ICADE, Universidad Pontificia de Comillas

ANÁLISIS DE DATOS BIOMÉTRICOS RECOGIDOS POR DISPOSITIVOS WEARABLE: DESARROLLO DE UN DASHBOARD INTERACTIVO Y UN MODELO PREDICTIVO DE MONITORIZACIÓN FISIOLÓGICA

Autor: Samuel Corcoba Isorna
Clave: 202106366

Director: Carlos Miguel Vallez Fernández

MADRID | Abril 2026

Resumen

El envejecimiento de la población y la presión sobre los sistemas sanitarios están empujando la salud hacia modelos más continuos y preventivos. En ese contexto, los wearables recogen señales fisiológicas durante todo el día y generan un volumen masivo de datos biométricos. El problema es que, en bruto, esos datos no son directamente útiles porque son ruidosos, densos y difíciles de interpretar directamente. El reto está en convertir esas señales en información útil.

De ahí parte este TFG, planteado como un ejercicio de Business Analytics. El objetivo es diseñar un pipeline completo que procese datos biométricos brutos y los convierta en métricas, predicciones y visualizaciones comprensibles, cubriendo toda la cadena analítica de principio a fin. Como base experimental se utiliza el dataset WESAD, una referencia consolidada que recoge señales multimodales de 15 sujetos mediante un dispositivo de pecho y una pulsera.

El trabajo se estructura en tres bloques: un proceso ETL que limpia y transforma las señales brutas en variables fisiológicas por ventanas temporales, un modelo predictivo que compara tres algoritmos de machine learning para la clasificación de estados afectivos, y un dashboard interactivo desplegado en la nube que traslada el pipeline a una interfaz visual.

El modelo final, un Random Forest, alcanza un 87,9% de accuracy en la clasificación de los estados afectivos del experimento. La actividad electrodérmica se identifica como la señal más discriminante, al concentrar el 42,0% de la importancia del modelo. Además, se construye un Stress Index como indicador resumen interpretable por un usuario no técnico.

El trabajo se presenta como prototipo de la capa analítica de Kora, una startup desarrollada en paralelo como TFG de ADE. Se cierra reconociendo sus limitaciones, principalmente el tamaño reducido de la muestra, y plantea como siguiente paso su validación con datos propios recogidos en un entorno real.

Palabras clave: Business Analytics, pipeline de datos, ETL, wearables, machine learning, dashboard interactivo, Random Forest, WESAD.

Abstract

Population ageing and growing pressure on healthcare systems are pushing healthcare toward more continuous and preventive models. In this context, wearables collect physiological signals throughout the day and generate a massive volume of biometric data. The problem is that, in raw form, these data are not directly useful because they are noisy, dense, and difficult to interpret. The real challenge is in transforming these signals into useful information.

This is the starting point of this thesis, conceived as a Business Analytics project. Its aim is to design a complete pipeline that processes raw biometric data and converts them into understandable metrics, predictions, and visualizations, covering the entire analytical chain from end to end. As an experimental basis, the study uses the WESAD dataset, a well-established benchmark that contains multimodal signals from 15 subjects collected through a chest-worn device and a wristband.

The project is structured into three main blocks: an ETL process that cleans and transforms raw signals into physiological variables computed over time windows, a predictive model that compares three machine learning algorithms for affective state classification, and an interactive dashboard deployed in the cloud that translates the pipeline into a visual interface.

The final model, a Random Forest, achieves 87.9% accuracy in classifying the affective states of the experiment. Electrodermal activity is identified as the most discriminative signal, accounting for 42.0% of the model's total importance. In addition, a Stress Index is developed as a summary indicator that can be interpreted by a non-technical user.

This thesis is presented as a prototype of the analytical layer of Kora, a startup developed in parallel as a Business Administration bachelor's thesis. It concludes by acknowledging its main limitations, particularly the small sample size, and proposes as a next step its validation with proprietary data collected in a real-world setting.

Keywords: Business Analytics, data pipeline, ETL, wearables, machine learning, dashboard, Random Forest, WESAD.

Índice

Índice de ilustraciones

Índice de tablas

1. Introducción

1.1. Contexto general

El envejecimiento de la población está cambiando las sociedades europeas. El porcentaje de adultos mayores continúa creciendo cada vez más rápido. En España, las personas mayores de 65 años ya representan un 20,1% de la población total (9,69 millones en 2023), y proyecciones oficiales estiman que habrá 14,6 millones para 2041 (Pérez Díaz et al., 2024). Este patrón no es único en España, sino que se replica en varios países de Europa. En Portugal e Italia, la población +65 ya alcanza el 24% del total (Pérez Díaz et al., 2024).

A su vez, el envejecimiento incrementa la dependencia y necesidad de cuidados intensivos para estas personas, vivir más años viene acompañado de mayor aparición de enfermedades crónicas y pérdida de la capacidad cognitiva. Las pirámides poblacionales se invierten y esto da lugar a una mayor presión en los sistemas sanitarios y cuidadores.

La dimensión social de este problema es igual de relevante. Los modelos de convivencia están cambiando y cada vez son más las personas mayores (especialmente mujeres) que viven solas. En muchos casos, los familiares, condicionados principalmente por motivos laborales, se ven obligados a trasladarse a otras ciudades y a vivir lejos de sus mayores. Esto los expone a situaciones de vulnerabilidad menos visibles y difíciles de detectar o reaccionar. En España la situación ya es preocupante, más de 1,7 millones de personas mayores de 70 años viven solas (Pérez Díaz et al., 2024).

Las necesidades de cuidado son cada vez más elevadas. En España existen alrededor de 638.000 cuidadores profesionales que atienden a adultos dependientes de 70 años o más dentro del hogar y 920.000 fuera del hogar (Pérez Díaz et al., 2024).

Este contexto demográfico y social está empujando el cuidado de la salud hacia modelos más preventivos y continuos. Las atenciones puntuales y reactivas son difíciles de sostener en sociedades envejecidas. Ser capaz de detectar señales tempranas y monitorizar tendencias puede ser tan importante como actuar cuando surgen problemas. En este escenario, los wearables y tecnologías de monitorización de salud se convierten en un apoyo útil y relevante. Permiten apoyar decisiones médicas e identificar anomalías en etapas tempranas. Un mejor seguimiento implica decisiones más informadas por parte de

cuidadores, médicos y familiares, siempre que las señales brutas sean interpretables por estas personas.

1.2. Dispositivos *wearables* como fuente masiva de datos biométricos

Los dispositivos wearable son pequeños sensores que se pueden llevar pegados al cuerpo (normalmente en la muñeca o el pecho) para recoger señales fisiológicas de forma continua durante el día. No son invasivos en el día a día y permiten capturar datos para después procesarlos e interpretarlos. Son principalmente el origen del dato, pero sin limpieza y procesamiento no aportan valor.

Los wearables han evolucionado de simples medidores de actividad a dispositivos multi-sensor que permiten capturar señales diferentes al mismo tiempo. Los primeros modelos a la venta se centraban en medir pasos y movimiento. Hoy existen opciones que combinan sensores ópticos, eléctricos, de temperatura y de movimiento, capaces de recoger un rango mayor de reacciones del cuerpo. Dependiendo del dispositivo y donde se coloca se pueden capturar señales cardíacas, oxígeno en sangre, de conductancia de la piel (sudoración), de temperatura, patrones de respiración y movimiento físico entre otros.

Una ventaja clave de esta tecnología es que monitoriza continuamente en vez de hacer mediciones puntuales en momentos específicos. Los datos prolongados en el tiempo permiten observar tendencias, detectar desviaciones y entender cómo el cuerpo reacciona ante situaciones diferentes. Esto es especialmente relevante para fenómenos que cambian en el tiempo como el estrés, la fatiga o la recuperación, que trataremos de medir en este trabajo.

1.3. El reto: del dato bruto a la información útil

El hecho de que los wearables generen grandes volúmenes de datos biométricos no implica que se genere valor automáticamente. Las señales brutas suelen ser ruidosas y no son interpretables directamente por los usuarios. El movimiento, la colocación del sensor o la actividad que esté haciendo la persona puede introducir distorsiones.

Para aprovechar los datos al máximo debemos introducir un buen pipeline. Comienza con la limpieza para filtrar ruido, tratar datos perdidos o sincronizar sensores. Continúa con la transformación para generar variables comprensibles por el humano (por ejemplo, frecuencia cardíaca a partir de electrocardiograma) o agregar la información en ventanas.

Después de este proceso se puede resumir la información en indicadores clave e interpretables.

Aquí toma relevancia la visualización, que se utiliza para convertir datos complejos en diseños que muestran tendencias y anomalías o cambios a lo largo del tiempo. Además, los datos limpios y transformados pueden ser interpretados por modelos capaces de predecir situaciones que aportan valor.

Por este motivo, este proyecto va más allá y no se centra solo en la recogida de datos del dispositivo sino en el proceso analítico que convierte los datos recogidos en información que puede ayudar en la toma de decisiones.

1.4. Objetivos del trabajo

El objetivo principal de este trabajo es desarrollar un sistema de análisis y predicción de datos biométricos que pueda integrarse en un ecosistema capaz de transmitir información real y útil al usuario.

Los objetivos específicos son:

- Analizar la estructura y composición del dataset WESAD.
- Diseñar y ejecutar un proceso ETL para limpiar, transformar y preparar los datos biométricos.
- Unificar los datos en una base única, estandarizada y lista para su análisis.
- Desarrollar un dashboard de visualizaciones interactivo para explorar e interpretar las variables fisiológicas.
- Entrenar un modelo predictivo en Python capaz de identificar estados fisiológicos a partir de las señales procesadas.
- Evaluar el potencial del sistema desarrollado para su integración futura en el ecosistema de la startup Kora.

2. Enfoque del trabajo

Este trabajo está planteado principalmente como un ejercicio de Business Analytics. Su mayor aportación es diseñar y poner a prueba un proceso completo que transforma datos biométricos complejos en resultados consistentes, interpretables y útiles en monitorización. El objetivo final no es construir un producto terminado, por lo que el trabajo se desarrolla como un experimento.

El alcance se centra en la capa analítica: limpieza de datos, generación y transformación de variables, visualización y modelos predictivos. Temas como el diseño de hardware, la experiencia de usuario, el desarrollo clínico o cualquier interpretación médica quedan fuera del alcance.

El diseño experimental parte de un dataset público que se utiliza para probar ideas en un entorno controlado, con condiciones establecidas y conocidas. Sin embargo, el número de participantes y el tiempo de monitorización son limitados. Además, se recogen datos siguiendo un protocolo específico que no refleja la vida cotidiana de los individuos. En consecuencia, los resultados de este proyecto deben entenderse como una demostración de viabilidad técnica y capacidad de detectar patrones en los datos. El objetivo es desarrollar un proceso replicable y comparable, que en el futuro podrá extrapolarse a nuevos conjuntos de datos recogidos con sensores y procedimientos similares y, a partir de ellos, mejorar los modelos.

El resultado final será un prototipo funcional que conecta toda la cadena de los datos, desde que se recogen hasta que llegan como información al usuario. Se analizará el potencial de las señales biométricas para detectar patrones, tendencias, cambios de estado y generar métricas fáciles de interpretar.

El prototipo servirá como fundamento para una posible aplicación futura. No obstante, para poder convertirse en un producto real debería probarse a un nivel superior. Probarlo con datasets nuevos y de mayor tamaño, recogidos en condiciones reales y mediante diferentes dispositivos para poder conseguir la validación de profesionales del sector y confirmar su relevancia y confianza en la interpretación de los datos. Concluyendo, la contribución del trabajo será crear una base estructurada y extensible capaz de ser integrada en el futuro en sistemas más grandes y complejos una vez se valide.

3. Estado del Arte

3.1. De la medición puntual a la monitorización continua

Durante mucho tiempo, el diagnóstico médico se ha basado en mediciones puntuales hechas en consulta. El médico tomaba la tensión, escuchaba el corazón o pedía un análisis de sangre, y a partir de ese momento concreto sacaba conclusiones sobre el estado de salud del paciente. Es un enfoque que ha funcionado durante décadas, aunque tiene una limitación importante. El cuerpo no se comporta igual durante todo el día, la presión arterial sube y baja según lo que estés haciendo, el ritmo cardíaco cambia con el estrés o el ejercicio, y muchas alteraciones aparecen y desaparecen sin que un médico llegue a verlas en una visita de quince minutos.

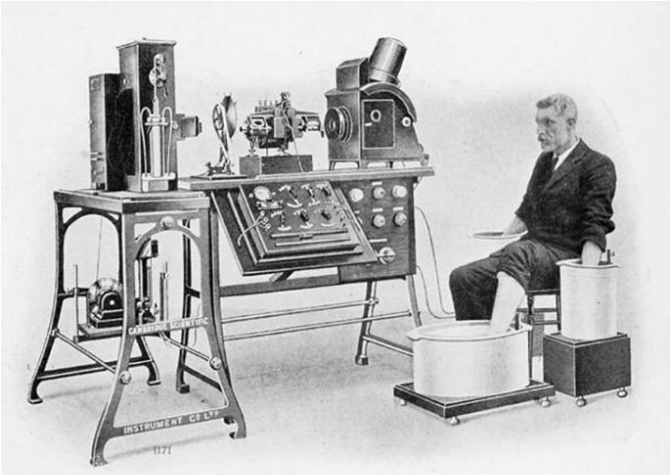
El cambio real empezó cuando aparecieron los primeros instrumentos capaces de recoger señales fisiológicas de forma sistemática. Uno de los pasos más importantes fue el electrocardiograma, que permitió por primera vez convertir la actividad eléctrica del corazón en ondas que podían interpretarse de manera consistente. Cuando Willem Einthoven desarrolló el primer sistema de ECG a principios del siglo XX, descubrió una forma totalmente nueva de entender la fisiología humana, basada en patrones medibles y no solo en síntomas observados.

Su trabajo, que le dio el Premio Nobel de Medicina en 1924, sentó las bases de toda la electrocardiografía moderna y supuso el primer ejemplo claro de cómo los datos biométricos podían analizarse buscando patrones objetivos. Gracias al ECG se pudieron empezar a detectar arritmias y problemas en la conducción eléctrica del corazón con una fiabilidad que antes era impensable (Vincent, 2022).

Con el ECG vino también un problema que sigue siendo relevante hoy. Medir una señal no sirve de nada si cada hospital lo hace de una forma distinta. Si los electrodos se colocan en sitios diferentes, si las escalas no coinciden, si las máquinas calibran de forma distinta, los resultados no son comparables entre pacientes ni entre centros. Por eso uno de los grandes avances posteriores fue más metodológico que tecnológico, la estandarización del proceso de medición. Establecer dónde van los electrodos, cómo se calibran los equipos y cómo se representan las señales fue lo que hizo posible que el ECG se convirtiese en una herramienta clínica de verdad (Vincent, 2022). Esta misma necesidad

de estandarización aparece hoy en los wearables: si dos dispositivos miden lo mismo de forma diferente, sus datos no se pueden comparar.

Ilustración 1. Primer modelo de ECG desarrollado por Einthoven.



PHOTOGRAPH OF A COMPLETE ELECTROCARDIOGRAPH, SHOWING THE MANNER IN WHICH THE ELECTRODES ARE ATTACHED TO THE PATIENT, IN THIS CASE THE HANDS AND ONE FOOT BEING IMMERSED IN JARS OF SALT SOLUTION

Fuente: Vincent (2022, Fig. 3).

Aun con un ECG estandarizado, seguía habiendo un problema. La medición se hacía en la consulta, en un momento concreto, y eso dejaba fuera todo lo que pasaba el resto del día. Para muchos problemas cardíacos esto era especialmente grave, porque hay alteraciones del ritmo que aparecen de forma esporádica y que es prácticamente imposible captar en una visita médica. La solución llegó de la mano de Norman Holter, que en los años cincuenta desarrolló el primer monitor portátil capaz de registrar la actividad cardíaca de forma continua durante 24 horas o más, mientras el paciente seguía con su vida normal (Vincent, 2022).

La idea detrás del monitor Holter era sorprendentemente moderna. Además de recoger datos durante periodos largos también los almacenaba para poder analizarlos a posteriori, para que el médico pudiese revisarlos con calma. Se convirtió en el primer ejemplo de un dispositivo que cubría toda la cadena: recogida, almacenamiento, procesamiento e interpretación posterior (Vincent, 2022). Con el tiempo los Holter pasaron de ser aparatos voluminosos a sensores compactos y digitales, mucho más cómodos para el paciente y con mayor capacidad de almacenamiento.

En paralelo aparecieron otras formas de monitorización ambulatoria. La MAPA (monitorización ambulatoria de la presión arterial) es un buen ejemplo. Hasta entonces,

la presión arterial se medía en consulta, y los médicos sabían que esa medición a menudo era engañosa. Algunos pacientes mostraban presiones altas solo por el nerviosismo de estar en el médico (lo que se conoce como síndrome de la bata blanca), y otros tenían picos de tensión durante el día que pasaban completamente desapercibidos. La MAPA permitió sacar el tensiómetro de la consulta y llevarlo al día a día del paciente, registrando la presión en distintos momentos y dando una imagen mucho más realista de su estado (Lee, 2024).

Todos estos avances mejoraron mucho las decisiones médicas, pero también trajeron nuevos problemas. Cuando mides en condiciones reales, fuera del entorno controlado de un hospital, aparece ruido, errores de medición e intervalos de tiempo en los que la señal se corta y no es válida. Estos problemas no son fallos del dispositivo, son consecuencias inevitables de medir en la vida real. Y son exactamente los mismos problemas a los que se enfrentan hoy los wearables modernos.

3.2. Aparición y evolución de los wearables

Durante años, llevar la monitorización continua al gran público fue prácticamente imposible por razones puramente técnicas. Los dispositivos eran grandes, pesados, gastaban mucha batería, almacenaban poco y resultaban incómodos de llevar durante largos periodos. Solo se utilizaban en hospitales o en estudios clínicos muy concretos, y no había forma de integrarlos en la rutina de una persona (Escobar-Linero et al., 2023). Lo que diferencia un wearable de un dispositivo ambulatorio clásico no es solo que sea portátil, sino que está pensado para usarse de forma continua, en cualquier contexto y sin supervisión médica (Escobar-Linero et al., 2023).

Lo que cambió el panorama fue una combinación de varios avances que ocurrieron casi al mismo tiempo. Los sensores se hicieron mucho más pequeños y baratos, las baterías mejoraron, los chips de procesamiento ganaron potencia, y la conectividad inalámbrica permitió enviar los datos al móvil sin necesidad de cables. Todo esto junto convirtió la idea del wearable en algo viable a gran escala. Por primera vez tener un dispositivo médico encima dejó de ser un incordio y pasó a ser cómodo y casi invisible (Escobar-Linero et al., 2023).

Los pioneros del mercado de consumo fueron empresas como Fitbit, que en 2009 lanzó uno de los primeros dispositivos pensados para el público general. Era un sensor sencillo,

centrado en contar pasos y monitorizar la actividad física diaria, pero tuvo un éxito enorme porque ofrecía algo que hasta entonces era muy difícil de tener, información personal sobre la salud, sin pasar por un médico. Poco después aparecieron otros competidores que trataban de hacerse un hueco en lo que empezaba a ser un mercado nuevo.

El gran salto vino en 2015, cuando Apple lanzó el primer Apple Watch. Hasta entonces los wearables eran sobre todo pulseras de actividad, pero el Apple Watch los convirtió en algo mucho más ambicioso: un reloj inteligente con sensores fisiológicos, conectividad permanente y un ecosistema de aplicaciones detrás (Apple, 2015). A partir de ese momento el sector creció a una velocidad enorme. Garmin y Polar se posicionaron con fuerza en el deporte de alto rendimiento y el entrenamiento profesional y aparecieron empresas nuevas como Whoop y Oura, que se centraron más en métricas de sueño, esfuerzo y carga fisiológica diaria que en el conteo simple de pasos.

Hoy en día las grandes marcas no compiten tanto por tener más sensores, porque a nivel de hardware todas integran prácticamente lo mismo (acelerómetro, giroscopio, sensor óptico de pulso, sensor de temperatura, a veces ECG y SpO2). La diferencia real está en la capa analítica, en cómo cada empresa convierte los datos brutos en información útil. Apple ha orientado su análisis hacia la detección clínica con validación regulatoria. El Watch procesa los datos del sensor cardíaco para identificar patrones de ritmos irregulares (fibrilación auricular) o detectar apneas del sueño y alerta al usuario en tiempo real (Apple, 2018, 2025). Whoop y Oura se han especializado en métricas de recuperación. Sintetizan los datos de sueño, frecuencia cardíaca y variabilidad en una puntuación diaria que refleja el estado del usuario, y han dado un paso más allá incorporando agentes de inteligencia artificial (WHOOP Coach y Oura Advisor, respectivamente) capaces de generar recomendaciones personalizadas sobre hábitos, entrenamiento y recuperación (WHOOP, 2026; Oura Health, 2025). Por otro lado, Fitbit, ahora propiedad de Google, ha incorporado puntuaciones de manejo del estrés basadas en la actividad electrodérmica y la variabilidad de la frecuencia cardíaca (Google, n.d.).

Sin embargo, hay que tener cuidado porque estos dispositivos siguen siendo productos de consumo. Sus métricas son estimaciones algorítmicas, no diagnósticos clínicos, y su

precisión depende mucho del contacto con la piel, del movimiento del usuario y de las características individuales de cada persona (Jamieson et al., 2025).

3.3. Sensores clave en el affective computing

El affective computing es el campo que estudia cómo las máquinas pueden detectar e interpretar estados emocionales humanos, hay tres tipos de sensores que aparecen prácticamente en todos los estudios sobre detección de estrés.

El primero es la fotopletismografía, más conocida como PPG. Es la técnica que usa casi cualquier reloj inteligente actual para medir el pulso. Funciona con un LED que proyecta luz sobre la piel y un fotodetector que mide cuánta luz vuelve, que cambia ligeramente con cada latido del corazón porque la cantidad de sangre en los vasos varía a lo largo del ciclo cardíaco. A partir de la onda PPG se pueden calcular tanto la frecuencia cardíaca como su variabilidad, dos métricas muy útiles para detectar activación fisiológica (Jamieson et al., 2025). El PPG es atractivo para la industria porque los componentes son pequeños, baratos y fáciles de integrar en cualquier dispositivo, pero como ya se comentaba antes su gran debilidad es la sensibilidad al movimiento (Jamieson et al., 2025).

El acelerómetro no es un sensor fisiológico en sí mismo, pero es importante porque permite saber qué está haciendo el usuario en cada momento. Si el ritmo cardíaco sube, el acelerómetro ayuda a distinguir si es porque la persona está corriendo o porque está pasando por una situación emocionalmente intensa. Sin esa información sería casi imposible interpretar correctamente las señales fisiológicas. Los modelos modernos de detección de estrés y emociones casi siempre incluyen el acelerómetro como variable de contexto, ya que permite filtrar los falsos positivos y ajustar las predicciones.

El tercer sensor clave es la actividad electrodérmica, normalmente abreviada como EDA. La EDA mide cambios en la conductancia de la piel, que están directamente relacionados con la activación del sistema nervioso simpático. Esto es importante porque el simpático es la rama del sistema nervioso autónomo que se activa en situaciones de estrés, miedo o excitación, y lo hace de forma involuntaria. Por eso la EDA es una de las señales más utilizadas en estudios de detección de estrés, ya que es muy difícil de manipular conscientemente (Posada-Quintero et al., 2018).

Medir bien la EDA tiene sus complicaciones. La señal varía mucho según dónde se coloque el sensor, y las zonas con mayor densidad de glándulas sudoríparas (como la palma de la mano o las yemas de los dedos) son las que dan mejores resultados. En los wearables comerciales esto supone un reto, porque el típico reloj de muñeca no está en una zona ideal para medir EDA con precisión. Aun así, dispositivos como la Empatica E4 o algunos modelos de Fitbit han incorporado sensores de EDA con buenos resultados (Schmidt et al., 2018; Jamieson et al., 2025).

3.4. Modelos predictivos en affective computing

El enfoque más habitual en affective computing parte de extraer un conjunto de variables descriptivas (features) a partir de cada señal fisiológica. Estas features incluyen estadísticos básicos como medias, máximos, mínimos y desviaciones, métricas de variabilidad como el RMSSD o el SDNN, y a veces también descriptores en el dominio frecuencial como la potencia espectral en distintas bandas. Una vez extraídas las features, se utilizan como entrada de un algoritmo de clasificación supervisada, que aprende a relacionar los patrones fisiológicos con etiquetas de estado afectivo previamente conocidas (Schmidt et al., 2018).

Los modelos más utilizados en este campo siguen siendo los clásicos del machine learning supervisado, sobre todo Random Forest, Support Vector Machines (SVM) y distintos métodos ensemble basados en boosting, especialmente XGBoost. Cada uno tiene sus ventajas. Random Forest ha sido uno de los más utilizados porque suele ofrecer una precisión alta, resiste bastante bien el overfitting y funciona de forma sólida incluso cuando los datos no están perfectamente equilibrados. SVM también aparece con frecuencia en este tipo de experimentos, aunque con un papel más secundario. En paralelo, algunos métodos ensemble se utilizan como alternativa competitiva, pero su rendimiento depende más del ajuste fino del modelo. En este tipo de problemas, la búsqueda de hiperparámetros es importante, porque puede afectar de forma significativa al resultado final y cambiar la comparación entre modelos (Pinge et al., 2024).

El propio dataset WESAD reportó accuracies en torno al 80% para clasificación de tres clases (Baseline, Stress y Amusement) usando Random Forest sobre las señales del dispositivo de pecho. Desde su publicación se ha convertido en una referencia común

para comparar nuevos modelos en este campo, y muchos trabajos posteriores han usado WESAD como benchmark para validar enfoques alternativos (Schmidt et al., 2018).

En los últimos años también han ganado terreno los enfoques basados en deep learning, especialmente las redes neuronales aplicadas directamente sobre las señales brutas o sobre representaciones intermedias. Estos modelos prometen detectar patrones más complejos, pero suelen beneficiarse de datasets más grandes y variados para rendir bien. Esto hace que las empresas con grandes bases de datos fisiológicos recogidos de sus usuarios tengan, en principio, mejores condiciones para aprovechar modelos de redes neuronales. Sin embargo, en los datasets típicos del affective computing, que en muchos casos siguen siendo reducidos, los modelos clásicos continúan siendo competitivos y a menudo resultan preferibles por su mayor interpretabilidad (Pinge et al., 2024; Faust et al., 2018).

