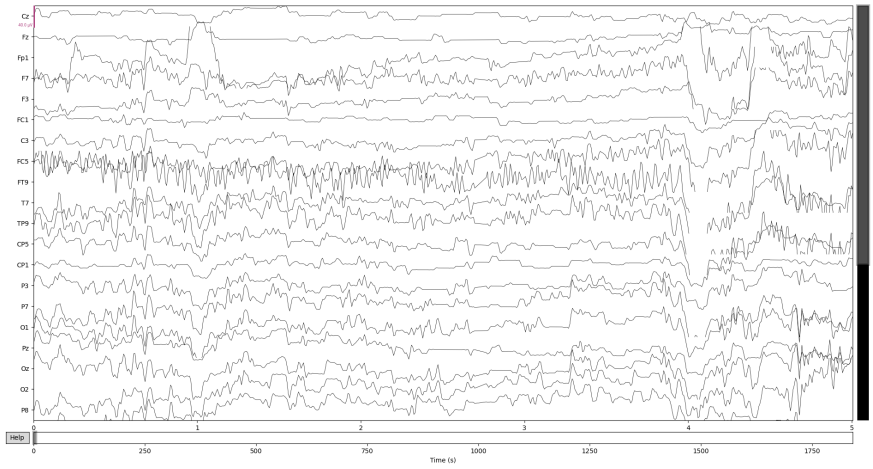
Tabla 3.5: Carga de trabajo autoreportada por dificultad del bloque en el experimento de evaluación UX

	Carga de trabajo	
Dificultad	% Baja	% Alta
Baja	83	17
Alta	23	77

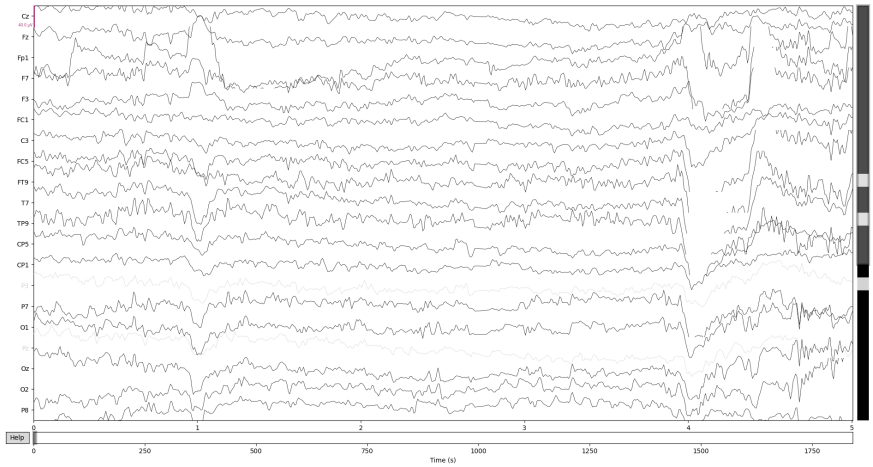
De igual forma, la duración total del experimento fue en promedio de 53 minutos (duración mínima de 40 min. y máxima de 67 min.), reduciendo en 19 %, 20 % y 23 %, el promedio, mínimo y máximo, respectivamente, en comparación con el estudio piloto. En particular, la sesión cognitiva tuvo una duración promedio de 29 min. (mínimo de 22 min. y máximo de 35 min.), variando debido al desempeño de cada participante, ya que la aplicación PruebasCognitivas avanza en automático una vez que se termina una tarea exitosamente.

El flujo de preprocesamiento PREP para datos EEG permitió limpiar las señales y descartar canales que fueron afectados por diversas condiciones y que fueron irrecuperables. La Figura 3.21a muestra los primeros 5s de los datos EEG del participante 01 de la sesión cognitiva. La Figura 3.21b muestra el mismo fragmento de datos después del flujo PREP, marcando en tono gris los canales detectados como malos (P3, CP2 y Pz).

La Tabla 3.6 muestra la validez de los canales EEG de ambas sesiones considerando solo los participantes con suficientes canales válidos. Se observa un gran número de datos EEG no útiles debido a la cantidad de ruido y artefactos en las señales. Durante la prueba se les pedía a los



(a) Datos EEG originales



(b) Datos EEG posteriores a PREP

Figura 3.21: Primeros 5s de datos EGG del participante 01, sesión cognitiva

participantes limitar sus movimientos corporales a lo necesario, sin embargo, se tuvieron algunas observaciones en cuanto a la calidad del gorro utilizado ya que presentaba un notable desgaste y en ocasiones la sujeción no era la ideal, marcando un buen contacto según los parámetros del software EmotivPro pero con intermitencias en la calidad reportada de la señal.

El análisis ICA permitió remover artefactos oculares considerando los canales de referencia. La Figura 3.22 muestra los componentes detectados en el participante 01 de la sesión cognitiva, denotados por el nombre con tonalidad gris. La Figura 3.23 muestra un fragmento de la señal marcando con rojo los componentes oculares y la Figura 3.24 muestra los primeros 5s de la señal tras su remoción.

Tabla 3.6: Validez de canales EEG por participante en sesiones cognitiva y de evaluación UX

	Participantes en sesión cognitiva								Participantes en sesión de evaluación UX								
Canales	p01	p02	p05	p07	p08	p14	p18	p19	p01	p02	p04	p05	p07	p08	p14	p19	Conteo
F3[1,2,3]	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
P7[1,2,3]	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
FC1[1,2,3]	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
Fz[1,2,3]	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
FC2[1,2,3]	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	16
Fp1[1]	X	X	X	X	X	X	X	X	X	X	-	X	X	X	X	X	15
Fp2[1]	X	X	X	X	X	X	X	X	X	X	-	X	X	X	X	X	15
F7[1]	X	X	X	X	X	X	X	X	X	X	X	X	X	-	X	X	15
Oz[2]	X	-	X	X	X	X	X	X	X	X	X	X	X	X	X	X	15
P4[2]	X	X	X	X	X	-	X	X	X	X	X	X	X	X	X	X	15
F8[1]	X	X	X	X	X	X	X	X	X	X	-	X	X	X	X	X	15
TP10[2]	X	X	X	-	X	-	X	X	X	X	X	X	X	X	X	X	14
T7	X	X	X	X	-	X	X	X	X	X	-	X	X	X	X	X	14
FT9[2]	X	-	X	X	-	X	X	X	X	X	X	X	X	X	X	X	14
P8[2]	X	-	X	X	X	-	X	X	X	X	X	X	X	X	X	X	14
T8	X	X	X	X	X	-	X	X	X	X	-	X	X	X	X	X	14
O2[2]	X	-	X	X	X	X	-	X	X	X	X	X	X	X	X	X	14
TP9	X	-	X	X	-	X	X	X	X	X	-	X	X	X	X	X	13
C4	X	X	X	X	-	-	X	X	X	X	X	X	X	-	X	X	13
F4	X	X	X	-	X	-	X	X	X	X	X	X	X	X	-	X	13
CP6	X	X	X	X	-	-	X	X	X	X	X	X	X	-	X	X	13
O1	X	X	-	X	X	X	X	X	X	-	X	X	X	-	X	X	13
Cz	X	X	X	X	-	-	-	X	X	X	X	X	X	-	X	X	12
FC5	X	-	X	-	-	X	-	X	X	X	X	X	X	-	X	X	11
FT10	X	-	X	-	-	X	X	X	X	X	-	X	X	-	X	X	11
FC6	X	X	X	-	-	-	X	X	X	X	-	X	X	-	X	X	11
CP5	X	X	X	-	X	-	-	X	X	-	X	X	X	X	-	-	10
C3	X	X	X	-	-	-	-	-	X	X	X	X	X	-	-	X	9
P3	-	X	X	-	X	-	-	X	X	-	X	X	X	-	-	X	9
CP2	-	-	X	-	-	-	X	X	X	X	-	X	X	-	X	X	9
CP1	X	-	X	-	-	-	-	X	X	-	X	X	X	-	-	X	8
Pz	-	-	X	-	-	-	X	-	-	X	X	X	-	-	X	X	7
Conteo	29	22	31	21	19	17	25	30	31	28	23	32	31	19	27	31	

Solo se muestran los participantes con suficientes canales EEG válidos por sesión; X canal válido, - canal no válido
[1] Canales válidos compartidos entre participantes de sesión cognitiva: Fz, Fp2, P7, Fp1, F7, F3, FC1, F8, FC2
[2] Canales válidos compartidos entre participantes de sesión evaluación: Fz, P4, P7, F3, FC1, FT9, O2, TP10, FC2, Oz, P8
[3] Canales válidos compartidos entre participantes en ambas sesiones: Fz, P7, F3, FC1, FC2

Como ejemplo de los métodos para obtener la densidad espectral, la Figura 3.25 muestra el resultado de la primera ventana de tiempo calculada utilizando los métodos de Welch y *multitaper* sobre el canal Fp1 del participante 01, con una ventana de 4s y considerando la banda del ritmo cerebral *delta*.

En cuanto a las pruebas de clasificación binaria de carga de trabajo. La Tabla 3.7 muestra un resumen de los resultados considerando el enfoque de entrenamiento y prueba con datos de la sesión cognitiva, obteniendo la mejor exactitud de 65.4 % con un modelo RF de 135 árboles, conjunto normalizado y ajuste gtMedia. Utilizando el conjunto de datos sin preprocesamiento EEG la mejor exactitud fue del 57.03 % con un modelo RF de 100 árboles y el conjunto estandarizado.

El resto de resultados descritos se relacionan con el enfoque de entrenamiento con datos de la

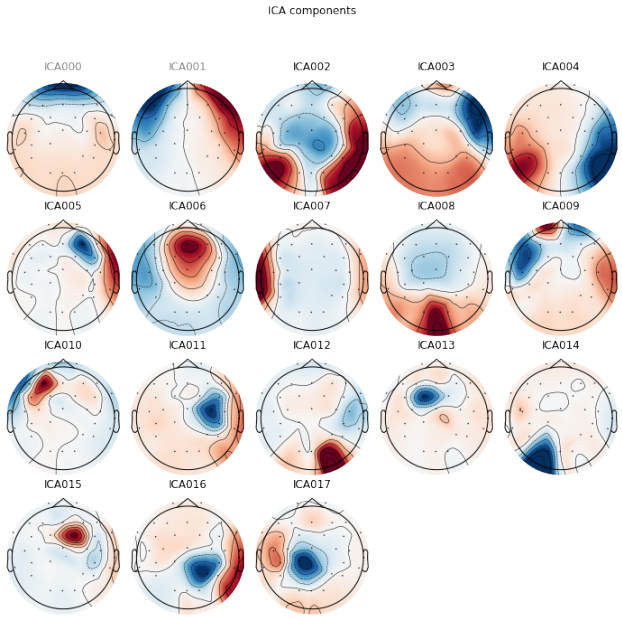


Figura 3.22: Componentes ICA detectados en la señal EEG del participante 01, sesión cognitiva

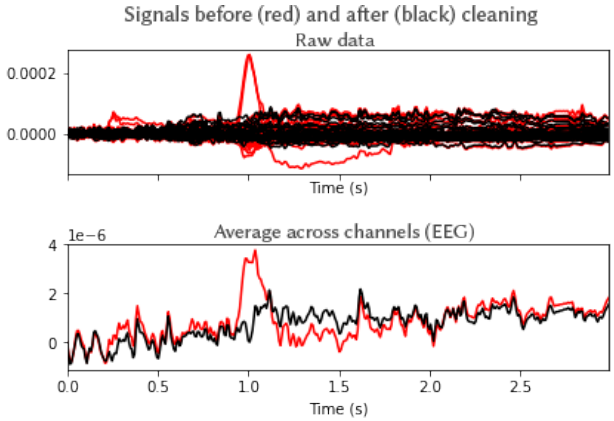


Figura 3.23: Fragmento de la señal con componentes oculares ICA identificados, participante 01, sesión cognitiva

sesión cognitiva y prueba con datos de la sesión web o de evaluación UX.

Dado que el conjunto de datos se construyó descartando los usuarios que no contaran con suficientes canales EEG válidos, los usuarios por sesión fueron los siguientes:

- *Sesión cognitiva (sesión 01)*: 1, 2, 5, 7, 8, 14, 18 y 19.
- *Sesión web (sesión 02)*: 1, 2, 4, 5, 7, 8, 14 y 19.