Uno de los retos principales de cualquier modelo en este campo es la enorme variabilidad entre personas. La misma situación de estrés puede generar respuestas fisiológicas muy distintas en dos sujetos diferentes, dependiendo de su edad, su forma física, su nivel basal de ansiedad, el momento del día o incluso de su estado emocional previo. Esto hace que entrenar un modelo que generalice bien sea mucho más difícil de lo que parece a primera vista. Un modelo que funciona perfectamente con los datos de un sujeto puede fallar estrepitosamente con otro (Vos et al., 2023).

Un último aspecto a tener en cuenta es la interpretabilidad. En aplicaciones de salud, especialmente en contextos donde los datos pueden influir en decisiones sobre el bienestar de una persona, no basta con tener un modelo que acierte. Hace falta entender por qué acierta y qué variables está utilizando para tomar sus decisiones. Esto descarta de entrada muchos enfoques de deep learning, que funcionan como cajas negras, y favorece a modelos como Random Forest, que permiten extraer la importancia de cada variable y entender qué señales fisiológicas son las que mejor discriminan entre estados (Pinge et al., 2024).

4. Motivación

4.1. Motivación académica

La base fundamental del Business Analytics es la capacidad de transformar conjuntos de datos complejos en información útil y fácil de entender. Además de trabajar con los datos como tal, la clave está en comprender el recorrido completo, desde el momento en que el dato se recoge hasta que acaba apoyando una decisión real. La monitorización de datos fisiológicos puede ser muy relevante en este contexto.

Algunas señales como el ritmo cardíaco, la actividad electrodérmica o la respiración requieren procesos de limpieza, transformación y agregación antes de poder interpretarse. Este tipo de información se sitúa en un entorno idóneo para aplicar conceptos clave de Business Analytics como los procesos ETL, la visualización de datos y el desarrollo de modelos predictivos.

El verdadero reto e interés de este TFG está en convertir las señales fisiológicas brutas en visualizaciones y respuestas que aporten valor para el usuario final. Este trabajo surge, por tanto, como una oportunidad de aplicar los conocimientos adquiridos durante el grado a un problema real, exigente en lo analítico y con un claro potencial de aplicación práctica.

Trabajar con datos de salud supone una dificultad adicional. No son datos estructurados, sino señales que reflejan procesos fisiológicos complejos a lo largo del tiempo. Por ello, hay que ser cuidadoso en cada decisión analítica y justificarlas en todo momento.

4.2. Conexión con mi TFG de ADE y visión de integración futura

Este trabajo está ligado a mi TFG desarrollado para el grado de Administración y Dirección de Empresas. Ambos parten del mismo problema: el envejecimiento de la población, el aumento de situaciones de aislamiento y dependencia, y la limitada integración tecnológica en la vida diaria de muchas personas mayores.

Desde el TFG de ADE, esta problemática se trata desde una perspectiva de negocio. Se propondrá un ecosistema basado en un dispositivo wearable y una plataforma digital que aportará valor a familias y cuidadores. El proyecto Kora surge como una propuesta orientada a la prevención y a la tranquilidad. Para ello se entrenarán modelos con bases de datos específicas para personas mayores.

El presente TFG de Analytics sirve como base analítica de ese proyecto más amplio. El sistema de análisis y predicción que se desarrolla en este trabajo será un primer prototipo de la capa analítica que, en un futuro, podría integrarse en un producto real. El dashboard y el modelo predictivo serán la demostración de cómo los datos biométricos pueden estructurarse, analizarse y presentarse de forma útil.

El uso del dataset académico WESAD permite trabajar sin depender todavía de datos de usuarios reales, ya que estos no están disponibles en fases tempranas del desarrollo de proyectos de este tipo. Trabajando con un dataset consolidado y muy utilizado en el entorno científico, se podrá diseñar el sistema analítico y realizar una interpretación de resultados. Además, de esta forma, se evitan condiciones legales, éticas o de disponibilidad de datos.

Ambos TFGs se complementan mutuamente. Mientras que el trabajo de ADE se orienta a definir el problema, el mercado y la propuesta de valor, este TFG de Analytics se centra en demostrar la viabilidad técnica y analítica capaz de apoyar esa visión. La idea es construir un modelo base sólido, escalable y coherente sobre el que seguir trabajando en el futuro si llegamos a disponer de datos reales.

5. Metodología

5.1. Estructura del dataset original

5.1.1. Introducción al dataset WESAD

El dataset WESAD nace como experimento diseñado dentro del *affective computing*, un campo de investigación que busca mejorar el entendimiento de las máquinas a los humanos detectando su estado afectivo y adaptando su comportamiento en consecuencia (Schmidt et al., 2018). Antes de WESAD, los experimentos en esta materia eran de muestras pequeñas, en su mayoría cerrados y poco reutilizables, lo que hacía difícil comparar resultados entre trabajos y limitaba el desarrollo de modelos sobre una base común.

WESAD fue creado como un dataset público y académico que surge de una investigación sobre la detección automática de estrés y estados afectivos a partir de datos recogidos por dispositivos *wearable*. Fue desarrollado por investigadores académicos e industriales de la University of Siegen y Robert Bosch GmbH, y presentado en 2018 en la conferencia internacional ACM ICMI (referencia en la interacción multimodal y *affective computing*).

No es solo un conjunto de datos creado de forma ad hoc para un experimento aislado, sino para ser reutilizado, analizado y comparado dentro de la comunidad científica de este sector. Sirve como base común para el desarrollo, evaluación y *benchmarking* de algoritmos de detección de estrés (Schmidt et al., 2018).

WESAD es un dataset multimodal ya que recoge datos fisiológicos y movimiento de dos dispositivos *wearables* diferentes, uno en el pecho y otro en la muñeca, sincronizados durante el experimento. De esta forma se ofrece una visión más completa del estado del sujeto durante el experimento que el que aportan estudios basados en un solo sensor.

El dispositivo del pecho (*RespiBAN*) recoge señales de alta resolución mientras el de la muñeca (*Empatica E4*) recoge señales más parecidas a las de los wearables comerciales actuales. Aunque las señales de la muñeca tienen menor resolución que las del pecho, es una solución más realista y menos intrusiva para el usuario.

Otro elemento clave es la diferenciación entre varios estados afectivos. En lugar de limitarse a clasificar binariamente estrés y no estrés, WESAD incluye estados neutrales,

de estrés y de diversión. Así refleja mejor la complejidad real del comportamiento afectivo humano.

Finalmente, WESAD destaca por ser una base reproducible y de uso extendido en la investigación. El hecho de que sea público, accesible y al haber sido utilizado como referencia en otros trabajos, refuerza su validez y su potencial desde una perspectiva analítica (Schmidt et al., 2018).

5.1.2. Diseño del experimento

Participantes y perfil de la muestra

El estudio se realizó con participantes reclutados en el propio entorno de investigación y en concreto se buscó un perfil que pudiese seguir un protocolo relativamente exigente. Por eso, los autores indican que trabajaron con estudiantes de posgrado vinculados a su centro (Schmidt et al., 2018).

En total participaron 17 personas, pero dos registros tuvieron que descartarse por fallos en los sensores durante la experimentación, por tanto, el dataset final se construye con 15 sujetos válidos. La muestra final tiene una edad media de 27,5 años con una desviación estándar de 2,4 años. En total 12 hombres y 3 mujeres (Schmidt et al., 2018).

Para reducir factores de riesgo y evitar que ciertas condiciones médicas sesgaran las señales fisiológicas, el protocolo estableció los siguientes criterios de exclusión: embarazo, tabaquismo intenso, trastornos mentales y enfermedades crónicas o cardiovasculares (Schmidt et al., 2018). Esto refuerza el propósito del experimento, ya que permite que las diferencias entre las condiciones (estrés, neutral y diversión) se reflejen en las señales fisiológicas con la menor interferencia posible.

Dispositivos, sensores y colocación

Como se ha mencionado anteriormente, se utilizaron dos wearables ubicados en diferentes partes del cuerpo para recoger los datos y poder obtener una combinación de señales eficiente a la vez que se mantenía coherencia con los dispositivos más habituales en la práctica.

El *RespiBAN Professional* es el dispositivo torácico que se utilizó como fuente principal de señales de alta resolución. Utiliza sensores que registran:

- Electrocardiograma (ECG): mide el ritmo cardíaco y cómo varía en el tiempo.

- Actividad electrodérmica (EDA): mide cambios en la conductancia eléctrica de la piel, que dependen de las glándulas sudoríparas. Está directamente relacionada con la activación del sistema nervioso simpático (muy útil para medir estrés o excitación).
- Electromiograma (EMG): mide la tensión muscular (suele aumentar en situaciones de estrés).
- Temperatura de la piel (TEMP): en contextos de estrés suele haber descensos leves en la temperatura cutánea.
- Respiración: la frecuencia respiratoria y cambios en la regularidad del ritmo suelen alterarse en situaciones de estrés.
- Acelerómetro: mide movimiento y actividad física del sujeto en distintos ejes. Permite diferenciar cambios debidos a movimiento corporal de aquellos asociados al estrés u otras respuestas emocionales.

Todas las señales recogidas por este dispositivo se muestrean a 700 Hz lo que da una granularidad muy alta para analizar cambios rápidos y sutiles (Schmidt et al., 2018).

El *RespiBAN* se ajusta alrededor del pecho. La respiración se mide mediante un sensor inductivo. El ECG se registra con un montaje estándar de tres puntos. El EMG se recoge en el trapecio superior y la temperatura se coloca en el esternón. En el caso de la EDA, se registra en el abdomen (recto abdominal), una zona que los autores justifican como adecuada por su densidad de glándulas sudoríparas (Schmidt et al., 2018).

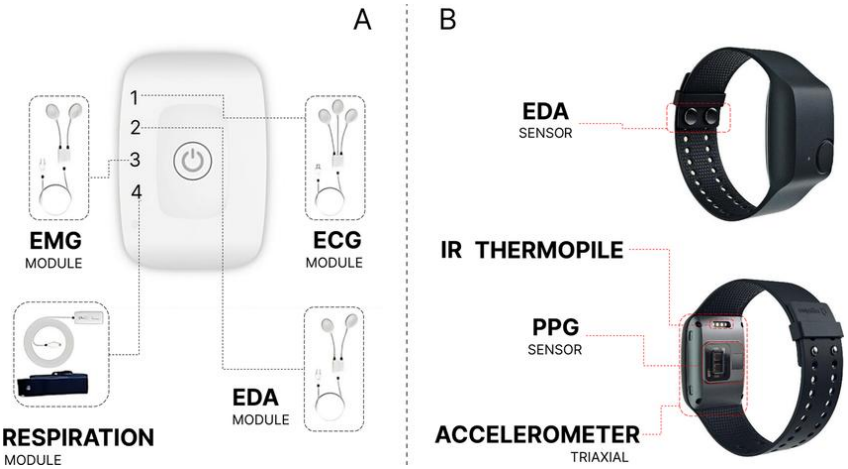
La *Empatica E4* es una pulsera que se coloca en la mano no dominante y registra señales más típicas de *wearables* comerciales. Al igual que *RespiBAN*, registra actividad electrodérmica (EDA), temperatura de la piel e integra acelerómetro. Además, aporta la medición de:

- Volumen del pulso sanguíneo (BVP): registra cambios en la cantidad de sangre que circula por los vasos sanguíneos. A partir de esta señal se puede estimar la frecuencia cardíaca y su variabilidad.

La dualidad de dispositivos es importante para el dataset porque permite recoger señales de alta calidad (pecho) y a la vez explorar qué se podría conseguir en escenarios más

realistas y comerciales (muñeca). La sincronización entre los dispositivos se realizó manualmente mediante un gesto de "doble toque" (Schmidt et al., 2018).

Ilustración 2. Dispositivos utilizados para la recogida de datos: RespiBAN Professional (A) y Empatica E4 (B)



Fuente: La Porta et al. (2025, Fig. 1).

Protocolo experimental y duración total

El estudio está diseñado para inducir tres estados afectivos principales: neutral, estrés y diversión. Además, se incluyen fases intermedias para reducir la activación fisiológica de cada sujeto después de condiciones intensas, con el objetivo de evitar que el "arrastre" de una condición contamine la siguiente (Schmidt et al., 2018).

Antes de empezar, se exigió a los participantes evitar cafeína y tabaco en la hora previa, además de evitar ejercicio físico intenso el día previo al estudio. A su llegada, firmaron el consentimiento, se colocaron los sensores y se realizó una prueba breve para comprobar que todo funcionaba correctamente (Schmidt et al., 2018).

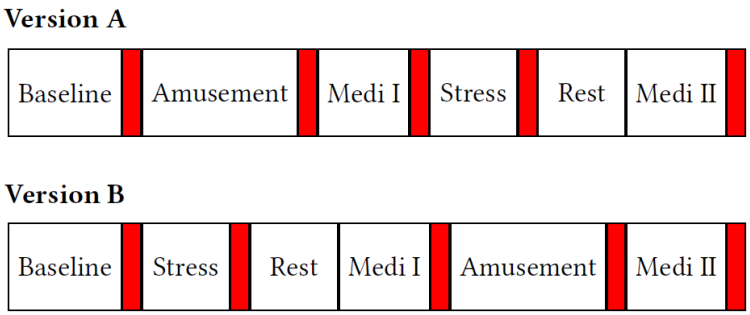
A partir de ahí, el protocolo se organiza en bloques (Schmidt et al., 2018):

- *Baseline* (estado neutral de referencia): se registra un período de 20 minutos. Durante esta fase, los participantes permanecieron sentados o de pie en una mesa y se les proporcionaron revistas para leer. El objetivo era capturar un estado lo más cercano posible a la neutralidad en un contexto controlado.
- *Amusement* (diversión): los sujetos observaron 11 vídeos graciosos con el objetivo de generar diversión. Entre clips se introdujeron secuencias neutrales breves de 5 segundos. La duración total de esta condición es de 392 segundos.

- *Stress* (estrés): se utilizó el *Trier Social Stress Test* (TSST), un protocolo común para inducir estrés en laboratorio generando una elevada carga mental en el sujeto. La versión WESAD incluye una tarea de exposición (hablar en público) y una tarea aritmética mental. Primero el participante prepara un discurso sobre sus fortalezas y debilidades durante 3 minutos y tiene 5 para presentarlo ante tres personas. A continuación, realiza una resta seriada (contar hacia atrás desde 2023 en pasos de 17) durante 5 minutos, reiniciando cuando comete un error. En total, la fase de TSST dura aproximadamente 10 minutos, seguida de un descanso de otros 10 minutos.
- *Meditation* (de-excitación): tras las pruebas de estrés y diversión se introdujo una meditación guiada basada en respiración controlada con los ojos cerrados y sentados en una postura cómoda. Dura 7 minutos y su objetivo es reducir la activación y devolver al participante a un estado más neutro.

El protocolo se aplicaría en dos versiones para los diferentes sujetos, alternando el orden de las pruebas de estrés y diversión. De esta forma se minimizarían los efectos de orden. Además, para introducir variabilidad en la postura, aproximadamente la mitad de los participantes realizó las pruebas sentados y la otra mitad de pie.

Ilustración 3. Versiones del protocolo WESAD.



Fuente: Schmidt et al. (2018, Fig. 2)

El estudio completo tendría una duración aproximada de dos horas por participante incluyendo preparación y cierre. Sin embargo, la recogida efectiva de datos por sujeto dura de media unos 95 minutos, de los que se etiquetan aproximadamente 36 minutos, el resto del tiempo forma parte del contexto experimental.

Etiquetado del estado afectivo

Para asociar cada segmento de los datos recogidos con un estado afectivo, el protocolo combina las condiciones definidas con autoinformes recogidos por los sujetos durante el estudio. Cada participante completó cinco autoinformes recogidos a lo largo del experimento: cuestionarios estandarizados *PANAS, STAI* y escalas *Self-Assessment Manikin* (Schmidt et al., 2018). Estos autoinformes están recogidos en el dataset y aportan una interpretación subjetiva extra que podría ser útil en el análisis posterior.

5.1.3. Estructura del ecosistema WESAD

El dataset WESAD no se presenta como un único archivo, sino como un ecosistema de datos completo y estructurado, diseñado para facilitar el análisis de señales fisiológicas recogidas durante el experimento. Sigue una estructura lógica, clara y coherente con el diseño experimental, que permite trabajar sobre él de forma ordenada.

Los datos siguen una organización jerárquica por sujetos. Cada sujeto tiene su propia carpeta con su número (S2, S3, …, SX) que contiene exclusivamente la información recogida sobre ese sujeto. Esta separación facilita el análisis individual y la comparación entre participantes (Ubiquitous Computing Lab, 2018).

Dentro de cada carpeta de participante se distinguen varios archivos y subcarpetas con funciones diferentes (Ubiquitous Computing Lab, 2018):

- *SX_E4_Data*: carpeta que contiene todos los datos recogidos por la pulsera *Empatica E4*. Dentro de ella se almacenan distintos *".csv"* correspondientes a cada tipo de señal registrada por el dispositivo. Además, se incluye otro *".csv"* con las etiquetas temporales y un fichero *"info.txt"* que documenta las unidades de medida y codificación utilizada en cada uno de los demás archivos. Esta separación facilita el acceso directo a los datos del wearable de la muñeca y permite trabajar con ellos de manera independiente si se requiere.
- *SX_readme.txt*: archivo de texto que aporta información contextual relevante sobre el perfil del participante. Se recogen datos personales básicos como edad, altura, peso, sexo y mano dominante, así como la confirmación de que cumple los prerrequisitos establecidos para participar en el estudio. Además, se incluyen notas cualitativas sobre situaciones concretas ocurridas durante el experimento.

Esta información resulta útil para contextualizar los registros fisiológicos según el perfil de cada sujeto.

- *SX_respiban.txt*: contiene la información completa de todas las variables recogidas por el dispositivo torácico *RespiBAN*, organizada en formato tabular.
- *SX_quest.csv*: recoge los resultados de los cuestionarios cubiertos por el participante a lo largo del experimento. En ellos se evalúa el estado afectivo y nivel de estrés en momentos concretos del estudio. Sirve como fuente adicional de información subjetiva que complementa los datos fisiológicos.
- *SX.pkl*: contenedor unificado del conjunto de datos. Archivo en formato *pickle*[1] de Python. Integra de forma estructurada toda la información recogida de las señales (tanto para *RespiBAN* como *Empatica E4*, sincronizadas), las etiquetas asociadas a las condiciones experimentales (estrés, diversión, neutral) y la información auxiliar necesaria para el análisis.

Se trabajará a partir del archivo *".pkl"* ya que permitirá cargar el dataset de forma directa en Python, contiene las señales asociadas a cada etiqueta y facilita su manipulación sin la necesidad de procesar manualmente múltiples archivos independientes para cada señal o sensor.

La estructura de WESAD está pensada para ser clara y reproducible. La organización por sujetos, separación por dispositivos y la coexistencia de archivos brutos, metadatos y contenedores estructurados proporcionan una base sólida para el análisis posterior y refuerzan el valor del dataset como referencia académica dentro del *affective computing* (Schmidt et al., 2018).

5.1.4. Variables recogidas (señales brutas)

El dataset WESAD recoge un conjunto amplio de variables fisiológicas y de contexto, registradas mediante dispositivos wearable, con el objetivo de capturar de forma multimodal la respuesta del organismo humano ante distintos estados afectivos. Estas variables abarcan señales cardiovasculares, electrodérmicas, respiratorias, musculares, térmicas y de movimiento, así como información de contexto (perfil del sujeto) e

[1] Un archivo *".pkl"* es un fichero serializado mediante el módulo *pickle* de Python, utilizado para almacenar y recuperar estructuras de datos complejas (diccionarios, *arrays* o tablas) conservando su organización interna.

información subjetiva procedente de cuestionarios. En conjunto, permiten analizar la respuesta fisiológica asociada al estrés y a otras emociones de forma integral.

La interpretación de estas variables debe realizarse teniendo en cuenta tanto su significado fisiológico como sus limitaciones debidas al uso de dispositivos portables. Por ello, en este apartado se describen en profundidad las principales variables recogidas en su estado bruto, explicando qué miden, en qué unidades se recogen. Estas señales corresponden directamente a magnitudes físicas o eléctricas medidas por los sensores, por lo que deben considerarse señales brutas que requieren procesamiento posterior para obtener variables de interés clínico.

Señales recogidas por *RespiBAN:*

El dispositivo torácico es la principal fuente de señales fisiológicas de alta resolución del dataset. Toda la información recogida por *RespiBAN* se registra a una frecuencia de muestreo de 700 Hz (Ubiquitous Computing Lab, 2018), siendo más preciso, lo que permite capturar variaciones rápidas y sutiles en la actividad fisiológica del sujeto.

- Electrocardiograma (ECG): recoge la actividad eléctrica del corazón medida en milivoltios (mV). Representa las variaciones de potencial eléctrico generadas por la despolarización y repolarización del corazón a lo largo del tiempo (Noble et al., 1990). Es una señal fundamental para el estudio de la actividad cardiovascular, contiene la información necesaria para identificar los latidos y analizar el ritmo cardíaco. En su forma bruta no aporta una interpretación clínica directa, pero de aquí derivan la frecuencia cardíaca y su variabilidad. Estas variables se calcularán posteriormente mediante algoritmos de detección de picos e intervalos entre latidos.
- Actividad electrodérmica (EDA): mide la conductancia de la piel, se registra en microsiemens (μS). Refleja la actividad de las glándulas sudoríparas, controladas directamente por el sistema nervioso simpático (Posada-Quintero et al., 2018). A diferencia de otras señales, la EDA ya ofrece información directamente interpretable. Incrementos en la conductancia de la piel suelen asociarse a estados de estrés, excitación o alerta.
- Electromiograma (EMG): mide la actividad eléctrica asociada a la contracción muscular medida en milivoltios (mV). Como se ha mencionado anteriormente,

este sensor se posiciona en el trapecio superior, una zona muy sensible a la acumulación de tensión en situaciones de estrés. En su forma bruta, la señal EMG oscila alrededor de cero y presenta alta variabilidad. Para obtener medidas más robustas de activación muscular es necesario aplicar técnicas de procesamiento como el cálculo de la envolvente o métricas de energía, que se abordarán en fases posteriores del análisis.

- Respiración: se obtiene mediante un sensor inductivo que mide los cambios en la expansión y contracción del tórax. Esta señal se registra como una magnitud continua en unidades arbitrarias propias del sensor y no representa directamente variables clínicas como la frecuencia respiratoria (Schmidt et al., 2018). La señal respiratoria bruta refleja el patrón respiratorio del sujeto, incluyendo la amplitud y la periodicidad de los ciclos respiratorios. Cambios en este patrón suelen estar asociados a estados de estrés o activación, aunque la frecuencia respiratoria debe calcularse posteriormente a partir de la señal procesada.
- Temperatura de la piel: se registra en grados Celsius (°C) y refleja cambios en la circulación periférica del sujeto. En condiciones normales, la temperatura cutánea puede variar lentamente en función del entorno y del estado fisiológico. En situaciones de estrés, la activación del sistema nervioso simpático puede provocar vasoconstricción periférica, lo que supone descensos leves de la temperatura de la piel (Herborn et al., 2015). Sin embargo, al tratarse de una señal recogida por sensores portables, pueden aparecer valores atípicos que no representan temperaturas fisiológicamente factibles. Por ello, esta señal requiere una fase de limpieza y validación previa a cualquier interpretación clínica.

Señales recogidas por *Empatica E4*

El dispositivo de la muñeca registra señales a una frecuencia de muestreo inferior a la del RespiBAN, pero ofrece información que se aproxima más a lo que es viable recoger con un wearable comercial. Las señales de este dispositivo tienen una menor resolución temporal y mayor susceptibilidad al movimiento.

- Volumen del pulso sanguíneo (BVP): se obtiene mediante un sensor óptico (PPG) y representa cambios relativos en el volumen de sangre en los vasos periféricos (Stuyck et al., 2022). Se expresa en unidades arbitrarias propias del sensor a una

frecuencia de 64 Hz. Al igual que ECG, el BVP bruto no proporciona valor clínico directamente. También sirve para estimar la frecuencia cardíaca y su variabilidad mediante procesamiento posterior.

- Actividad electrodérmica (EDA): también se registra en microsiemens (μS) pero a una frecuencia inferior (4 Hz). Aunque recoge la misma variable fisiológica que la EDA torácica, la ubicación del sensor puede hacer que difieran los valores absolutos.
- Temperatura de la piel: también se registra en grados Celsius (ºC), pero a una frecuencia inferior (4 Hz). Presenta en general una mayor estabilidad que la medida en el pecho.

Acelerometría

La acelerometría recoge la aceleración del cuerpo del sujeto en distintos ejes espaciales y se registra en el dataset en unidades de aceleración gravitatoria (g). Tanto el dispositivo del pecho como el de la muñeca integran acelerómetro. Aunque no constituye una variable afectiva en sí misma, es imprescindible como variable de contexto.

El registro del movimiento del sujeto permite identificar periodos de actividad física. Es clave para contextualizar los datos y evitar interpretaciones erróneas de incrementos fisiológicos que puedan deberse al movimiento y no a cambios en el estado afectivo.

Información de perfil del sujeto

Además de las señales fisiológicas, el dataset incluye información de contexto relativa a cada participante, almacenada en archivos de texto asociados a cada sujeto. Estos ficheros recogen variables de perfil básicas, como edad, sexo, altura, peso y mano dominante, así como la verificación de los prerrequisitos del estudio y observaciones cualitativas realizadas durante el experimento (Ubiquitous Computing Lab, 2018). Aunque estas variables no se utilizan directamente como señales fisiológicas, aportan contexto relevante para la interpretación de los datos y permiten caracterizar el perfil de la muestra.

Variables subjetivas (cuestionarios)

El dataset también incluye variables subjetivas procedentes de cuestionarios estandarizados. Estas variables recogen la percepción del propio sujeto sobre su estado afectivo y nivel de estrés en distintos momentos del experimento (Schmidt et al., 2018).

Aunque no constituyen una medición fisiológica directa, aportan una referencia adicional que permite contrastar e interpretar los datos objetivos.

En conjunto, las señales brutas recogidas en WESAD sirven como materia prima sobre la que se construirán las variables fisiológicas de interés clínico y analítico. Estas señales, registradas con diferentes frecuencias de muestreo y unidades según el sensor y el dispositivo, requieren un procesamiento posterior para extraer métricas interpretables y comparables.

La correcta comprensión de estas señales brutas es un paso imprescindible antes de abordar la fase de transformación y creación de variables derivadas, que se describe en el siguiente apartado.

5.1.5. Calidad y limitaciones del dataset

WESAD presenta diversas fortalezas que lo convierten en una base de datos sólida y adecuada para el análisis de señales fisiológicas asociadas al estrés u otros estados afectivos.

En primer lugar, los datos fueron recogidos en un entorno experimental controlado, lo que permite reducir el ruido externo y garantizar condiciones homogéneas entre participantes. Sin embargo, esto también puede ser una limitación desde ciertos puntos de vista, ya que al tratarse de un entorno de laboratorio no refleja completamente situaciones cotidianas de la vida real.

Otra fortaleza relevante es que el protocolo experimental está rigurosamente definido. Las etiquetas de estado afectivo se asignan con precisión en función de las condiciones inducidas durante el estudio y reforzadas por autoinformes de los propios participantes.

La información gana valor al ser recogida desde dos dispositivos diferentes, uno de alta resolución y precisión (*RespiBAN*) y otro con una aplicación práctica más realista (*Empatica E4*). Además, las señales de los dos dispositivos están sincronizadas temporalmente, esto facilita el análisis multimodal.

Por último, se trata de un dataset público, académico y reutilizado en el sector. Esto refuerza su validez como referencia y permite comparar resultados con trabajos previos y posteriores.

Sin embargo, el dataset presenta limitaciones a tener en cuenta. El tamaño muestral es reducido, el número de participantes es limitado y esto restringe la capacidad de generalizar los resultados. Por otro lado, la duración del experimento es relativamente corta y sigue un protocolo específico, esto impide monitorizar la variabilidad fisiológica a largo plazo. Estas limitaciones no invalidan el uso del dataset pero delimitan su alcance.

5.1.6. Adecuación del dataset al objetivo del trabajo

Este análisis en profundidad del dataset permite confirmar que su estructura se alinea con los objetivos del presente trabajo. Se hará un análisis exploratorio de señales biométricas recogidas por wearables partiendo de una de las bases más completas y consolidadas del affective computing.

Para el desarrollo del dashboard interactivo, WESAD permitirá explorar de forma visual la relación entre distintas señales fisiológicas (desde distintos dispositivos) y los estados afectivos registrados. Además, la disponibilidad de datos etiquetados y sincronizados proporciona una base sólida para construir un modelo predictivo orientado a monitorizar el estrés.

El uso de señales recogidas mediante wearables reales (especialmente *Empatica E4,* un dispositivo con sensores comunes en el mercado wearable) refuerza la relevancia del dataset para este trabajo. Aunque los datos proceden de un entorno experimental controlado, los tipos de sensores, y las señales que recogen, son comparables a los que se pretenden implementar.

Por último, recalcar la escalabilidad del dataset, ya que el código y los procesos definidos pueden aplicarse a conjuntos de datos de mayor tamaño o datos recogidos en entornos reales.

5.2. Herramientas empleadas

Para todo el desarrollo del trabajo se ha utilizado Python como lenguaje principal. Principalmente porque es uno de los lenguajes más utilizados hoy en día en el ámbito del análisis de datos y el machine learning. Además, tiene un ecosistema enorme de librerías especializadas que cubren prácticamente cualquier necesidad del proyecto, desde la lectura de archivos hasta el entrenamiento de modelos predictivos. Es un lenguaje sencillo que permite trabajar de forma muy ágil, algo importante en un proyecto que va construyendo cada fase sobre la anterior.

El trabajo se ha desarrollado en Jupyter Notebook, que permite escribir código y combinarlo con markdowns de texto explicativo y resultados visuales en un mismo documento. Es una herramienta especialmente útil para proyectos analíticos, donde lo importante no es solo que el código funcione sino también poder enseñar paso a paso cómo se ha llegado a cada resultado. Cada fase del trabajo (extracción, transformación y modelo predictivo) tiene su propio notebook, lo que facilita seguir la lógica del proceso completo.

Dentro del entorno de Python se han utilizado varias librerías clave, cada una con un papel concreto. Pandas y NumPy son la base de toda la manipulación de datos. Pandas se ha usado para cargar, filtrar, agrupar y exportar los datos en formato tabular (McKinney, 2010), mientras que NumPy se encarga de los cálculos numéricos sobre arrays cuando hace falta más rendimiento (Harris et al., 2020). Para el procesamiento específico de las señales fisiológicas se ha utilizado NeuroKit2, una librería especializada en biosignals que incluye algoritmos validados para detectar picos cardíacos, calcular variabilidad respiratoria o procesar la señal de actividad electrodérmica (Makowski et al., 2021). Usar una librería diseñada para esto, en lugar de implementar los algoritmos desde cero, aporta fiabilidad al proceso y ahorra mucho tiempo.

Para el modelo predictivo se ha empleado scikit-learn, que es la librería de referencia para machine learning en Python. Incluye implementaciones de los algoritmos más utilizados (Random Forest, Gradient Boosting, Regresión Logística, entre otros) y también herramientas para la búsqueda de hiperparámetros y la validación cruzada, que se han usado para entrenar y comparar los modelos del trabajo de forma rigurosa (Pedregosa et al., 2011).

En la parte visual, se han combinado Matplotlib y Seaborn para los gráficos exploratorios (boxplots, matrices de correlación, gráficos de importancia de variables) (Hunter, 2007; Waskom, 2021) y Plotly para los gráficos interactivos del dashboard (Plotly Technologies Inc., 2015). Plotly tiene la ventaja de que sus gráficos se pueden manipular directamente desde el navegador (hacer zoom, pasar el cursor sobre los datos, ocultar series), lo que da una experiencia mucho más rica al usuario.

Para el dashboard la elección fue Streamlit, una librería de Python que permite crear aplicaciones web interactivas escribiendo solo código Python, sin necesidad de

saber HTML, CSS ni JavaScript (Snowflake Inc., n.d.). La idea inicial era usar Power BI o Tableau, que son las herramientas más típicas para hacer dashboards en el mundo del Business Analytics, pero al investigar opciones se descubrió que Streamlit encajaba mejor con el resto del proyecto. La razón principal es que el dashboard se conecta directamente con los datos generados por el resto del código, sin necesidad de exportar a otro formato ni cambiar de entorno. Todo el flujo (extracción, transformación, modelo y visualización) queda dentro del mismo ecosistema Python, lo que facilita la coherencia del proyecto y permite defender cada decisión técnica sin saltar entre tecnologías. Otra ventaja importante es que Streamlit se puede desplegar de forma abierta en la nube a través de Streamlit Community Cloud, lo que permite compartir el dashboard con cualquier persona simplemente enviándole un enlace, sin que tenga que instalar nada.

Por último, todo el código del proyecto se ha alojado en GitHub, una plataforma de repositorios que permite mantener un historial ordenado del trabajo, hacer cambios sin perder versiones anteriores y compartir el repositorio de forma pública. Esto no solo facilita la entrega y revisión del proyecto, sino que también deja todo el código disponible para futuras consultas o ampliaciones.

6. Procesos ETL

6.1. Enfoque general

Una vez entendido el dataset WESAD y todo lo que recoge, el siguiente paso es preparar los datos para poder trabajar con ellos. Esta es la parte más larga y posiblemente más importante de todo el proyecto porque si los datos no están bien procesados, no se podrán construir buenos modelos ni visualizaciones a partir de ellos.

El proceso ETL (Extracción, Transformación y Carga) cubre todo lo que pasa entre los archivos originales del dataset y los datos limpios y listos para usar. La extracción consiste en cargar los datos brutos en Python y entender qué hay dentro. La transformación es donde se concentra la mayor parte del trabajo: se limpian las anomalías, se calculan las variables fisiológicas relevantes, se segmentan las señales en ventanas de tiempo y se construye el indicador compuesto de estrés. Por último, la carga genera los dos datasets finales en formato CSV que alimentarán el modelo predictivo y el dashboard. A lo largo de los siguientes apartados se explica cada fase del proceso con detalle, mostrando qué decisiones se han tomado en cada momento y por qué.

Para llevarlo a cabo se ha trabajado con un enfoque bottom-up. Se parte de las señales fisiológicas en su forma más bruta (los registros directos de los sensores) y a partir de ahí se van construyendo, paso a paso, variables más interpretables. En lugar de coger métricas ya calculadas y empaquetadas, se ha preferido procesar las señales desde el principio, lo que da más control sobre el resultado final y permite entender exactamente qué representa cada variable.

6.2. Extracción

6.2.1. Carga de los datos

Lo primero es cargar los archivos. Como ya se vio en la metodología, el dataset WESAD viene organizado en una carpeta por sujeto, y en este trabajo se utiliza el archivo SX.pkl de cada uno, que es un contenedor único que ya tiene todas las señales sincronizadas y etiquetadas.

Para abrir estos archivos en Python se ha usado la librería pickle, que está pensada justamente para leer este tipo de formatos. Junto a ella se han importado otras librerías básicas como pandas y numpy para el manejo de datos, y matplotlib para los gráficos exploratorios.

Una vez cargado el primer archivo, lo primero que se descubre es que cada .pkl es un diccionario jerárquico donde cada sujeto tiene tres bloques principales: signal, que guarda todas las señales fisiológicas separadas por dispositivo (pecho y muñeca); label, que contiene las etiquetas que indican en qué estado afectivo estaba el sujeto en cada momento (Baseline, Stress, Amusement, etc.); y subject, que es simplemente el identificador del participante (S2, S3, S10, ...).

6.2.2. Análisis del sujeto de ejemplo

Antes de aplicar nada al dataset completo, se analiza un solo sujeto (S10) para entender bien su estructura interna.

Dentro del bloque signal, los sensores del pecho (RespiBAN) son seis: ECG, EDA, EMG, respiración, temperatura y acelerómetro. Todos están muestreados a 700 Hz, lo que significa que registran 700 valores por segundo. Esto da arrays muy largos, en torno a 3,8 millones de muestras por señal y por sujeto. Los sensores de la muñeca (Empatica E4) vienen muestreados a frecuencias mucho más bajas (32 Hz, 64 Hz o 4 Hz dependiendo del sensor), porque están pensados para un dispositivo comercial donde la duración de la batería y el tamaño importan.

Tabla 1. Estructura de señales por dispositivo y sensor.

Dispositivo	Sensor	Tipo	Dimensiones
chest	ACC	ndarray	(3.847.200, 3)
chest	ECG	ndarray	(3.847.200, 1)
chest	EMG	ndarray	(3.847.200, 1)
chest	EDA	ndarray	(3.847.200, 1)
chest	Temp	ndarray	(3.847.200, 1)
chest	Resp	ndarray	(3.847.200, 1)
wrist	ACC	ndarray	(175.872, 3)
wrist	BVP	ndarray	(351.744, 1)
wrist	EDA	ndarray	(21.984, 1)
wrist	TEMP	ndarray	(21.984, 1)

Elaboración propia.

6.2.3. Análisis de las etiquetas

El bloque label es una larga lista de números, uno por cada muestra registrada, que indica en qué estado afectivo estaba el sujeto en cada momento. Los códigos van del 0 al 7, pero

en la documentación de WESAD solo cuatro de ellos tienen un significado claro: 1 corresponde a Baseline, 2 a Stress, 3 a Amusement y 4 a Meditation. Los demás (0, 5, 6 y 7) corresponden a fases de transición o periodos no etiquetados como ningún estado afectivo concreto.

Para este trabajo, los estados que realmente interesan son los tres principales (Baseline, Stress y Amusement), porque son los que el modelo predictivo intenta clasificar. La fase de meditación se mantiene en el dashboard como contexto, pero no entra en el modelo. El resto de etiquetas se descartan más adelante en la fase de transformación.

Tabla 2. Distribución de estados afectivos del sujeto S10.

Código	Estado	Nº muestras	Duración (min)	Porcentaje (%)
0	Not defined	1.589.000	37,83	41,30
1	Baseline	826.000	19,67	21,47
2	Stress	507.500	12,08	13,19
3	Amusement	260.400	6,20	6,77
4	Meditation	557.200	13,27	14,48
5	Not defined	35.700	0,85	0,93
6	Not defined	31.500	0,75	0,82
7	Not defined	39.900	0,95	1,04

Elaboración propia.

Al calcular cuánto tiempo del experimento está realmente etiquetado con alguno de los tres estados válidos, se ve que en S10 representa solo un 38% del total. El resto del tiempo corresponde a las fases de transición entre condiciones, la meditación y los periodos de preparación, que no aportan información útil para la clasificación pero sí ocupan tiempo dentro del experimento.

6.2.4. Análisis global de todos los sujetos

Una vez entendida la estructura interna de un sujeto, se aplica el mismo análisis al dataset completo para tener una visión general. Esto se ha hecho recorriendo automáticamente todos los archivos .pkl disponibles y calculando para cada sujeto su duración total, la duración de cada estado afectivo y otras métricas básicas.

Tabla 3. Resumen de disponibilidad por sujeto.

Sujeto	Duración total (min)	Baseline (min)	Stress (min)	Amusement (min)	Estados válidos (min)	% válido
S2	101,32	19,07	10,25	6,03	35,35	34,89
S3	108,22	19,00	10,67	6,25	35,92	33,19
S4	107,05	19,30	10,58	6,20	36,08	33,71
S5	104,30	19,97	10,75	6,23	36,95	35,43
S6	117,85	19,67	10,83	6,20	36,70	31,14
S7	87,30	19,77	10,67	6,20	36,63	41,96
S8	91,10	19,48	11,17	6,17	36,82	40,41
S9	87,05	19,67	10,75	6,20	36,62	42,06
S10	91,60	19,67	12,08	6,20	37,95	41,43
S11	87,22	19,67	11,33	6,13	37,13	42,58
S13	92,28	19,67	11,07	6,37	37,10	40,20
S14	92,47	19,67	11,25	6,20	37,12	40,14
S15	87,53	19,58	11,43	6,20	37,22	42,52
S16	93,85	19,67	11,22	6,13	37,02	39,44
S17	98,67	19,68	12,05	6,20	37,93	38,45

Elaboración propia.

Los resultados confirman que el dataset está formado por 15 sujetos (S2, S3, ..., S17, sin S1 ni S12 porque sus registros se descartaron por problemas en los sensores). La duración media del experimento por sujeto es de unos 96 minutos en total, con un rango que va de los 87 a los 118 minutos. De ese tiempo, una media de 36,8 minutos corresponde a estados etiquetados como válidos (alrededor del 38%), y el resto son transiciones, meditación y momentos no clasificados.

Tabla 4. Duración de los estados (estadísticas globales, en minutos).

	Duración total	Baseline	Stress	Amusement	Estados válidos
Media	96,52	19,57	11,07	6,19	36,84
Std	9,37	0,26	0,52	0,07	0,69
Mín	87,05	19,00	10,25	6,03	35,35
Máx	117,85	19,97	12,08	6,37	37,95

Elaboración propia.

La distribución por estado es consistente entre sujetos. Baseline ocupa de media unos 19,6 minutos, Stress unos 11,1 minutos y Amusement alrededor de 6,2 minutos. Esto significa que el dataset está desbalanceado: hay aproximadamente el doble de tiempo etiquetado como Baseline que como Stress, y casi tres veces más que como Amusement. Es un

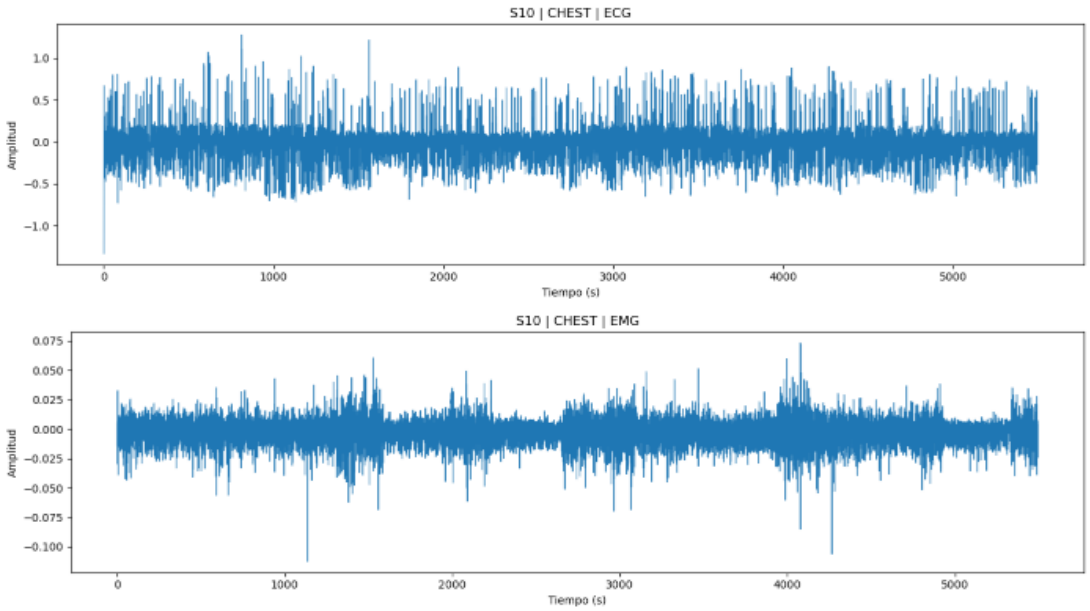
detalle importante, porque condicionará algunas decisiones a la hora de entrenar el modelo predictivo más adelante.

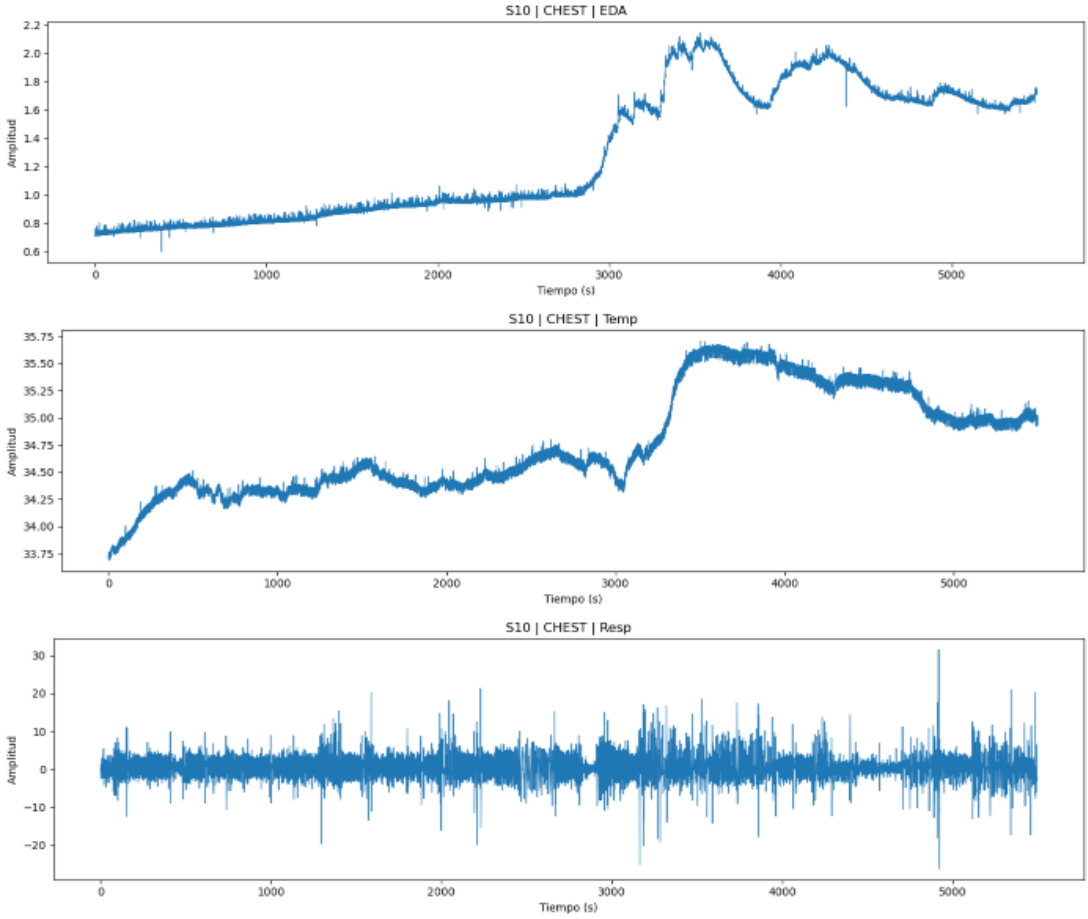
6.2.5. Visualización de las señales brutas

Para terminar de entender qué tipo de información contienen las señales, se han representado gráficamente todas las variables del sujeto S10. Esto permite ver de forma intuitiva cómo se comporta cada señal a lo largo del experimento, qué amplitudes manejan, si presentan ruido y si se aprecian cambios visibles entre las distintas fases.

Antes de los gráficos hay un detalle técnico que conviene explicar. Las señales del pecho tienen 3,8 millones de puntos cada una, así que dibujarlas todas sería inviable: el gráfico tardaría una eternidad y se vería como un bloque negro completamente ilegible. Para evitar esto se ha aplicado un downsampling, que consiste en reducir el número de puntos representados quedándose con uno de cada N (200 para señales de alta frecuencia, 10 para frecuencias medias y 4 para las más bajas). Esto no modifica los datos reales, solo afecta a la visualización, los datos originales siguen intactos y se utilizan tal cual en todas las fases posteriores.

Ilustración 4. Gráficos de las señales brutas recogidas por RespiBAN para el Sujeto 10.

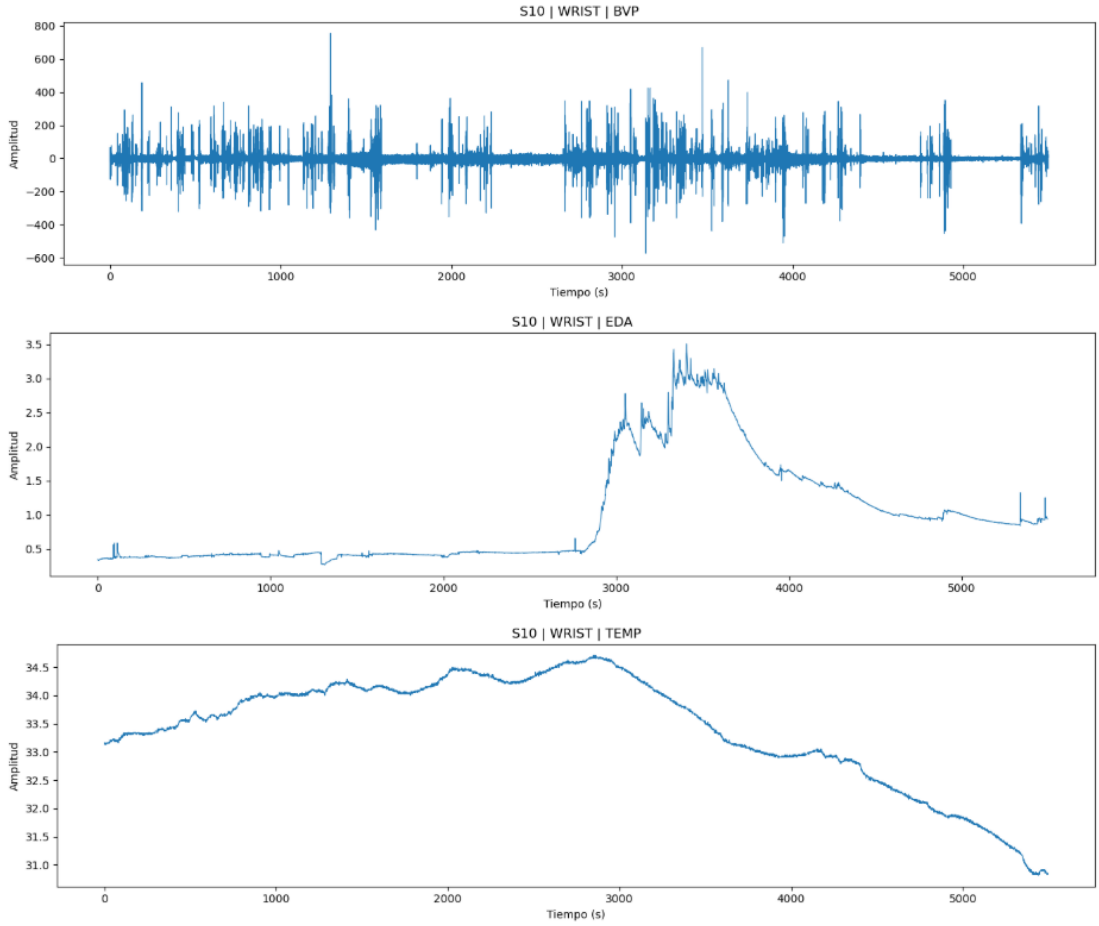




Elaboración propia.

En las señales del pecho se observan patrones coherentes con lo esperado. El ECG es una señal oscilatoria centrada en cero, con mucho ruido de alta frecuencia, tal y como se espera de una señal cardíaca sin filtrar. El EMG tiene una amplitud muy baja con picos esporádicos que probablemente correspondan a momentos de mayor tensión muscular. La EDA muestra una tendencia ascendente progresiva a lo largo del experimento, lo que es coherente con un aumento de la activación simpática durante las fases de estrés. La temperatura sube de forma suave a lo largo del registro, pasando de unos 33,7 °C a 35,7 °C, algo normal según se va calentando la zona de contacto del sensor. Y la respiración tiene un patrón oscilatorio claro que refleja los ciclos respiratorios del sujeto.