Considerando las diferencias y la cantidad de registros entre los participantes de cada sesión (ver Tabla 3.8), se realizaron pruebas con algunos filtros por usuario:

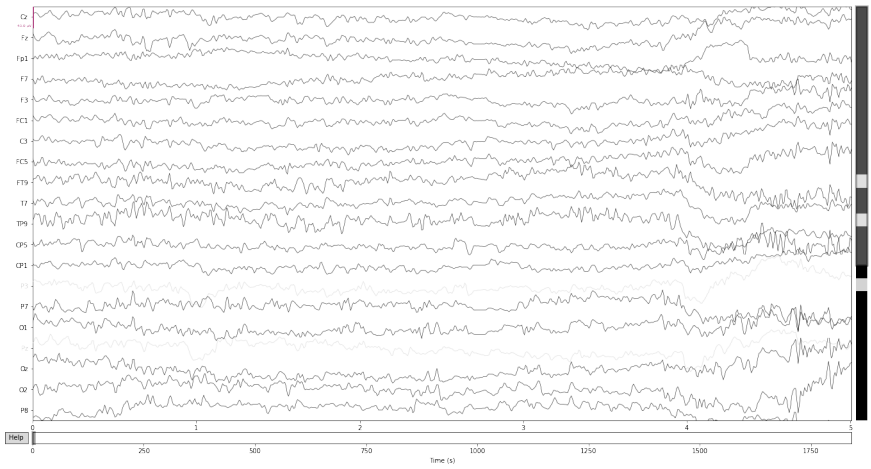


Figura 3.24: Primeros 5s de datos después de remover componentes oculares ICA, participante 01, sesión cognitiva

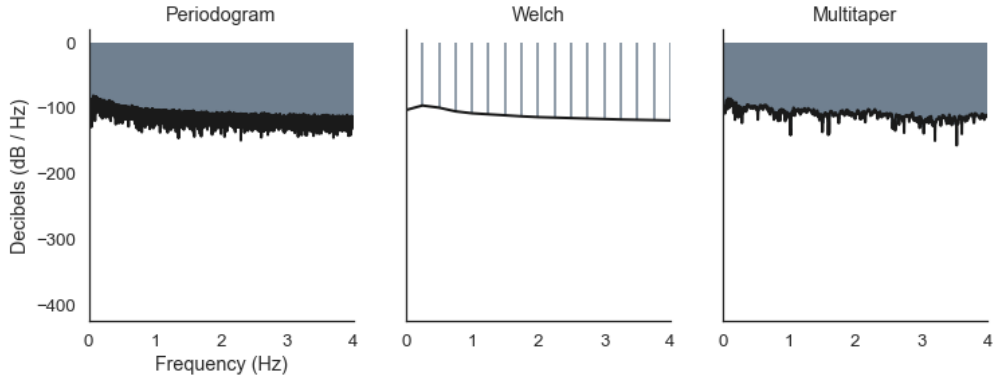


Figura 3.25: Densidad espectral de poder calculada con los métodos de Welch y *multitaper*

- *sin p04s02*: Sin los datos del participante 04 de la sesión 02, debido a que no se encuentra en la sesión 01 y a que por problemas en los sensores se obtuvo una cantidad de registros menor al promedio.
- *sin p19s01*: Sin los datos del participante 19 de la sesión 01, debido a que por problemas en los sensores se obtuvo una cantidad de registros menor al promedio.
- *mismos usuarios*: Sin los datos de los participantes 18 y 19 de la sesión 01 y de los participantes 04 y 19 de la sesión 02, para realizar pruebas con los mismos participantes en ambas sesiones.

La Tabla 3.9 muestra un resumen de las pruebas realizadas y los resultados, obteniendo la mejor exactitud de 82 % con un modelo XGB[1], conjunto normalizado, considerando los mismos usuarios para entrenamiento y validación así como ajustes statEEG, gtMedia y segP.

[1]Parámetros básicos: objective binary:logistic, n_estimators 150, max_depth 10, colsample_bytree 0.7 y scale_pos_weight 0.86

Tabla 3.7: Resultados de pruebas con sesión cognitiva

Modelo	Conjunto	Ajustes	ID Parámetros	% Acc
XGB	Base	-	0	39.3
XGB	Base	-	1	38.5
XGB	Normalizado	-	0	52.1
XGB	Normalizado	statEEG	0	58.2
XGB	Normalizado	statEEG	1	59.3
XGB	Normalizado	gtMedia	0	63.4
XGB	Normalizado	gtMedia	1	62.6
XGB	Normalizado	statEEG, gtMedia	0	65.2
XGB	Normalizado	statEEG, gtMedia	1	64.8
XGB	Estandarizado	-	0	63.6
XGB	Estandarizado	-	1	64.0
RF	Base	-	0	33.9
RF	Base	-	1	38.0
RF	Base	-	2	41.8
RF	Base	-	3	31.0
RF	Normalizado	-	0	64.3
RF	Normalizado	gtMedia	0	64.3
RF	Normalizado	gtMedia	1	62.4
RF	Normalizado	gtMedia	2	60.5
RF	Normalizado	gtMedia	3	65.4[1]
RF	Normalizado	statEEG, gtMedia	0	64.4
RF	Normalizado	statEEG, gtMedia	1	63.0
RF	Normalizado	statEEG, gtMedia	2	60.6
RF	Normalizado	statEEG, gtMedia	3	64.5
RF	Estandarizado	-	0	62.4
RF	Estandarizado	-	1	64.0
RF	Estandarizado	-	2	62.2
RF	Estandarizado	-	3	59.7

[1] Exactitud más alta

Tabla 3.8: Cantidad de registros por participante

Participante	Registros	Duración aprox. (segundos)
Sesión cognitiva		
p1	137953	1078
p2	154347	1206
p5	128472	1004
p7	158071	1235
p8	142203	1111
p14	143118	1118
p18	128216	1002
p19	28839	225
Sesión de evaluación UX		
p1	67418	527
p2	77531	606
p4	9382	73
p5	53271	416
p7	79733	623
p8	91021	711
p14	54913	429
p19	88413	691

Con el mejor modelo encontrado y solo con los participantes con datos válidos para las dos sesiones, se obtiene la carga de trabajo de cada segmento de evaluación UX a partir de las predicciones mayoritarias de cada participante. La Figura 3.26 presenta los resultados para cada una de las tareas de interacción con el sitio web, denotando que en la mayoría se encontró una

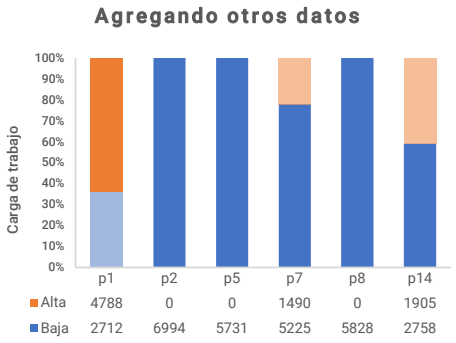
Tabla 3.9: Resultados de las pruebas entrenando con sesión cognitiva y validando con sesión web

Modelo	Conjunto	Ajustes	Fitro usuario	ID Parámetros	% Acc
XGB	Normalizado	-	sin p04s02	0	66.1
XGB	Normalizado	-	sin p19s01, sin p04s02	0	60.0
XGB	Normalizado	-	mismos usuarios	0	64.8
XGB	Normalizado	statEEG, gtMedia	mismos usuarios	0	79.8
XGB	Normalizado	statEEG, gtMedia, sin20s	mismos usuarios	0	78.1
XGB	Normalizado	statEEG, gtMedia, segP	mismos usuarios	0	80.8
XGB	Normalizado	statEEG, gtMedia, segP	mismos usuarios	1	82.0[1]
XGB	Normalizado	sinEEG, gtMedia, segP	mismos usuarios	0	79.9
XGB	Normalizado	sinEEG, gtMedia, segP	mismos usuarios	1	79.3
RF	Normalizado	-	-	0	69.4
RF	Normalizado	-	sin p04s02	0	69.2
RF	Normalizado	-	sin p19s01, sin p04s02	0	65.2
RF	Normalizado		mismos usuarios	0	66.0
RF	Normalizado	-	mismos usuarios	1	65.5
RF	Normalizado	-	mismos usuarios	2	66.5
RF	Normalizado	-	mismos usuarios	3	46.4
RF	Normalizado	statEEG, gtMedia	mismos usuarios	0	78.8
RF	Normalizado	statEEG, gtMedia	mismos usuarios	1	78.9
RF	Normalizado	statEEG, gtMedia	mismos usuarios	2	73.4
RF	Normalizado	statEEG, gtMedia	mismos usuarios	3	70.6
RF	Normalizado	statEEG, gtMedia, sin20s	mismos usuarios	0	80.0
RF	Normalizado	statEEG, gtMedia, sin20s	mismos usuarios	1	81.0
RF	Normalizado	statEEG, gtMedia, sin20s	mismos usuarios	2	75.0
RF	Normalizado	statEEG, gtMedia, sin20s	mismos usuarios	3	70.5
RF	Normalizado	statEEG, gtMedia, segP	mismos usuarios	0	79.7
RF	Normalizado	statEEG, gtMedia, segP	mismos usuarios	1	79.6
RF	Normalizado	statEEG, gtMedia, segP	mismos usuarios	2	73.6
RF	Normalizado	statEEG, gtMedia, segP	mismos usuarios	3	72.1
RF	Normalizado	sinEEG, gtMedia, segP	mismos usuarios	0	80.6
RF	Normalizado	sinEEG, gtMedia, segP	mismos usuarios	1	80.0
RF	Normalizado	sinEEG, gtMedia, segP	mismos usuarios	2	75.1
RF	Normalizado	sinEEG, gtMedia, segP	mismos usuarios	3	77.2

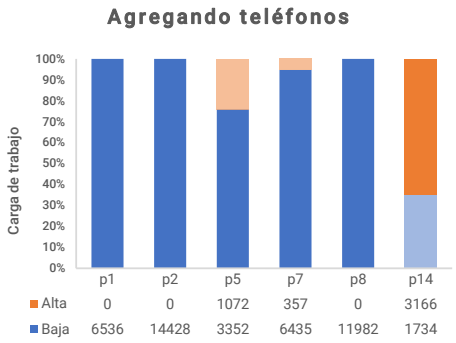
[1] Exactitud más alta.

carga de trabajo baja y solo equilibrada en el caso de la tarea de modificación de datos personales.

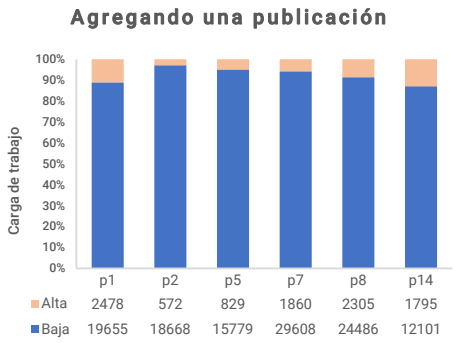
En resumen, se utilizaron los datos de la sesión cognitiva y se realizaron pruebas de clasificación binaria de carga de trabajo (clases baja y alta) con variación en algunos ajustes, posteriormente, utilizando el mejor modelo evaluado con la métrica de exactitud (enfoque LOSO), se predijo la carga de trabajo en los segmentos de evaluación UX utilizando un criterio de mayoría entre los participantes, obteniendo cuatro segmentos con mayoría de carga de trabajo baja y un segmento con carga de trabajo equilibrada (tres participantes con mayoría baja y tres con mayoría alta).



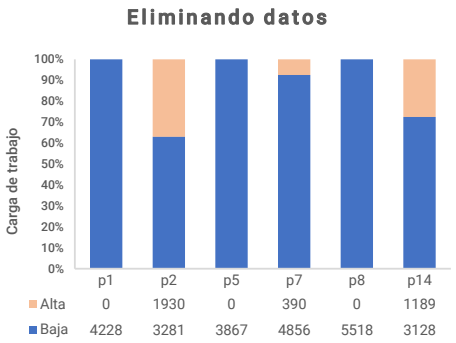
Carga de trabajo mayoritaria: **BAJA**



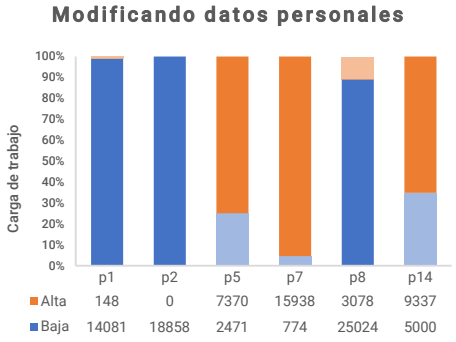
Carga de trabajo mayoritaria: **BAJA**



Carga de trabajo mayoritaria: **BAJA**



Carga de trabajo mayoritaria: **BAJA**



Carga de trabajo mayoritaria: **EQUILIBRADA**

Figura 3.26: Resultados de la carga de trabajo detectada en cada tarea de evaluación UX

4 | Metodología propuesta

La metodología para complementar la evaluación de UX interpretando la carga de trabajo a partir de datos fisiológicos de los usuarios, se diseñó con base en un flujo de trabajo o *pipeline* tradicional de aprendizaje automático [133], con adaptaciones acorde a los hallazgos encontrados a lo largo de la investigación y los objetivos planteados.

En este capítulo se presenta su estructura, la descripción de los componentes y el experimento para evaluar su desempeño.

4.1. Estructura y descripción de los componentes

La Figura 4.1 muestra la estructura del diagrama de bloques, visualmente se identifica: un componente inicial de "Constructo", componentes de "Estímulo" y "Adquisición" que se relacionan con el experimento para adquisición de datos, una separación conceptual entre el "Preprocesamiento" y la "Extracción de características", y componentes finales de "Entrenamiento" y "Predicción" para la obtención de resultados. Las siguientes secciones describen con mayor detalle cada uno de ellos.

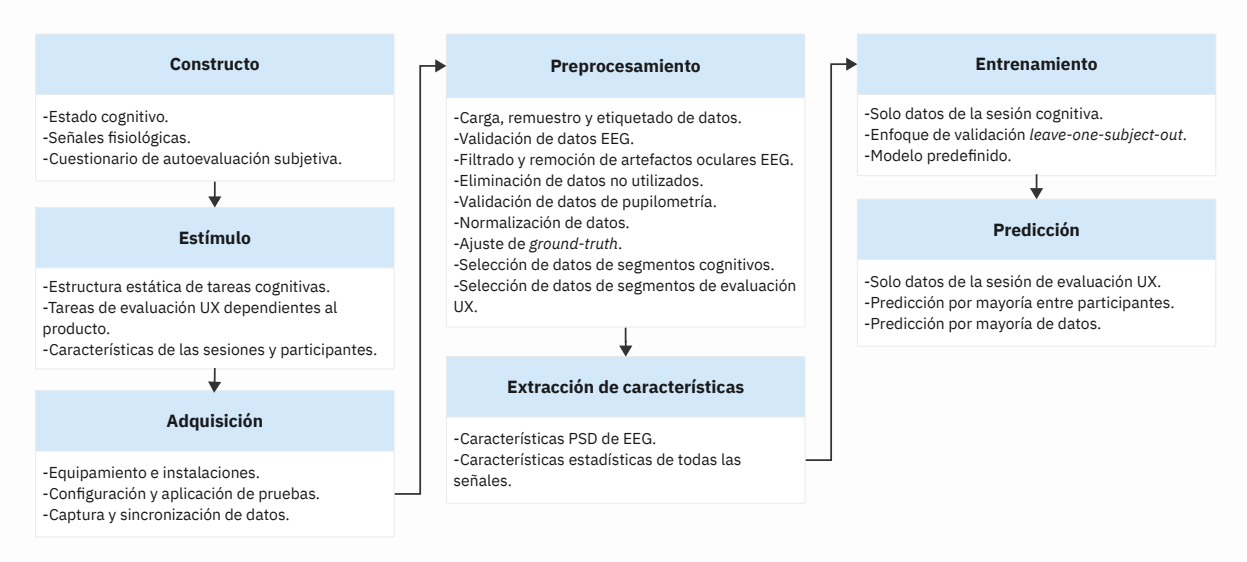


Figura 4.1: Diagrama de bloques de la metodología

4.1.1. Constructo

Esta etapa representa un componente conceptual dentro de la metodología, implica la elección del estado cognitivo de carga de trabajo (ver sección 2.1.3) como complemento a las evaluaciones de experiencia de usuario (ver sección 1.3), a NASA-TLX como el cuestionario estandarizado para su valoración subjetiva (ver sección 2.1.5) y a las señales fisiológicas de EEG, GSR, PPG y pupilometría como los datos inherentes al usuario para poder reconocerlo (ver sección 2.1.4).