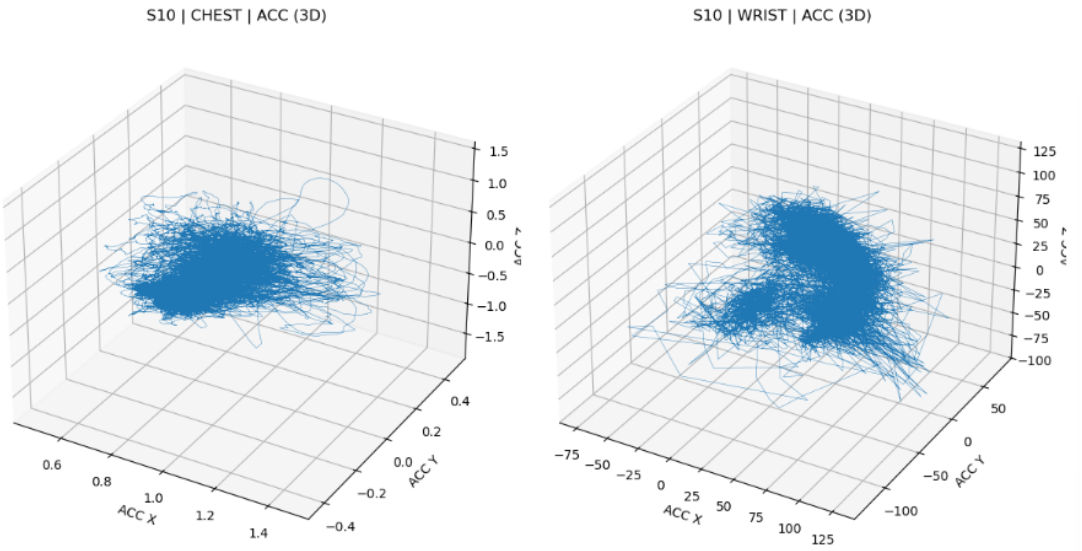
Ilustración 5. Gráficos de las señales brutas recogidas por Empatica E4 para el Sujeto 10.



Elaboración propia.

Las señales de la muñeca tienen un comportamiento parecido pero con algunas diferencias. El BVP es muy oscilatorio y con una amplitud bastante variable, lo cual es esperable porque el sensor óptico es muy sensible al movimiento del brazo. La EDA de muñeca muestra una fase muy estable durante los primeros minutos seguida de un cambio brusco, probablemente coincidiendo con alguna de las fases de estrés. La temperatura de muñeca evoluciona de forma similar a la del pecho pero más lentamente.

Ilustración 6. Gráficos 3D del acelerómetro de RespiBAN y Empatica E4 para el Sujeto 10.



Elaboración propia.

Para los acelerómetros se ha utilizado una representación en 3D, ya que cada sensor mide en tres ejes simultáneamente (X, Y, Z). El acelerómetro del pecho muestra una nube de puntos muy concentrada, lo que indica que el dispositivo apenas se mueve durante el experimento. Esto tiene sentido porque va sujeto al tórax con una banda y queda fijo. En cambio, el acelerómetro de la muñeca muestra una nube de puntos mucho más dispersa, con una variabilidad mucho mayor en los tres ejes. Esto es coherente con el hecho de que el brazo se mueve constantemente durante las distintas tareas del experimento (hablar en público, hacer cálculos mentales, ver vídeos, etc.).

6.2.6. Estadísticas descriptivas de las señales

Por último, se han calculado las estadísticas básicas (mínimo, máximo, rango, media y desviación típica) para cada señal y cada sujeto. Estas métricas son útiles para validar que los datos están dentro de rangos esperables y para detectar posibles anomalías que requieran tratamiento posterior.

Se han generado dos tablas. La primera muestra las estadísticas por sujeto, calculadas individualmente para cada combinación de sensor y participante. La segunda calcula las mismas métricas pero a nivel global, concatenando todos los valores de todos los sujetos. Esta segunda tabla es más útil para entender el comportamiento general de cada señal en el dataset completo, porque ofrece una visión agregada y permite ver, por ejemplo, en qué rangos se mueve la temperatura de la piel a lo largo de todos los participantes.

Tabla 5. Estadísticas descriptivas de las señales brutas del sujeto S10.

Dispositivo	Sensor	Hz	Mín	Máx	Rango	Media	Std	Muestras
chest	ECG	700	-1,460	1,479	2,939	0,001	0,150	3.847.200
chest	EDA	700	0,416	2,258	1,842	1,282	0,455	3.847.200
chest	EMG	700	-0,159	0,142	0,302	-0,003	0,010	3.847.200
chest	Resp	700	-26,930	31,896	58,826	0,051	3,451	3.847.200
chest	Temp	700	33,659	35,778	2,119	34,774	0,480	3.847.200
wrist	BVP	64	-609,230	779,690	1.388,920	0,000	46,252	351.744
wrist	EDA	4	0,266	3,508	3,241	0,982	0,767	21.984
wrist	TEMP	4	30,810	34,710	3,900	33,430	0,972	21.984

Elaboración propia.

Tabla 6. Estadísticas descriptivas globales de las señales brutas (todos los sujetos).

Dispositivo	Sensor	Hz	Mín	Máx	Rango	Media	Std	Muestras
chest	ECG	700	-1,500	1,500	3,000	0,001	0,269	60.807.600
chest	EDA	700	0,000	22,411	22,411	4,888	3,531	60.807.600
chest	EMG	700	-1,500	1,464	2,964	-0,003	0,018	60.807.600
chest	Resp	700	-50,000	38,800	88,800	0,054	4,099	60.807.600
chest	Temp	700	-273,150	35,778	308,928	33,905	1,217	60.807.600
wrist	BVP	64	-1.773,760	1.789,000	3.562,760	0,002	65,166	5.559.552
wrist	EDA	4	0,045	15,922	15,876	1,802	2,343	347.472
wrist	TEMP	4	28,970	35,970	7,000	32,598	1,498	347.472

Elaboración propia.

Los rangos observados son coherentes con lo esperable a nivel fisiológico. El ECG se mueve en torno a ±1,5 mV, la EDA en niveles que van de 0 a unos 20 µS, la temperatura cutánea entre 28 y 36 °C, y los acelerómetros tienen rangos muy distintos según el dispositivo. No se detectan valores claramente anómalos a nivel global, salvo algún pico aislado en la temperatura de un sujeto concreto, que se trata específicamente en la fase de transformación.

6.2.7. Exportación de los resultados

Para cerrar la fase de extracción, se han exportado tres archivos CSV con los resultados del análisis exploratorio: el resumen del dataset completo (dataset_summary.csv), las estadísticas por sujeto (subject_level_stats.csv) y las estadísticas globales (global_signal_stats.csv). Estos archivos no son inputs del modelo ni del dashboard, pero sirven como referencia documentada del estado inicial de los datos antes de empezar a transformarlos. Tener esta información guardada también permite consultarla en cualquier momento sin tener que volver a recorrer todos los archivos .pkl desde cero.

Con todo esto, se ha completado la fase de extracción. Llegados a este punto se sabe exactamente qué hay en el dataset, cómo está organizado, en qué unidades vienen las señales, cuánto dura cada experimento y cómo están distribuidos los estados afectivos. Esa información es la que da la base para abordar la fase de transformación, donde sí se empieza a manipular y procesar los datos para construir las variables fisiológicas que se utilizarán en el modelo y en el dashboard.

6.3. Transformación

La transformación es la parte más larga y la más importante de todo el proceso ETL. Es donde realmente se construye el valor del trabajo, porque las señales brutas que vienen del dataset no se pueden usar directamente ni para entrenar un modelo ni para mostrar nada útil en un dashboard. Lo que hay que hacer es limpiarlas primero y después convertirlas en variables que tengan un significado claro a nivel fisiológico.

A grandes rasgos, esta fase tiene tres bloques. Primero se limpia el dataset corrigiendo cualquier anomalía que se haya detectado en los datos. Después se calculan las variables derivadas a partir de las señales brutas, agrupando la información en ventanas de tiempo. Y por último se hace una inspección global del resultado para asegurarse de que todo está bien antes de pasar a la fase de carga.

Antes de entrar en el detalle, conviene aclarar dos decisiones de partida que afectan a todo lo que viene después.

6.3.1. Decisiones de partida

¿Por qué no se usa el acelerómetro?

En el dataset WESAD hay señales de movimiento recogidas tanto en el pecho como en la muñeca, pero en este trabajo no se utilizan. El motivo es que el experimento se realizó en un entorno controlado donde el movimiento no varía libremente, sino que está marcado por el propio protocolo (estar sentado leyendo, hablar de pie, ver vídeos, etc.). Aunque en producción el acelerómetro sí sería imprescindible, en este entorno controlado no aporta información directa sobre la activación fisiológica del estrés, sino más bien sobre la postura y la actividad física, que ya están condicionadas por el diseño del estudio. Incluirlo añadiría ruido sin aportar valor real al análisis.

¿Por qué solo se utilizan las señales del pecho?

Para este trabajo se han descartado las señales de muñeca y se han usado solo las del pecho. La razón principal es la calidad de la señal. El ECG torácico permite detectar los latidos con mucha más precisión que el sensor óptico de muñeca, que mide el volumen sanguíneo (BVP) y es muy sensible al movimiento y al ruido. Esto permite obtener estimaciones mucho más fiables de la frecuencia cardíaca y de su variabilidad. Además, hay variables fundamentales para este análisis (como la respiración o la EMG) que solo están disponibles en el dispositivo del pecho. Trabajar con una sola fuente de datos también simplifica el proceso y evita mezclar señales con niveles de precisión muy distintos, lo que podría introducir sesgos en los resultados.

Se trabaja con datos de RespiBAN porque es la mejor fuente disponible en WESAD. Sin embargo, los wearables de alta gama actuales ya integran ECG, sensores de EDA y estimaciones de variabilidad cardíaca con una fiabilidad que antes solo se conseguía con un electrodo torácico. Todo lo que se construye en este trabajo se podría trasladar a un wearable de muñeca moderno sin cambiar la lógica.

¿Por qué se usan ventanas de 60 segundos con paso de 1 segundo?

Las señales fisiológicas no se analizan punto por punto, sino por tramos. En este caso se trabaja con ventanas de 60 segundos que van avanzando de segundo en segundo.

La duración de 60 segundos es lo bastante larga como para que dé tiempo a recoger suficientes latidos o ciclos respiratorios para calcular bien la variabilidad. Si es demasiado larga, se pierde la capacidad de detectar cambios en el estado del sujeto.

El paso de 1 segundo, mucho menor que la ventana, hace que las ventanas se solapen entre sí y permite obtener una serie temporal continua, casi como si fuese instantánea. Por ejemplo, una persona que ha estado dos minutos conectada al sensor. En lugar de partir el registro en dos bloques fijos (del segundo 0 al 60 y del 60 al 120), lo que se hace es ir deslizando la ventana segundo a segundo: del 0 al 60, del 1 al 61, del 2 al 62, y así sucesivamente. El resultado es un registro por cada segundo del experimento, con cada valor resumiendo lo que ha pasado en el minuto anterior. Esto beneficia tanto al modelo, que puede seguir los cambios con mucha resolución, como al dashboard, donde todo se visualiza de forma mucho más natural.

6.3.2. Limpieza de datos

Antes de calcular nada, hay que asegurarse de que los datos brutos estén en condiciones. Para eso se han hecho dos comprobaciones: buscar valores nulos y detectar valores anómalos.

Valores nulos

Se ha recorrido todo el dataset comprobando si hay celdas vacías en alguna de las señales y el resultado ha sido que no existe ningún valor nulo. Esto tiene sentido porque los sensores están grabando de forma continua y siempre devuelven algún valor numérico, aunque a veces ese valor pueda no ser válido, de ahí la siguiente comprobación.

Detección de anomalías

En el análisis exploratorio de la fase de extracción se detectó que la temperatura del pecho tenía un mínimo de -273 ºC en algún sujeto, lo cual es físicamente imposible para un cuerpo humano. Para evitar que el error afecte al modelo, se ha definido un rango fisiológicamente válido para la temperatura cutánea entre 20 y 45 ºC y se ha contado cuántas muestras quedan fuera de ese rango en cada sujeto.

El resultado es que solo el sujeto S3 presenta anomalías, con un total de 83 muestras fuera de rango (apenas un 0,0018% del total de sus datos). Lo curioso es que estas 83 muestras no están dispersas, sino concentradas en un tramo continuo (entre los índices 1771710 y 1771792), donde el sensor registra repetidamente el valor -273,15 ºC. Esto sugiere que hubo un fallo puntual del sensor, probablemente una desconexión momentánea o una interferencia.

Corrección de anomalías

Para arreglar estos valores se ha aplicado interpolación lineal únicamente sobre las muestras anómalas. Este método consiste en trazar una línea recta entre el último valor válido antes del fallo y el primer valor válido después, y rellenar los huecos con los puntos de esa línea. Es una solución que encaja muy bien aquí porque la temperatura es una señal de variación lenta, no cambia bruscamente de un instante a otro, así que asumir una progresión suave entre dos valores válidos es perfectamente razonable.

En el caso concreto de S3, los valores corregidos quedan entre 33,618 y 33,611 ºC, una transición suave y coherente con el comportamiento esperado de la señal. Después de la

corrección, se ha vuelto a comprobar el rango y ya no quedan valores fuera de los límites válidos.

Una vez corregidos los datos del sujeto S3, se ha guardado una versión limpia del dataset completo en una carpeta separada (data/cleaned) para que el resto del proceso trabaje siempre sobre los datos ya validados.

6.3.3. Estimación de variables derivadas por ventanas

Esta es la parte central de la transformación. Aquí se cogen las señales brutas (ECG, EDA, respiración, EMG y temperatura) y se convierten en variables que sí tienen un significado fisiológico claro.

El procedimiento es el mismo para todas las señales y sigue siempre la misma lógica. Primero se prepara un dataframe base por sujeto con dos columnas iniciales: time_s, que es el segundo en el que termina cada ventana (empieza en el segundo 60 porque la primera ventana cubre del 0 al 60), y LABEL, que es la etiqueta del estado fisiológico asignada a esa ventana. La etiqueta se calcula por votación mayoritaria, es decir, si en los 60 segundos de la ventana hay mayoría de muestras etiquetadas como Stress, esa ventana se queda con el label de Stress.

En esta etapa se mantienen todas las etiquetas (incluidas las de transición y meditación). El filtrado de los estados válidos (1, 2 y 3) se hace al final, justo antes de exportar los datasets.

A partir de ese dataframe base, se van añadiendo columnas nuevas conforme se van calculando las variables derivadas de cada señal.

<u>Variables derivadas del ECG (HR y HRV)</u>

La señal de ECG está muestreada a 700 Hz, lo que significa que tiene 700 valores por cada segundo. Lo importante de esta señal no es la forma de la onda en sí, sino el momento exacto en el que ocurre cada latido. Una vez se sabe cuándo late el corazón, se pueden calcular la frecuencia cardíaca (cuántos latidos hay por minuto) y la variabilidad cardíaca (cuánto cambia el tiempo entre un latido y el siguiente).

Para detectar los latidos se ha usado NeuroKit2, que es una librería especializada en señales biomédicas. Esta librería identifica los picos R del ECG, que son los puntos más

altos de cada latido. Después se calculan los intervalos RR, que es simplemente el tiempo que pasa entre dos picos R consecutivos (Makowski et al., 2021).

A partir de los intervalos RR se obtienen dos variables por cada ventana:

- HR_bpm: la frecuencia cardíaca media en latidos por minuto.
- HRV_RMSSD: la variabilidad cardíaca, calculada con la métrica RMSSD, que mide la diferencia cuadrática media entre intervalos RR sucesivos. Está expresada en milisegundos.

Antes de aplicar este proceso a todos los sujetos, se hizo una prueba con S10 para validar que el método funcionaba bien. Los valores obtenidos son fisiológicamente coherentes (HR media en torno a 75-80 bpm, RMSSD en valores típicos de adultos jóvenes), y las series resultantes son suaves y continuas en el tiempo, como cabía esperar dado el solapamiento entre ventanas (cada ventana comparte 59 segundos con la anterior).

Variables derivadas de la EDA (SCL, SCR_count, SCR_amplitude)

Como se explica en apartados anteriores, la EDA mide la conductancia de la piel, que cambia según la actividad de las glándulas sudoríparas. Es una de las señales más útiles para detectar estrés porque responde de forma muy directa a la activación emocional.

La señal EDA se puede descomponer en dos componentes que aportan información distinta. El componente tónico (SCL) es el nivel basal de conductancia, una señal lenta que refleja el estado de activación general del sujeto. El componente fásico (SCR) son picos rápidos y transitorios que aparecen como respuesta a estímulos puntuales.

A partir de la descomposición de la señal EDA se calculan tres variables por ventana:

- SCL: nivel tónico medio en la ventana, en microsiemens.
- SCR_count: número de respuestas fásicas detectadas en la ventana.
- SCR_amplitude: amplitud media de esas respuestas. Si no hay respuestas detectadas, este valor es 0.

En el sujeto S10, el SCL se mueve entre 0,7 y 2,1 μS, lo que refleja cambios claros entre fases de menor y mayor activación. El SCR_count tiene una distribución asimétrica con mediana 0, lo que es totalmente normal ya que no en todas las ventanas hay respuestas fásicas, y cuando las hay suelen agruparse en periodos concretos.

Variables derivadas de la respiración (Resp_Rate y Resp_RMSSD)

La señal de respiración mide la expansión del tórax y permite reconstruir el patrón respiratorio del sujeto.

El método para calcular las variables respiratorias es prácticamente idéntico al del ECG, pero cambian dos cosas. En lugar de detectar picos R, se detectan ciclos respiratorios (inspiración/espiración). Y en lugar de calcular intervalos RR, se calculan intervalos entre respiraciones (IBI, Inter-Breath Interval).

A partir de ahí se obtienen dos variables por ventana:

- Resp_Rate: frecuencia respiratoria en respiraciones por minuto.
- Resp_RMSSD: variabilidad del ritmo respiratorio, también con la métrica RMSSD.

En el sujeto S10, la frecuencia respiratoria media ronda las 14-15 respiraciones por minuto, que es exactamente lo que se espera en un adulto en condiciones normales. La variabilidad respiratoria también se mantiene en valores plausibles a lo largo del experimento.

Variable derivada del EMG (EMG_RMS)

La señal de EMG mide la actividad eléctrica de los músculos. Es una señal de muy alta variabilidad instantánea, así que su forma bruta tampoco se puede interpretar directamente. Lo que se hace es calcular el RMS (Root Mean Square) dentro de cada ventana. El RMS siempre da valores positivos y representa cuánta activación muscular hay en ese intervalo. Esto da una sola variable por ventana:

- EMG_RMS: intensidad media de activación muscular, en milivoltios.

En S10, los valores se mueven en un rango muy pequeño (entre 0,004 y 0,008 mV), con una media de unos 0,006 mV. Son valores bajos pero estables, coherentes con una señal EMG del trapecio en reposo con momentos puntuales de mayor tensión.

Variables derivadas de la temperatura (Temp_mean y Temp_slope)

La temperatura es una señal que no cambia de forma brusca, así que las estadísticas instantáneas no aportan demasiado. Lo que sí tiene sentido es resumir el nivel medio en cada ventana y ver si la temperatura está subiendo o bajando.

De aquí salen dos variables por ventana:

- Temp_mean: temperatura media en la ventana, en grados Celsius.
- Temp_slope: pendiente o tendencia de la temperatura dentro de la ventana, que indica si está subiendo (valores positivos) o bajando (valores negativos).

En S10, la temperatura media está en torno a los 34,8 ºC, con un rango entre 33,8 y 35,6 ºC. La pendiente está casi siempre cercana a cero, lo que confirma que la temperatura cambia de forma muy gradual a lo largo del experimento.

Stress Index

A diferencia de las variables anteriores, el Stress_index no se calcula a partir de una señal bruta, sino que es un indicador compuesto que combina las variables derivadas que ya se han calculado. Su objetivo es resumir el nivel de activación fisiológica del sujeto en un único número fácil de interpretar, especialmente pensando en un usuario no técnico (un familiar o un cuidador).

El cálculo se hace en dos pasos. Primero se seleccionan las variables que reflejan activación simpática, es decir, las que aumentan cuando hay estrés: HR, SCL, SCR_count, Resp_Rate y EMG_RMS. Después, para cada una de estas variables, se calcula el z-score normalizado por sujeto.

¿Por qué estas 5 variables? Son las que reflejan activación del sistema nervioso simpático (todas aumentan con el estrés): la frecuencia cardíaca se acelera, la conductancia de la piel sube, las respuestas electrodérmicas se hacen más frecuentes, la respiración se acelera y la tensión muscular aumenta. No se incluyen HRV_RMSSD ni Resp_RMSSD (que disminuyen con el estrés) porque mezclar variables que suben y bajan complicaría la interpretación sin aportar información adicional al promedio.

Un z-score mide cuántas desviaciones típicas se aleja un valor concreto de la media del sujeto. Por ejemplo, si un sujeto tiene una HR media de 75 bpm con desviación típica de 8, una ventana con HR de 91 bpm tendría un z-score de (91-75)/8 = 2,0. Esto significa que en esa ventana la frecuencia cardíaca está dos desviaciones típicas por encima de su nivel habitual.

La normalización se hace por sujeto y no a nivel global, y esto es muy importante. Cada persona tiene un nivel fisiológico basal diferente, hay quien tiene una HR media de 65

bpm y quien la tiene de 85 bpm. Comparar valores absolutos entre personas no tiene sentido, porque alguien con HR alta no necesariamente está más estresado que otro con HR baja, simplemente tiene un basal distinto. Al normalizar por sujeto, lo que se mide es cuánto se desvía cada persona respecto a sí misma, lo cual sí permite comparar niveles de activación entre individuos de forma justa.

Una vez calculados los z-scores de cada variable, se promedian y el resultado es el Stress_index. Valores positivos indican mayor activación de la habitual, valores negativos indican menor activación.

En el sujeto S10, el Stress_index medio es de +0,482 durante las fases de Stress, -0,151 durante Baseline y -0,651 durante Amusement. La separación clara entre Stress y los otros dos estados confirma que el indicador captura bien la activación inducida por el protocolo experimental. Que Amusement quede por debajo de Baseline es coherente con este sujeto concreto, ya que durante la fase de diversión se ven vídeos de forma pasiva y sentado, mientras que el Baseline incluye lectura activa de revistas, lo que puede generar un nivel de activación basal ligeramente superior.

Una decisión importante sobre el Stress_index es que no se incluye como variable de entrada del modelo predictivo. La razón es que sería redundante, como es una combinación lineal directa de variables que ya están en el modelo (HR, SCL, SCR_count, Resp_Rate y EMG_RMS), añadirlo no aportaría información nueva e introduciría problemas de colinealidad. El Stress_index se utiliza únicamente en el dashboard, donde su valor está en facilitar la interpretación visual del nivel de estrés.

6.3.4. Inspección final del dataset

Una vez calculadas todas las variables derivadas, hay que validar el resultado antes de pasar a la fase de carga. Esta inspección final tiene tres objetivos: comprobar que la estructura del dataset es la esperada, analizar los valores nulos y detectar valores extremos que puedan ser artefactos.

Estructura del dataset

El dataframe global tiene 85.968 ventanas (sumando todos los sujetos) y 13 columnas: identificadores (subject_id, time_s, LABEL), las 10 variables derivadas y el Stress_index. Cada fila representa una ventana de 60 segundos.

Tabla 7. Estadísticas descriptivas globales del dataset transformado.

Variable	count	mean	std	min	max	n_nan
time_s	85.968	2.950,810	1.700,548	60,000	7.070,000	0
LABEL	85.968	1,347	1,629	0,000	7,000	0
HR	85.968	76,946	15,462	47,795	148,844	0
HRV_RMSSD	85.968	58,004	50,443	2,010	1.162,169	0
SCL	85.968	4,887	3,527	0,391	21,948	0
SCR_count	85.968	1,444	2,399	0,000	16,000	0
SCR_amplitude	85.968	0,048	0,140	0,000	2,374	0
Resp_Rate	85.875	14,500	3,887	1,391	26,897	93
Resp_RMSSD	85.875	1.887,078	2.698,722	137,139	79.407,143	93
EMG_RMS	85.968	0,008	0,004	0,003	0,046	0
Temp_mean	85.968	33,908	1,159	28,202	35,632	0
Temp_slope	85.968	0,000	0,004	-0,090	0,104	0
Stress_index	85.968	-0,000	0,634	-1,323	3,884	0

Elaboración propia.

Análisis de valores nulos

Se ha comprobado en qué columnas y en qué sujetos aparecen valores nulos. El resultado es que solo hay NaNs en dos columnas (Resp_Rate y Resp_RMSSD) y solo en un sujeto (S8). En total son 93 ventanas afectadas, lo que representa apenas un 0,108% del dataset.

Tabla 8. Distribución de valores nulos por sujeto y variable.

Sujeto	Variable	n_nan	pct_nan (%)
S8	**Resp_Rate**	93	0,108
S8	**Resp_RMSSD**	93	0,108

Elaboración propia.

En el sujeto S8, hay 93 ventanas en las que el algoritmo no detecta suficientes ciclos respiratorios para calcular la Resp_Rate o el Resp_RMSSD de forma fiable. Para calcular el RMSSD se necesitan al menos dos intervalos entre respiraciones, y cuando hay menos de dos respiraciones detectadas en la ventana el cálculo no es posible ya que se necesitan diferencias.

¿Qué se hace con estos NaNs?

Para el modelo predictivo, se eliminan. 93 filas sobre 86.000 es una proporción totalmente despreciable y no afecta a la distribución ni a la capacidad predictiva. Además, no tiene

sentido entrenar un modelo con ventanas en las que no se conoce el valor real de una variable.

Para el dashboard, se mantienen. El dashboard tiene que reflejar el comportamiento real del dato, así que, si en una ventana concreta no se pudo calcular la respiración, lo correcto es mostrar ese hueco, no inventarse un valor.

Detección de valores extremos

Para identificar posibles artefactos, se han comparado los máximos reales de cada variable con sus percentiles 99,5. La idea es que, si el máximo está muy por encima del percentil 99,5, probablemente se trate de un valor anómalo y no de un valor fisiológico real.

Tabla 9. Comparación de máximos reales y percentiles 99,5 por variable.

Variable	min	p 0,5	p 99,5	max	n_extreme	pct_extreme
HR	47,80	50,85	134,29	148,84	860	1,000
HRV_RMSSD	2,01	4,91	243,98	1.162,17	860	1,000
SCL	0,39	0,58	20,26	21,95	860	1,000
EMG_RMS	0,0032	0,0037	0,0350	0,0458	860	1,000
Temp_mean	28,20	28,43	35,58	35,63	860	1,000
Temp_slope	-0,0903	-0,0122	0,0120	0,1040	860	1,000
Stress_index	-1,32	-1,15	1,88	3,88	860	1,000
Resp_RMSSD	137,14	224,56	8.546,32	79.407,14	856	0,997
Resp_Rate	1,39	4,66	23,78	26,90	855	0,996
SCR_amplitude	0,00	0,00	0,70	2,37	421	0,490
SCR_count	0,00	0,00	11,00	16,00	299	0,348

Elaboración propia.

La mayoría de variables tienen máximos coherentes con sus percentiles. Por ejemplo, la HR llega a 148,8 bpm, que es perfectamente plausible durante el estrés agudo del TSST. La SCL llega a 21,9 µS y la Resp_Rate a 26,9 brpm, valores también coherentes con activación fisiológica intensa.

Pero hay dos variables que tienen máximos muy por encima de sus percentiles, lo que indica artefactos de cálculo:

- Resp_RMSSD: el máximo real es de 79.407 ms frente a un percentil 99,5 de 8.546 ms (casi 10 veces superior).

- HRV_RMSSD: el máximo real es de 1.162 ms frente a un percentil 99,5 de 244 ms (casi 5 veces superior).

Ambos casos tienen el mismo origen. Son ventanas en las que el algoritmo detectó muy pocos eventos (muy pocos picos R en el caso del ECG, o muy pocos ciclos respiratorios en el caso de la respiración). Con tan pocos intervalos, el cálculo del RMSSD se vuelve inestable y produce valores excesivamente altos que no reflejan la variabilidad fisiológica real del sujeto. Los valores normales de RMSSD en adultos jóvenes están entre 20 y 200 ms, así que valores superiores a 1.000 ms simplemente no son fisiológicamente posibles.

Limpieza de outliers por clipping al percentil 99

Para corregir estos artefactos sin perder filas, se aplica un clipping al percentil 99. Esto significa que cualquier valor por encima del percentil 99 se sustituye por el valor de ese percentil. Es una solución conservadora porque no elimina ninguna observación, solo recorta los valores extremos que sabemos que son artefactos. El percentil 99 deja fuera el 1% más alto de la distribución, que es justo donde se concentran los valores anómalos identificados.

Después del clipping, las dos variables quedan con máximos mucho más razonables y dentro de rangos fisiológicos plausibles, sin afectar al resto de la distribución.

6.4. Carga del dataset final

Con todas las variables calculadas y validadas, el último paso es exportar los datos en formato CSV para que se puedan usar en las siguientes fases. Aquí se ha tomado una decisión importante: en lugar de generar un único dataset, se han creado dos datasets independientes, cada uno pensado para un uso concreto.

dashboard_dataset.csv. Está pensado para el dashboard interactivo. Incluye todas las variables derivadas, el Stress_index y una columna adicional con el nombre legible del estado (Baseline, Stress, Amusement). Mantiene los 93 NaNs del sujeto S8 porque en visualización lo correcto es reflejar el dato real, sin rellenar huecos artificialmente. Solo se conservan las ventanas etiquetadas con los tres estados válidos (Baseline, Stress y Amusement); se descartan las fases de transición (label 0) y la meditación (label 4) porque no forman parte del análisis del estrés.

model_dataset.csv. Está pensado para entrenar el modelo predictivo. Tiene las mismas variables fisiológicas pero sin Stress_index (sería redundante como input, ya que es una combinación lineal de variables que ya están dentro) y sin filas con NaN (el modelo necesita datos completos para funcionar correctamente). La pérdida de 93 filas sobre unas 33.000 (un 0,28%) es completamente despreciable y no afecta a la capacidad predictiva.

Ambos datasets incluyen una columna subject_id que identifica al sujeto. Esto es imprescindible para dos cosas: filtrar por usuario en el dashboard y aplicar la validación cruzada por sujetos (LOSO) en el entrenamiento del modelo, que se explicará en la siguiente sección.

Los resultados finales son:

- dashboard_dataset.csv: 33.149 filas y 15 columnas. Distribución por estado: Baseline 17.608, Stress 9.966, Amusement 5.575.
- model_dataset.csv: 33.056 filas y 13 columnas (las 93 filas con NaN eliminadas).

Con esta exportación queda cerrada la fase de transformación y, con ella, todo el proceso ETL. A partir de aquí ya se puede empezar a trabajar con los datos limpios y estructurados, tanto para el desarrollo del modelo predictivo como para la construcción del dashboard.

7. Desarrollo

7.1. Modelo predictivo

Con el dataset ya procesado y las variables derivadas calculadas, la siguiente fase del trabajo es construir un modelo capaz de predecir el estado afectivo de un sujeto a partir de sus señales fisiológicas. Este apartado recoge el proceso completo: cómo se ha planteado el problema, qué decisiones se han tomado para evaluar los modelos de forma honesta, qué algoritmos se han probado y qué resultados se han obtenido.

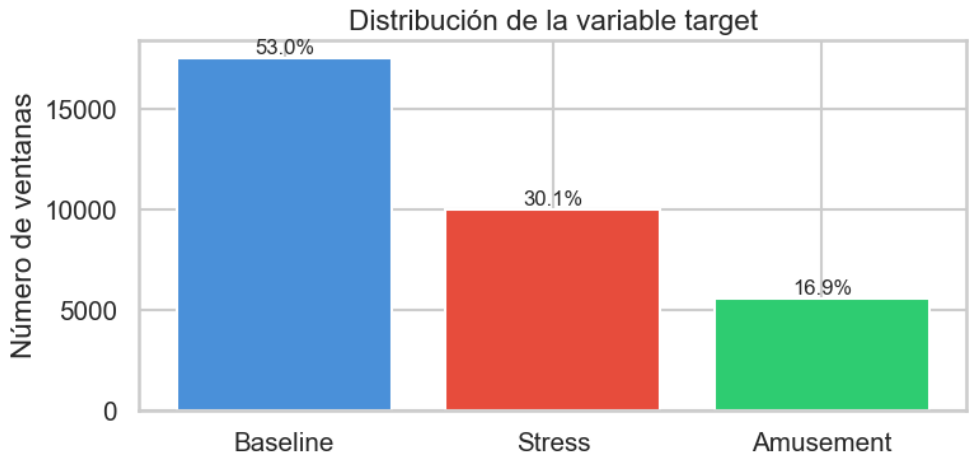
7.1.1. Definición del problema

El objetivo es clasificar cada ventana temporal del experimento en uno de tres estados: Baseline (reposo), Stress (estrés inducido) y Amusement (diversión). Se trata, por tanto, de un problema de clasificación multiclase con tres categorías.

Los datos de entrada son las 10 variables fisiológicas calculadas en la fase anterior: HR, HRV_RMSSD, SCL, SCR_count, SCR_amplitude, Resp_Rate, Resp_RMSSD, EMG_RMS, Temp_mean y Temp_slope. Cada fila del dataset corresponde a una ventana de 60 segundos de un sujeto concreto, con su etiqueta de estado asociada. En total se trabaja con 33.056 ventanas repartidas entre 15 sujetos (model_dataset.csv).

La distribución de las clases no es uniforme. El protocolo experimental dedica más tiempo a algunas fases que a otras (Baseline dura en torno a 20 minutos, Stress unos 11 y Amusement unos 6), y el dataset hereda ese desbalanceo.

Ilustración 7. Distribución de la variable objetivo.



Elaboración propia.

El desbalanceo no es extremo (ninguna clase está por debajo del 5%) pero sí lo suficiente como para tenerlo en cuenta. Para que el modelo no se sesgue hacia Baseline simplemente porque es mayoritaria, se utiliza el parámetro *class_weight="balanced"* en los algoritmos que lo soportan (Random Forest y Regresión Logística), y *sample_weight* calculado manualmente en Gradient Boosting. Ambos mecanismos hacen lo mismo, asignan a cada clase un peso inversamente proporcional a su frecuencia, de forma que un error sobre Amusement penalice más que un error sobre Baseline.

Normalización z-score por sujeto

La variabilidad entre personas vuelve a ser el problema principal antes de entrenar el modelo (no todos los sujetos tienen los mismos niveles fisiológicos basales). Si se entrena con los valores absolutos de las señales, el modelo acaba aprendiendo a distinguir personas en lugar de estados afectivos, y al llegar un sujeto nuevo falla.

La solución es la misma que se aplicó para construir el Stress_index en la fase de transformación, una normalización z-score por sujeto. Para cada variable y cada persona se resta su propia media y se divide por su propia desviación típica, de forma que los valores pasan a medirse en desviaciones respecto al nivel habitual de cada individuo. La diferencia es que aquí se aplica sobre las diez features que entran directamente al modelo. El objetivo es el mismo, eliminar las diferencias basales entre personas y hacer que todos los sujetos compartan una escala relativa común.

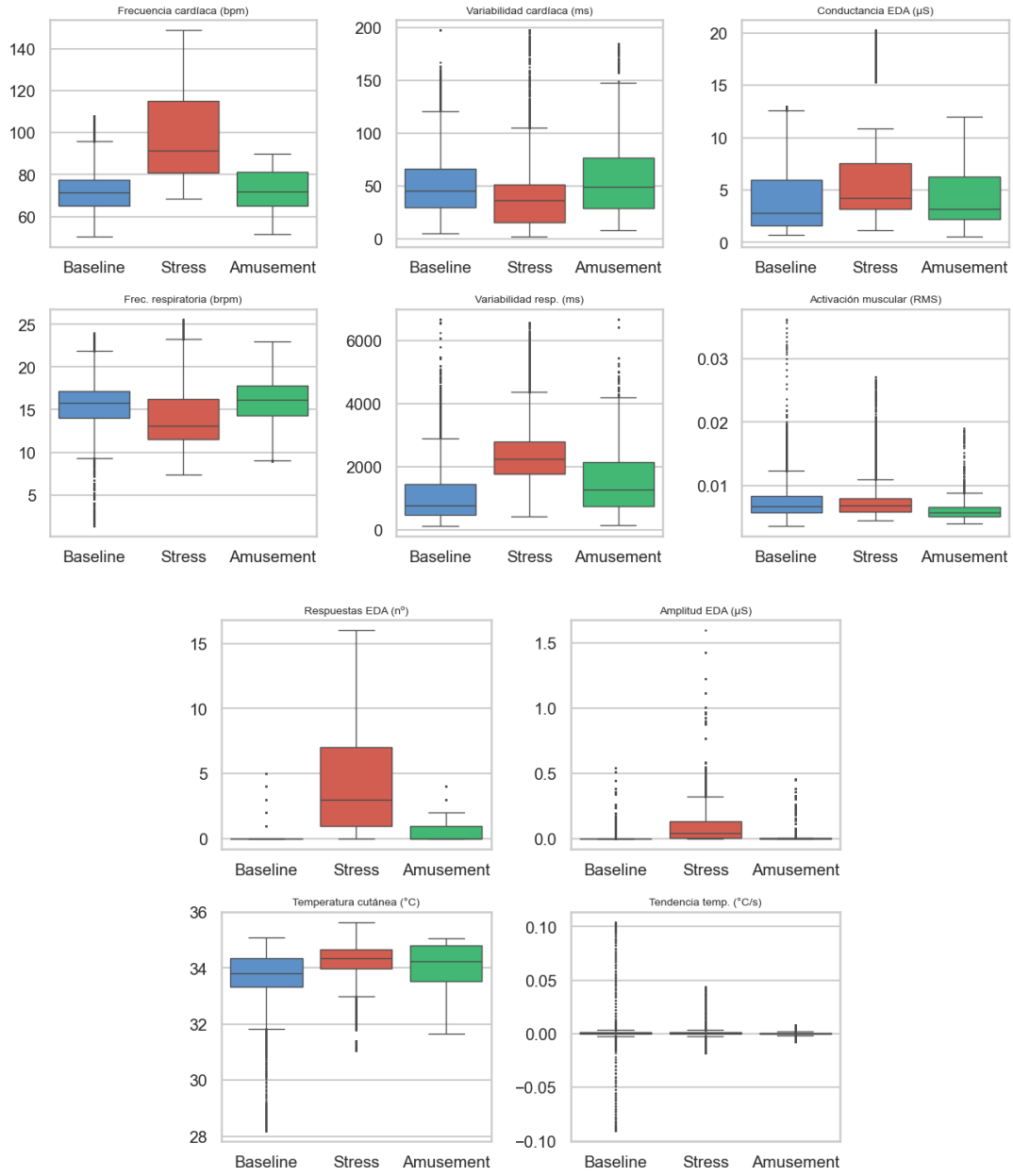
La normalización se ha calculado usando todos los datos de cada sujeto. Esta decisión introduce una pequeña impureza metodológica. En un sistema real, cuando un usuario se pone el wearable por primera vez, el modelo no conoce su media ni su desviación típica desde el minuto cero, necesita una fase de calibración. Al usar aquí la serie completa del sujeto se está aprovechando información que en la realidad no estaría disponible, lo que puede elevar ligeramente el rendimiento del modelo. La alternativa (calcular la normalización solo con el conjunto de entrenamiento) no tiene sentido fisiológico, porque los sujetos de train y de test son personas distintas y los estadísticos de unas no sirven para normalizar a otras. En un producto real lo adecuado sería sustituir esta normalización por una fase inicial de unos minutos u horas de calibración, o por estadísticas basales estimadas a partir de un tramo previo del usuario y que se actualicen según llegan datos. Esta cuestión se retoma en el apartado de limitaciones.

El análisis exploratorio de las variables se hace sobre los datos sin normalizar para conservar la interpretabilidad fisiológica (bpm, µS, °C), pero todos los modelos se entrenan con los datos normalizados.

7.1.2. Análisis exploratorio de las variables

Antes de entrenar el modelo, conviene entender cómo se comportan las 10 variables de entrada y qué relación tienen con los tres estados afectivos. Esto permite anticipar qué variables serán útiles para la predicción y detectar posibles redundancias entre ellas.

Ilustración 8. Distribución de variables fisiológicas por estado afectivo.



Elaboración propia.

Los boxplots permiten identificar visualmente qué variables separan mejor los tres estados afectivos. Estas son las que probablemente tendrán más importancia en el modelo:

- *HR* (Frecuencia cardíaca): es la variable con mayor separación. Stress presenta una mediana cercana a 100 bpm, por encima de Baseline y Amusement. La caja de Stress apenas se solapa con las otras dos.
- *SCL* (Conductancia EDA): incremento progresivo desde Baseline hasta Stress, con Amusement en niveles intermedios. Es una de las señales más informativas porque refleja directamente la activación simpática.
- *SCR_count y SCR_amplitude* (Respuestas y Amplitud EDA): durante Stress aparecen más respuestas y de mayor amplitud. En Baseline y Amusement la mayoría de ventanas tienen valor 0 (sin respuestas fásicas), lo que es coherente con lo observado en los estadísticos descriptivos.
- *Resp_RMSSD* (Variabilidad respiratoria): Stress muestra una mediana (cerca de 2000 ms) claramente superior a Baseline (aprox. 800 ms) y Amusement (aprox. 1000 ms). Bajo estrés la variabilidad entre ciclos respiratorios aumenta de forma notable, lo que convierte a esta variable en un buen discriminante.
- *HRV_RMSSD* (Variabilidad cardíaca): Stress presenta una mediana más baja (aprox. 35 ms) que Baseline y Amusement (~50 ms), reflejando la reducción de variabilidad cardíaca bajo estrés. La separación entre la caja de Stress y las otras dos es visible.
- *EMG_RMS* (Activación muscular): ligero aumento en Stress respecto a Baseline y Amusement, aunque las diferencias son pequeñas. La tensión muscular del trapecio sube con el estrés pero de forma sutil.
- *Temp_mean* (Temperatura media): Stress y Amusement muestran medianas ligeramente superiores a Baseline (~34.5°C vs ~34°C). La diferencia es modesta pero consistente.

Las variables con menor capacidad discriminante entre estados son:

- *Resp_Rate* (Frecuencia respiratoria): distribuciones muy similares entre los tres estados, lo que sugiere que la frecuencia respiratoria por sí sola no discrimina bien entre estados en este dataset.

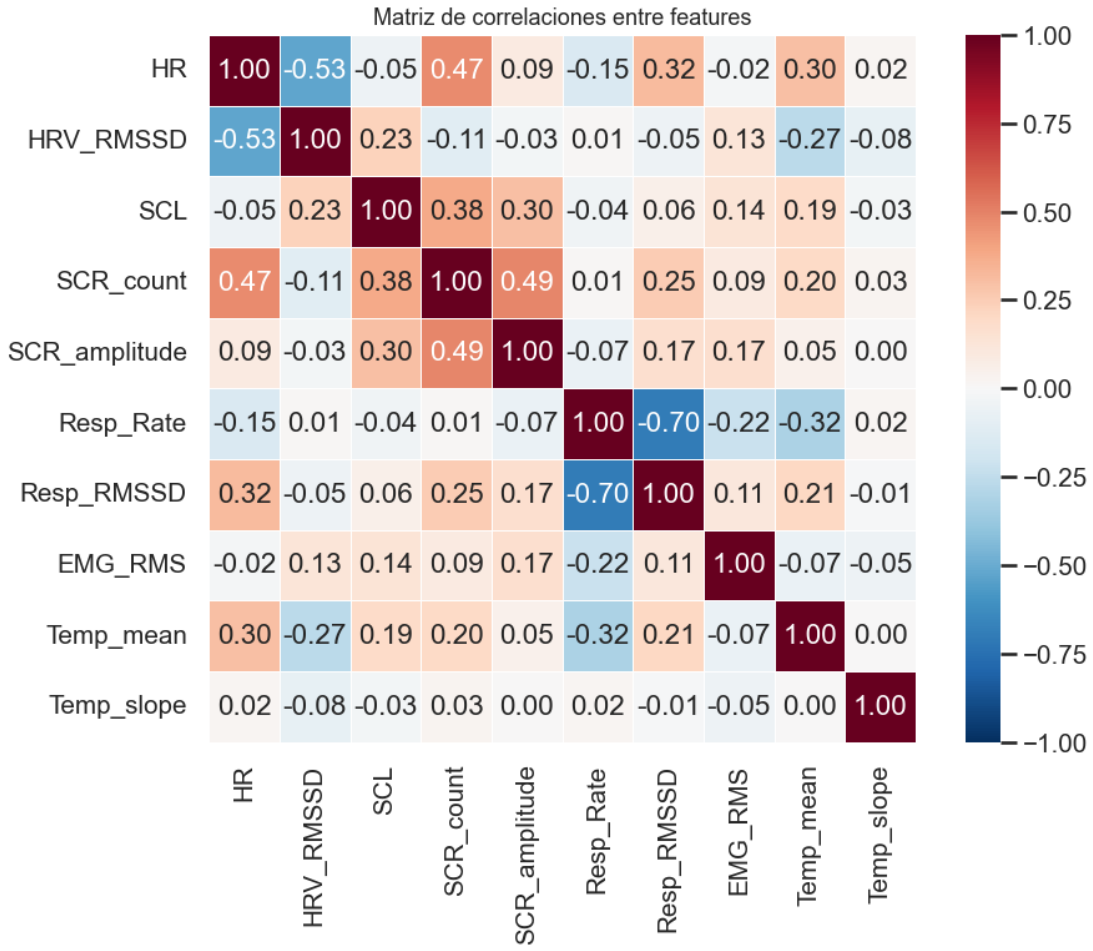
- *Temp_slope* (Tendencia de la temperatura): valores muy concentrados en torno a 0 para los tres estados, con poca separación. La tendencia no varía de forma significativa entre condiciones.

Un patrón importante es que Amusement y Baseline presentan distribuciones similares en la mayoría de variables, lo que anticipa que serán los dos estados más difíciles de distinguir para cualquier modelo.

Matriz de correlaciones

La matriz de correlaciones sirve para detectar variables que aporten información redundante. Si dos features están muy correlacionadas ($|r| > 0{,}80$), una de ellas apenas aporta información adicional al modelo.

Ilustración 9. Matriz de correlaciones entre features.



Elaboración propia.

La correlación más alta del dataset es entre Resp_Rate y Resp_RMSSD (-0,70). Es esperable, ambas se calculan a partir de la misma señal respiratoria, y una respiración acelerada implica ciclos más cortos y regulares (es decir, menor variabilidad). La segunda más alta es entre HR y HRV_RMSSD (-0,53), misma lógica fisiológica, cuando el corazón se acelera, la variabilidad disminuye. Los dos pares quedan por debajo del umbral habitual de 0,80 y se mantienen ambas variables en cada caso porque miden cosas distintas, una captura velocidad y la otra regularidad. El resto de correlaciones son moderadas (0,3-0,5) y reflejan que el estrés activa varios sistemas simultáneamente. Temp_slope es la variable más independiente del conjunto.

No hay redundancia real entre las features, así que se mantienen las 10 originales como entrada del modelo.

7.1.3. Estrategia de validación (LOSO)

En los modelos de clasificación convencionales se suelen dividir los datos aleatoriamente en un conjunto de entrenamiento (70-80%) y otro de prueba (20-30%). Sin embargo, en este caso esa estrategia no sería correcta.

¿Por qué no se puede hacer un train/test split aleatorio?

El dataset contiene ventanas temporales con un paso de 1 segundo. Esto significa que la ventana del segundo 340 comparte 59 de sus 60 segundos con la ventana del segundo 341. Si hiciésemos un split aleatorio, podrían caer ventanas casi idénticas en train y en test, inflando artificialmente el accuracy. Además, las señales fisiológicas son individuales. Si los datos de un mismo sujeto aparecen a la vez en train y en test, el modelo puede reconocer patrones propios de esa persona concreta en lugar de aprender el estrés como fenómeno general. Eso da resultados muy buenos en evaluación pero completamente inútiles con un usuario nuevo.

Leave-One-Subject-Out (LOSO)

Leave-One-Subject-Out (LOSO) resuelve ambos problemas, en cada iteración se entrena con los datos de 14 sujetos y se evalúa con el sujeto restante. El modelo nunca ha visto ninguna ventana del sujeto de test. Esto se repite 15 veces (una por sujeto) y al final se agregan todas las predicciones.

LOSO tiene tres ventajas importantes. Primero, simula el escenario real del producto. Cuando un sistema como Kora empieza a funcionar con un usuario nuevo, no tiene datos históricos de esa persona, solo sabe lo que ha aprendido de otros. LOSO reproduce esa situación. Segundo, elimina por completo la fuga de información (*data leakage*) entre ventanas solapadas, porque ninguna ventana del sujeto de test puede aparecer en train. Y tercero, es el estándar de evaluación en la literatura de WESAD (Schmidt et al., 2018), lo que permite comparar resultados con el paper original más adelante.

Validación cruzada anidada (nested CV)

La búsqueda de hiperparámetros también debe respetar la separación de sujetos. Si se buscan los mejores hiperparámetros una sola vez usando todos los sujetos y después evaluásemos con LOSO, el sujeto de test habría participado indirectamente en la selección de hiperparámetros (por lo que habría otro *data leakage*).

Para evitarlo, se utiliza validación cruzada anidada (nested CV). Dentro de cada iteración LOSO (que deja fuera al sujeto k), la búsqueda de hiperparámetros se realiza exclusivamente con los 14 sujetos restantes usando *GroupKFold*. Así, el sujeto k no influye ni en el entrenamiento ni en la selección de hiperparámetros. Esto significa que cada iteración LOSO puede seleccionar hiperparámetros ligeramente diferentes, lo cual es correcto y refleja mejor el rendimiento real del pipeline completo.

7.1.4. Algoritmos evaluados

Para este problema de clasificación se prueban tres algoritmos de diferente naturaleza, con el objetivo de comparar su rendimiento y seleccionar el más adecuado en base a los resultados:

- *Random Forest*: método de ensemble que construye múltiples árboles de decisión independientes y combina sus predicciones por votación mayoritaria. Es robusto, resiste bien el *overfitting* y no necesita que los datos estén escalados.
- *Gradient Boosting*: otro método de ensemble de árboles, pero secuencial. Cada árbol corrige los errores del anterior. Puede capturar patrones más complejos, aunque es más sensible a la configuración de hiperparámetros y más lento al no paralelizar el entrenamiento.

- *Regresión Logística*: modelo lineal que busca fronteras rectas entre clases. Se incluye como referencia base para evaluar si los patrones fisiológicos son linealmente separables o si se necesitan modelos más complejos.

La selección cubre tres niveles de complejidad: un modelo lineal, un ensemble de bagging y un ensemble de boosting. Esto permite comparar rendimientos y entender la naturaleza del problema.

¿Por qué estos tres y no otros?

Se descartan redes neuronales porque con 33.000 filas y 10 variables el dataset es demasiado pequeño para que aporten ventaja real, tienen muchos hiperparámetros que configurar y son difíciles de interpretar (no permiten extraer importancia de variables de forma directa). Se descarta *K-Nearest Neighbours (KNN)* porque con ventanas solapadas de 1 segundo las ventanas "más parecidas" serían las del mismo sujeto en segundos adyacentes, lo que sesgaría la evaluación. Además, es sensible a la escala de las variables y no proporciona *feature importance*. Estas dos razones (tamaño del dataset y ventanas solapadas) hacen que *KNN* y redes neuronales no sean opciones prácticas en este contexto.

7.1.5. Búsqueda de hiperparámetros

Los hiperparámetros son las configuraciones que controlan cómo se construye el modelo: cuántos árboles hacer, cuánto pueden crecer, cuántas variables mirar en cada partición. No se aprenden de los datos, los define el analista, y una mala elección puede condicionar el resultado. Por eso hay que buscarlos de forma sistemática.

Para Random Forest y Gradient Boosting se ha utilizado RandomizedSearchCV, que prueba combinaciones aleatorias del espacio de búsqueda y se queda con la que mejor rendimiento da en validación cruzada interna. Se han probado 15 combinaciones en Random Forest y 10 en Gradient Boosting, con GroupKFold(3) como validación interna, es decir, tres folds que respetan la separación por sujeto.

En Random Forest la búsqueda se realiza dentro del loop LOSO, con nested CV completa. Cada una de las 15 iteraciones lanza su propia búsqueda usando solo los 14 sujetos de train. Esto garantiza que el sujeto de test no influye en ninguna decisión del modelo. Como consecuencia, los hiperparámetros seleccionados pueden variar entre iteraciones.