Cada uno de los elementos mencionados ya han sido presentados con anterioridad, por lo que se omite su descripción.

4.1.2. Estímulo

En esta etapa se describe el estímulo cognitivo para inducir carga de trabajo así como la estructura y características de las dos sesiones de prueba y de los participantes.

El estímulo cognitivo se conforma de las seis tareas cognitivas con una barra de tiempo límite descritas en el estudio piloto (ver sección 3.2.4).

El experimento se divide en dos sesiones de pruebas:

- Sesión cognitiva (primera sesión). Presenta dos bloques con las seis tareas cognitivas cada uno y una tarea de relajación intermedia (respiración guiada). Las actividades de cada bloque tienen una misma dificultad, iniciando con dificultad baja y luego alta. Antes de cada tarea se muestran las instrucciones para llevarla a cabo y después de cada tarea se solicita contestar el cuestionario de autoevaluación NASA-TLX. Previo a finalizar la prueba se solicita contestar también el cuestionario de comparaciones entre subescalas NASA-TLX para identificar las fuentes de carga de trabajo La Figura 4.2 muestra la estructura de forma gráfica.
- Sesión de evaluación UX (segunda sesión). Consiste de tareas guiadas en el producto digital a evaluar (p. ej., sitio web). Se muestran las instrucciones antes de cada tarea, con un ejercicio de relajación entre ellas. Las tareas se realizan en un navegador web o aplicación de escritorio (dependiendo del producto) y la descripción de cada paso se muestra en una ventana con configuración “siempre al frente” posicionada en la esquina superior derecha.

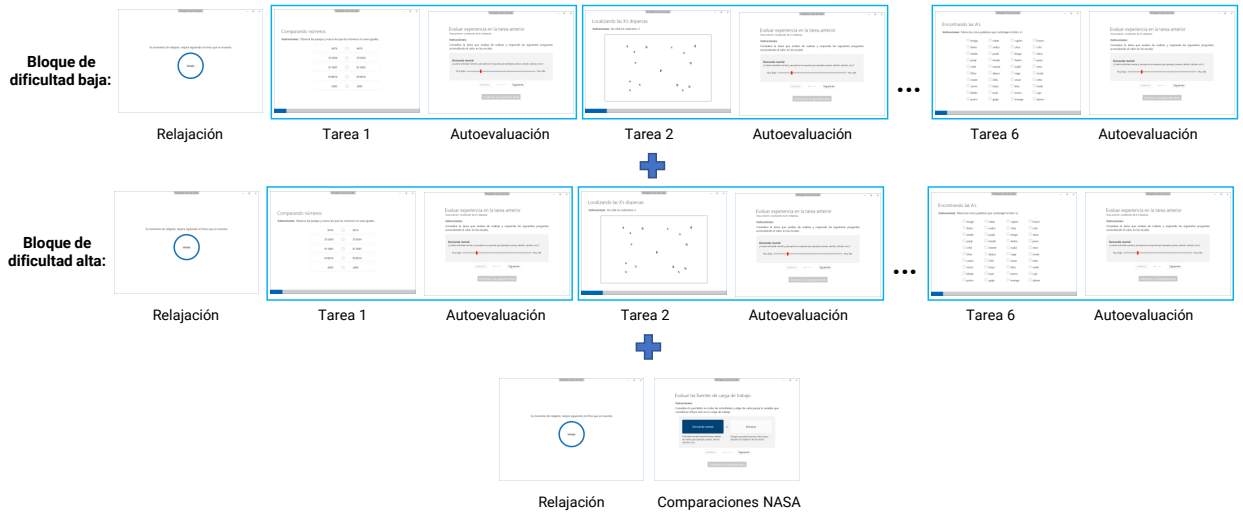


Figura 4.2: Estructura de la sesión cognitiva

La sesión cognitiva tiene una duración aproximada de 30 min. considerando que el software utilizado para su aplicación avanza en automático una vez que se termina exitosamente una tarea cognitiva o se alcanza el tiempo límite. Se ha observado una duración máxima de 35 min. en participantes que ocupan un mayor tiempo al realizar los autoreportes o leer las instrucciones.

Se recomienda que la sesión de evaluación UX se planee con una duración similar de aproximadamente 30 min., esto para evitar el cansancio de los participantes y otras situaciones, como por ejemplo la degradación del gel electroconductivo, contemplando además que se requiere tiempo extra para preparar el software e instalar, calibrar y/o revisar los dispositivos de captura de datos.

En cuanto al número de participantes, en la literatura se encontraron estudios relacionados con estados cognitivos que presentan experimentos con un número de participantes muy diverso, contrario al contexto práctico de evaluación de experiencia de usuario, donde algunos análisis indican que cinco participantes podría ser suficiente (ver sección 2.3.1).

En acuerdo con estas observaciones, se requieren al menos cinco participantes con datos validados en la etapa de preprocesamiento, considerando que es posible descartar participantes debido a corrupción de datos, interrupción de pruebas por fallos eléctricos o de los dispositivos, pérdida de contacto de electrodos, entre otros problemas.

4.1.3. Adquisición

Para la captura de datos fisiológicos se utilizan los dispositivos comerciales Shimmer3 GSR+, Gazepoint GP3 y Emotiv EPOC Flex (ver Figura 4.3); los cuales fueron descritos en la sección 3.2.2.

Figura 4.3: Dispositivos para captura de datos fisiológicos

El proceso de captura de datos se fundamenta en el sistema LSL y se utilizan varias herramientas de software (descritas a detalla en la sección 3.2.3): PruebasCognitivas, ShimmerCapture, ConsensysBASIC, Gazepoint Control, GazepointGP3-LSL, EmotivPro y LabRecorder.

Consensys, Gazepoint Control, EmotivPro[1] y LabRecorder se utilizan tal cual son proporcionadas por sus desarrolladores. PruebasCognitivas y GazepointGP3-LSL se desarrollaron para esta investigación y ShimmerCapture fue modificada para transmitir los datos capturados utilizando el sistema LSL.

Todos los datos son transmitidos y recolectados dentro de una misma computadora con modo de pantalla extendida. El participante utiliza un monitor, teclado y ratón independientes para realizar las tareas. El investigador utiliza otra combinación de pantalla, teclado y ratón para configurar, iniciar las sesiones de grabación y supervisar algunos parámetros durante su desarrollo. La Figura 4.4 muestra la vista del experimentador, un participante en acción y una representación de los datos capturados.

Para minimizar la presencia de ruido en las señales se recomienda que durante la prueba se solicite a los participantes limitar sus movimientos corporales a lo necesario. En el caso de EEG, otros factores como la calidad y desgaste del gorro así como su incorrecta sujeción, pueden generar intermitencias en la calidad de la señal, lo cual puede verificarse en tiempo real con el software EmotivPro.

[1] Esta versión tiene un costo de suscripción mensual.

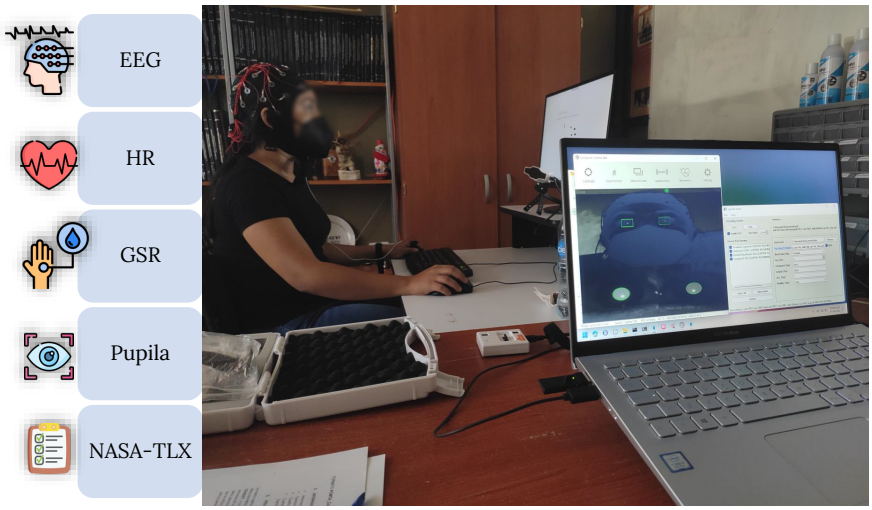


Figura 4.4: Vista del experimentador y representación de los datos capturados

4.1.4. Preprocesamiento y extracción de características

Conceptualmente los componentes de preprocesamiento y extracción de características se presentan por separado, como se observa en el diagrama de la figura 4.1, sin embargo, en la práctica se desarrollan en un único flujo de trabajo[1] que inicia con la lectura de los archivos de datos de cada participante y finaliza con la generación del conjunto de datos empleado para el entrenamiento y predicción.

Este flujo de trabajo recibe como parámetros de entrada las rutas de las carpetas de sesión, donde se encuentran los archivos XDF de datos de cada participante[2], y las rutas de los archivos CSV de etiquetado, exportados desde la herramienta PruebasCognitivas. Presenta como salida archivos tipo *pickle* (formato de serialización de Python) con los datos procesados por participante, por sesión y por ambas sesiones, todos en sus versiones base y normalizada. La Figura 4.5 muestra el flujo de las tareas[3] que se realizan, añadiendo otras del procesamiento final y almacenamiento de resultados. Estas tareas fueron definidas a partir de la experimentación previa, por lo que la descripción y justificación de los métodos elegidos puede consultarse en la sección 3.3.3.

Algunos aspectos a destacar son los siguientes:

- Se ajusta el *ground-truth* por usuario en ambas sesiones, tomando como referencia la media

[1] Implementado en Jupyter *notebooks* con un kernel Python.
[2] Los archivos tiene un nombre con formato específico, p. ej., p01s01.xdf, participante 01 en la sesión 01.
[3] Por simplificación no se muestran las entradas y salidas de cada tarea.

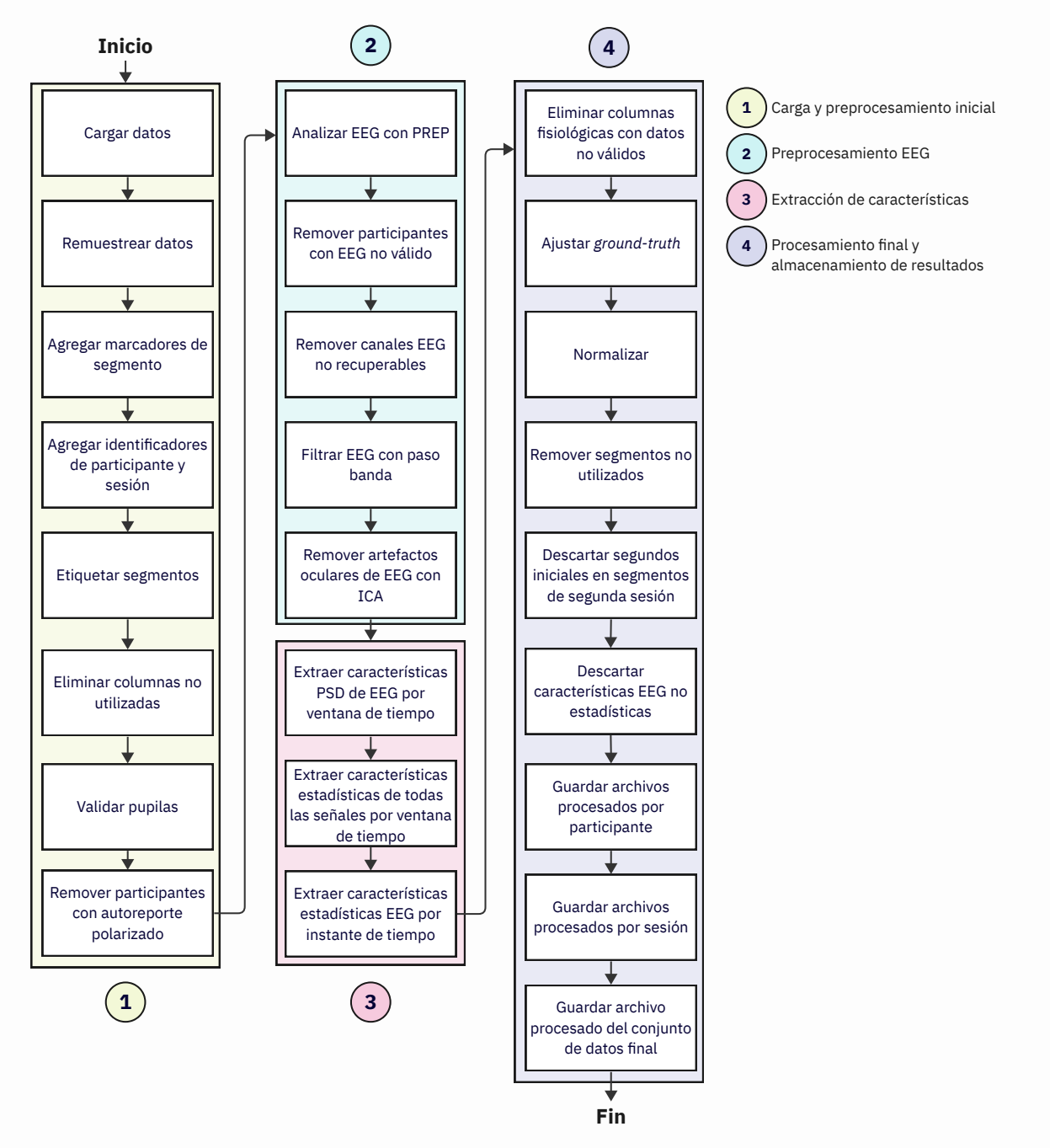


Figura 4.5: Flujo de tareas desde la carga de datos hasta el guardado del conjunto final procesado.

de la carga reportada por el usuario en la sesión cognitiva; esto debido a que se asume que los segmentos de dificultad más alta generan una carga alta y viceversa, independientemente del autoreporte.