Al revisar la tabla de hiperparámetros por fold se ve que n_estimators se estabiliza en 100 en la mayoría de casos, mientras que max_depth, min_samples_split y max_features oscilan entre distintos valores. Esta variabilidad indica que el modelo no es muy sensible a estos parámetros, diferentes configuraciones dan resultados similares. Esto en realidad es una buena señal ya que significa que la señal útil está suficientemente presente en los datos normalizados como para que el modelo rinda bien bajo diferentes configuraciones.

En Gradient Boosting no ha sido posible aplicar nested CV completa. Como los árboles se construyen de forma secuencial, el entrenamiento no es paralelizable y cada iteración LOSO con su propia búsqueda interna habría superado las seis horas de ejecución sin llegar a completarse. Por eso la búsqueda se ha hecho una sola vez, usando todos los sujetos con GroupKFold(3), y los mejores hiperparámetros se han aplicado después en la evaluación LOSO. El resultado de esa búsqueda es n_estimators=200, max_depth=3 y learning_rate=0,05, lo que apunta a un modelo conservador (árboles pequeños con contribuciones suaves). Hay que reconocer que este enfoque introduce una pequeña impureza. El sujeto de test ha participado indirectamente en la elección de hiperparámetros, aunque no en el entrenamiento ni en la predicción. En la práctica el impacto es mínimo porque los hiperparámetros son configuraciones generales del modelo y no codifican información específica de ningún sujeto, pero se menciona explícitamente como limitación frente al enfoque más estricto aplicado en Random Forest.

La Regresión Logística se utiliza con la configuración estándar de scikit-learn (C=1,0, solver=lbfgs), sin búsqueda de hiperparámetros. La idea es usarla como baseline puro. Si un modelo lineal sin optimizar ya rinde bien, significa que los patrones en los datos normalizados son fáciles de interpretar y relativamente sencillos. Aunque las features ya están normalizadas por sujeto, se aplica además un StandardScaler dentro de un Pipeline antes del clasificador. Los modelos lineales sí son sensibles a la escala de las variables, por lo que este paso se aplica para facilitar un entrenamiento estable y correcto.

7.1.6. Resultados y comparación

Los tres modelos se han evaluado con la misma estrategia LOSO y las mismas métricas. La siguiente tabla resume los resultados:

Tabla 10. Comparativa de modelos.

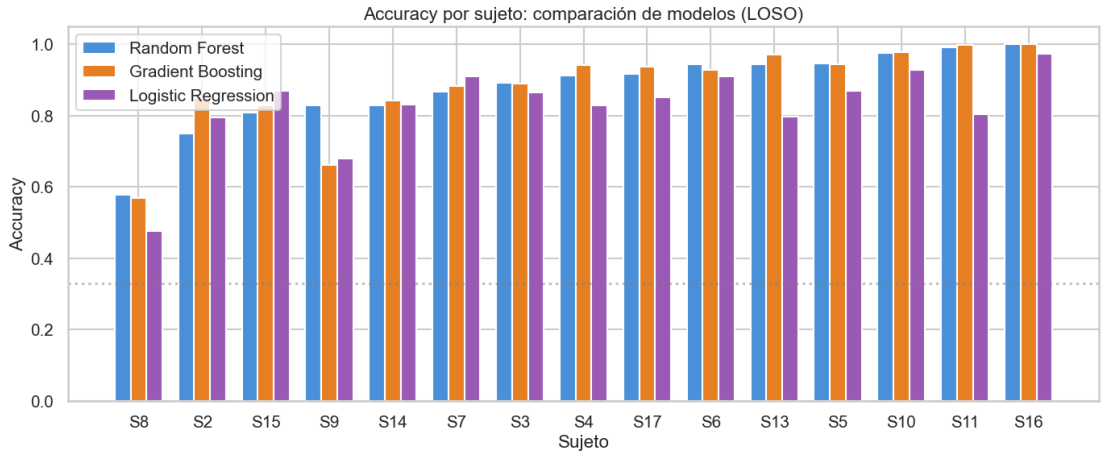
Modelo	Accuracy global	Accuracy medio ± std	Macro F1	F1 Baseline	F1 Stress	F1 Amusement
Random Forest	0,880	0,879 ± 0,110	0,840	0,920	0,910	0,690
Gradient Boosting	0,883	0,881 ± 0,122	0,847	0,930	0,890	0,720
Regresión Logística	0,827	0,826 ± 0,119	0,793	0,870	0,890	0,620

Elaboración propia.

Random Forest y Gradient Boosting obtienen rendimientos prácticamente idénticos: 87,9% y 88,1% de accuracy medio respectivamente. La diferencia de 0,2 puntos porcentuales es despreciable en la práctica, sobre todo teniendo en cuenta que Gradient Boosting se evaluó con una búsqueda de hiperparámetros menos estricta. Los dos coinciden también en el patrón por clases: Baseline y Stress se clasifican con alto rendimiento (F1 > 0,89) y Amusement es la clase más difícil (F1 0,69 en RF y 0,72 en GB).

La Regresión Logística queda unos cinco puntos por debajo (82,73%), que es un resultado más relevante de lo que parece. Un modelo lineal sin optimización de hiperparámetros alcanzando el 83% de accuracy dice algo importante sobre los datos: los patrones fisiológicos del estrés, una vez eliminadas las diferencias basales entre personas, son en gran medida linealmente separables. Una frontera recta ya captura la mayor parte de la información. La diferencia entre RF/GB y LR (unos cinco puntos) es la mejora que aportan los ensembles al capturar relaciones no lineales.

Ilustración 10. Accuracy por sujeto: comparación de los tres modelos.



Elaboración propia.

El gráfico comparativo por sujeto confirma el patrón. Las barras de RF y GB son prácticamente iguales en todos los sujetos. La Regresión Logística queda por debajo de forma sistemática, con mayor distancia en los sujetos más difíciles. S8 aparece como el peor en los tres modelos (0,58 RF, 0,57 GB, 0,48 LR), lo que indica que no es un problema del algoritmo sino un sujeto con respuestas fisiológicas atípicas incluso respecto a su propia línea base.

7.1.7. Modelo final: Random Forest

A pesar de que Gradient Boosting gana por 0,26 puntos porcentuales, se ha seleccionado Random Forest como modelo final. Hay cuatro razones que justifican esta decisión.

La primera es metodológica. Random Forest se ha evaluado con nested CV completa, sin ninguna fuga de información en la búsqueda de hiperparámetros. Gradient Boosting ha necesitado una búsqueda global única y la pequeña ventaja que obtiene podría estar precisamente explicada por esa impureza. Cuando la diferencia entre dos modelos es del orden de 0,26 puntos, lo correcto es quedarse con el que ofrece mayores garantías.

La segunda es de eficiencia computacional. Random Forest paraleliza la construcción de árboles y permite evaluaciones completas en tiempos razonables. Gradient Boosting es secuencial y multiplica el tiempo de ejecución de forma considerable. De cara a una posible re-ejecución del modelo con nuevos datos, o a una integración futura con pipelines automáticos, RF es claramente más práctico y rápido.

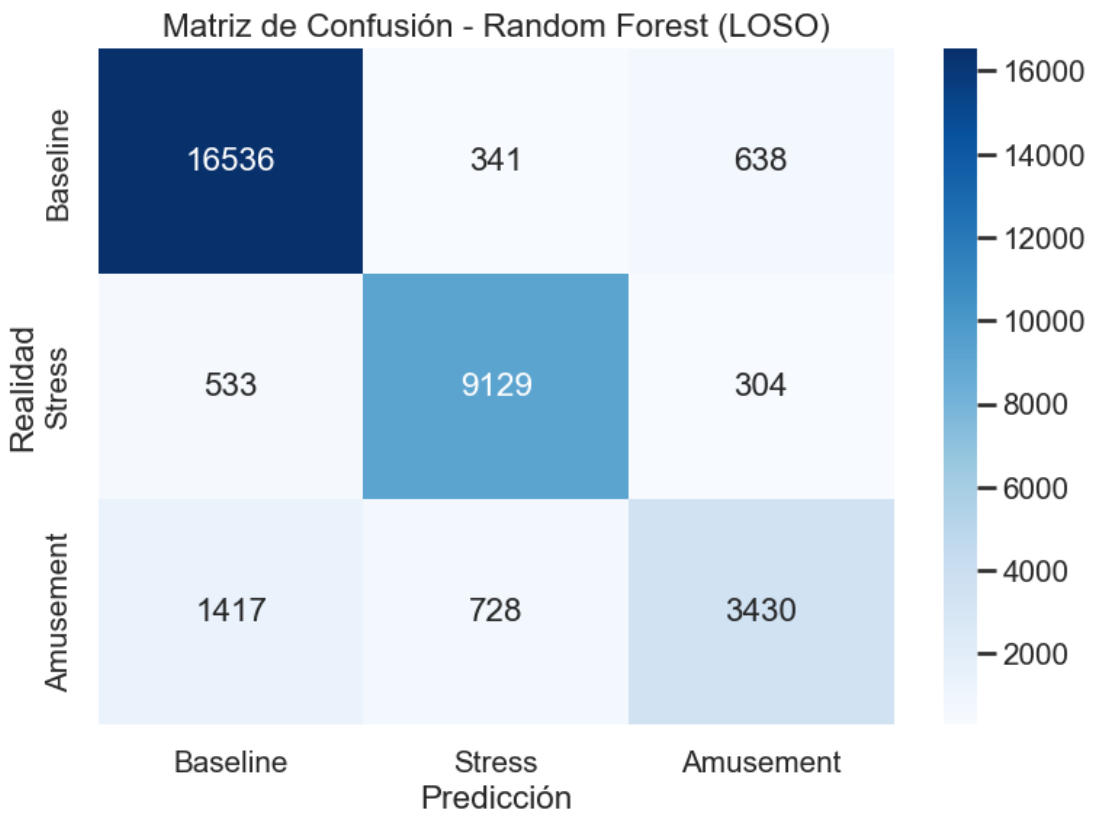
La tercera es la interpretabilidad. Random Forest permite extraer directamente la importancia de cada variable, una información relevante para entender qué señales fisiológicas discriminan mejor los estados. En un producto real esta interpretabilidad es un requisito, especialmente en contextos de salud donde las decisiones del sistema pueden influir en el bienestar de una persona.

Y la cuarta es la estabilidad. RF presenta una desviación típica por sujeto de ±0,110 frente a ±0,122 de GB. Es una diferencia pequeña, pero junto con las anteriores refuerza la decisión.

Una vez seleccionado Random Forest, conviene entrar en el detalle de sus predicciones. Dónde acierta, dónde se equivoca y qué variables le están dando más información.

Matriz de confusión

Ilustración 11. Matriz de confusión Random Forest.



Elaboración propia.

La diagonal principal recoge los aciertos: 16.536 ventanas de Baseline, 9.129 de Stress y 3.430 de Amusement clasificadas correctamente.

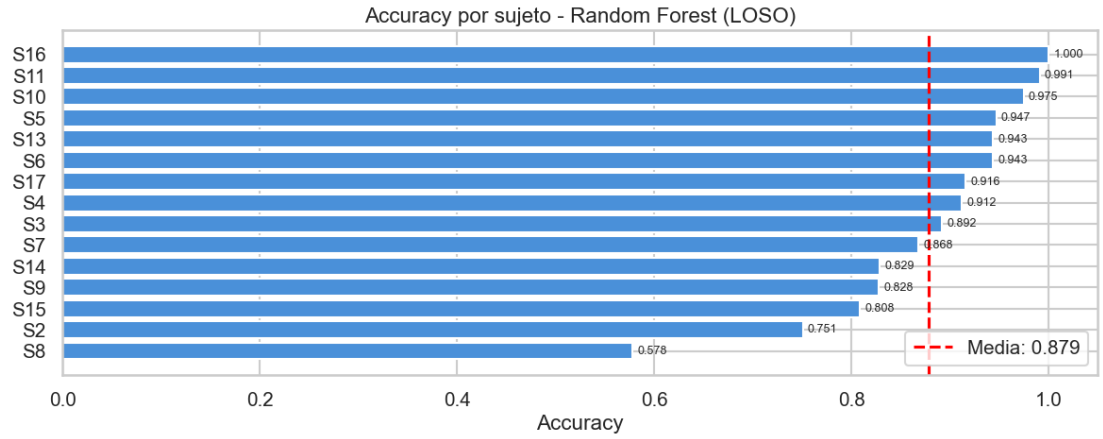
El error más frecuente, con diferencia, es la confusión entre Amusement y Baseline. 1.417 ventanas reales de Amusement acaban predichas como Baseline (un 25% del total de Amusement, y un 66% de sus errores). No es un fallo sorprendente: durante la fase de diversión los sujetos están sentados viendo vídeos de forma pasiva, con un nivel de activación fisiológica parecido al de reposo. Esta confusión es el cuello de botella del modelo y el motivo por el que Amusement queda siempre con el F1 más bajo (0,69).

El resto de errores se reparten sin un patrón claro. Stress se clasifica con un 91,6% de aciertos (9.129 de 9.966), y los 837 errores se dividen entre Baseline (533) y Amusement

(304). Baseline también funciona muy bien, con 16.536 de 17.515 ventanas correctas (94,4%). Solo 341 se confunden con Stress y 638 con Amusement.

Accuracy por sujeto

Ilustración 12. Accuracy por sujeto en Random Forest.



Elaboración propia.

El rendimiento del modelo varía mucho entre sujetos. Once de los quince superan el 80% de accuracy y seis sobrepasan el 90%. En el extremo superior, S16 alcanza un accuracy de 1,00, lo que significa que el modelo clasifica correctamente todas sus ventanas. En el extremo inferior, S8 se queda en 0,578 y S2 en 0,751.

S16 es un caso que merece atención. Un accuracy de 1,00 con evaluación LOSO, donde el modelo no vio nada de S16 durante el entrenamiento, parece demasiado bueno para ser cierto. Hay tres elementos que pueden hacerlo creíble. El primero es que los tres modelos (Random Forest, Gradient Boosting y Regresión Logística) coinciden en clasificarlo casi perfectamente, lo que descarta un artefacto específico de un algoritmo. El segundo es la propia validación LOSO, que garantiza que S16 nunca participó en el entrenamiento de su fold. Y el tercero, que es la explicación fisiológica más probable, S16 parece ser un sujeto con respuestas fisiológicas extremadamente diferenciadas entre estados, alguien que responde al estrés de forma muy marcada. En una muestra de 15 personas es perfectamente posible que uno o dos individuos presenten este perfil.

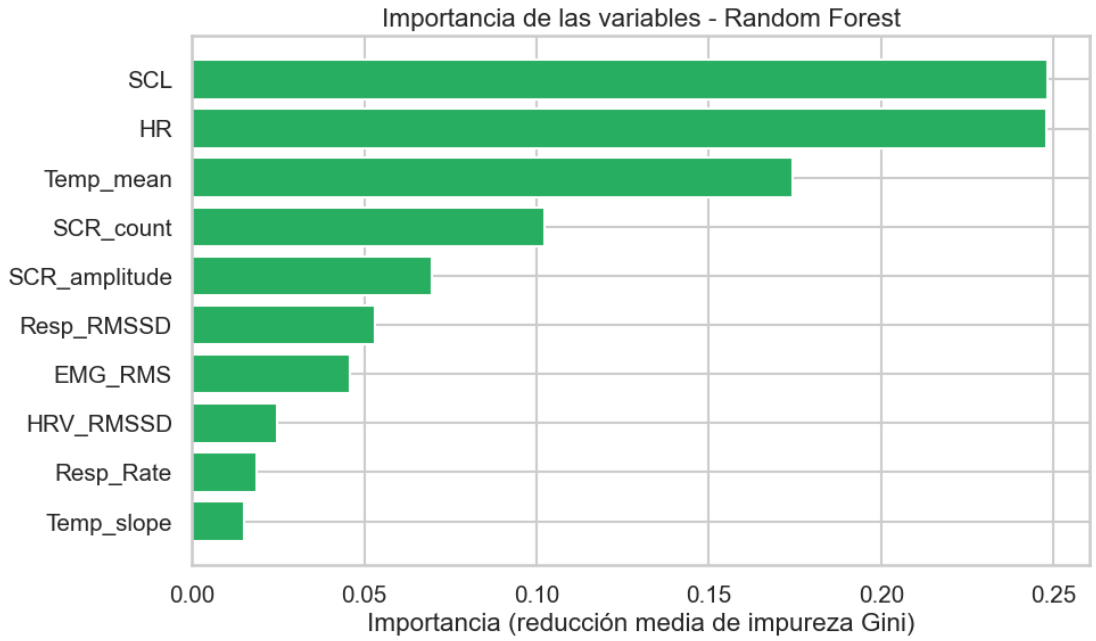
S8 es el caso opuesto. Sus respuestas fisiológicas al estrés son atípicas incluso en términos relativos a su propia línea base, y por eso ningún modelo (ni lineal ni no lineal) consigue clasificarlo bien. Este tipo de variabilidad es habitual en estudios con señales biométricas y representa un límite inherente al enfoque. Un modelo genérico no se adapta igual a

todos los individuos, lo que subraya la necesidad de personalización cuando se quiera llevar este tipo de sistema a un producto real.

Todos los sujetos superan ampliamente el umbral del azar (33%), lo que confirma que el modelo aporta valor predictivo incluso en los casos más difíciles.

Importancia de las variables

Ilustración 13. Importancia de variables Random Forest.



Elaboración propia.

Las dos variables más importantes son SCL (0,248) y HR (0,248), prácticamente empatadas. Juntas explican casi el 50% de la capacidad predictiva del modelo. La conductancia electrodérmica refleja directamente la activación del sistema nervioso simpático a través de la sudoración, y la frecuencia cardíaca mide la respuesta cardiovascular al estrés. Que ambas tengan una importancia casi idéntica indica que el modelo se apoya de forma equilibrada en dos ejes fisiológicos complementarios.

En tercer lugar, Temp_mean aporta un 17% adicional. Los cambios en la temperatura cutánea reflejan la vasoconstricción periférica que provoca el estrés, una señal más lenta que HR y SCL pero que captura información diferente. Estas tres variables juntas suman el 67% de la importancia total.

Las variables electrodérmicas en su conjunto (SCL, SCR_count y SCR_amplitude) acumulan un 42% de importancia. Eso convierte a la EDA en el eje fisiológico más relevante para detectar estrés en este modelo, por encima incluso de las variables cardiovasculares.

En el otro extremo, HRV_RMSSD (0,025), Resp_Rate (0,019) y Temp_slope (0,015) quedan en posiciones muy bajas. Coincide con lo observado en los boxplots del análisis exploratorio, donde estas tres variables mostraban la menor separación visual entre estados. Su contribución no es nula (el modelo las sigue usando en combinación con el resto para resolver casos ambiguos), pero su peso individual es pequeño.

El ranking tiene todo el sentido fisiológico. El estrés activa el sistema nervioso simpático, lo que se manifiesta principalmente en aumento de sudoración (SCL, SCR), aceleración cardíaca (HR) y vasoconstricción periférica (Temp_mean). Esas son precisamente las variables que el modelo ha identificado como más informativas.

7.1.8. Comparación con Schmidt et al. (2018)

El paper original de WESAD no plantea exactamente el mismo enfoque que este trabajo. Schmidt et al. construyen un benchmark amplio con el objetivo de comparar muchas combinaciones posibles de sensores y algoritmos. Evalúan por separado señales de muñeca, de pecho, combinaciones de ambas y distintos subconjuntos fisiológicos, y sobre cada configuración prueban cinco clasificadores: Decision Tree, Random Forest, AdaBoost, LDA y KNN. La estrategia de validación es la misma que se ha usado aquí (LOSO), lo que sí permite una comparación razonable aunque no completamente directa (Schmidt et al., 2018).

En el planteamiento de Schmidt et al. la segmentación se hace con ventanas de 60 segundos y desplazamiento de 0,25 segundos, y a partir de ahí extraen un conjunto muy amplio de variables por señal. No se limitan a medidas simples como frecuencia cardíaca o conductancia media, sino que añaden bandas frecuenciales de HRV, percentiles, rangos, pendientes, relaciones entre componentes, energía espectral y otros descriptores estadísticos y frecuenciales para ECG, EDA, EMG, respiración, temperatura y acelerometría. Su enfoque da mucho peso a la amplitud del conjunto de features y a la comparación entre modalidades (Schmidt et al., 2018).

Sus mejores resultados para la tarea de tres clases se obtienen con AdaBoost sobre todas las modalidades del dispositivo de pecho, alcanzando un 80,34% de accuracy. En ese mismo escenario, Random Forest se queda en 71,37%. Cuando se combinan pecho y muñeca, Random Forest sube hasta el 74,97%. La conclusión del paper es que las señales torácicas son las que mejor funcionan dentro de WESAD y que Amusement es la clase más difícil de separar (Schmidt et al., 2018).

El enfoque de este trabajo es distinto. No se ha intentado montar un benchmark del dataset, sino un pipeline más dirigido y más limpio desde el punto de vista analítico. Se trabaja solo con RespiBAN, la fuente más precisa de WESAD. A partir de esas señales se genera solo un conjunto reducido de 10 variables fisiológicas con un paso de ventana de 1 segundo que genera 33.056 ventanas para entrenar.

La diferencia más importante entre ambos trabajos está en cómo se preparan los datos antes de entrenar. Aquí las diez variables se normalizan por sujeto con z-score. Este paso reduce el peso de las diferencias basales entre individuos y hace que el clasificador se centre mejor en los cambios fisiológicos asociados al estado afectivo. Schmidt et al. no aplican una normalización de este tipo en su benchmark.

Hay también otras diferencias metodológicas que ayudan a explicar la mejora. Aquí se corrige el desbalanceo entre clases con class_weight="balanced" en Random Forest y Regresión Logística, y con sample_weight en Gradient Boosting. Además, en Random Forest la búsqueda de hiperparámetros se realiza dentro de cada iteración LOSO con validación anidada, de modo que el sujeto de test no influye en ninguna decisión del modelo. Schmidt et al., en cambio, usan configuraciones fijas y estándar, con un enfoque más orientado a comparar métodos que a optimizar uno solo.

Con este pipeline, el Random Forest final alcanza un 87,9% de accuracy global, por encima del 80,34% del benchmark original para la tarea de tres clases. La comparación no es perfectamente directa, porque los dos trabajos no usan las mismas ventanas, ni las mismas features, ni la misma estrategia de ajuste. Lo que se consigue demostrar con este trabajo es que la forma en que se representan y se normalizan las variables puede ser incluso más decisiva que aumentar el número de ellas, y que un conjunto pequeño de variables bien elegidas y preprocesadas puede rendir mejor que uno más amplio tratado de forma genérica.

7.2. Dashboard interactivo

7.2.1. Objetivo y contexto

La última pieza del proyecto es un dashboard interactivo que permite explorar de forma visual todos los datos procesados durante el ETL y los resultados del modelo predictivo. Antes de entrar en los detalles, conviene aclarar qué es y qué no es este dashboard dentro del contexto del trabajo.

Este dashboard es un prototipo funcional, no un producto final. Está construido con los datos del dataset WESAD y su objetivo principal es demostrar que el pipeline analítico completo funciona de inicio a fin, desde las señales brutas hasta una visualización interactiva que un usuario pueda explorar. En un escenario real, un dashboard de este tipo sería solo una de las capas de presentación del sistema, y su contenido cambiaría significativamente respecto a lo que se muestra aquí.

En concreto, la versión actual incluye elementos que tienen sentido en un contexto académico pero que no aparecerían en un producto dirigido a familias o cuidadores. Los porcentajes de precisión del modelo, por ejemplo, están aquí porque permiten evaluar y defender el modelo, pero en la vida real no se etiquetaría externamente al usuario como se hace en WESAD, así que no habría con qué comparar las predicciones. Lo mismo ocurre con los gráficos detallados de cada variable fisiológica, que son muy útiles para el análisis pero que a un usuario sin formación médica no le transmitirían demasiada información.

Lo que sí tendría sentido en una versión final orientada al usuario son las tarjetas de indicadores principales (frecuencia cardíaca, temperatura, nivel de estrés) y sobre todo el gráfico de evolución del Stress Index, que resume el estado del sujeto en una escala de 0 a 10 fácil de interpretar. Pero todo eso se aborda con más detalle en el apartado de trabajo futuro.

El dashboard está desplegado de forma pública y se puede consultar en cualquier momento a través de la URL incluida en el anexo del trabajo. Las capturas que se incluyen en los siguientes apartados muestran el estado del dashboard en el momento de la entrega, pero si se quiere interactuar con los datos (cambiar de sujeto, mover el slider, seleccionar distintas señales) es mucho más útil acceder directamente al enlace.

7.2.2. Arquitectura del dashboard

El dashboard está construido con Streamlit, y parte de dos archivos CSV generados en la fase de carga del ETL: el dashboard_dataset.csv (con todas las variables derivadas, el Stress Index y los estados afectivos) y el model_predictions.csv (con las predicciones del modelo para cada ventana). Los combina en una sola tabla por sujeto y ventana. De esta forma cada fila tiene, para una ventana concreta, tanto los valores fisiológicos reales como la predicción del modelo, lo que permite comparar ambos en la misma interfaz.

Uno de los pasos más relevantes de la preparación de los datos para el dashboard es la transformación del Stress Index a una escala comprensible para el usuario. El Stress Index original es un z-score que no resulta intuitivo para alguien que no está familiarizado con el concepto. Para hacerlo más accesible, en el dashboard se reescala a una escala de 0 a 10 mediante un proceso de dos pasos. Primero se recorta el valor original al rango [-2, +2] y después se aplica un mapeo lineal donde -2 pasa a ser 0, el 0 pasa a ser 5 y +2 pasa a ser 10. El Stress Index, al ser una media de z-scores, tiene por media 0 y una distribución aproximadamente normal. En cualquier distribución normal, el 95% de los valores cae entre -2 y +2 desviaciones típicas. Recortar ahí significa que la escala 0-10 cubre prácticamente toda la variación real del indicador, y los pocos valores que quedan fuera se asignan al extremo más cercano (0 o 10) sin perder información relevante. De esta forma, un Stress Index de 5/10 representa el nivel medio de activación del sujeto, valores por encima indican más activación de lo habitual y valores por debajo indican menos.

Es importante aclarar que esta transformación es exclusivamente visual. No modifica los datos originales ni afecta a ningún cálculo, simplemente traduce el z-score a un formato que un usuario cualquiera pueda interpretar sin explicaciones adicionales. Los umbrales elegidos son ilustrativos y coherentes con la distribución estadística de los datos, pero en un producto real necesitarían validación clínica. Dispositivos como Whoop u Oura aplican transformaciones similares para ofrecer escalas fisiológicas simplificadas.

Otro aspecto que relevante es la gestión del color. Todos los colores del dashboard están definidos en un diccionario centralizado al principio del código, de manera que cada estado afectivo tiene siempre el mismo color en todos los gráficos (azul para Baseline, rojo para Stress, verde para Amusement, gris para las transiciones y la meditación). Esto facilita la lectura visual y da coherencia a toda la interfaz.

Los perfiles de los sujetos (edad, género, altura, peso) se han extraído manualmente de los archivos readme del dataset y se han incluido directamente en el código como un diccionario. Dado que son solo 15 participantes y los datos nunca cambian, almacenarlos así evita tener que incluir archivos adicionales en el proyecto y simplifica el despliegue.

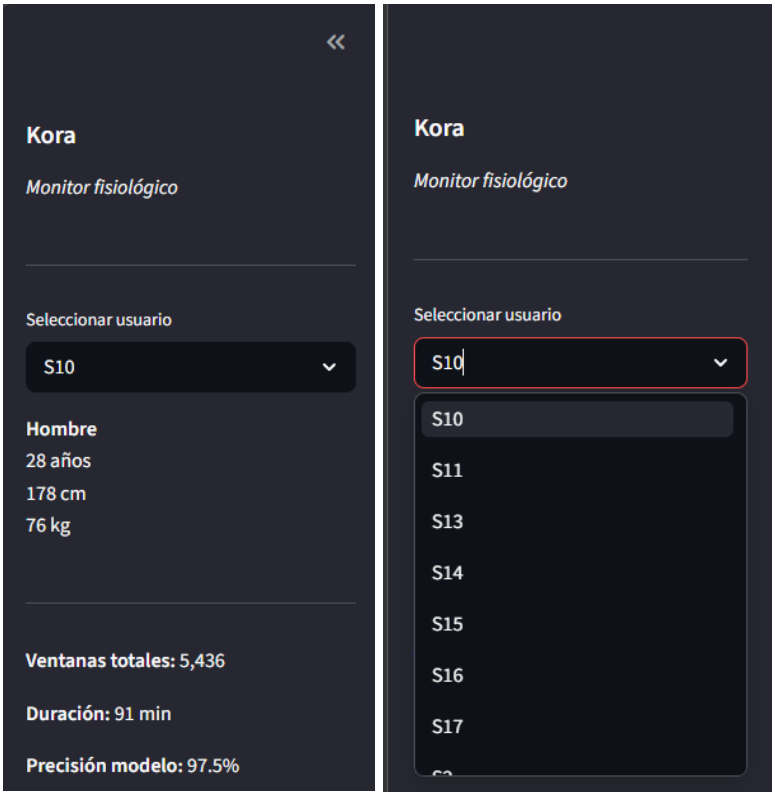
7.2.3. Diseño y componentes del dashboard

El dashboard se organiza en una barra lateral de navegación y un área principal de contenido. A continuación se describe cada bloque en el orden en el que aparece en la interfaz.

Panel lateral: selector de usuario y perfil del sujeto

La barra lateral muestra el nombre del proyecto (Kora) y un desplegable que permite seleccionar cualquiera de los 15 sujetos del dataset. Al elegir un sujeto, se carga automáticamente su perfil (género, edad, altura, peso) extraído de los archivos readme del dataset original. Debajo del perfil aparecen tres métricas generales del sujeto seleccionado: el número total de ventanas disponibles, la duración del registro en minutos y la precisión que el modelo predictivo ha obtenido para ese sujeto concreto.

Ilustración 14. Panel desplegable del dashboard.



Elaboración propia.

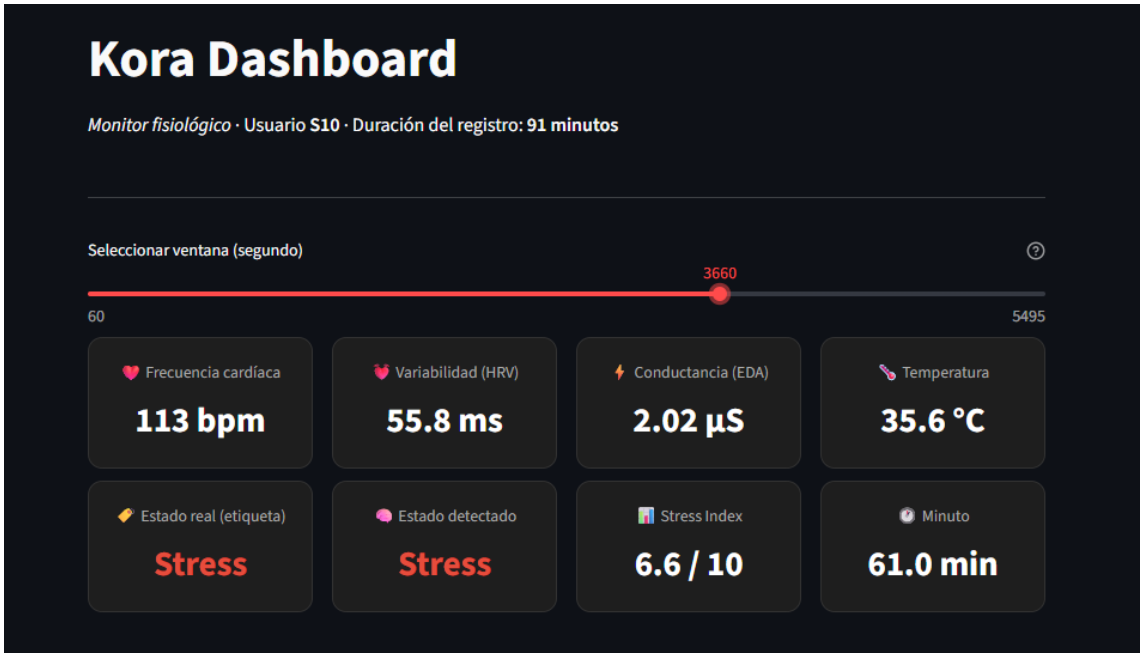
Cabecera y tarjetas de indicadores

La parte superior del área principal muestra el título del dashboard, el sujeto seleccionado y la duración de su registro. Justo debajo hay un slider que permite navegar segundo a segundo por todo el experimento, seleccionando la ventana temporal que se quiere inspeccionar.

A continuación, aparecen ocho tarjetas organizadas en dos filas. La primera fila muestra los cuatro indicadores fisiológicos principales de la ventana seleccionada: frecuencia cardíaca (en bpm), variabilidad cardíaca (HRV en milisegundos), conductancia electrodérmica (EDA en microsiemens) y temperatura cutánea (en grados Celsius). Estos valores cambian en tiempo real conforme se mueve el slider, lo que permite observar cómo evolucionan las variables a lo largo del experimento.

La segunda fila de tarjetas muestra información de contexto: el estado afectivo real (la etiqueta del protocolo experimental), el estado detectado por el modelo predictivo, el Stress Index en una escala de 0 a 10 y el minuto del experimento correspondiente a la ventana seleccionada. Las tarjetas de estado se colorean según la condición (azul para Baseline, rojo para Stress, verde para Amusement), lo que permite ver de un vistazo si el modelo coincide con la etiqueta real.

Ilustración 15. Cabecera del dashboard con slider y tarjetas KPI (Sujeto 10).



Elaboración propia.

Evolución del Stress Index y estados afectivos

Este es probablemente el gráfico más importante del dashboard y el que tendría más sentido en una versión final del producto. Muestra tres capas de información en un solo panel.

Ilustración 16. Gráfico de evolución del Stress Index (Sujeto 10).

Elaboración propia.

En la parte central hay una línea blanca que representa la evolución del Stress Index a lo largo del tiempo, en una escala de 0 a 10. Valores altos indican mayor activación fisiológica y valores bajos indican un estado más relajado. La línea discontinua gris marca el punto medio (5/10) como referencia visual.

En la parte superior del gráfico hay una franja horizontal de color que muestra el estado afectivo real (la etiqueta que el protocolo del experimento asigna a cada momento). En la parte inferior aparece otra franja con el estado que el modelo predictivo ha detectado. Donde los colores de ambas franjas coinciden, el modelo ha acertado. Donde difieren, ha fallado. Esto permite evaluar visualmente la calidad de las predicciones sin necesidad de mirar números.

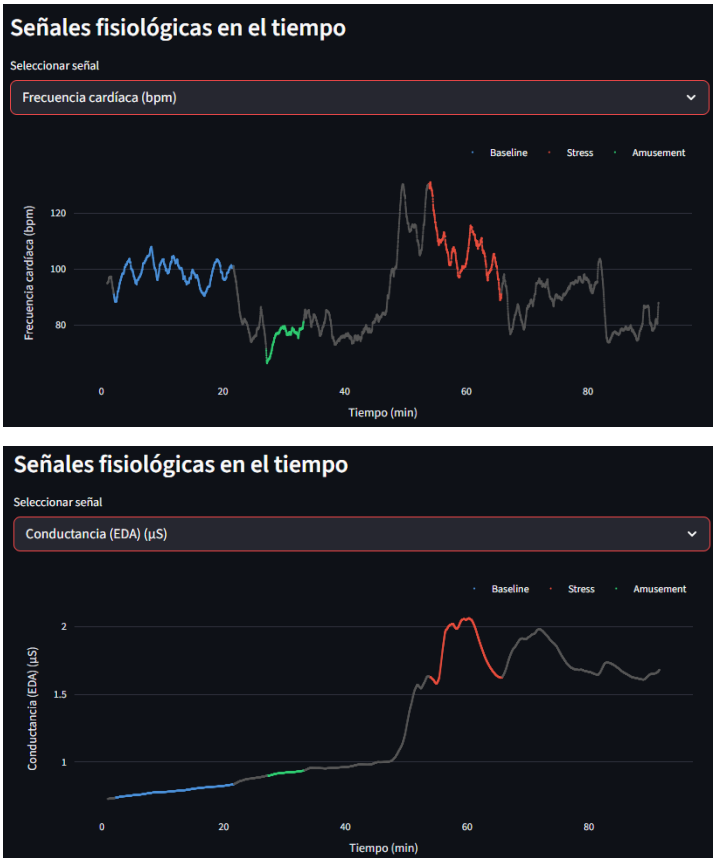
Una línea vertical roja marca la posición de la ventana seleccionada con el slider, lo que permite localizar en qué punto del experimento se encuentra el usuario en cada momento.

Al explorar el gráfico se observa que Stress Index tiende a ser más alto durante las ventanas etiquetadas como Stress y más bajo durante las ventanas de Baseline y Amusement. Esto confirma que el indicador cumple su función y captura correctamente la diferencia de activación fisiológica entre estados, lo que lo convierte en una buena herramienta de síntesis para el usuario final. No es un clasificador en sí mismo (para eso está el modelo), pero ofrece una lectura continua y sencilla del nivel de estrés que resulta mucho más fácil de interpretar que cualquier variable fisiológica por separado.

Señales fisiológicas en el tiempo

A continuación, hay un gráfico de líneas interactivo que permite explorar la evolución temporal de cualquier variable fisiológica a lo largo del experimento. Un desplegable permite seleccionar la señal que se quiere visualizar (frecuencia cardíaca, variabilidad, conductancia EDA, frecuencia respiratoria, temperatura, etc.).

Ilustración 17. Gráficos de señales fisiológicas (Frecuencia cardíaca y Conductancia EDA; Sujeto 10).



Elaboración propia.

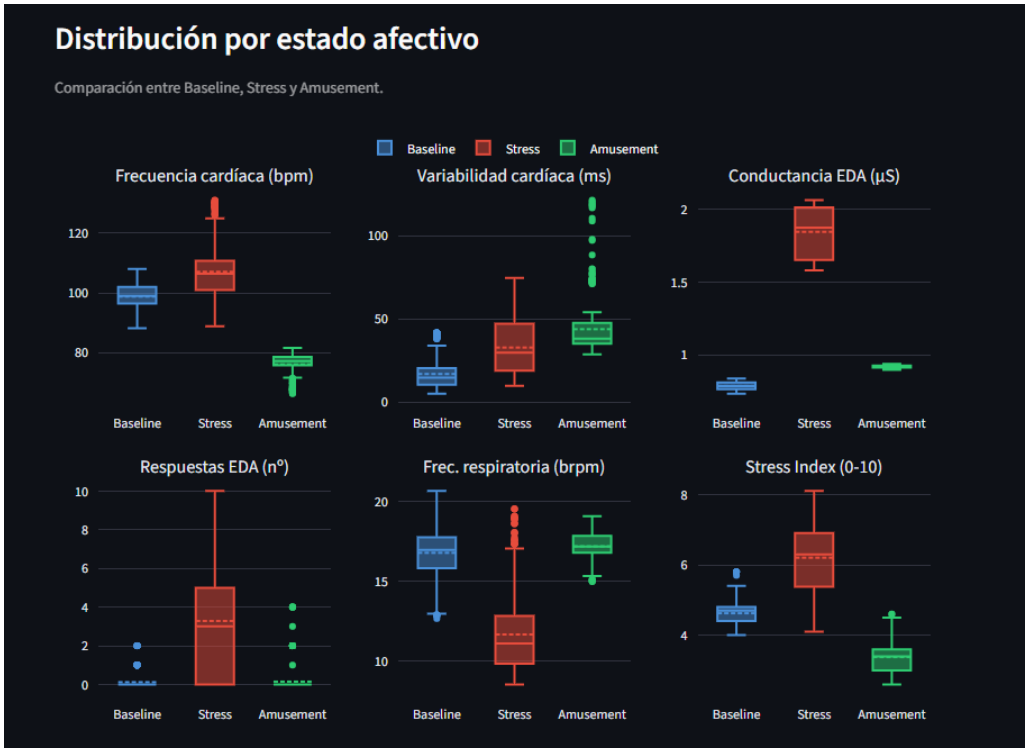
La señal aparece representada con puntos coloreados según el estado afectivo de cada ventana (azul para Baseline, rojo para Stress, verde para Amusement). Los tramos correspondientes a transiciones o meditación se muestran en gris. De fondo se dibuja una línea gris continua que conecta todos los puntos para dar continuidad visual y evitar las rectas artificiales que aparecerían al unir directamente puntos de estados distintos.

Distribución por estado afectivo

Este bloque muestra seis boxplots que comparan la distribución de las principales variables fisiológicas entre los tres estados afectivos (Baseline, Stress y Amusement) para el sujeto seleccionado. Las variables representadas son la frecuencia cardíaca, la variabilidad cardíaca, la conductancia EDA, el número de respuestas electrodérmicas, la frecuencia respiratoria y el Stress Index.

Los boxplots son especialmente útiles para ver de un vistazo si las distribuciones se separan bien entre estados. En un sujeto donde el modelo funciona bien, como S10, se observan diferencias claras entre los tres estados en la mayoría de variables. La frecuencia cardíaca, la EDA y el Stress Index presentan distribuciones claramente separadas, lo que explica por qué el modelo consigue un 97,5% de precisión en ese sujeto.

Ilustración 18. Boxplots de distribución por estado afectivo (Sujeto 10).



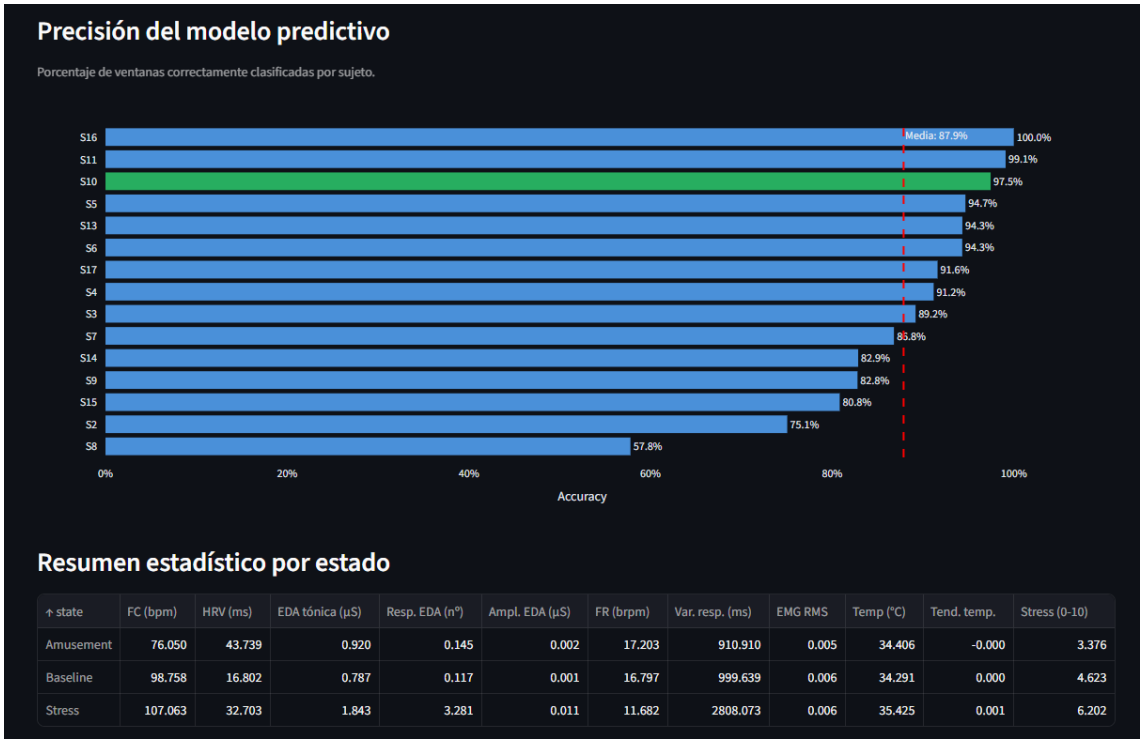
Elaboración propia.

Precisión del modelo predictivo y resumen estadístico

El penúltimo bloque del dashboard muestra un gráfico de barras horizontales con la precisión del modelo para cada uno de los 15 sujetos, ordenados de mayor a menor accuracy. El sujeto seleccionado aparece resaltado en verde para que sea fácil localizarlo. Una línea roja discontinua marca la media global del modelo (87,9%).

Debajo del gráfico de precisión aparece una tabla resumen con los valores medios de todas las variables fisiológicas agrupados por estado afectivo. Esta tabla permite comparar numéricamente los niveles medios de cada métrica entre Baseline, Stress y Amusement para el sujeto seleccionado.

Ilustración 19. Gráfico de precisión por sujeto y tabla resumen (Sujeto 10).



Resumen estadístico por estado

state	FC (bpm)	HRV (ms)	EDA tónica (µS)	Resp. EDA (nº)	Ampl. EDA (µS)	FR (brpm)	Var. resp. (ms)	EMG RMS	Temp (°C)	Tend. temp.	Stress (0-10)
Amusement	76.050	43.739	0.920	0.145	0.002	17.203	910.910	0.005	34.406	-0.000	3.376
Baseline	98.758	16.802	0.787	0.117	0.001	16.797	999.639	0.006	34.291	0.000	4.623
Stress	107.063	32.703	1.843	3.281	0.011	11.682	2808.073	0.006	35.425	0.001	6.202

Elaboración propia.

7.2.4. Casos de uso y observaciones

El dashboard no solo sirve para mostrar datos, también permite descubrir patrones y anomalías que no se ven fácilmente mirando solo los números. A lo largo de la exploración de los datos se han identificado tres casos de uso que ilustran bien el valor de la visualización interactiva.

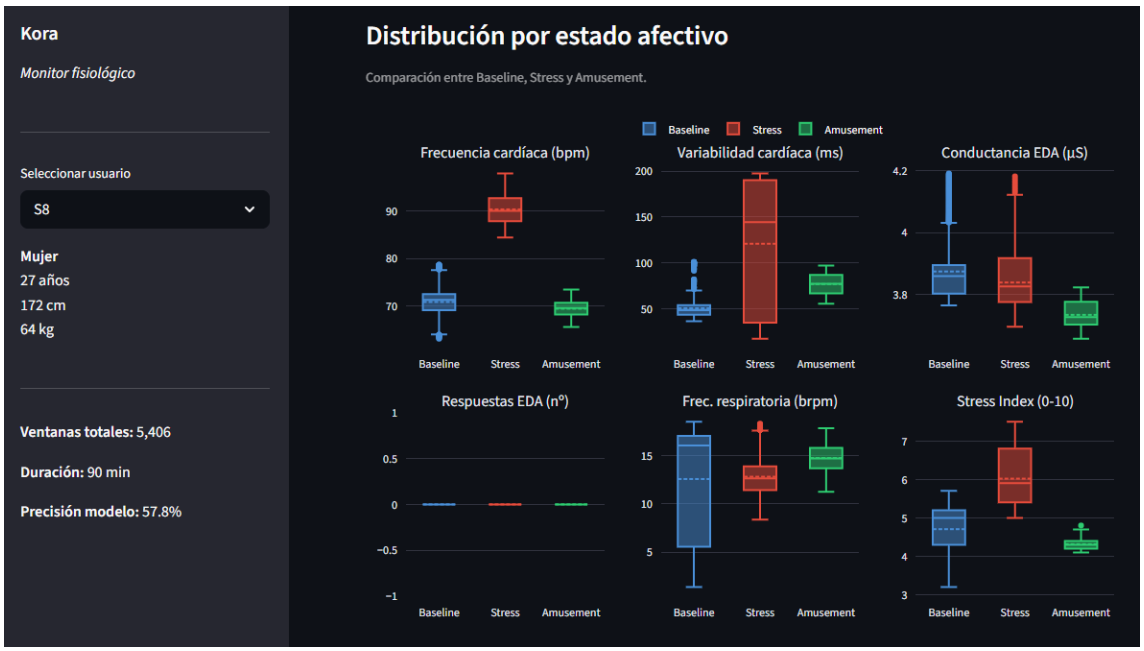
S10: el caso representativo.

El sujeto S10 es un buen ejemplo de cómo se comporta el sistema cuando todo funciona correctamente. Sus variables fisiológicas muestran diferencias claras entre los tres estados afectivos, el Stress Index (Ilustración 16). sube de forma evidente durante las fases de estrés y el modelo consigue una precisión del 97,5%. Los boxplots confirman que la frecuencia cardíaca, la EDA y la respiración se separan bien entre estados, dando al modelo suficiente información para clasificar correctamente casi todas las ventanas (Ilustración 18).

S8: el caso problemático

El sujeto S8 es el que peor accuracy tiene de todo el dataset (57,8%) y la exploración visual permite entender por qué. Al seleccionar S8 en el dashboard, lo primero que llama la atención es que los boxplots de respuestas EDA (SCR_count y SCR_amplitude) son completamente planos. Además, el sujeto no muestra ni una sola respuesta fásica de conductancia durante los tres estados válidos. La tabla resumen confirma que SCR_count y SCR_amplitude son exactamente 0 en Baseline, Stress y Amusement.

Ilustración 20. Boxplots de distribución por estado afectivo (Sujeto 8).



Elaboración propia.

Ilustración 21. Tabla resumen (Sujeto 8).

Resumen estadístico por estado

state	FC (bpm)	HRV (ms)	EDA tónica (µS)	Resp. EDA (nº)	Ampl. EDA (µS)	FR (brpm)	Var. resp. (ms)	EMG RMS	Temp (°C)	Tend. temp.	Stress (0-10)
Amusement	69.427	76.777	3.735	0.000	0.000	14.738	1724.436	0.006	35.012	0.000	4.314
Baseline	70.854	50.383	3.875	0.000	0.000	12.566	2625.202	0.007	34.793	0.001	4.705
Stress	90.301	120.738	3.839	0.000	0.000	12.803	2742.186	0.009	35.104	0.001	6.020

Elaboración propia.

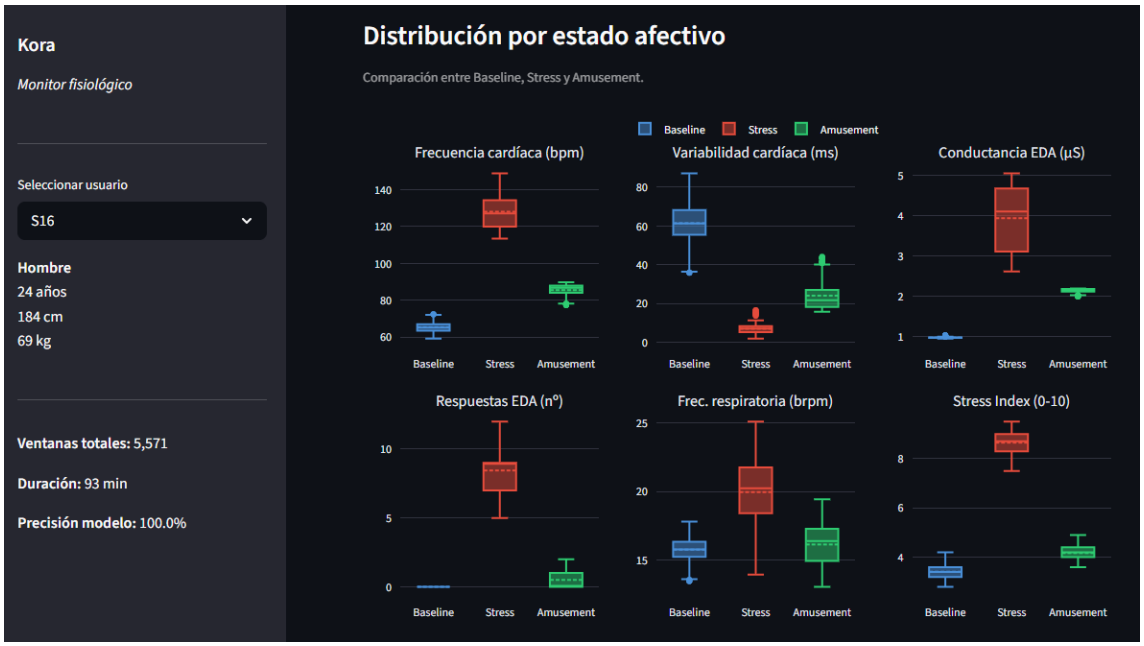
Esto no es un error del código ni del procesamiento. El archivo readme del sujeto S8 en el dataset original documenta que el participante reportó haber sentido frío durante el experimento. El frío inhibe la actividad de las glándulas sudoríparas, que son las responsables de generar las respuestas fásicas de la EDA. Sin sudoración no hay respuestas, y sin respuestas el modelo pierde una de sus herramientas de discriminación más potentes (la EDA en conjunto acumula un 42,0% de la importancia total del modelo según el análisis de feature importance). Es un caso real de variabilidad fisiológica individual que afecta directamente al rendimiento del sistema, y que demuestra lo importante que es poder inspeccionar los datos sujeto a sujeto.