- Se descartan los segmentos de datos de relajación, instrucciones y llenado de cuestionarios.
- Se descartan los datos de los participantes con la totalidad de su autoreporte polarizado

hacia carga baja o carga alta.

- Se normalizan los datos, por participante, por sesión.
- De la sesión cognitiva finalmente solo se consideran los tres segmentos[1] con mejor relación carga de trabajo esperada/reportada para cada nivel.
- De la sesión de evaluación UX se descartan los primeros 20s de datos de cada segmento, considerando mantengan un tamaño mínimo de datos de 30s.

4.1.5. Entrenamiento

Se definen las características del modelo de aprendizaje automático acorde a la configuración de mejor rendimiento encontrada en la experimentación previa. Esto corresponde a un clasificador XGB cuyos valores de principales parámetros se pueden observar en la Tabla 4.1.

Tabla 4.1: Parámetros del clasificador XGB

Parámetro	Valor	Parámetro	Valor
objective	binary:logistic	max_depth	10
base_score	0.5	max_leaves	0
booster	gbtree	min_child_weight	1
colsample_bylevel	1	num_parallel_tree	1
colsample_bynode	1	predictor	auto
colsample_bytree	0.7	random_state	11
eval_metric	None	reg_alpha	0
gamma	0	reg_lambda	1
gpu_id	-1	n_estimators	150
grow_policy	depthwise	scale_pos_weight	0.86
learning_rate	0.300000012	max_bin	256
max_cat_to_onehot	4	max_delta_step	0

El modelo fue validado con un enfoque *leave-one-subject-out* (LOSO) con los datos de la sesión cognitiva, obteniendo un rendimiento del 82 % de exactitud en el experimento de evaluación (sección 3.3.5). Se utiliza un enfoque LOSO porque representa una validación independiente del participante, lo que es más adecuado para proyectos que pretendan una aplicación práctica [91] debido a que se evalúa la generalización del modelo con datos completamente desconocidos.

El entrenamiento se realiza con el modelo descrito y los datos de la sesión cognitiva del conjunto resultante de la etapa previa, que en resumen:

- Contiene los valores capturados y características estadísticas de GSR, HR y diámetro de pupilas.

[1] Estos son: XsDispersas_baja, gestalt_baja, encontrandoAs_baja, gestalt_alta, encontrandoAs_alta y patron_alta.

- Incluye las características estadísticas de los valores EEG y PSD absolutos y relativos de los canales válidos por instante de tiempo.
- Contiene solo los segmentos con mejor autoreporte esperado de los participantes con EEG válido y autoreporte no totalmente polarizado.
- Tiene el *ground-truth* ajustado por usuario y está normalizado por usuario.

4.1.6. Predicción

Se utiliza el modelo entrenado en la etapa previa para predecir la carga de trabajo de cada segmento o tarea[1] de la sesión de evaluación UX (segunda sesión). Se determina la carga preponderante por participante y la carga final del segmento a través de un criterio de mayoría; los segmentos con carga de trabajo alta se relacionan con un detrimento en la experiencia del usuario debido a la demanda de recursos requerida.

La Figura 4.6 muestra un ejemplo conceptual del cálculo de la carga mayoritaría que genera una tarea de evaluación UX. La carga mayoritaria resultante puede ser “Alta”, “Baja” o “Equilibrada”[2].

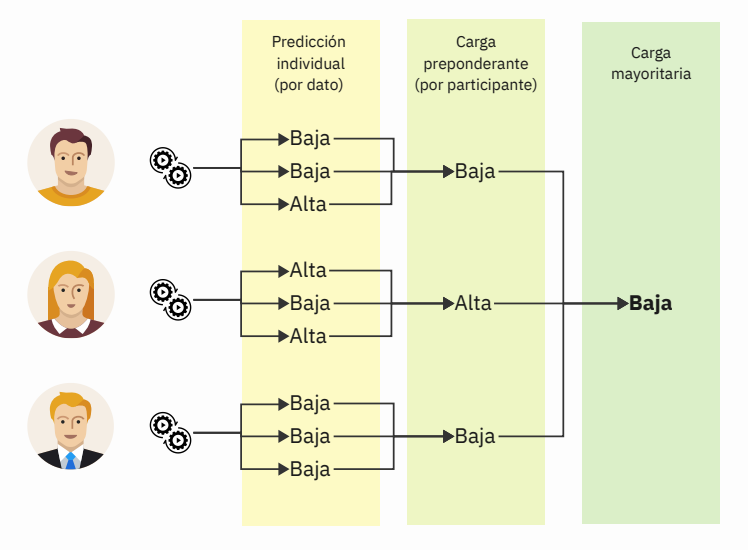


Figura 4.6: Cálculo de carga de trabajo que genera una tarea de evaluación UX

[1] Se descartan los primeros 20s de cada tarea considerando el impacto inicial en los niveles de carga de trabajo.
[2] Posible solo en el caso de un número par de participantes.

4.2. Evaluación

Para evaluar la metodología se realizó un nuevo experimento donde el producto digital a analizar fue el editor de texto en línea denominado Documentos de Google[1].

Las siguientes secciones describen las características del experimento y los resultados obtenidos. Se omiten algunos aspectos, por ejemplo el estímulo cognitivo, por ya estar definidos por la propia metodología.

4.2.1. Participantes

Se capturaron datos de 11 participantes, 5 mujeres y 6 hombres, con una media de edad de 30.7 años; todos estudiantes del TecNM Cenidet. Fueron descartados los datos de uno de los participantes debido a que se presentó un "apagón" eléctrico cuando estaba por iniciar la última tarea de la segunda sesión. El experimento se realizó en un cubículo dentro del edificio de Computación atendiendo las recomendaciones de higiene y seguridad. La Figura 4.7 muestra a un participante realizando una tarea cognitiva.

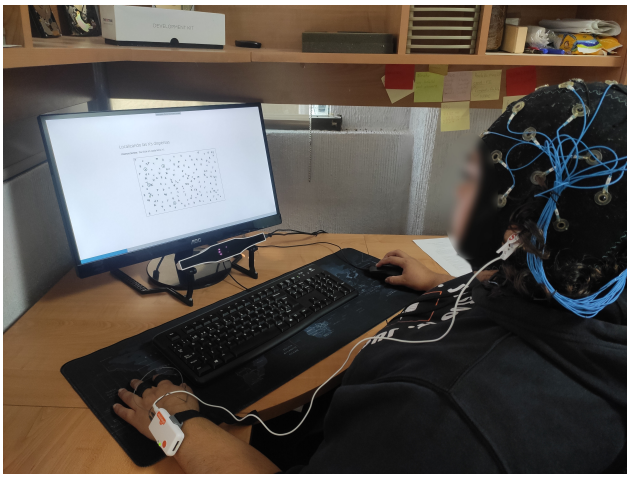


Figura 4.7: Participante realizando una tarea cognitiva

Una de la personas ya había participado en el experimento de evaluación UX previo, por lo que se reutilizaron sus datos de la sesión cognitiva y solo realizó la sesión de evaluación del producto digital. Uno de los objetivos de este experimento era evaluar el desempeño de la metodología bajo esta perspectiva de reutilización de datos previos, sin embargo, no hubo respuesta favorable a la solicitud realizada a otros cinco participantes.

[1]Disponible en: `https://docs.google.com/`

4.2.2. Procedimiento y estímulos

A continuación se describen las tareas correspondientes a la sesión de evaluación UX:

1. Edición de documento de texto (complejidad baja). Ingresar a Google Docs y realizar tareas de edición de texto en la copia de un documento específico, tales como: cambiar fuente, formato y colores de letra; insertar imágenes y cambiar tamaño; configurar interlineado de párrafos; entre otras.
2. Colaboración y versiones (complejidad alta). Continuar desde la tarea previa y trabajar con opciones de colaboración y versiones como: agregar comentarios y asignar tareas, cambiar configuración de notificaciones, intercambiar entre modos de trabajo del documento, rechazar sugerencias, asignar nombres a diferentes versiones del documento, entre otras.
3. Compartir y descargar documento (complejidad media). Continuar desde la tarea previa y publicar el documento en la Web, enviar el archivo por correo electrónico a los colaboradores y descargar el documento en un formato específico.

Cada tarea se diseñó con diferente complejidad considerando el análisis heurístico propio y los publicados por Pranjal Jain [134] y Mayank Khandelwal [135], posteriormente esta complejidad fue confirmada para las tareas de complejidad baja y alta en la entrevista post-prueba realizada a cada participante. La complejidad alta se percibe por la poca familiaridad que se tiene con esas opciones –corroborada también con una encuesta previa–; caso opuesto con la tarea de complejidad baja, donde las opciones de edición básica son similares en ubicación y simbología a las del editor Microsoft Word.

4.2.3. Resultados

La sesión cognitiva tuvo una duración aproximada de 32 min., con un máximo de 36 min. y un mínimo de 24 min., similar a los tiempos del experimento previo (promedio 29 min., máximo de 35 min. y mínimo de 22 min.). La Tabla 4.2 muestra las estampas de tiempo y duración de la sesión de cada participante.

Se descartaron los datos de cuatro participantes por no contar con suficientes canales EEG válidos en ambas sesiones, resultado del análisis PREP. La Tabla 4.3 muestra los canales válidos por participante en ambas sesiones. Se observa que al considerar los datos de todos los participantes se

Tabla 4.2: Duración de las sesiones cognitivas

Participante	Inicio	Fin	Duración
2	12:34:42 p. m.	01:10:07 p. m.	00:35:25
3	02:49:55 p. m.	03:21:40 p. m.	00:31:45
4	08:19:01 a. m.	08:54:55 a. m.	00:35:54
5	10:43:27 a. m.	11:07:33 a. m.	00:24:06
7	12:11:30 p. m.	12:40:14 p. m.	00:28:44
8	10:14:51 a. m.	10:50:23 a. m.	00:35:32
9	12:45:46 p. m.	01:21:07 p. m.	00:35:21
10	04:19:14 p. m.	04:51:22 p. m.	00:32:08
11	09:22:41 a. m.	09:54:01 a. m.	00:31:20
12	12:49:51 p. m.	01:20:49 p. m.	00:30:58

encontraron siete canales “buenos” compartidos en sesión cognitiva, cinco en sesión de evaluación UX y tres compartidos en ambas sesiones.

Tabla 4.3: Validez de canales EEG por participante en sesiones cognitiva y de evaluación UX

	Participantes en sesión cognitiva						Participantes en sesión de evaluación UX						
Canales	p01	p04	p05	p08	p11	p12	p01	p04	p05	p08	p11	p12	Conteo
O1[1,2,3]	X	X	X	X	X	X	X	X	X	X	X	X	12
FT9[1,2,3]	X	X	X	X	X	X	X	X	X	X	X	X	12
F7[1,2,3]	X	X	X	X	X	X	X	X	X	X	X	X	12
TP9[1]	X	X	X	X	X	X	X	X	X	X	X	-	11
Fp2[2]	X	-	X	X	X	X	X	X	X	X	X	X	11
P7[1]	X	X	X	X	X	X	-	X	X	X	X	X	11
F3[2]	X	X	X	X	X	-	X	X	X	X	X	X	11
FC2[1]	X	X	X	X	X	X	X	X	X	X	X	-	11
FC1	X	X	X	X	X	-	X	X	X	X	X	-	10
Fp1[1]	X	X	X	X	X	X	-	X	X	X	-	X	10
Oz	X	X	-	X	X	X	X	X	-	X	X	X	10
O2	X	X	-	X	X	X	X	X	-	X	X	X	10
F8	X	-	X	X	X	X	-	X	X	X	X	X	10
FT10	X	X	X	X	-	X	-	X	X	X	-	X	9
F4	X	X	X	X	X	-	X	X	X	X	-	-	9
Fz	X	X	X	-	X	-	X	X	X	X	X	-	9
FC5	X	-	-	X	X	-	X	X	-	X	X	X	8
P8	X	X	-	X	X	X	-	X	-	-	X	X	8
Cz	X	X	-	X	X	-	X	X	-	X	X	-	8
TP10	X	X	-	X	-	X	-	X	X	-	X	-	7
T7	X	-	-	X	X	-	X	X	-	X	-	X	7
T8	X	X	X	-	-	-	-	X	X	-	-	X	6
FC6	X	-	X	X	-	-	-	X	-	-	-	X	5
CP5	X	-	-	X	-	-	X	X	-	-	-	X	5
C3	X	-	-	X	X	-	-	-	-	-	X	-	4
CP6	X	-	-	X	-	X	-	X	-	-	-	-	4
CP2	-	X	-	X	-	-	-	X	-	X	-	-	4
C4	X	-	-	X	-	-	-	X	-	-	-	-	3
CP1	X	-	-	X	-	-	-	X	-	-	-	-	3
P4	X	-	-	X	-	-	-	X	-	-	-	-	3
P3	-	X	-	X	-	-	-	X	-	-	-	-	3
Pz	-	X	-	-	-	-	-	X	-	-	-	-	2
Conteo	29	21	16	29	20	15	16	31	16	20	18	17	

Solo los participantes con suficientes canales EEG válidos por sesión; X canal válido, - canal no válido
[1] Canales válidos compartidos entre participantes de sesión cognitiva: Fp1, F7, O1, FT9, P7, TP9, FC2
[2] Canales válidos compartidos entre participantes de sesión evaluación: Fp2, F7, O1, FT9, F3
[3] Canales válidos compartidos entre participantes en ambas sesiones: F7, O1, FT9

Como se observa en la Tabla 4.4, la sesión cognitiva mantiene un autoreporte de carga de

trabajo acorde a lo esperado dada su dificultad; los porcentajes representan la proporción en cuanto a la totalidad de datos por dificultad y por cada segmento.

Tabla 4.4: Proporción de autoreporte de carga de trabajo en segmentos cognitivos

Segmento	Proporción	
	Carga baja	Carga alta
Segmentos de dificultad baja	86 %	14 %
XsDispersas_baja	100 %	0 %
busqueda_baja	100 %	0 %
gestalt_baja	100 %	0 %
patron_baja	79 %	21 %
encontrandoAs_baja	76 %	24 %
comparacion_baja	73 %	27 %
Segmentos de dificultad alta	24 %	76 %
encontrandoAs_alta	11 %	89 %
gestalt_alta	11 %	89 %
patron_alta	15 %	85 %
busqueda_alta	23 %	77 %
comparacion_alta	25 %	75 %
XsDispersas_alta	45 %	55 %

Se encontraron variaciones en el orden de los segmentos con mejor proporción de autoreporte esperado de carga de trabajo baja, en comparación con el experimento previo:

- Experimento previo: XsDispersas, gestalt, encontrandoAs, comparacion, busqueda, patron.
- Nuevo experimento: XsDispersas, busqueda, gestalt, patron, encontrandoAs, comparacion.

No hubo variaciones significativas en cuanto al orden de los segmentos con mejor proporción de autoreporte esperado de carga de trabajo alta:

- Experimento previo: gestalt, encontrandoAs, patron, busqueda, comparacion, XsDispersas.
- Nuevo experimento: encontrandoAs, gestalt, patron, busqueda, comparacion, XsDispersas.

En cuanto al desempeño del modelo de aprendizaje automático, se obtuvo una exactitud promedio del 89 % en el enfoque de validación LOSO (ver Tabla 4.5), conservando una exactitud alta aún con nuevos participantes y diferentes condiciones de datos EEG.

Tabla 4.5: Resultado de la validación LOSO del clasificador XGB

Usuario de prueba	acc train	acc test	precision		recall		f1-score	
			Carga baja	Carga alta	Carga baja	Carga alta	Carga baja	Carga alta
4	1.0	0.97	0.97	0.98	0.96	0.98	0.97	0.98
5	1.0	0.98	1.0	0.96	0.94	1.0	0.97	0.98
8	1.0	0.76	0.78	0.75	0.52	0.91	0.63	0.82
11	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
12	1.0	0.65	0.43	0.83	0.68	0.63	0.53	0.72
1	1.0	0.99	1.0	0.98	0.97	1.0	0.98	0.99
Promedio:	1.0	0.89	0.86	0.92	0.85	0.92	0.85	0.92

En cuanto a la predicción de la carga de trabajo en los segmentos de evaluación UX, se puede observar en la Figura 4.8 que para el segmento de complejidad baja (web_edicionBasica) se obtiene una predicción esperada, para el segmento de complejidad alta (web_permisosVersiones) una predicción equilibrada cuando se esperaría ser alta y para el segmento de complejidad media (web_compartirDescargar) una predicción alta.

Para fines de evaluar la metodología se consideraron los segmentos de complejidad alta y baja y se esperaría una carga de trabajo similar predicha por el modelo. Del segmento de complejidad media solo se observa su comportamiento pero no se considera debido a que los segmentos cognitivos provocan una carga de trabajo polarizada.

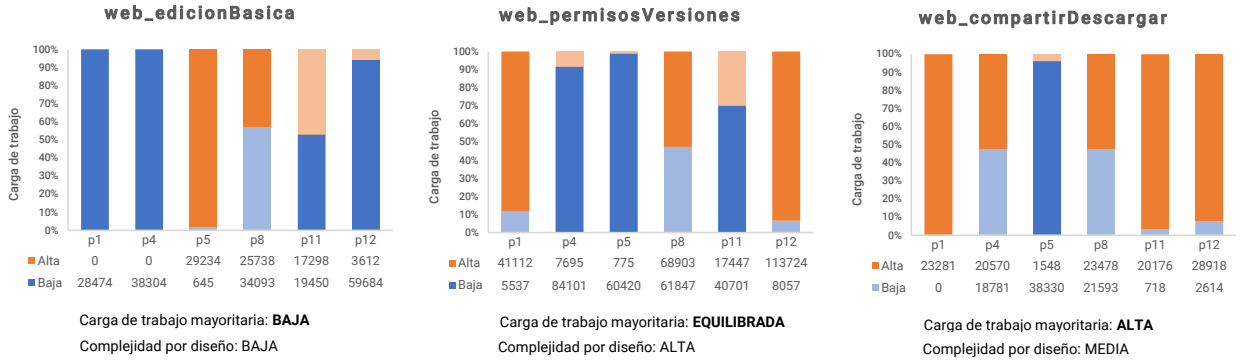


Figura 4.8: Predicción de carga de trabajo en segmentos de evaluación UX

Debido a lo anterior, se analizó el autoreporte de cada uno de los participantes (ver Tabla 4.6), observando que el participante 05 reportó una carga de trabajo baja en todos los segmentos cognitivos, contrario a lo esperado en los segmentos de dificultad alta. El participante 05 terminó las sesiones cognitivas en el menor tiempo (ver Tabla 4.2) y con el mejor desempeño, además, en la entrevista manifestó ser un jugador exhaustivo de videojuegos y haberse tomado las tareas como un reto de esa índole.

En la Tabla 4.6 se observa también que el participante 01 presenta un mayor autoreporte de carga baja en los segmentos de mayor dificultad, pero no para todos los datos como en el caso del participante 05.

Considerando estos aspectos, se descartaron los datos del participante 05 y se realizó de nuevo el entrenamiento y la predicción de carga de trabajo para analizar los cambios entre un resultado y otro, obteniendo una exactitud promedio menor, de 86.8 % con la misma configuración de modelo

Tabla 4.6: Autoreporte de carga de trabajo por participante

Participante	Segmentos de dificultad baja		Segmentos de dificultad alta	
	Carga baja	Carga alta	Carga baja	Carga alta
p1	100 %	0 %	57 %	43 %
p4	75 %	25 %	0 %	100 %
p5	100 %	0 %	100 %	0 %
p8	59 %	41 %	0 %	100 %
p11	100 %	0 %	0 %	100 %
p12	100 %	0 %	17 %	83 %

de aprendizaje automático, pero una predicción más acertada en cuanto a lo esperado dada la complejidad por diseño (ver Figura 4.9).

El comportamiento de autoreporte del participante 05 no se presentó en los participantes de experimentos previos, dado a que es una anomalía de lo que se pretende inducir se determinó ajustar la metodología para contemplar esta situación, este ajuste se refleja en la descripción presentada en la sección 4.1.4 en el bloque denominado “Remover participantes con autoreporte polarizado”.

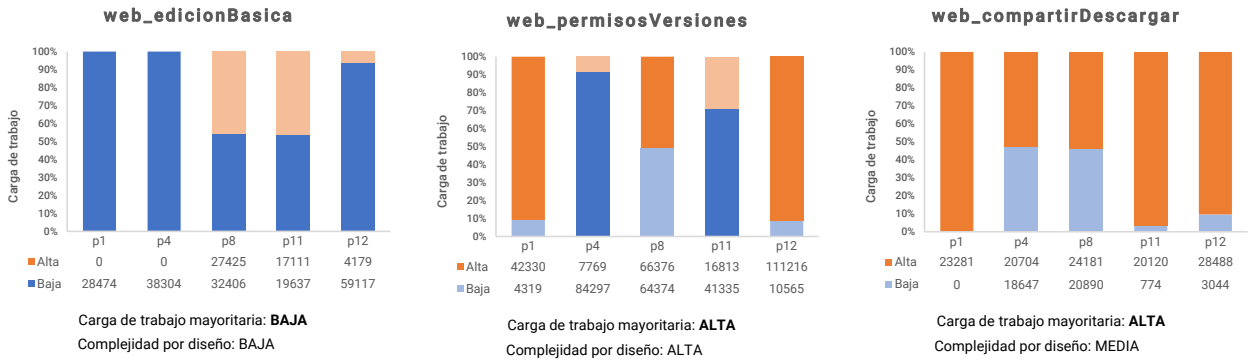


Figura 4.9: Predicción de carga de trabajo en segmentos de evaluación UX, sin considerar participante 05.

5 | Discusión final

5.1. Conclusiones

El trabajo de tesis cumplió los objetivos de la investigación: 1) identificando las características de las señales fisiológicas que se emplean en el reconocimiento de carga de trabajo y las correlaciones con métricas de UX (ver sección 2.3), 2) construyendo varios conjuntos de datos con estímulos relacionados con UX y seleccionando el *dataset* CogLoad, evaluando modelos de aprendizaje automático (ver capítulo tres) y 3) definiendo y evaluando la metodología propuesta (ver capítulo cuatro).

Se comprobó la hipótesis, ya que al aplicar la metodología se complementa la evaluación tradicional de UX dentro del plano cuantitativo-objetivo, detectando la carga de trabajo de los usuarios utilizando un modelo entrenado con datos fisiológicos de actividad cerebral, ritmo cardíaco, conductancia de la piel y seguimiento ocular. Sin embargo, persiste un componente subjetivo, a través del cuestionario estandarizado NASA-TLX, que es base para el etiquetado de la sesión cognitiva.

La experimentación reafirma que gestionar las diferencias individuales de los participantes y obtener un buen rendimiento en la clasificación de carga de trabajo con un enfoque LOSO continúa siendo un reto en el estado del arte. Por esta razón, la metodología define el entrenamiento con un subconjunto de datos donde el estímulo cognitivo induce una carga de trabajo polarizada, para luego realizar predicciones en datos desconocidos de las tareas con el producto digital, obteniendo el resultado para cada segmento a partir de la predicción mayoritaria entre participantes.

Este enfoque brinda a su vez la posibilidad de reutilizar los datos validados de usuarios que realizaron la sesión cognitiva en otros procesos de evaluación UX, utilizándolos en la mejora del modelo de aprendizaje o incluso convocando de nuevo a estos usuarios pero solo para realizar las tareas de interacción con otro producto digital a evaluar. Esta flexibilidad es consecuencia de la estructura modular de la metodología, la cual le brinda una mayor adaptabilidad y permite efectuar ajustes posteriores en su diseño y aplicación.

Los experimentos fueron realizados con participantes dentro del perfil de usuario potencial de cada software, condición básica en un proceso de investigación en UX, por lo que la metodología ha sido construida y verificada a partir de los datos de 40 y 11 usuarios, respectivamente, cuyas edades van desde los 20 a los 47 años, con una media de 26.9 años, la mayoría estudiantes de licenciatura o posgrado.

Finalmente, aunque la metodología no define detalles de las condiciones físicas de las instalaciones y del equipo para desarrollar los experimentos –se obvia–, esto es prioritario para capturar datos con mejor calidad.

5.2. Aportaciones

La contribución más relevante de esta tesis se da en la definición de la metodología para complementar la evaluación de UX con base en el reconocimiento de carga de trabajo, modelos de aprendizaje y datos fisiológicos; estableciendo: características de los estímulos; condiciones y recomendaciones para la experimentación; flujo de trabajo de procesamiento de datos; configuración y elección de modelos de clasificación y obtención y presentación de resultados.

A continuación se enlistan otros productos de la investigación.

Publicación en revista JCR[1]:

- E. Bañuelos-Lozoya, G. González-Serna, N. González-Franco *et al.*, "A Systematic Review for Cognitive State-Based QoE/UX Evaluation," *Sensors*, vol. 21, n.º 10, 2021. DOI: 10.3390/s21103439

Publicaciones en congresos nacionales:

- E. O. Bañuelos Lozoya, J. G. González Serna, N. González Franco *et al.*, *Metodología para complementar la evaluación de UX interpretando la carga de trabajo a partir de datos fisiológicos del usuario*, Cuernavaca, Morelos, México, 2023. X JCyTA. Artículo corto y póster.
- E. O. Bañuelos Lozoya, J. G. González Serna y N. González Franco, *Experimento para capturar datos fisiológicos e interpretar estados cognitivos en evaluaciones UX*, Cuernavaca, Morelos, México, 2022. V JCyTA.

[1] Se redactaron otros artículos de este tipo, no se enlistan por no estar publicados a la fecha de la última revisión del documento.

- E. O. Bañuelos Lozoya, J. G. González Serna y N. González Franco, *Evaluación de conjuntos de datos para reconocimiento de carga cognitiva y estrés*, Cuernavaca, Morelos, México, 2021. IV JCyTA.

Software desarrollados/adaptado:

- PruebasCognitivas, con registro ante INDAUTOR[1].
- GazepointGP3-LSL.
- ShimmerCapture, solo adaptación LSL.

5.3. Trabajos futuros

La metodología facilita su mejora continua en varios aspectos, siendo el rendimiento en la clasificación de la carga de trabajo uno de los primordiales, por lo que se recomienda realizar pruebas con los datos de las sesiones cognitivas experimentando con modelos de aprendizaje mas robustos, como los de tipo profundo, ajustando los componentes necesarios acorde a los hallazgos encontrados.

El aplicar la metodología con participantes recurrentes y reutilizar sus datos cognitivos en varias pruebas de evaluación UX permite, por lo menos, tiempos más bajos en la experimentación, procesamiento de los datos y entrega de resultados, sin embargo, dada la poca disponibilidad de participantes no fue posible evaluar este enfoque a detalle.

El software PruebasCognitivas permite aplicar cuestionarios estandarizados con preguntas tipo Likert usando una configuración basada en plantillas, no obstante solo procesa las respuestas del cuestionario NASA-TLX acorde a la metodología del instrumento, entregando el resto de las respuestas “tal cual”, por lo que es necesario implementar las funciones necesarias para procesar otros cuestionarios que así lo requieran.

[1] Número de registro: 03-2023-041315011100-01

Referencias

[1] E. L. Law y P. Van Schaik, "Modelling user experience - An agenda for research and practice," *Interacting with Computers*, vol. 22, n.º 5, págs. 313-322, 2010, ISSN: 09535438. DOI: 10.1016/j.intcom.2010.04.006.

[2] B. van de Laar, H. Gürkök, D. P.-O. Bos, F. Nijboer y A. Nijholt, "Brain-Computer Interfaces and User Experience Evaluation," en *Towards Practical Brain-Computer Interfaces: Bridging the Gap from Research to Real-World Applications*, B. Z. Allison, S. Dunne, R. Leeb, J. Del R. Millán y A. Nijholt, eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, págs. 223-237, ISBN: 978-3-642-29746-5. DOI: 10.1007/978-3-642-29746-5_11.

[3] E. L.-C. Law, P. van Schaik y V. Roto, "Attitudes towards user experience (UX) measurement," *International Journal of Human-Computer Studies*, vol. 72, n.º 6, págs. 526-541, 2014, ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2013.09.006.

[4] G. Lasa, D. Justel y A. Retegi, "Eyeface: A new multimethod tool to evaluate the perception of conceptual user experiences," *Computers in Human Behavior*, vol. 52, págs. 359-363, 2015, ISSN: 0747-5632. DOI: 10.1016/j.chb.2015.06.015.

[5] O. Matthews, A. Davies, M. Vigo y S. Harper, "Unobtrusive arousal detection on the web using pupillary response," *International Journal of Human-Computer Studies*, vol. 136, pág. 102 361, 2020, ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2019.09.003.

[6] J. Posner, J. A. Russell y B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and Psychopathology*, vol. 17, n.º 3, págs. 715-734, 2005. DOI: 10.1017/S0954579405050340.

[7] M. A. Vidulich y P. S. Tsang, "Mental workload and situation awareness," en *Hanbook of Human Factors and Ergonomics*, G. Salvendy, ed., Cuarta edi, John Wiley & Sons, Inc., 2012, cap. 8, págs. 243-273, ISBN: 978-0-470-52838-9.

[8] S. Yang, Z. Yin, Y. Wang, W. Zhang, Y. Wang y J. Zhang, "Assessing cognitive mental workload via EEG signals and an ensemble deep learning classifier based on denoising autoencoders," *Computers in Biology and Medicine*, vol. 109, págs. 159-170, 2019, ISSN: 18790534. DOI: 10.1016/j.compbiomed.2019.04.034.

[9] A. Jimenez-Molina, C. Retamal y H. Lira, "Using Psychophysiological Sensors to Assess Mental Workload During Web Browsing," *Sensors*, vol. 18, n.º 2, pág. 458, 2018, ISSN: 1424-8220. DOI: 10.3390/s18020458.

[10] J. Frey, M. Daniel, J. Castet, M. Hachet y F. Lotte, "Framework for Electroencephalography-Based Evaluation of User Experience," en *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ép. CHI '16, New York, NY, USA: Association for Computing Machinery, 2016, págs. 2283-2294, ISBN: 9781450333627. DOI: 10.1145/2858036.2858525.

[11] H. Yokoyama, K. Eihata, J. Muramatsu e Y. Fujiwara, "Prediction of Driver's Workload from Slow Fluctuations of Pupil Diameter," en *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, págs. 1775-1780. DOI: 10.1109/ITSC.2018.8569279.

[12] P. Aricò, G. Borghini, G. Di Flumeri *et al.*, "Adaptive automation triggered by EEG-based mental workload index: A passive brain-computer interface application in realistic air traffic control environment," *Frontiers in Human Neuroscience*, vol. 10, n.º OCT2016, 2016, ISSN: 16625161. DOI: 10.3389/fnhum.2016.00539.

[13] S. Federici, M. L. Mele, M. Bracalenti, A. Buttafuoco, R. Lanzilotti y G. Desolda, "Bio-behavioral and Self-Report User Experience Evaluation of a Usability Assessment Platform (UTAssistant)," en *14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, 2019.

[14] F. Putze y T. Schultz, *Adaptive cognitive technical systems*, 2014. DOI: 10.1016/j.jneumeth.2014.06.029.

[15] V. Meza-Kubo, A. L. Morán, I. Carrillo, G. Galindo y E. García-Canseco, "Assessing the user experience of older adults using a neural network trained to recognize emotions from brain signals," *Journal of Biomedical Informatics*, vol. 62, págs. 202-209, 2016, ISSN: 15320464. DOI: 10.1016/j.jbi.2016.07.004.

[16] V. Georges, F. Courtemanche, S. Sénécal, P. M. Léger, L. Nacke y R. Pourchon, "The adoption of physiological measures as an evaluation tool in UX," en *HCI in Business, Government and Organizations. Interacting with Information Systems*, F. F.-H. Nah y C.-H. Tan, eds., Springer International Publishing, 2017, págs. 90-98, ISBN: 978-3-319-58481-2.

[17] J. Hussain, W. A. Khan, T. Hur *et al.*, "A multimodal deep log-based user experience (UX) platform for UX evaluation," *Sensors*, vol. 18, n.º 5, pág. 1622, 2018, ISSN: 14248220. DOI: 10.3390/s18051622.

[18] C. H. Chuang, C. S. Huang, L. W. Ko y C. T. Lin, "An EEG-based perceptual function integration network for application to drowsy driving," *Knowledge-Based Systems*, vol. 80, págs. 143-152, 2015, ISSN: 09507051. DOI: 10.1016/j.knosys.2015.01.007.

[19] M. Lohani, B. R. Payne y D. L. Strayer, "A Review of Psychophysiological Measures to Assess Cognitive States in Real-World Driving," *Frontiers in Human Neuroscience*, vol. 13, pág. 57, 2019, ISSN: 1662-5161. DOI: 10.3389/fnhum.2019.00057.

[20] P.-M. Léger, F. Courtemanche, M. Fredette y S. Sénécal, "A Cloud-Based Lab Management and Analytics Software for Triangulated Human-Centered Research," en *Information Systems and Neuroscience*, F. D. Davis, R. Riedl, J. vom Brocke, P.-M. Léger y A. B. Randolph, eds., Cham: Springer International Publishing, 2019, págs. 93-99, ISBN: 978-3-030-01087-4. DOI: 10.1007/978-3-030-01087-4_11.

[21] U. Engelke, D. P. Darcy, G. H. Mulliken *et al.*, *Psychophysiology-Based QoE Assessment: A Survey*, 2017. DOI: 10.1109/JSTSP.2016.2609843.

[22] S. G. Charlton, "Measurement of Cognitive States in Test and Evaluation," en *Handbook of Human Factors Testing and Evaluation*, S. G. Charlton y T. G. O'Brien, eds., Segunda ed, Lawrence Erlbaum Associates, 2002, cap. 6, págs. 97-126, ISBN: 1-4106-0380-6.

[23] M. X. Huang, J. Li, G. Ngai y H. V. Leong, "StressClick: Sensing Stress from Gaze-Click Patterns," en *Proceedings of the 24th ACM International Conference on Multimedia*, ép. MM '16, New York, NY, USA: Association for Computing Machinery, 2016, págs. 1395-1404, ISBN: 9781450336031. DOI: 10.1145/2964284.2964318.

[24] S. Pritam, K. Ross, A. J. Ruberto, D. Rodenburg, P. Hungler y A. Etemad, "Classification of Cognitive Load and Expertise for Adaptative Simulation using Deep Multitask Learning," *arXiv e-prints*, 2019. arXiv: 1908.00385.

[25] F. Courtemanche, P. M. Léger, A. Dufresne, M. Fredette, É. Labonté-Lemoyne y S. Sénécal, "Physiological heatmaps: A tool for visualizing users' emotional reactions," *Multimedia Tools and Applications*, vol. 77, n.º 9, págs. 11 547-11 574, 2018, ISSN: 15737721. DOI: 10.1007/s11042-017-5091-1.

[26] ISO, "ISO 9241-210:2010 Ergonomics of human-system interaction - Part 210: Human-centred design for interactive systems," International Organization for Standardization, inf. téc., 2010.

[27] M. Hassenzahl y N. Tractinsky, "User experience - A research agenda," *Behaviour and Information Technology*, vol. 25, n.º 2, págs. 91-97, 2006, ISSN: 0144929X. DOI: 10.1080/01449290500330331.

[28] I. Wechsung y K. De Moor, "Quality of Experience Versus User Experience," en *Quality of Experience: Advanced Concepts, Applications and Methods*, S. Möller y A. Raake, eds., Switzerland: Springer, 2014, cap. 3, págs. 35-54. DOI: 10.1007/978-3-319-02681-7_3.

[29] A. Raake y S. Egger, "Quality and Quality of Experience," en *Quality of Experience: Advanced Concepts, Applications and Methods*, S. Möller y A. Raake, eds., Switzerland: Springer, 2014, cap. 2, págs. 11-33. DOI: 10.1007/978-3-319-02681-7_2.

[30] F. Hammer, S. Egger-Lampl y S. Möller, "Quality-of-user-experience: a position paper," *Quality and User Experience*, vol. 3, n.º 1, pág. 9, 2018, ISSN: 2366-0147. DOI: 10.1007/s41233-018-0022-0.

[31] C. D. Salzman y S. Fusi, "Emotion, cognition, and mental state representation in amygdala and prefrontal cortex," eng, *Annual review of neuroscience*, vol. 33, págs. 173-202, 2010, ISSN: 1545-4126. DOI: 10.1146/annurev.neuro.051508.135256.

[32] M. D. Robinson, E. R. Watkins y E. Harmon-Jones, "Cognition and Emotion: An Introduction," en *Handbook of Cognition and Emotion*, M. D. Robinson, E. R. Watkins y E. Harmon-Jones, eds., The Guilford Press, 2013, cap. 1, págs. 3-16, ISBN: 978-1-4625-0999-7.

[33] L. Pessoa, "On the relationship between emotion and cognition," *Nature Reviews Neuroscience*, vol. 9, n.º 2, págs. 148-158, 2008, ISSN: 1471-0048. DOI: 10.1038/nrn2317.

[34] J. Heard, C. E. Harriott y J. A. Adams, "A survey of workload assessment algorithms," *IEEE Transactions on Human-Machine Systems*, vol. 48, n.º 5, págs. 434-451, 2018. DOI: 10.1109/THMS.2017.2782483.

[35] T. C. Dolmans, M. Poel, J.-W. J. R. van 't Klooster y B. P. Veldkamp, "Perceived Mental Workload Classification Using Intermediate Fusion Multimodal Deep Learning," *Frontiers in Human Neuroscience*, vol. 14, 2021, ISSN: 1662-5161. DOI: 10.3389/fnhum.2020.609096.

[36] D. Jaiswal, D. Chatterjee, R. Gavas, R. K. Ramakrishnan y A. Pal, "Effective Assessment of Cognitive Load in Real-World Scenarios Using Wrist-Worn Sensor Data," en *Proceedings of the Workshop on Body-Centric Computing Systems*, ép. BodySys'21, New York, NY, USA: Association for Computing Machinery, 2021, págs. 7-12, ISBN: 9781450386005. DOI: 10.1145/3469260.3469666.

[37] E. Debie, R. F. Rojas, J. Fidock *et al.*, "Multimodal Fusion for Objective Assessment of Cognitive Workload: A Review," *IEEE Transactions on Cybernetics*, págs. 1-14, 2019, ISSN: 2168-2275. DOI: 10.1109/TCYB.2019.2939399.

[38] J. Zhou, K. Yu, F. Chen, Y. Wang y S. Z. Arshad, "Multimodal Behavioral and Physiological Signals as Indicators of Cognitive Load," en *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition - Volume 2*. Association for Computing Machinery y Morgan & Claypool, 2018, págs. 287-329, ISBN: 9781970001716. DOI: 10.1145/3107990.3108002.

[39] D. Cernea y A. Kerren, "A survey of technologies on the rise for emotion-enhanced interaction," *Journal of Visual Languages and Computing*, vol. 31, págs. 70-86, 2015, ISSN: 1045926X. DOI: 10.1016/j.jvlc.2015.10.001.

[40] A. J. Schall y J. Romano Bergstrom, "Introduction to Eye Tracking," en *Eye Tracking in User Experience Design*, J. Romano Bergstrom y A. J. Schall, eds., Boston: Morgan Kaufmann, 2014, cap. 1, págs. 3-26, ISBN: 978-0-12-408138-3. DOI: 10.1016/B978-0-12-408138-3.00001-7.

[41] F. Onorati, R. Barbieri, M. Mauri, V. Russo y L. Mainardi, "Characterization of affective states by pupillary dynamics and autonomic correlates," *Frontiers in Neuroengineering*, vol. 6, pág. 9, 2013, ISSN: 1662-6443. DOI: 10.3389/fneng.2013.00009.

[42] A. Naït-Ali y P. Karasinski, "Biosignals: Acquisition and General Properties," en *Advanced Biosignal Processing*, A. Naït-Ali, ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, págs. 1-13, ISBN: 978-3-540-89506-0. DOI: 10.1007/978-3-540-89506-0_1.

[43] P. A. Abhang, B. W. Gawali y S. C. Mehrotra, "Technological Basics of EEG Recording and Operation of Apparatus," en *Introduction to EEG- and Speech-Based Emotion Recognition*, P. A. Abhang, B. W. Gawali, S. C. B. T. .-. I. t. E. Mehrotra y S.-B. E. Recognition, eds., Academic Press, 2016, cap. 2, págs. 19-50, ISBN: 978-0-12-804490-2. DOI: 10.1016/B978-0-12-804490-2.00002-6.

[44] S. Okutucu y A. Oto, "Fundamentals of ECG," en *Interpreting ECGs in Clinical Practice*. Cham: Springer International Publishing, 2018, cap. 1, págs. 1-18, ISBN: 978-3-319-90557-0. DOI: 10.1007/978-3-319-90557-0_1.

[45] British Heart Foundation, "Heart rhythms," inf. téc., 2013. dirección: https://www.bhf.org.uk/%7B~%7D/media/files/publications/large-print/his14lp%7B%5C_%7Dheart-rhythms%7B%5C_%7D0512.pdf.

[46] M. Z. Baig y M. Kavakli, "A Survey on Psycho-Physiological Analysis & Measurement Methods in Multimodal Systems," *Multimodal Technologies and Interaction*, vol. 3, n.º 2, 2019. DOI: 10.3390/mti3020037.

[47] D. Biswas, N. Simões-Capela, C. Van Hoof y N. Van Helleputte, "Heart Rate Estimation From Wrist-Worn Photoplethysmography: A Review," *IEEE Sensors Journal*, vol. 19, n.º 16, págs. 6560-6570, 2019, ISSN: 1558-1748. DOI: 10.1109/JSEN.2019.2914166.

[48] C. Zeagler, "Where to Wear It: Functional, Technical, and Social Considerations in on-Body Location for Wearable Technology 20 Years of Designing for Wearability," en *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, ép. ISWC '17, New York, NY, USA: Association for Computing Machinery, 2017, págs. 150-157, ISBN: 9781450351881. DOI: 10.1145/3123021.3123042.

[49] S. Koelstra, C. Mühl, M. Soleymani *et al.*, "DEAP: A Database for Emotion Analysis Using Physiological Signals," *IEEE Transactions on Affective Computing*, vol. 3, n.º 1, págs. 18-31, 2012. DOI: 10.1109/T-AFFC.2011.15.

[50] S. G. Hart y L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," en *Human Mental Workload*, P. A. Hancock y N. B. T. .-. A. i. P. Meshkati, eds., vol. 52, North-Holland, 1988, págs. 139-183. DOI: 10.1016/S0166-4115(08)62386-9.

[51] F. Chollet, *Deep Learning with Python*. Manning Publications Co., 2018, ISBN: 9781617294433.

[52] J. Hurwitz y D. Kirsch, *Machine Learning For Dummies*. John Wiley & Sons, Inc., 2018, ISBN: 978-1-119-45495-3.

[53] I. Goodfellow, Y. Bengio y A. Courville, *Deep Learning*. MIT Press, 2016, ISBN: 0262035618.

[54] T. M. Mitchell, *Machine Learning*. McGraw-Hill Education, 1997, ISBN: 0070428077.

[55] C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. 8, n.º 4, págs. 283-298, 1978, ISSN: 0001-2998. DOI: 10.1016/S0001-2998(78)80014-2.

[56] H. O. Alejandres Sánchez, "Evaluación centrada en el usuario de sistemas de recomendación sensibles al contexto: efecto de interfaces multimodales interactivas y esquemas de explicación en la experiencia del usuario," Tesis de doctorado, TecNM/Cenidet, 2017.

[57] N. González Franco, "Metodología UXEeg para la evaluación de la Experiencia del Usuario en personas con discapacidad a partir de Interfaces Cerebro Computadora," Tesis de doctorado, TecNM/CENIDET, 2017.

[58] D. E. Fouilloux Quiroz, "Método para integrar y sincronizar datos EEG y mutimedia para su aplicación en el proceso de evaluación de la experiencia del usuario," Tesis de maestría, TecNM/Cenidet, 2018.

[59] G. A. García Pinzón, "Procesamiento de datos fisiológicos para detectar estados afectivos en el proceso de evaluación de la experiencia de usuario," Tesís de maestría, TecNM/Cenidet, 2020.

[60] J. Soriano Terrazas, "Metodología para caracterizar e inducir estados mentales a través de realidad virtual inmersiva e interfaz cerebro computadora," Tesis de maestría, TecNM/Cenidet, 2018.

[61] D. A. Lagunes Ramírez, "Método para analizar el movimiento ocular de usuarios para generar métricas y correlación con estados emocionales y cognitivos," Tesis de maestría, TecNM/Cenidet, 2020.

[62] J. A. Morales Morante, "Metodología para el preprocesamiento y clasificación de datos fisiológicos multimodales basado en el modelo Valencia-Excitación," Tesis de maestría, TecNM/Cenidet, 2021.

[63] B. Kitchenham y S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering (Technical Report)," Universidad de Durham, inf. téc., 2007.

[64] S. Lallé, C. Conati y G. Carenini, "Predicting Confusion in Information Visualization from Eye Tracking and Interaction Data," en *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ép. IJCAI'16, AAAI Press, 2016, págs. 2529-2535, ISBN: 9781577357704.

[65] J. Salminen, M. Nagpal, H. Kwak, J. An, S.-g. Jung y B. J. Jansen, "Confusion Prediction from Eye-Tracking Data: Experiments with Machine Learning," en *Proceedings of the 9th International Conference on Information Systems and Technologies (ICIST 2019)*, New York, New York, USA: ACM Press, 2019, págs. 1-9, ISBN: 9781450362924. DOI: 10.1145/3361570.3361577.

[66] N. V. Chawla, K. W. Bowyer, L. O. Hall y W. P. Kegelmeyer, "SMOTE: Synthetic minority oversampling technique," *Journal of Artificial Intelligence Research*, vol. 16, págs. 321-357, 2002, ISSN: 10769757. DOI: 10.1613/jair.953.

[67] A. Mathur, N. D. Lane y F. Kawsar, "Engagement-Aware Computing: Modelling User Engagement from Mobile Contexts," en *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, ép. UbiComp '16, New York, NY, USA: Association for Computing Machinery, 2016, págs. 622-633, ISBN: 9781450344616. DOI: 10.1145/2971648.2971760.

[68] M. Gjoreski, T. Kolenik, T. Knez *et al.*, "Datasets for Cognitive Load Inference Using Wearable Sensors and Psychological Traits," *Applied Sciences*, vol. 10, n.º 11, pág. 3843, 2020, ISSN: 2076-3417. DOI: 10.3390/app10113843.

[69] L. H. Du, W. Liu, W. L. Zheng y B. L. Lu, "Detecting driving fatigue with multimodal deep learning," en *International IEEE/EMBS Conference on Neural Engineering, NER*, 2017, págs. 74-77, ISBN: 9781538619162. DOI: 10.1109/NER.2017.8008295.

[70] F. Li, G. Zhang, W. Wang *et al.*, "Deep Models for Engagement Assessment with Scarce Label Information," *IEEE Transactions on Human-Machine Systems*, vol. 47, n.º 4, págs. 598-605, 2017, ISSN: 21682291. DOI: 10.1109/THMS.2016.2608933.

[71] A. Qayyum, I. Faye, A. S. Malik y M. Mazher, "Assessment of Cognitive Load using Multimedia Learning and Resting States with Deep Learning Perspective," en *2018 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, 2018, págs. 600-605. DOI: 10.1109/IECBES.2018.8626702.

[72] S. Siddharth, T.-P. Jung y T. J. Sejnowski, "Utilizing Deep Learning Towards Multi-modal Bio-sensing and Vision-based Affective Computing," *IEEE Transactions on Affective Computing*, págs. 1-1, 2019, ISSN: 2371-9850. DOI: 10.1109/taffc.2019.2916015.

[73] Y. LeCun, Y. Bengio y G. Hinton, "Deep learning," *Nature*, vol. 521, n.º 7553, págs. 436-444, 2015, ISSN: 1476-4687. DOI: 10.1038/nature14539.

[74] H. He, Y. Bai, E. A. Garcia y S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," en *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 2008, págs. 1322-1328. DOI: 10.1109/IJCNN.2008.4633969.

[75] I. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*, "Generative Adversarial Nets," en *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence y K. Q. Weinberger, eds., vol. 27, Curran Associates, Inc., 2014, págs. 2672-2680, ISBN: 9781510800410.

[76] K. Nikolaidis, S. Kristiansen, V. Goebel, T. Plagemann, K. Liestøl y M. Kankanhalli, "Augmenting Physiological Time Series Data: A Case Study for Sleep Apnea Detection," en *Machine Learning and Knowledge Discovery in Databases*, U. Brefeld, E. Fromont, A. Hotho, A. Knobbe, M. Maathuis y C. Robardet, eds., Cham: Springer International Publishing, 2020, págs. 376-399, ISBN: 978-3-030-46133-1.

[77] F. Fahimi, Z. Zhang, W. B. Goh, K. K. Ang y C. Guan, "Towards EEG Generation Using GANs for BCI Applications," en *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 2019, págs. 1-4. DOI: 10.1109/BHI.2019.8834503.

[78] T. Zhou, J. S. Cha, G. Gonzalez, J. P. Wachs, C. P. Sundaram y D. Yu, "Multimodal Physiological Signals for Workload Prediction in Robot-Assisted Surgery," *J. Hum.-Robot Interact.*, vol. 9, n.º 2, 2020. DOI: 10.1145/3368589.

[79] Z. Sun, B. Li, F. Duan *et al.*, "WLnet: Towards an Approach for Robust Workload Estimation Based on Shallow Neural Networks," *IEEE Access*, vol. 9, págs. 3165-3173, 2021. DOI: 10.1109/ACCESS.2020.3044732.

[80] A. Qayyum, M. K. A. A. Khan, M. Mazher y M. Suresh, "Classification of EEG Learning and Resting States using 1D-Convolutional Neural Network for Cognitive Load Assesment," en *2018 IEEE Student Conference on Research and Development (SCOReD)*, 2018, págs. 1-5. DOI: 10.1109/SCORED.2018.8711150.

[81] W. L. Lim, O. Sourina y L. P. Wang, "STEW: Simultaneous Task EEG Workload Data Set," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, n.º 11, págs. 2106-2114, 2018. DOI: 10.1109/TNSRE.2018.2872924.

[82] S. Han, J. Kim y S. Lee, "Recognition of Pilot's Cognitive States based on Combination of Physiological Signals," en *2019 7th International Winter Conference on Brain-Computer Interface (BCI)*, 2019, págs. 1-5. DOI: 10.1109/IWW-BCI.2019.8737317.

[83] D. Das Chakladar, S. Dey, P. P. Roy y D. P. Dogra, "EEG-based mental workload estimation using deep BLSTM-LSTM network and evolutionary algorithm," *Biomedical Signal Processing and Control*, vol. 60, pág. 101 989, 2020, ISSN: 1746-8094. DOI: 10.1016/j.bspc.2020.101989.

[84] S. Arndt, K. Brunnström, E. Cheng, U. Engelke, S. Möller y J.-N. Antons, "Review on using physiology in quality of experience," *Electronic Imaging*, vol. 2016, n.º 16, págs. 1-9, 2016. DOI: 10.2352/ISSN.2470-1173.2016.16.HVEI-125.

[85] N. Momeni, F. Dell'Agnola, A. Arza y D. Atienza, "Real-Time Cognitive Workload Monitoring Based on Machine Learning Using Physiological Signals in Rescue Missions," en *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2019, págs. 3779-3785, ISBN: 978-1-5386-1312-2. DOI: 10.1109/EMBC.2019.8857501.

[86] B. Mahesh, E. Prassler, T. Hassan y J. U. Garbas, "Requirements for a Reference Dataset for Multimodal Human Stress Detection," en *2019 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2019*, 2019, págs. 492-498, ISBN: 9781538691519. DOI: 10.1109/PERCOMW.2019.8730884.

[87] M. Shojaeizadeh, S. Djamasbi, R. C. Paffenroth y A. C. Trapp, "Detecting task demand via an eye tracking machine learning system," *Decision Support Systems*, vol. 116, págs. 91-101, 2019, ISSN: 01679236. DOI: 10.1016/j.dss.2018.10.012.

[88] M. Shojaeizadeh, S. Djamasbi, R. C. Paffenroth y A. C. Trapp, *Eye-tracking system for detection of cognitive load*, 2020.

[89] M. Erins, O. Minejeva, R. Kivlenieks y J. Lauznis, "Feasibility study of physiological parameter registration sensors for non-intrusive human fatigue detection system," en *Engineering for Rural Development*, vol. 18, Latvia University of Life Sciences y Technologies, 2019, págs. 827-832. DOI: 10.22616/ERDev2019.18.N363.

[90] A. Momin, S. Bhattacharya, S. Sanyal y P. Chakraborty, "Visual Attention, Mental Stress and Gender: A Study Using Physiological Signals," *IEEE Access*, vol. 8, págs. 165 973-165 988, 2020, ISSN: 2169-3536 VO - 8. DOI: 10.1109/ACCESS.2020.3022727.

[91] P. Schmidt, A. Reiss, R. Dürichen y K. V. Laerhoven, "Wearable-Based Affect Recognition—A Review," *Sensors*, vol. 19, n.º 19, 2019, ISSN: 1424-8220. DOI: 10.3390/s19194079.

[92] E. Aydemir, S. Dogan, M. Baygin *et al.*, *CGP17Pat: Automated Schizophrenia Detection Based on a Cyclic Group of Prime Order Patterns Using EEG Signals*, 2022. DOI: 10.3390/healthcare10040643.

[93] M. Aljalal, S. A. Aldosari, M. Molinas, K. AlSharabi y F. A. Alturki, "Detection of Parkinson's disease from EEG signals using discrete wavelet transform, different entropy measures, and machine learning techniques," *Scientific Reports*, vol. 12, n.º 1, pág. 22 547, 2022, ISSN: 2045-2322. DOI: 10.1038/s41598-022-26644-7.

[94] A. Apraiz Iriarte, G. Lasa Eele y M. Mazmela Etxabe, "Evaluating User Experience with Physiological Monitoring: A Systematic Literature Review," *DYNA New Technologies*, vol. 8, n.º 1, 2021. DOI: 10.6036/nt10072.

[95] J. Nielsen, *How Many Test Users in a Usability Study?* 2012. dirección: https://www.nngroup.com/articles/how-many-test-users/.

[96] V. Georges, F. Courtemanche, S. Sénécal, T. Baccino, M. Fredette y P. M. Léger, "UX heatmaps: Mapping user experience on visual interfaces," en *Conference on Human Factors in Computing Systems - Proceedings*, 2016, págs. 4850-4860, ISBN: 9781450333627. DOI: 10.1145/2858036.2858271.

[97] J. Chai, Y. Ge, Y. Liu *et al.*, "Application of frontal EEG asymmetry to user experience research," en *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8532 LNAI, 2014, págs. 234-243, ISBN: 9783319075143. DOI: 10.1007/978-3-319-07515-0_24.

[98] L. Yao, Y. Liu, W. Li *et al.*, "Using physiological measures to evaluate user experience of mobile applications," en *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8532 LNAI, 2014, págs. 301-310, ISBN: 9783319075143. DOI: 10.1007/978-3-319-07515-0_31.

[99] T. McMahan, I. Parberry y T. D. Parsons, "Modality specific assessment of video game player's experience using the Emotiv," *Entertainment Computing*, vol. 7, págs. 1-6, 2015, ISSN: 1875-9521. DOI: 10.1016/j.entcom.2015.03.001.

[100] T. McMahan, I. Parberry y T. D. Parsons, "Evaluating Player Task Engagement and Arousal Using Electroencephalography," *Procedia Manufacturing*, vol. 3, págs. 2303-2310, 2015, ISSN: 2351-9789. DOI: 10.1016/j.promfg.2015.07.376.

[101] S. Mirhoseini, P.-M. Leger y S. Senecal, "Investigating the Effect of Product Sorting and Users' Goal on Cognitive load," en *SIGHCI 2017 Proceedings*, 2017, pág. 3. dirección: https://aisel.aisnet.org/sighci2017/3.

[102] K. Tzafilkou y N. Protogeros, "Diagnosing user perception and acceptance using eye tracking in web-based end-user development," *Computers in Human Behavior*, vol. 72, págs. 23-37, 2017, ISSN: 07475632. DOI: 10.1016/j.chb.2017.02.035.

[103] C. Juanéda, S. Sénécal y P.-M. Léger, "Product Web Page Design: A Psychophysiological Investigation of the Influence of Product Similarity, Visual Proximity on Attention and Performance," en *HCIBGO 2018: HCI in Business, Government, and Organizations*, F. F.-H. Nah y B. S. Xiao, eds., Cham: Springer International Publishing, 2018, págs. 327-337, ISBN: 978-3-319-91716-0.

[104] C. Desrochers, P.-M. Léger, M. Fredette, S. Mirhoseini y S. Sénécal, "The arithmetic complexity of online grocery shopping: the moderating role of product pictures," *Industrial Management & Data Systems*, vol. 119, n.º 6, págs. 1206-1222, 2019, ISSN: 0263-5577. DOI: 10.1108/IMDS-04-2018-0151.

[105] S. Schneegass, B. Pfleging, N. Broy, F. Heinrich y A. Schmidt, "A Data Set of Real World Driving to Assess Driver Workload," en *Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, ép. AutomotiveUI '13, New York, NY, USA: Association for Computing Machinery, 2013, págs. 150-157, ISBN: 9781450324786. DOI: 10.1145/2516540.2516561.

[106] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx y W. Kraaij, "The SWELL Knowledge Work Dataset for Stress and User Modeling Research," en *Proceedings of the 16th International Conference on Multimodal Interaction*, ép. ICMI '14, New York, NY, USA: Association for Computing Machinery, 2014, págs. 291-298, ISBN: 9781450328852. DOI: 10.1145/2663204.2663257.

[107] C. Mühl, C. Jeunet y F. Lotte, "EEG-based workload estimation across affective contexts," *Frontiers in Neuroscience*, vol. 8, pág. 114, 2014, ISSN: 1662-453X. DOI: 10.3389/fnins.2014.00114.

[108] N. E. Haouij, J.-M. Poggi, S. Sevestre-Ghalila, R. Ghozi y M. Jaïdane, "AffectiveROAD System and Database to Assess Driver's Attention," en *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, ép. SAC '18, New York, NY, USA: Association for Computing Machinery, 2018, págs. 800-803, ISBN: 9781450351911. DOI: 10.1145/3167132.3167395.

[109] I. Zyma, S. Tukaev, I. Seleznov *et al.*, "Electroencephalograms during Mental Arithmetic Task Performance," *Data*, vol. 4, n.º 1, 2019. DOI: 10.3390/data4010014.

[110] M. Gjoreski, M. Luštrek y V. Pejović, "My Watch Says I'm Busy: Inferring Cognitive Load with Low-Cost Wearables," en *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, ép. UbiComp '18, New York, NY, USA: Association for Computing Machinery, 2018, págs. 1234-1240, ISBN: 9781450359665. DOI: 10.1145/3267305.3274113.

[111] M. Gjoreski, B. Mahesh, T. Kolenik *et al.*, "Cognitive Load Monitoring With Wearables–Lessons Learned From a Machine Learning Challenge," *IEEE Access*, vol. 9, págs. 103 325-103 336, 2021. DOI: 10.1109/ACCESS.2021.3093216.

[112] V. Borisov, E. Kasneci y G. Kasneci, "Robust cognitive load detection from wrist-band sensors," *Computers in Human Behavior Reports*, vol. 4, pág. 100 116, 2021, ISSN: 2451-9588. DOI: 10.1016/j.chbr.2021.100116.

[113] J. Tervonen, K. Pettersson y J. Mäntyjärvi, "Ultra-Short Window Length and Feature Importance Analysis for Cognitive Load Detection from Wearable Sensors," *Electronics*, vol. 10, n.º 5, 2021, ISSN: 2079-9292. DOI: 10.3390/electronics10050613.

[114] A. Salfinger, "Deep Learning for Cognitive Load Monitoring: A Comparative Evaluation," en *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, ép. UbiComp-ISWC '20, New York, NY, USA: Association for Computing Machinery, 2020, págs. 462-467, ISBN: 9781450380768. DOI: 10.1145/3410530.3414433.

[115] J. Gwizdka, "Email Task Management Styles: The Cleaners and the Keepers," en *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, ép. CHI EA '04, New York, NY, USA: Association for Computing Machinery, 2004, págs. 1235-1238, ISBN: 1581137036. DOI: 10.1145/985921.986032.

[116] L. Turpin, D. Kelly y J. Arguello, "To Blend or Not to Blend? Perceptual Speed, Visual Memory and Aggregated Search," en *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ép. SIGIR '16, New York, NY, USA: Association for Computing Machinery, 2016, págs. 1021-1024, ISBN: 9781450340694. DOI: 10.1145/2911451.2914739.

[117] S. Lallé, C. Conati y G. Carenini, "Impact of Individual Differences on User Experience with a Real-World Visualization Interface for Public Engagement," en *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, ép. UMAP '17, New York, NY, USA: Association for Computing Machinery, 2017, págs. 369-370, ISBN: 9781450346351. DOI: 10.1145/3079628.3079634.

[118] E. Haapalainen, S. Kim, J. F. Forlizzi y A. K. Dey, "Psycho-Physiological Measures for Assessing Cognitive Load," en *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, ép. UbiComp '10, New York, NY, USA: Association for Computing Machinery, 2010, págs. 301-310, ISBN: 9781605588438. DOI: 10.1145/1864349.1864395.

[119] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine Learning in {P}ython," *Journal of Machine Learning Research*, vol. 12, págs. 2825-2830, 2011.

[120] 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*, ép. KDD '16, New York, NY, USA: Association for Computing Machinery, 2016, págs. 785-794, ISBN: 9781450342322. DOI: 10.1145/2939672.2939785.

[121] S. Galli, "Performing Feature Scaling," en *Python Feature Engineering Cookbook*, Packt Publishing, 2020, cap. 8, págs. 241-259, ISBN: 978-1-78980-631-1.

[122] S. A. Hossein Aqajari, E. K. Naeini, M. A. Mehrabadi, S. Labbaf, N. Dutt y A. M. Rahmani, "pyEDA: An Open-Source Python Toolkit for Pre-processing and Feature Extraction of Electrodermal Activity," en *Procedia Computer Science*, vol. 184, 2021, págs. 99-106. DOI: 10.1016/j.procs.2021.03.021.

[123] M. Jas, D. A. Engemann, Y. Bekhti, F. Raimondo y A. Gramfort, "Autoreject: Automated artifact rejection for MEG and EEG data," *NeuroImage*, vol. 159, págs. 417-429, 2017, ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2017.06.030.

[124] A. Mognon, J. Jovicich, L. Bruzzone y M. Buiatti, "ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features," *Psychophysiology*, vol. 48, n.º 2, págs. 229-240, 2011, ISSN: 0048-5772. DOI: 10.1111/j.1469-8986.2010.01061.x.

[125] V. Lawhern, W. D. Hairston y K. Robbins, "DETECT: A MATLAB Toolbox for Event Detection and Identification in Time Series, with Applications to Artifact Detection in EEG Signals," *PLOS ONE*, vol. 8, n.º 4, e62944, 2013. DOI: 10.1371/journal.pone.0062944.

[126] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K.-M. Su y K. A. Robbins, "The PREP pipeline: standardized preprocessing for large-scale EEG analysis," *Frontiers in Neuroinformatics*, vol. 9, 2015, ISSN: 1662-5196. DOI: 10.3389/fninf.2015.00016.

[127] 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 Transactions on Audio and Electroacoustics*, vol. 15, n.º 2, págs. 70-73, 1967, ISSN: 1558-2582 VO - 15. DOI: 10.1109/TAU.1967.1161901.

[128] D. J. Thomson, "Spectrum estimation and harmonic analysis," *Proceedings of the IEEE*, vol. 70, n.º 9, págs. 1055-1096, 1982, ISSN: 1558-2256 VO - 70. DOI: 10.1109/PROC.1982.12433.

[129] M. J. Prerau, R. E. Brown, M. T. Bianchi, J. M. Ellenbogen y P. L. Purdon, "Sleep Neurophysiological Dynamics Through the Lens of Multitaper Spectral Analysis," *Physiology*, vol. 32, n.º 1, págs. 60-92, 2016, ISSN: 1548-9213. DOI: 10.1152/physiol.00062.2015.

[130] B. Babadi y E. N. Brown, "A Review of Multitaper Spectral Analysis," *IEEE Transactions on Biomedical Engineering*, vol. 61, n.º 5, págs. 1555-1564, 2014, ISSN: 1558-2531 VO - 61. DOI: 10.1109/TBME.2014.2311996.

[131] I. Winkler, S. Debener, K. Müller y M. Tangermann, "On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP," en *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2015, págs. 4101-4105. DOI: 10.1109/EMBC.2015.7319296.

[132] W. Skrandies, "Global field power and topographic similarity," *Brain Topography*, vol. 3, n.º 1, págs. 137-141, 1990, ISSN: 1573-6792. DOI: 10.1007/BF01128870.

[133] F. Courtemanche, A. Dufresne y É. L. LeMoyne, "Multiresolution Feature Extraction During Psychophysiological Inference: Addressing Signals Asynchronicity," en *Physiological Computing Systems*, H. P. da Silva, A. Holzinger, S. Fairclough y D. Majoe, eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, págs. 43-56, ISBN: 978-3-662-45686-6.

[134] P. Jain, *UX Critique — Google Docs*, 2019. dirección: https://uxplanet.org/ux-critique-google-docs-770c3b13b3a3.

[135] M. Khandelwal, *Usability Review: Google Docs*, 2019. dirección: https://uxplanet.org/usability-review-google-docs-e99a2ac93cdd.

[136] E. Bañuelos-Lozoya, G. González-Serna, N. González-Franco, O. Fragoso-Diaz y N. Castro-Sánchez, "A Systematic Review for Cognitive State-Based QoE/UX Evaluation," *Sensors*, vol. 21, n.º 10, 2021. DOI: 10.3390/s21103439.

[137] E. O. Bañuelos Lozoya, J. G. González Serna, N. González Franco *et al.*, *Metodología para complementar la evaluación de UX interpretando la carga de trabajo a partir de datos fisiológicos del usuario*, Cuernavaca, Morelos, México, 2023.

[138] E. O. Bañuelos Lozoya, J. G. González Serna y N. González Franco, *Experimento para capturar datos fisiológicos e interpretar estados cognitivos en evaluaciones UX*, Cuernavaca, Morelos, México, 2022.

[139] E. O. Bañuelos Lozoya, J. G. González Serna y N. González Franco, *Evaluación de conjuntos de datos para reconocimiento de carga cognitiva y estrés*, Cuernavaca, Morelos, México, 2021.

Lista de figuras

Lista de tablas

Siglario

ADASYN *Adaptive Synthetic*
ECG Electrocardiograma
EDA *Electrodermal Activity*
EEG Electroencefalografía
ET *Eye Tracking*
FFT *Fast Fourier transformation*
GAN *Generative Adversarial Networks*
GC *Gradient Boosting*
GFP *Global Field Power*
GSR *Galvanic Skin Response*
HR *Heart Rate*
HRV *Heart Rate Variability*
ICA *Independent Component Analysis*
kNN *k-Nearest Neighbor*
LDA *Linear Discriminant Analysis*
LOSO *Leave One Subject Out*
LSL *Lab Streaming Layer*
MLP *Multilayer Perceptron*
NASA-TLX *NASA Task Load Index*
PPG *Photoplethysmography*
PSD *Power Spectral Density*
RF *Random Forest*
SCR *Skin Conductance Response*
SLR *Systematic Literature Review*
SVM *Support Vector Machine*
UX *User Experience*
XDF *Extensible Data Format*

TECNOLÓGICO
NACIONAL DE MÉXICO

cenidet
Centro Nacional de Investigación
y Desarrollo Tecnológico