Además, S8 también es el único sujeto con NaNs en las variables respiratorias (93 ventanas donde no se detectaron suficientes ciclos respiratorios), lo que refuerza la idea de que las condiciones de este participante durante el experimento no fueron del todo favorables para la recogida de datos.

S16: el caso ideal

En el extremo opuesto, el sujeto S16 alcanza un 100% de accuracy. Al explorar sus datos en el dashboard, se entiende rápidamente la razón. Los boxplots muestran una separación enorme entre los tres estados en prácticamente todas las variables. La frecuencia cardíaca de Stress está muy por encima de la de Baseline y Amusement, la variabilidad cardíaca se dispara durante el estrés, la EDA tónica sube de forma clara y el Stress Index separa los tres estados sin solapamiento. En estas condiciones, el modelo no tiene ninguna dificultad para clasificar cada ventana en su estado correcto.

Ilustración 22. Boxplots de distribución por estado afectivo (Sujeto 16).



Elaboración propia.

La comparación entre estos tres sujetos sirve para demostrar que la variabilidad entre personas es enorme. El mismo protocolo experimental genera respuestas fisiológicas muy distintas según el sujeto, y eso afecta directamente a la capacidad del modelo para clasificar correctamente. Poder visualizar esas diferencias de forma interactiva es precisamente el tipo de valor que aporta un dashboard como este, y es la base sobre la que se construiría una versión más orientada al usuario final en el futuro.

8. Conclusiones

8.1. Resultados e insights finales

Este trabajo ha conseguido construir un sistema completo que transforma señales fisiológicas brutas en información comprensible. Desde el archivo .pkl hasta el dashboard desplegado en la nube, de inicio a fin, todo el pipeline funciona, está documentado y se puede reproducir. Pero más allá de eso, hay conclusiones concretas que aplican a este trabajo que conviene destacar.

Se ha demostrado que es posible clasificar con precisión estados afectivos inducidos a partir de señales fisiológicas procesadas.

En las condiciones de este experimento, el modelo Random Forest alcanza un 87,9% de accuracy usando solo variables temporales sencillas (medias, variabilidades, pendientes). No se han utilizado features frecuenciales, que son las que Schmidt et al. emplearon para llegar a su 80%. Esto deja margen de mejora claro, pero también confirma que con un procesamiento relativamente simple ya se obtienen resultados útiles.

La actividad electrodérmica es la señal más importante para detectar estrés.

Las tres variables derivadas de la EDA (SCL, SCR_count y SCR_amplitude) acumulan un 42,0% de la importancia del modelo. Más que la frecuencia cardíaca, más que la respiración y más que la temperatura. Esto quiere decir que si el sensor de EDA falla o no está, el sistema pierde su mejor herramienta. El caso de S8 lo demuestra perfectamente.

Cada sujeto tiene sus niveles fisiológicos propios.

El mismo modelo con las mismas señales da un 100% en S16 y un 57,8% en S8. No es un fallo, es la realidad de los datos fisiológicos. El sujeto S8 sintió frío durante el experimento y eso anuló su respuesta electrodérmica. S16, en cambio, reaccionó de forma muy marcada a cada condición y sus variables se separan sin solaparse entre estados. Cada persona parte de un estado basal diferente. Por eso en este trabajo se ha normalizado por sujeto (mediante z-scores), para que el sistema mida cuánto se desvía cada persona respecto a sí misma y no respecto a una media que no le representa. Un sistema de monitorización real tiene que convivir con esta variabilidad y adaptarse a ella.

El Stress Index se presenta como indicador resumen prometedor.

A pesar de ser una variable estadística sencilla (media de z-scores normalizados por sujeto), captura bien la activación inducida por el protocolo en casi todos los sujetos. Sube durante el estrés, baja en reposo y diversión. Demuestra que es posible condensar varias señales en un número fácil de interpretar, que es exactamente lo que necesitaría un usuario final.

El prototipo es viable, pero no es un producto.

Todo lo construido en este trabajo (ETL, modelo, dashboard) funciona con 15 sujetos jóvenes monitorizados durante 36 minutos en un laboratorio. Para que esto sirva en la vida real hace falta probarlo con datos recogidos durante semanas en personas mayores, en sus casas, haciendo su vida normal. El dataset WESAD ha cumplido su papel como base experimental. El siguiente paso es salir de él.

8.2. Limitaciones

Sin embargo, el trabajo tiene limitaciones que hay que tener presentes a la hora de interpretar los resultados.

La más evidente es el tamaño del dataset. 15 sujetos con aproximadamente 36 minutos de datos válidos cada uno es suficiente para construir y probar un prototipo, pero no para extraer conclusiones generalizables. La muestra, además, es joven (media de 27,5 años) y mayoritariamente masculina (12 hombres y 3 mujeres), lo cual no se parece en nada al perfil de usuario final de Kora, que serían personas mayores de 65 años con perfiles de salud mucho más diversos.

Otra limitación importante es el contexto experimental. Los datos se recogieron en un laboratorio, con un protocolo controlado que induce estrés de forma artificial. En la vida real el estrés no aparece en bloques de 10 minutos perfectamente delimitados, se mezcla con la actividad física, con el cansancio, con el dolor y con cien factores más que el modelo no ha visto nunca. No se sabe cómo se comportaría el sistema en esas condiciones, y averiguarlo requiere datos reales que todavía no existen.

También hay que ser honesto con la decisión de usar solo señales de pecho. Da mejores resultados que la muñeca, pero un producto comercial real probablemente llevaría el sensor en la muñeca, donde la calidad de la señal podría ser menor a pesar de los avances

en esta tecnología. Los resultados obtenidos aquí son un techo optimista de lo que se podría conseguir con un wearable de consumo.

Además, hay una limitación metodológica en la forma en que se ha realizado la normalización. En este trabajo, las variables se estandarizaron por sujeto utilizando la media y la desviación típica calculadas sobre toda su serie temporal. Esto mejora la capacidad del modelo para detectar cambios entre estados. Sin embargo, en una aplicación real este procedimiento no podría hacerse de la misma forma, ya que al analizar a un usuario nuevo no se dispondría desde el principio de toda su señal completa. Sería necesario contar con un periodo inicial de calibración o estimar sus valores basales de forma progresiva e ir actualizándolos con el uso. Por tanto, esta decisión puede contribuir a que la precisión obtenida por este modelo sea superior a la reportada por Schmidt et al. (2018), junto con otras diferencias de preprocesado y modelización. En consecuencia, los resultados deben interpretarse como válidos dentro de este entorno experimental, pero no como una simulación exacta de despliegue real.

9. Trabajo futuro

Este TFG demuestra que convertir señales brutas en información útil es posible, pero también deja claro todo lo que queda por hacer para que eso tenga impacto real. Las siguientes líneas resumen hacia dónde iría el proyecto si se continuase desarrollando.

9.1. Extracción de bases de datos reales

El roadmap de Kora, definido en el TFG de ADE complementario a este mismo proyecto, plantea una primera fase de validación en residencias geriátricas donde se desplegarían los wearables con entre 150 y 200 usuarios mayores de 65 años durante al menos seis meses. Esos datos serían el primer dataset propio de Kora con la población objetivo real, algo que el dataset WESAD (joven, homogéneo, 36 minutos por sujeto) no puede ofrecer.

Ese volumen de datos permitiría entrenar modelos específicos para población mayor, donde las respuestas fisiológicas son distintas (menor variabilidad cardíaca basal, señales PPG más débiles por menor perfusión cutánea, patrones de actividad diferentes). También permitiría explorar la personalización por usuario, es decir, entrenar un modelo base con datos generales y después ajustarlo con el historial de cada persona. Así el sistema aprendería lo que es normal para cada uno, y las alertas se dispararían cuando algo se desviase de ese patrón individual, no de una media poblacional.

9.2. Evolución del modelo

A nivel técnico, la mejora más directa sería extraer más información de las señales. En este trabajo se han calculado métricas de resumen por ventana (medias, variabilidades, pendientes), pero las señales también contienen patrones rítmicos que se pueden analizar descomponiéndolas en sus frecuencias, algo que Schmidt et al. sí hicieron en su estudio original y que contribuyó a su 80% de accuracy. Incorporar este tipo de análisis sería la vía más rápida para mejorar el rendimiento sin necesidad de cambiar de algoritmo.

A más largo plazo, si el volumen de datos crece lo suficiente, tendría sentido probar modelos de deep learning que trabajen directamente sobre las señales brutas. Pero con datasets pequeños los modelos clásicos siguen ganando en rendimiento e interpretabilidad, y la interpretabilidad importa mucho con datos de salud.

El modelo actual clasifica el estado del sujeto en un instante concreto, pero no detecta tendencias a largo plazo. Una frecuencia cardíaca en reposo que sube medio punto cada

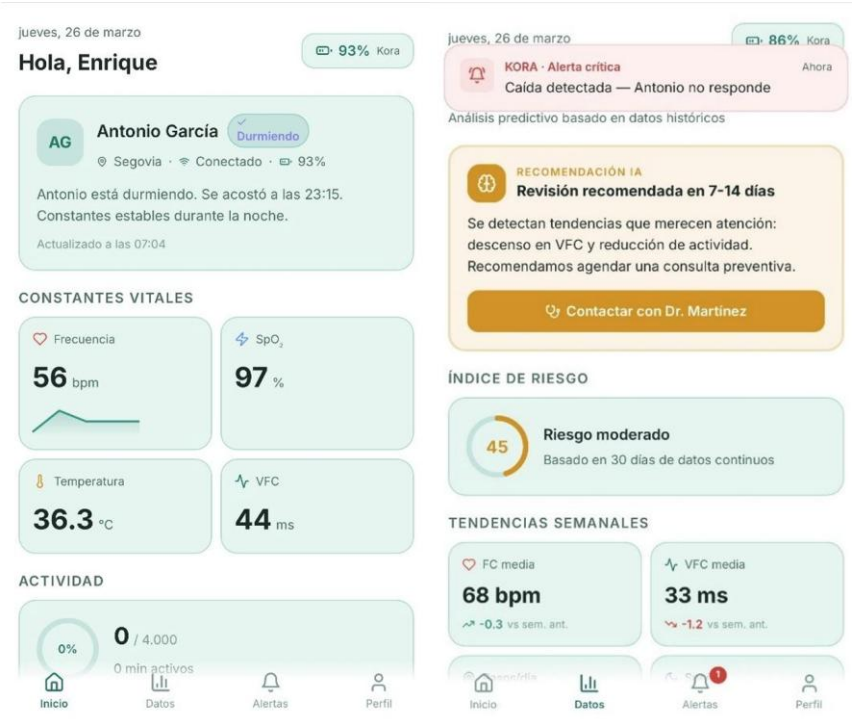
semana durante tres meses podría ser una señal temprana de un problema cardiovascular, y un sistema de monitorización debería ser capaz de captarla. Añadir esa capacidad de detección de patrones es clave e incrementaría mucho el valor del sistema.

9.3. De un dashboard técnico a un producto para familias

El dashboard actual muestra boxplots, porcentajes de accuracy y variables fisiológicas con nombres técnicos. Todo eso tiene sentido para evaluar el sistema en un contexto académico, pero no es lo que debería ver un familiar que quiere saber cómo está su padre.

En el TFG de ADE se ha diseñado cómo sería la interfaz real de Kora para familias. La pantalla principal no mostraría gráficos de conductancia electrodérmica ni valores de RMSSD. Mostraría un mensaje directo del tipo "Antonio está tranquilo. Constantes estables durante la noche", acompañado de unos pocos indicadores visuales (frecuencia cardíaca, temperatura, nivel de actividad) presentados con colores e iconos que cualquier persona pueda entender sin formación médica. El Stress Index desarrollado en este trabajo sí podría integrarse como un indicador de bienestar general, validado clínicamente y presentado de forma comprensible, similar al "45, Riesgo Moderado" que aparece en el prototipo de interfaz diseñado para el TFG de ADE.

Ilustración 23. Mock-up de la interfaz de la app de Kora para familias.



Elaboración propia.

El sistema de alertas también sería muy distinto. En lugar de mostrar datos y dejar que el familiar los interprete, Kora clasificaría las notificaciones en tres niveles. Las urgentes (una caída confirmada, una arritmia) requerirían acción inmediata. Las de atención (frecuencia cardíaca elevada de forma sostenida) indicarían que conviene estar pendiente. Las de observación (menos actividad durante varias semanas, sueño más fragmentado) serían información para comentar con el médico en la próxima visita, no para una llamada de pánico. Esa jerarquía evita que la familia viva en alerta permanente, que es uno de los mayores problemas de los sistemas de teleasistencia actuales.

Otra pieza clave serían los informes mensuales automatizados que se enviarían al familiar por correo, resumiendo la evolución del mayor durante el último mes. Estos informes se podrían compartir directamente con un médico, ofreciendo datos objetivos y continuos que complementan lo que normalmente se recoge en una consulta de quince minutos.

Toda la complejidad analítica desarrollada en este TFG (el ETL, el modelo, las variables derivadas) quedaría detrás, procesándose en la nube, invisible para el usuario. Lo que llegaría a la familia sería información limpia y accionable. Esa es la diferencia entre un prototipo académico y un producto de salud.

Este TFG ha sentado la base analítica de un proyecto que empieza aquí pero que no termina aquí. Las señales se pueden procesar, los modelos funcionan y la información se puede presentar de forma comprensible. Ahora falta lo más difícil, probarlo con datos reales, con personas reales y durante el tiempo suficiente como para que el sistema demuestre que puede aportar valor real al usuario. Esa es la siguiente etapa.

10. Bibliografía

Escobar-Linero, E., Muñoz-Saavedra, L., Luna-Perejón, F., Sevillano, J. L., & Domínguez-Morales, M. (2023). Wearable Health Devices for Diagnosis Support: Evolution and Future Tendencies. *Sensors, 23*(3), 1678. https://doi.org/10.3390/s23031678

Pérez Díaz, J., Castillo Belmonte, A. B., Aceituno Nieto, P., & Ramiro Fariñas, D. (2024). *Un perfil de las personas mayores en España, 2024: Indicadores estadísticos básicos* (Informes Envejecimiento en Red, nº 33). Consejo Superior de Investigaciones Científicas (CSIC). https://envejecimientoenred.csic.es/wp-content/uploads/2024/12/enred-indicadoresbasicos2024.pdf

Schmidt, P., Reiss, A., Dürichen, R., Van Laerhoven, K., & Plötz, T. (2018). Introducing WESAD, a multimodal dataset for wearable stress and affect detection. *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI 2018),* 400–408. https://doi.org/10.1145/3242969.3242985

Ubiquitous Computing Lab. (2018). *WESAD: Wearable Stress and Affect Detection Dataset* [Data set]. University of Siegen. https://ubi29.informatik.uni-siegen.de/usi/data_wesad.html

La Porta, N., Oldano, G., Puiatti, A., Leidi, T., & Papandrea, M. (2025). *Toward unbiased emotion recognition: overcoming user bias with siamese convolutional networks*. Signal, Image and Video Processing, 19, 875. https://doi.org/10.1007/s11760-025-04500-1

Python Software Foundation. (s. f.). *pickle - Python object serialization*. https://docs.python.org/3/library/pickle.html

McKinney, W. (2010). *Data structures for statistical computing in Python.* In S. van der Walt & J. Millman (Eds.), Proceedings of the 9th Python in Science Conference (pp. 56–61). https://doi.org/10.25080/Majora-92bf1922-00a

Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020*). Array programming with NumPy.* Nature, 585(7825), 357-362. https://doi.org/10.1038/s41586-020-2649-2

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. A. (2021). *NeuroKit2: A Python toolbox for neurophysiological*

signal processing. Behavior Research Methods, 53(4), 1689–1696. https://doi.org/10.3758/s13428-020-01516-y

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python.* Journal of Machine Learning Research, 12, 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment*. Computing in Science & Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Waskom, M. L. (2021). *seaborn: Statistical data visualization.* Journal of Open Source Software, 6(60), 3021. https://doi.org/10.21105/joss.03021

Plotly Technologies Inc. (2015). *Collaborative data science*. Plotly Technologies Inc. https://plotly.com/chart-studio-help/citations/

Snowflake Inc. (n.d.). Streamlit documentation. Streamlit. https://docs.streamlit.io/

Herborn, K. A., Graves, J. L., Jerem, P., Evans, N. P., Nager, R., McCafferty, D. J., & McKeegan, D. E. F. (2015). Skin temperature reveals the intensity of acute stress. *Physiology & Behavior, 152*(Pt A), 225–230. https://doi.org/10.1016/j.physbeh.2015.09.032

Stuyck, H., Dalla Costa, L., Cleeremans, A., & Van den Bussche, E. (2022). *Validity of the Empatica E4 wristband to estimate resting-state heart rate variability in a lab-based context* (Short version) [Technical report]. KU Leuven / Université libre de Bruxelles. https://lirias.kuleuven.be/retrieve/681505

Posada-Quintero, H. F., Florian, J. P., Orjuela-Cañón, A. D., & Chon, K. H. (2018). Electrodermal activity is sensitive to cognitive stress under water. *Frontiers in Physiology, 8*, 1128. https://doi.org/10.3389/fphys.2017.01128

Noble, R. J., Hillis, J. S., & Rothbaum, D. A. (1990). *Electrocardiography*. In H. K. Walker, W. D. Hall, & J. W. Hurst (Eds.), *Clinical Methods: The History, Physical, and Laboratory Examinations* (3rd ed.). Butterworths. NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK354/

Vincent, R. (2022). *From a laboratory to the wearables: A review on history and evolution of electrocardiogram*. *Iberoamerican Journal of Medicine, 4*(4), 248–255. https://iberoamjmed.com/article/doi/10.53986/ibjm.2022.0038

Lee, E. M. (2024). *When and how to use ambulatory blood pressure monitoring and home blood pressure monitoring for managing hypertension*. *Clinical Hypertension, 30*, 10. https://clinicalhypertension.org/DOIx.php?id=10.1186/s40885-024-00265-w

Jamieson, A., Chico, T. J. A., Jones, S., Chaturvedi, N., Hughes, A. D., & Orini, M. (2025). *A guide to consumer-grade wearables in cardiovascular clinical care and population health for non-experts*. *npj Cardiovascular Health, 2*, 44. https://www.nature.com/articles/s44325-025-00082-6

Apple. (2015). *Apple Watch available in nine countries on April 24.* Apple Newsroom. https://www.apple.com/newsroom/2015/03/09Apple-Watch-Available-in-Nine-Countries-on-April-24/

Apple. (2025). Receive sleep apnea notifications on Apple Watch. Apple Support. https://support.apple.com/en-us/120031

Apple. (2018). *ECG app and irregular heart rhythm notification available today on Apple Watch.* Apple Newsroom. https://www.apple.com/newsroom/2018/12/ecg-app-and-irregular-heart-rhythm-notification-available-today-on-apple-watch/

Google. (n.d.). How do I track & manage stress with my Fitbit device? Fitbit Help Center. https://support.google.com/fitbit/answer/14237928?hl=en#exp

Pinge, A., Gad, V., Jaisighani, D., Ghosh, S., & Sen, S. (2024). Detection and monitoring of stress using wearables: A systematic review. Frontiers in Computer Science, 6, 1478851. https://doi.org/10.3389/fcomp.2024.1478851

Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., & Acharya, U. R. (2018). *Deep learning for healthcare applications based on physiological signals: A review*. *Computer Methods and Programs in Biomedicine, 161*, 1–13. https://pubmed.ncbi.nlm.nih.gov/29852952/

Vos, G., Trinh, K., Sarnyai, Z., & Rahimi Azghadi, M. (2023). *Generalizable machine learning for stress monitoring from wearable devices: A systematic literature*

review. International Journal of Medical Informatics, 173, 105026. https://doi.org/10.1016/j.ijmedinf.2023.105026

WHOOP. (2026). Introducing WHOOP Coach: Your personal AI health and performance coach. WHOOP. https://support.whoop.com/s/article/How-to-Use-the-AI-Powered-WHOOP-Coach?language=en_US

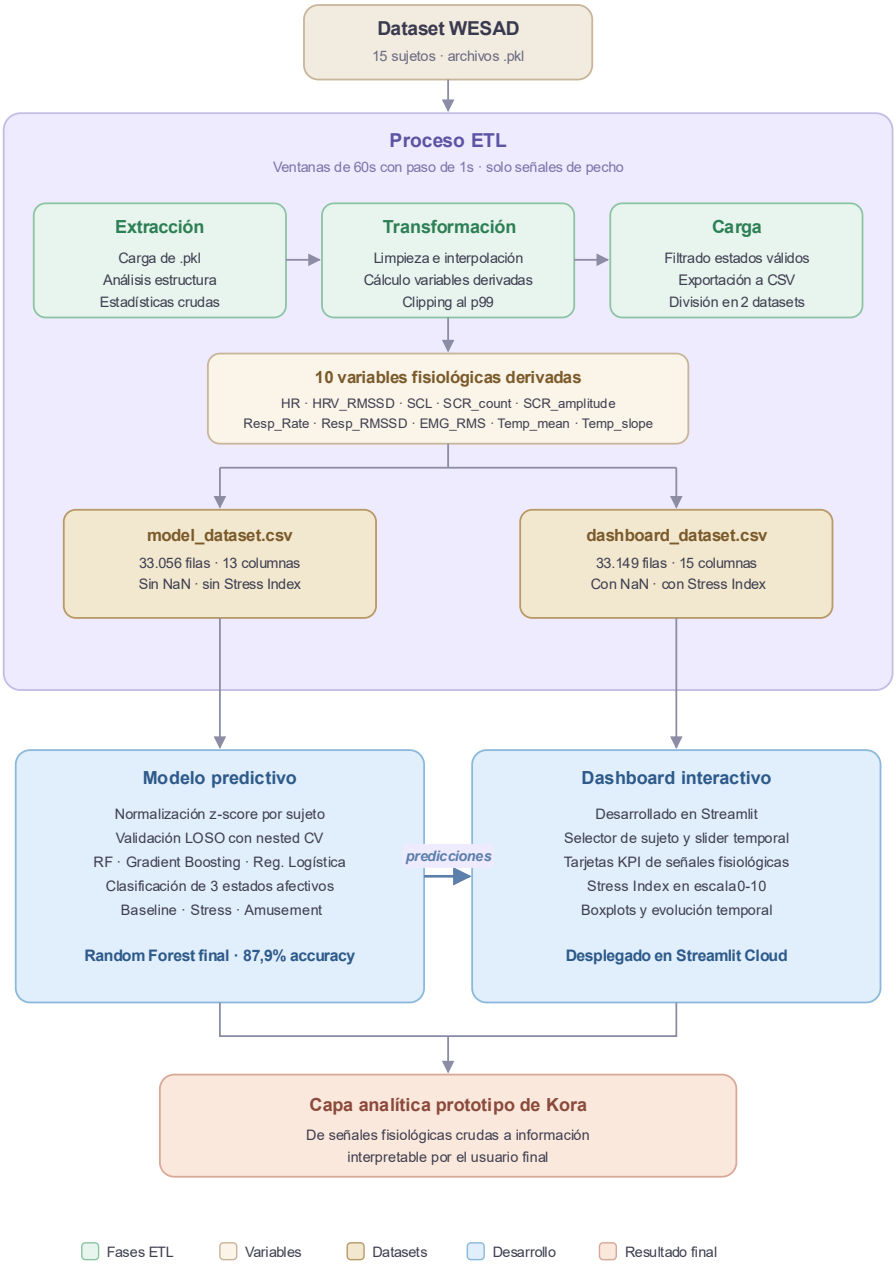
Oura Health. (2025). Introducing Oura Advisor: An AI-powered personal health companion. Oura Health. https://ouraring.com/blog/oura-advisor/

11. Anexos

Anexo 1. Repositorio de GitHub del proyecto (código y datasets necesarios): https://github.com/samuelcorcoba/kora-dashboard

Anexo 2. Dashboard interactivo online del proyecto desplegado en Streamlit Community Cloud: https://kora-dashboard.streamlit.app *(La página puede tardar unos minutos en cargar al abrirse por primera vez).*

Anexo 3. Diagrama global del pipeline de datos.



Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

Por la presente, yo, Samuel Corcoba Isorna, estudiante de Doble Grado en Administración y Dirección de Empresas y Business Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado "ANÁLISIS DE DATOS BIOMÉTRICOS RECOGIDOS POR DISPOSITIVOS WEARABLE: DESARROLLO DE UN DASHBOARD INTERACTIVO Y UN MODELO PREDICTIVO DE MONITORIZACIÓN FISIOLÓGICA" declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código solo en el contexto de las actividades descritas a continuación:

1. **Brainstorming de ideas de investigación:** Utilizado para idear y esbozar posibles áreas de investigación.
2. **Crítico:** Para encontrar contra-argumentos a una tesis específica que pretendo defender.
3. **Referencias:** Usado conjuntamente con otras herramientas, como Science, para identificar referencias preliminares que luego he contrastado y validado.
4. **Metodólogo:** Para descubrir métodos aplicables a problemas específicos de investigación.
5. **Interpretador de código:** Para realizar análisis de datos preliminares.
6. **Constructor de plantillas:** Para diseñar formatos específicos para secciones del trabajo.
7. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
8. **Generador previo de diagramas de flujo y contenido:** Para esbozar diagramas iniciales.
9. **Sintetizador y divulgador de libros complicados:** Para resumir y comprender literatura compleja.
10. **Generador de problemas de ejemplo:** Para ilustrar conceptos y técnicas.
11. **Revisor:** Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 20/04/2026

Firma: