UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INFORMÁTICA Y CIENCIAS DE LA COMPUTACIÓN

MODELO MULTIMODAL NO INVASIVO PARA LA ESTIMACIÓN DEL BIENESTAR ESTUDIANTIL

POR

Francisco Javier Cea Klapp

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

Profesor Patrocinante
Ricardo Antonio Flores Huenchullanca

Profesores Co-patrocinantes
Juan Carlos Caro Seguel
Jorge Ignacio Maluenda Albornoz

Septiembre 2025
Concepción (Chile)

Resumen

En el marco de un estudio piloto, se explora la viabilidad de un modelo multimodal no invasivo para estimar alteraciones asociadas al bienestar estudiantil en un entorno universitario. La propuesta integra, de forma experimental, datos biométricos obtenidos mediante dispositivos *wearables*, rasgos faciales y características acústicas extraídas de grabaciones audiovisuales, junto con representaciones semánticas de transcripciones textuales. El proceso metodológico incluye la aplicación de cuestionarios psicométricos estandarizados, la captura y procesamiento automatizado de cada modalidad y la construcción de vectores multimodales consolidados. Se implementan distintas estrategias de integración y selección de características con el fin de reducir el riesgo de sobreajuste derivado de la alta dimensionalidad y del tamaño reducido de la muestra. Los análisis realizados evidencian que la combinación de modalidades heterogéneas es técnicamente factible y que la selección de atributos relevantes contribuye a mejorar el rendimiento de los modelos, planteando un marco prometedor para futuros estudios de mayor alcance orientados al monitoreo no invasivo del bienestar estudiantil.

Índice General

Índice de Figuras

Índice de Tablas

1. Introducción

A nivel mundial, la carga de los trastornos depresivos y de ansiedad se ha consolidado como una de las principales amenazas para la salud pública, con tendencias que evidencian su persistencia y, en ciertos contextos, su aumento en las últimas décadas. Estimaciones recientes del *Global Burden of Disease* —un programa internacional de investigación que mide y compara de forma sistemática el impacto de enfermedades, lesiones y factores de riesgo en distintas poblaciones a partir de indicadores estandarizados— indican que, en 2019, más de 580 millones de personas vivían con alguna de estas condiciones, afectando de forma desproporcionada a mujeres, personas jóvenes y adultos mayores, y con marcadas disparidades entre regiones según su nivel socioeconómico [1].

La pandemia de COVID-19 exacerbó este escenario, especialmente en adolescentes y jóvenes, con incrementos cercanos al 30 % en los índices de depresión y ansiedad en países de altos ingresos, y ampliando brechas asociadas a desigualdad social, acceso a recursos y vulnerabilidad psicosocial [2]. Este panorama global subraya la necesidad urgente de desarrollar estrategias de evaluación y monitoreo más precisas, continuas y adaptadas a las realidades locales, que permitan detectar tempranamente alteraciones emocionales y orientar intervenciones oportunas.

El bienestar se entiende como un estado integral que combina factores externos —las condiciones necesarias para vivir bien— con factores internos, relacionados con el equilibrio físico, mental y emocional que favorece el desarrollo pleno de la persona [3]. Su opuesto, el malestar, se manifiesta en alteraciones que afectan este equilibrio, como la ansiedad, la depresión, la soledad o el estrés, las cuales repercuten en la estabilidad emocional y el funcionamiento cotidiano [4]. En este estudio, el término alteraciones al bienestar se emplea para referirse a manifestaciones de malestar

que, sin constituir necesariamente un diagnóstico clínico, interfieren con el desarrollo integral del estudiante.

En el contexto de la educación superior chilena, el bienestar estudiantil universitario se ha consolidado como un elemento clave para el rendimiento académico y la permanencia en los programas de estudio. Este enfoque reconoce al estudiante en su complejidad, considerando que emociones, relaciones sociales, capacidades cognitivas y valores personales inciden directamente en su salud mental y desempeño académico. Sin embargo, investigaciones recientes han evidenciado que la ansiedad, la depresión, la falta de sentido de pertenencia y la soledad representan barreras significativas para el éxito académico y afectan de forma directa el bienestar subjetivo [5, 6, 7].

Tradicionalmente, la estimación del bienestar se ha basado en métodos como encuestas extensas, entrevistas o registros diarios. Aunque útiles, estos presentan limitaciones importantes: sesgos de respuesta —como la deseabilidad social o el asentimiento automático—, errores de memoria y bajas tasas de participación [8]. Estas limitaciones han impulsado la búsqueda de alternativas más objetivas, continuas y escalables.

En este escenario, el uso de *proxies* o indicadores indirectos ha cobrado relevancia. Por un lado, instrumentos psicoemocionales validados, como el PHQ-4 [9], permiten obtener la autoevaluación subjetiva del estado emocional. Por otro, el análisis automatizado de registros faciales mediante técnicas de visión por computador permite identificar patrones de activación muscular y microexpresiones vinculadas a emociones específicas [10, 11]. La combinación de estos enfoques, junto con datos biométricos puede permitir una estimación más robusta y contextualizada de la ansiedad y el bienestar [12].

El desarrollo de tecnologías basadas en inteligencia artificial y aprendizaje automático (*machine learning*) ha potenciado el avance en el ámbito de la salud, con aplicaciones que abarcan desde la clasificación de imágenes médicas y el diagnóstico de enfermedades cardiovasculares, hasta la detección temprana de patologías neurodegenerativas [13, 14, 15].

Estos mismos principios pueden extrapolarse al campo de la salud mental, donde la capacidad de procesar grandes volúmenes de datos heterogéneos y encontrar correlaciones complejas abre la

puerta a estrategias más precisas y no invasivas de monitoreo. En este contexto, el enfoque multimodal —que combina diversas fuentes de datos— ha demostrado ser especialmente eficaz para capturar la complejidad de fenómenos como la ansiedad y la depresión, superando las limitaciones de herramientas aisladas [16, 17].

Considerando todo lo anterior, la presente Memoria de Título propone el diseño y la validación preliminar de un modelo multimodal para la estimación de alteraciones vinculadas al bienestar estudiantil, aplicado a un estudio piloto con estudiantes de pregrado en un entorno universitario. El pipeline experimental contempla la selección de participantes según criterios definidos, la aplicación de instrumentos psicométricos estandarizados [18, 9, 19, 20] y la caracterización sociodemográfica de la muestra.

Asimismo, se incorporará la recolección no invasiva de datos biométricos mediante dispositivos wearables —como Xiaomi Smart Band 9 Active, Fitbit Charge 5 o Garmin Vivosmart 5—, junto con el análisis de rasgos faciales y de voz obtenidos a partir de grabaciones audiovisuales en un entorno controlado.

La integración de estas modalidades permite explorar la construcción de un modelo de *machine learning* orientado a la detección temprana de estados emocionales adversos en estudiantes universitarios, sentando las bases para el desarrollo de herramientas escalables.

2. Objetivos

2.1. Objetivo general

Validar un modelo multimodal para la caracterización de patrones asociados al bienestar de estudiantes universitarios de pregrado, a partir de la recolección no invasiva de datos biométricos mediante dispositivos wearables, es decir, tecnologías portátiles que se llevan en el cuerpo y permiten monitorear variables fisiológicas en tiempo real.

2.2. Objetivos específicos

1. Seleccionar participantes para el estudio piloto a partir de los criterios definidos en un instrumento de caracterización sociodemográfica diseñado específicamente para este fin.
2. Capturar datos biométricos, faciales y de audio de forma sincronizada y no invasiva, a partir de un protocolo experimental previamente definido.
3. Evaluar la viabilidad del estudio piloto, analizando la calidad de los datos y el desempeño del modelo propuesto.
4. Desarrollar un modelo multimodal que integre datos biométricos, análisis de video y cuestionarios para estimar de forma no invasiva alteraciones en el bienestar.

2.3. Limitaciones

El estudio presenta diversas limitaciones que deben considerarse desde el inicio. En primer lugar, la disponibilidad limitada de dispositivos *wearables* para la recolección de datos biométricos representa una dificultad importante para la implementación de un estudio piloto a gran escala y el tiempo de uso por cada participante constituye otra limitación relevante. En segundo lugar, la participación voluntaria de estudiantes es un factor condicionante, ya que puede limitar el tamaño y la representatividad de la muestra. A todo esto se suma la necesidad de ajustar el diseño experimental a los tiempos académicos disponibles, lo que exige una planificación rigurosa que permita cumplir con los plazos establecidos.

Es importante recalcar que la metodología involucra la recolección y tratamiento de datos sensibles, lo cual impone restricciones éticas estrictas. Estas incluyen la protección de la privacidad, la confidencialidad de la información y la obtención de consentimiento informado (véase Anexo A.1) por parte de los participantes, esenciales para resguardar la integridad ética del estudio.

3. Marco teórico

Este capítulo presenta los fundamentos conceptuales y técnicos que respaldan el desarrollo y metodología de los siguientes capítulos. Para empezar, se revisan investigaciones previas sobre la evaluación del bienestar subjetivo, el uso de tecnologías no invasivas para recolectar datos y los enfoques computacionales que se han aplicado en estudios similares.

3.1. Antecedentes

La estimación de variables asociadas al bienestar y malestar, en particular de condiciones como la ansiedad y la depresión, continúa siendo un reto para la investigación en salud mental. Estas alteraciones afectan de manera significativa la calidad de vida y el desempeño académico, y su detección temprana es esencial para prevenir complicaciones y facilitar intervenciones oportunas.

En los últimos años, el uso de *proxies* o indicadores indirectos ha emergido como una estrategia prometedora para complementar o sustituir evaluaciones tradicionales basadas en autoinformes. Entre estos, los dispositivos *wearables* se han consolidado como herramientas relevantes para la recolección pasiva y continua de datos fisiológicos. En su revisión exploratoria, Abd-alrazaq et al. (2023) [21] recopilaron estudios que aplican inteligencia artificial a señales como la variabilidad de la frecuencia cardíaca (HRV), la actividad física, los patrones de sueño y la temperatura de la piel, con el objetivo de detectar ansiedad y depresión. Los resultados evidenciaron un rendimiento de moderado a alto, aunque los autores señalaron la necesidad de estandarizar protocolos y validar estos métodos en entornos reales.

En el ámbito de las expresiones faciales, Sharma et al. (2024) [11] desarrollaron un modelo basado en *facial action units* (AUs) para identificar microexpresiones vinculadas a emociones características de depresión, ansiedad y estrés. Mediante redes neuronales convolucionales y técnicas de *clustering*, alcanzaron precisiones superiores al 93 %, lo que respalda el análisis facial automatizado como un *proxy* objetivo y no invasivo para estimar estados emocionales.

Asimismo, las características acústicas de la voz se han posicionado como un indicador relevante en la evaluación de salud mental. Tlachac et al. (2025) [22] compararon grabaciones breves realizadas con dispositivos móviles frente a entrevistas clínicas tradicionales, analizando parámetros acústicos y prosódicos mediante *machine learning*. Sus resultados mostraron que las grabaciones móviles alcanzaban un rendimiento comparable al de las entrevistas clínicas, ofreciendo una alternativa escalable y menos invasiva para la detección temprana de depresión, ansiedad y estrés.

Por su parte, Thati et al. (2023) [23] propusieron un enfoque multimodal que combina patrones de uso de teléfonos inteligentes obtenidos mediante *mobile crowd sensing*[1] con análisis facial y acústico. En un estudio con 102 participantes, recopilaron dos semanas de datos de uso de teléfonos y extrajeron características faciales y acústicas. Las variables extraídas con OpenFace se representaron mediante descriptores estadísticos (media, desviación estándar, asimetría, curtosis, etc.), y se seleccionaron con el coeficiente de correlación de Pearson para conservar solo las más asociadas con las etiquetas depresión/no depresión. Esto redujo la redundancia y mejoró el rendimiento, logrando un 86 % de exactitud con *Support Vector Machines*.

En conjunto, estos estudios evidencian que la combinación de *proxies* fisiológicos (HRV, actividad, sueño), expresivos (análisis facial) y paralingüísticos (voz), junto con estrategias de representación y selección de variables sustentadas en descriptores estadísticos y correlación de Pearson, puede mejorar la precisión y continuidad en la estimación de estados de bienestar y malestar, sentando las bases para enfoques multimodales robustos.

[1] Técnica que implica la recopilación de datos de uso de dispositivos móviles para inferir patrones de comportamiento y estados emocionales.

3.2. Fundamentos conceptuales

Esta sección presenta los conceptos clave que sustentan la base teórica del presente estudio, con énfasis en aquellos relacionados con el aprendizaje automático y una forma de atacar el procesamiento de datos multimodales.

3.2.1. Descriptores estadísticos

Los descriptores estadísticos son fundamentales para analizar series temporales, como la frecuencia cardíaca, ya que permiten describirlas a través de indicadores de tendencia central, variabilidad, simetría, forma y complejidad. Entre los más utilizados se encuentran los indicadores presentados en la Tabla 3.1.

TABLA 3.1: Descriptores estadísticos relevantes para el análisis de señales fisiológicas.

Indicador	**Descripción**
Media (μ)	Promedio de los valores.
Mediana	Valor central cuando se ordenan. Más robusta ante valores atípicos.
Desviación estándar (σ)	Mide la dispersión de los datos respecto a la media.
Valor mínimo	Mínimo valor observado.
Valor máximo	Máximo valor observado.
Rango	Diferencia entre el valor máximo y mínimo.
RMS[a]	Valor cuadrático medio; refleja la magnitud general de la señal.
Skewness[b] (γ_1)	Mide la simetría de la distribución. Valores positivos o negativos indican sesgo hacia la derecha o izquierda, respectivamente.
Kurtosis[c] (γ_2)	Evalúa el grado de concentración de los valores respecto a la media.
Entropía	Cuantifica el grado de irregularidad o aleatoriedad presente.

[a] **RMS**: Raíz cuadrada media.
[b] **Skewness**: Asimetría.
[c] **Kurtosis**: Apuntamiento.

Este tipo de medidas resulta especialmente útil en el análisis de señales fisiológicas, que suelen presentar dinámicas no lineales, no estacionarias y alejadas de una distribución gaussiana. En este contexto, Chua et al. [24] señalan que métricas como la varianza, la asimetría (*Skewness*) y el apuntamiento (*Kurtosis*) ayudan a capturar aspectos clave de la distribución de estas señales, mientras

que la entropía ofrece una mirada más profunda sobre su irregularidad y complejidad.

3.2.2. Machine Learning

El aprendizaje automático (*Machine Learning*) [25] ees una rama de la inteligencia artificial orientada al desarrollo de algoritmos capaces de aprender patrones a partir de datos y realizar predicciones o clasificaciones sin ser programados explícitamente para cada tarea. En lugar de definir reglas fijas, estos modelos construyen representaciones estadísticas del comportamiento de los datos y las utilizan para generar respuestas ante nuevas entradas

El proceso típico de *machine learning* incluye: la recolección de datos, la selección y extracción de características relevantes, la partición del conjunto de datos (entrenamiento, validación y prueba), el entrenamiento del modelo, y su evaluación en función del rendimiento predictivo.

Algoritmos de aprendizaje supervisado

El aprendizaje supervisado es un paradigma de *machine learning* en el que el modelo se entrena con datos etiquetados, es decir, cada entrada está asociada a una salida conocida. La meta es aprender una función que mapee estas entradas a salidas, de modo que el modelo pueda predecir correctamente casos nuevos.

Dentro de los múltiples algoritmos que pueden emplearse, tres destacan por su aplicabilidad, fundamento teórico y diversidad de casos de uso: *Support Vector Machine* (SVM), *Logistic Regression* y *Random Forest*. Aunque difieren en sus principios y funcionamiento, comparten el objetivo común de encontrar patrones en los datos que permitan clasificar de forma precisa nuevas observaciones.

- **Support Vector Machine (SVM)**: Busca encontrar el hiperplano que mejor separa las clases, maximizando la distancia o margen entre los puntos más cercanos de cada clase (vectores de soporte). Mediante el uso de funciones *kernel*, puede proyectar los datos a espacios de mayor dimensión para resolver problemas no linealmente separables. Resulta especialmente útil en espacios de alta dimensión y en conjuntos de datos de tamaño reducido."
- **Logistic Regression**: Modelo estadístico utilizado principalmente para clasificación binaria, que estima la probabilidad de pertenencia a una clase a partir de una combinación lineal

de las características, transformada mediante la función logística o sigmoide. Destaca por su sencillez, bajo costo computacional e interpretabilidad de los coeficientes, siendo adecuado para relaciones aproximadamente lineales entre variables y salida.

- **Random Forest**: Algoritmo de ensamblaje basado en múltiples árboles de decisión, entrenados sobre subconjuntos aleatorios de datos y características. Combina las predicciones de todos los árboles mediante votación (clasificación) o promedio (regresión), lo que reduce el sobreajuste y mejora la capacidad de generalización. Es robusto ante ruido, maneja relaciones no lineales y proporciona medidas de importancia de variables.

La selección de uno u otro algoritmo depende de factores como la naturaleza y dimensionalidad de los datos, la necesidad de interpretabilidad del modelo y los recursos computacionales disponibles.

Métricas de evaluación

A continuación, en la Tabla 3.2 se presentan las métricas de evaluación más comunes para problemas de clasificación, cada métrica posee una interpretación y utilidad específicas, dependiendo del problema abordado y de la naturaleza de los datos analizados.

TABLA 3.2: Métricas comunes para evaluar modelos de clasificación

Métrica	**Descripción**	**Fórmula**
Accuracy	Proporción de predicciones correctas sobre el total de muestras.	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	Proporción de verdaderos positivos entre todas las predicciones positivas.	$\frac{TP}{TP + FP}$
Recall	Proporción de verdaderos positivos correctamente identificados entre todos los casos positivos.	$\frac{TP}{TP + FN}$
F1-score	Media armónica entre precisión y recall, útil en contextos de clases desbalanceadas.	$2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
AUC-ROC	Área bajo la curva ROC; evalúa la capacidad del modelo para distinguir entre clases.	No aplica[a]

[a] El AUC-ROC se calcula numéricamente como el área bajo la curva que relaciona la tasa de verdaderos positivos (TPR) con la tasa de falsos positivos (FPR), a distintos umbrales de clasificación.

TP: Verdaderos positivos. **TN**: Verdaderos negativos.

FP: Falsos positivos. **FN**: Falsos negativos.

Estas métricas permiten evaluar distintos aspectos del desempeño de un modelo de clasificación. La exactitud (*accuracy*) da una idea general del rendimiento, aunque puede resultar engañosa cuando las clases están desbalanceadas.

En esos casos, métricas como *precision* y *recall* ofrecen información más útil. *Precision* se centra en qué tan confiables son las predicciones positivas, mientras que *recall* mide qué tanto logra el modelo identificar correctamente los casos positivos.

El *F1-score* combina ambos enfoques, por lo que es especialmente útil cuando se busca un equilibrio entre precisión y sensibilidad.

Por último, la métrica *AUC-ROC* ofrece una perspectiva más amplia del rendimiento de un modelo, ya que evalúa su capacidad para distinguir entre clases en todos los posibles umbrales de decisión.

En muchos modelos, el umbral predeterminado es 0,5; es decir, una muestra se clasifica como positiva si su probabilidad estimada supera ese valor. Sin embargo, ese umbral no siempre es el más adecuado para todos los escenarios [26].

Lo valioso de *AUC-ROC* es que permite comparar modelos sin depender de un umbral específico, mostrando qué tan bien logra el modelo separar las clases en términos generales. Un valor cercano a 1 indica una excelente capacidad para discriminar entre clases, mientras que un valor cercano a 0,5 indica que el modelo no discrimina mejor que una clasificación aleatoria.

Embeddings

Los *embeddings* son representaciones vectoriales densas que permiten codificar datos no estructurados —como texto o lenguaje natural— en un formato numérico interpretable por modelos computacionales. Estas representaciones capturan relaciones semánticas y contextuales entre palabras, frases o documentos, facilitando tareas como clasificación, análisis de similitud o modelamiento semántico [27].

Una de las formas más utilizadas para generar *embeddings* lingüísticos es a través de modelos de lenguaje preentrenados, como BERT (*Bidirectional Encoder Representations from Transformers*). En el

caso del español, destaca BETO, una adaptación de BERT entrenada exclusivamente sobre corpus[2] en español [28].

Esto permite obtener *embeddings* contextuales, es decir, representaciones numéricas en las que el significado de una palabra se ajusta dinámicamente según el contexto en el que aparece.

Su entrenamiento exclusivo en español le permite capturar de forma más precisa las estructuras gramaticales, las ambigüedades semánticas y los matices contextuales propios del idioma, a diferencia de versiones multilingües como mBERT[3] [28].

Estas representaciones capturan tanto el contenido semántico como las relaciones entre las palabras dentro de una oración o documento, y pueden utilizarse como entrada en diversas tareas de procesamiento del lenguaje natural.

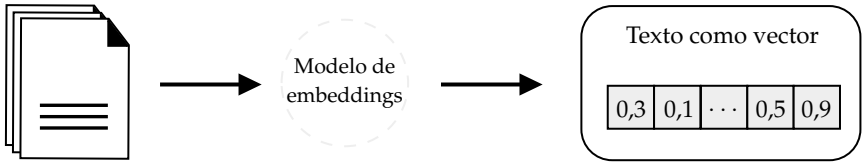


FIGURA 3.1: Representación esquemática del proceso de obtención de *embeddings* a partir de texto mediante un modelo de lenguaje preentrenado.

Reducción de dimensionalidad

La reducción de dimensionalidad consiste en transformar un conjunto de datos con numerosas variables en una representación más compacta que conserve la mayor parte de la información relevante. Este proceso contribuye a disminuir el sobreajuste, reducir el costo computacional y, en ciertos casos, facilitar la interpretación de patrones presentes en los datos.

Entre las técnicas más utilizadas se encuentra el *Principal Component Analysis* (PCA), que proyecta los datos en un nuevo espacio formado por componentes ortogonales que capturan la mayor varianza posible. Para ello, se calculan los vectores propios de la matriz de covarianza y se seleccionan aquellos asociados a los valores propios que explican un alto porcentaje de la variabilidad

[2]En procesamiento del lenguaje natural, un *corpus* es una colección estructurada de textos, usualmente de gran tamaño, utilizada para el análisis lingüístico o el entrenamiento de modelos computacionales.

[3]mBERT: Multilingual BERT, es una adaptación de BERT entrenada en múltiples idiomas realizada por Google.

total, lo que permite representar los datos con menos dimensiones sin pérdida significativa de información [29].

3.2.3. Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson (r) es una medida estadística que cuantifica la fuerza y la dirección de la relación lineal entre dos variables continuas. Se define como la covarianza de las variables normalizada por el producto de sus desviaciones estándar:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.1}$$

donde x_i y y_i corresponden a las observaciones de cada variable, y $\bar{x}$, $\bar{y}$ a sus medias respectivas.

El valor de r se encuentra en el rango $[-1, 1]$, interpretándose de la siguiente forma:

- $r > 0$: correlación lineal positiva, donde un aumento en una variable tiende a asociarse con un aumento en la otra.
- $r < 0$: correlación lineal negativa, donde un aumento en una variable tiende a asociarse con una disminución en la otra.
- $r \approx 0$: ausencia de correlación lineal significativa.

Este coeficiente se utiliza tanto en análisis exploratorio como en selección preliminar de variables dentro de modelos de aprendizaje automático. Sin embargo, su aplicación supone ciertos supuestos: la existencia de una relación lineal entre las variables, datos aproximadamente normales y ausencia de valores atípicos extremos.

3.2.4. Herramientas de extracción de características

La extracción de características constituye una etapa fundamental en el análisis de datos, ya que permite transformar datos en bruto —a menudo complejos y no estructurados— en representaciones numéricas que los modelos de aprendizaje automático pueden interpretar y procesar. Este

procedimiento facilita el aprovechamiento de diversos tipos de datos, como señales, audio, video o texto, convirtiéndolos en vectores estructurados que capturan sus atributos más relevantes para tareas como la clasificación o la predicción.

OpenFace: Herramienta de extracción de características faciales

Para la extracción de características faciales a partir de imágenes o secuencias de video, OpenFace [30] se presenta como una herramienta de código abierto robusta y eficiente. Diseñada para el análisis automático del comportamiento facial, permite capturar información clave del rostro en tiempo real.

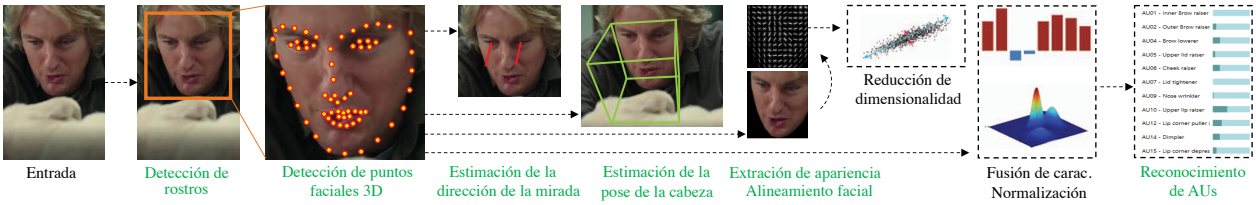


FIGURA 3.2: Adaptación al español del *pipeline* de análisis de comportamiento facial de OpenFace 2.0, que incluye: detección de puntos faciales (landmark detection), estimación de la pose de la cabeza (*head pose*), estimación de la dirección de la mirada (*eye gaze*) y reconocimiento de unidades de acción facial (*action unit*). Las salidas de cada uno de estos módulos (indicadas en verde) pueden guardarse en disco o enviarse por red en tiempo real. Adaptado de [30].

El funcionamiento de OpenFace se basa en una arquitectura modular que integra métodos de visión por computador con modelos de *machine learning*. La Figura 3.2 muestra de manera esquemática este *pipeline*. Este flujo de procesamiento permite comprender cómo cada etapa contribuye a generar un conjunto estructurado de características faciales que, en conjunto, proporcionan la base para el análisis automático de expresiones y estados emocionales. Para la detección de *landmarks*, OpenFace emplea un enfoque avanzado —denominado CE-CLM[4]— que permite identificar con precisión la estructura del rostro, incluso en escenarios complejos como poses no frontales u oclusiones parciales.

[4]El modelo CE-CLM (*Convolutional Experts Constrained Local Model*) fue propuesto por los mismos autores de OpenFace como parte de las mejoras en su última versión, OpenFace 2.0 [30].

A partir de esta detección de landmarks, se derivan otras variables faciales como:

- **La pose de la cabeza**, estimada mediante la proyección de un modelo facial 3D sobre la imagen 2D, lo que permite obtener los ángulos de orientación (yaw, pitch y roll)[5]
- **La dirección de la mirada**, calculada geométricamente a partir de la localización de las pupilas y la forma del globo ocular.
- **Las Unidades de Acción (Action Units)**, que representan activaciones musculares específicas del rostro y permiten describir expresiones faciales de forma cuantitativa.

El reconocimiento de *Action Units* (AUs) en OpenFace se realiza mediante clasificadores SVM entrenados con características extraídas del rostro previamente alineado [30]. Los resultados se expresan en dos modalidades: presencia (binaria) e intensidad (escalar), abarcando un conjunto amplio de AUs definidos por el *Facial Action Coding System* (FACS). Entre las más representativas se incluyen el levantamiento de cejas (AU1, AU2), la sonrisa (AU12) y el parpadeo (AU45), entre otras. Para más detalles, se puede consultar el Anexo A.2.

OpenSMILE: Herramienta de extracción de características acústicas

En el ámbito del procesamiento de audio, una herramienta que ha ganado protagonismo es OpenSMILE (Open Speech and Music Interpretation by Large-space Extraction) [31]. Se trata de un software de código abierto diseñado para extraer características acústicas de manera eficiente y flexible, ya sea en análisis por lotes o en tiempo real.

OpenSMILE convierte las señales de audio en representaciones numéricas estructuradas que capturan diversos aspectos relevantes del sonido, tales como el contenido espectral, la entonación, la calidad vocal o los patrones de energía [31]. Las cuales pueden ser utilizadas para entrenar modelos de aprendizaje automático, permitiendo detectar patrones, clasificar estados afectivos o realizar tareas de discriminación acústica, entre otras cosas.

[5] Los ángulos *yaw*, *pitch* y *roll* describen la orientación tridimensional del rostro: *yaw* representa la rotación horizontal, *pitch* la rotación vertical, y *roll* la inclinación lateral.

Las características extraídas suelen agruparse en categorías que describen propiedades físicas del sonido (como su forma espectral), rasgos vocales (como el tono o la estabilidad de la voz), y métricas estadísticas aplicadas a lo largo del tiempo, dando lugar a descriptores robustos y compactos que resumen la información contenida en una señal.

El alto grado de personalización permite definir flujos de procesamiento completos a través de archivos de configuración, en los cuales se especifican los módulos activos, sus parámetros y las relaciones entre ellos, lo que a su vez permite ajustar la ventana de análisis. Para más detalles, se puede consultar el Anexo A.2.

Whisper: Herramienta de transcripción de audio

Whisper [32] es un sistema de reconocimiento automático de voz (ASR) desarrollado por OpenAI, entrenado con un conjunto de datos multilingüe y multitarea de aproximadamente 680000 horas de audio y sus transcripciones [32, 33]. Este entrenamiento a gran escala y con datos diversos le permite alcanzar un alto nivel de robustez frente a variaciones de acento, ruido de fondo y condiciones de grabación, así como generalizar a contextos no observados previamente sin requerir un ajuste fino específico.

El proceso de transcripción de Whisper comienza con la segmentación del audio en fragmentos de hasta 30 segundos[6], que son transformados en un espectrograma log-Mel. Dicho espectrograma se procesa en un *encoder* basado en arquitectura Transformer, encargado de extraer representaciones de alto nivel del audio. Posteriormente, estas representaciones se conectan al decodificador (*decoder*) mediante un mecanismo de atención cruzada, el cual genera de forma autoregresiva secuencias de tokens que representan el texto transcrito y, opcionalmente, marcas de tiempo, tal como se ilustra en la Figura 3.3.

[6]Este límite responde a restricciones computacionales de la arquitectura Transformer y a la configuración usada durante el entrenamiento [32]. Audios más extensos se procesan de manera segmentada mediante ventanas consecutivas de 30 segundos.

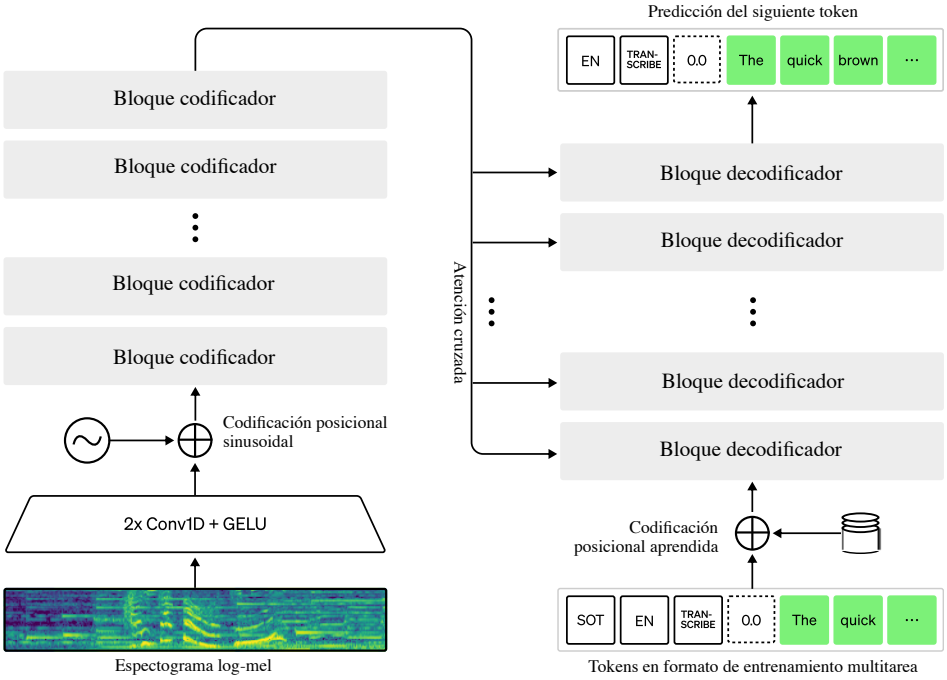


FIGURA 3.3: Arquitectura interna de Whisper. Adaptado de [33].

4. Desarrollo del proyecto

Este capítulo describe la metodología implementada para el diseño y validación preliminar del modelo multimodal propuesto, orientado a la estimación de alteraciones asociadas al bienestar estudiantil. El enfoque adoptado considera la integración de múltiples fuentes de información, priorizando métodos de recolección no invasivos que resguarden la comodidad y la privacidad de los participantes.

El proyecto se enmarca en la necesidad de recolectar datos provenientes de diversas fuentes, con el propósito de caracterizar el bienestar estudiantil desde una perspectiva integral. Para ello, se implementa una metodología sistemática que contempla tres principales canales de recolección:

- Instrumentos psicométricos estandarizados, aplicados mediante un formulario digital de caracterización.
- Datos biométricos obtenidos a partir del uso cotidiano de dispositivos *wearables*.
- Aplicación de una entrevista estructurada y grabada en un entorno controlado.

La estrategia de recolección se desarrolla en etapas previamente definidas (Figura 4.1), comenzando con la caracterización inicial de los participantes mediante un formulario que integra datos sociodemográficos y puntajes obtenidos en escalas psicométricas validadas, para luego avanzar hacia la captura de datos biométricos y audiovisuales.

La ejecución de cada una de estas etapas considera no solo aspectos técnicos y metodológicos, sino también criterios éticos orientados al resguardo de la información personal y a la protección de los participantes. Dado que el estudio involucra la recolección de datos sensibles y la participación

directa de personas, resulta esencial incorporar medidas éticas desde la etapa de diseño hasta la aplicación del protocolo experimental.

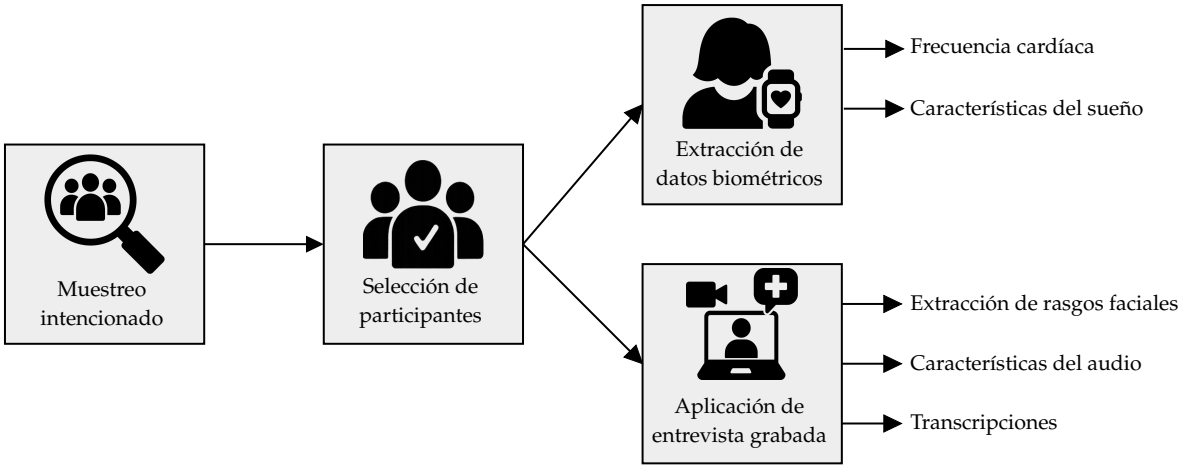


FIGURA 4.1: Diagrama de flujo de la recolección de datos.

4.1. Consideraciones éticas del estudio

El presente estudio contempla el diseño e implementación de un modelo multimodal orientado a la estimación de alteraciones asociadas al bienestar estudiantil, en el contexto de un estudio piloto desarrollado en un entorno universitario. Para ello, se recopilarán datos directamente de personas voluntarias, estudiantes de pregrado —principalmente de la Universidad de Concepción—, quienes serán asignadas a condiciones experimentales previamente definidas.

La información recolectada incluirá datos biométricos obtenidos de manera no invasiva mediante dispositivos *wearables*, así como grabaciones audiovisuales que se procesan mediante técnicas de visión por computador y análisis acústico, con el fin de extraer señales verbales y no verbales relevantes. Además, se aplican instrumentos psicométricos autoadministrados y validados internacionalmente, como el PHQ-4 (para la detección de sintomatología depresiva y ansiosa) [34, 9], la escala de soledad de UCLA (ULS) [18], la escala multidimensional de apoyo percibido [19] y la escala de estrés percibido (PSS-4) [20], entre otros. Estos instrumentos permitirán estimar el estado emocional de los participantes y servirán como referencia para el entrenamiento y validación.

Adicionalmente, se recopilan datos sociodemográficos con el objetivo de caracterizar adecuadamente a la muestra y mejorar el desempeño del modelo mediante la incorporación de variables contextuales. Si bien todos los datos se anonimizan durante el proceso de análisis, se mantiene una base de contacto segura y separada que permita la comunicación con los participantes en caso de requerir aclaraciones, seguimiento o retroalimentación. Esta estrategia busca resguardar la confidencialidad sin comprometer la integridad metodológica ni el acompañamiento ético a quienes participen en el estudio.

La investigación se adhiere a los principios establecidos en la Declaración de Singapur sobre la Integridad en la Investigación y en la Declaración de Helsinki sobre la Investigación Médica en Seres Humanos. Por ello, se contempla un proceso riguroso de consentimiento informado (véase Anexo A.1), que asegure la comprensión clara de los objetivos del estudio, los beneficios esperados, los posibles riesgos, y el derecho de los participantes a desistir en cualquier momento sin consecuencias negativas.

4.2. Selección y caracterización de participantes

Con el propósito de reclutar participantes para el estudio piloto, se diseña y distribuye un formulario digital de caracterización inicial, el cual se difunde de manera abierta entre estudiantes de pregrado de la Universidad de Concepción, sin excluir a estudiantes de otras instituciones, siempre que cumplan con el criterio de cursar estudios de pregrado. La participación es voluntaria y autoadministrada, enmarcada bajo principios éticos de confidencialidad y protección de datos personales.

En la invitación se informa que los datos recopilados se utilizan exclusivamente con fines de investigación, resguardando la privacidad de los participantes, y que quienes completan el instrumento pueden ser contactados posteriormente para continuar participando en etapas experimentales del estudio. Asimismo, se indica que, en caso de avanzar a dichas fases, se solicita la firma de un consentimiento informado formal (véase Anexo A.1), donde se detallan los procedimientos, alcances y derechos asociados a la participación, entre ellos la posibilidad de retirarse en cualquier momento sin necesidad de justificación.

Tras el periodo de difusión y recolección de respuestas, se obtienen un total de **93 formularios completos** por parte de estudiantes de pregrado interesados en participar en el estudio. De este conjunto, **66 participantes se identifican con el género masculino** y **27 con el género femenino**, tal como se muestra en la Figura 4.2. Esta distribución evidencia una asimetría de género significativa que se considera al momento de conformar la muestra final para las siguientes etapas del estudio, dado que un mayor equilibrio en este aspecto favorece la generalización de los resultados y contribuye a reducir sesgos de representación.

FIGURA 4.2: Distribución de identidad de género de quienes completaron el formulario de caracterización.

Además de las preguntas sociodemográficas, se aplican instrumentos psicométricos breves y validados internacionalmente, con el fin de estimar distintas dimensiones del bienestar subjetivo de los participantes. La Tabla 4.1 resume sus principales características, abarcando tanto indicadores emocionales como contextuales.

TABLA 4.1: Resumen de instrumentos psicométricos aplicados en la etapa de caracterización.

Instrumento	**Proxy**	**Nº ítems**	**Interpretación**	**Punto de corte[a]**
PHQ-4	Ansiedad y depresión	4	Puntajes altos indican mayor malestar emocional	≥ 6 (global)
PHQ-2	Depresión (subescala del PHQ-4)	2	Puntajes altos indican mayor sintomatología depresiva	≥ 3
GAD-2	Ansiedad (subescala del PHQ-4)	2	Puntajes altos indican mayor sintomatología ansiosa	≥ 3
UCLA-3	Soledad percibida	3	Puntajes altos indican mayor percepción de soledad	No definido[b]
PSS-4	Estrés percibido	4	Puntajes altos indican mayor percepción de estrés	No definido[b]
MSPSS	Apoyo social percibido	12	Puntajes altos indican mayor apoyo emocional	No clínico

[a] Los puntos de corte son referenciales y se aplican comúnmente en estudios de detección temprana [9, 34].
[b] Estas escalas no poseen puntos de corte clínicos establecidos; se interpretan de forma relativa.

Entre los instrumentos aplicados, el PHQ-4 [34] permite identificar síntomas de ansiedad y depresión a través de sus subescalas PHQ-2 y GAD-2 [9], lo que posibilita una clasificación preliminar del estado emocional de la muestra. Por su parte, escalas como la PSS-4, la versión abreviada de la UCLA y la MSPSS aportan información complementaria sobre estrés percibido, soledad y apoyo social. Aunque estas últimas no cuentan con puntos de corte clínicos definidos, sus puntajes permiten realizar comparaciones relativas dentro del grupo estudiado.

Las Figuras 4.3 y 4.4 presentan la distribución de puntajes obtenidos en las distintas escalas aplicadas, lo que permite caracterizar de forma preliminar el perfil emocional y psicosocial de los participantes. A partir de estos resultados, se asigna a cada estudiante una etiqueta preliminar según su nivel de malestar emocional o percepción de apoyo social, lo que posibilita una categorización inicial pertinente para la posterior selección de la muestra que avanza a las siguientes fases del estudio.

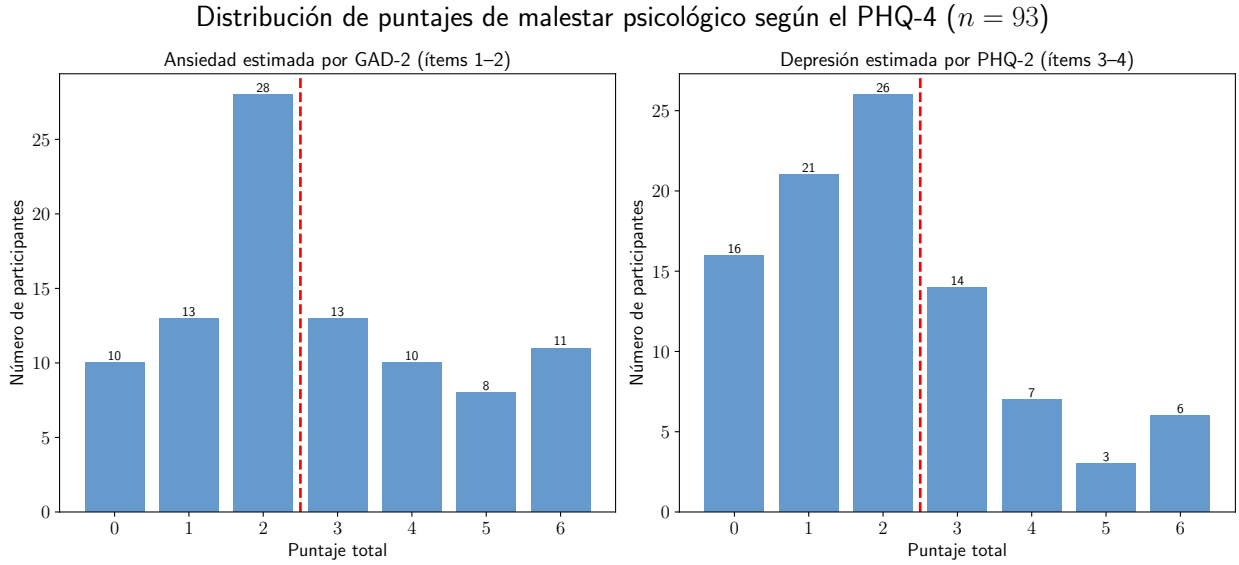


FIGURA 4.3: Distribución de los puntajes obtenidos en las subescalas del PHQ-4.

4.2.1. Criterio de selección

El principal criterio de selección corresponde a la voluntad de participar en el estudio, complementada por la disponibilidad para asistir de manera presencial a las instancias necesarias para su correcta implementación. Esta presencialidad resulta fundamental para garantizar la adecuada ejecución del protocolo experimental, ya que permite configurar el dispositivo *wearable* y aplicar la entrevista grabada en un entorno controlado y homogéneo para los participantes.

Además de estos aspectos logísticos, se incorporan criterios adicionales orientados a favorecer la representatividad de la muestra y reducir posibles sesgos de selección. Entre ellos, el género adquiere especial relevancia, dado que, como se observa en la Figura 4.2, la muestra inicial presenta una distribución desigual, con una menor proporción de personas que se identifican con el género femenino. Para contrarrestar esta asimetría, se prioriza su inclusión en las fases siguientes, a fin de favorecer una mayor equidad de género en la muestra final.

Esta decisión obedece tanto a un compromiso ético con la representación inclusiva como a una necesidad metodológica: un grupo equilibrado permite evaluar posibles diferencias entre géneros y evita que los resultados estén condicionados por una distribución desbalanceada.

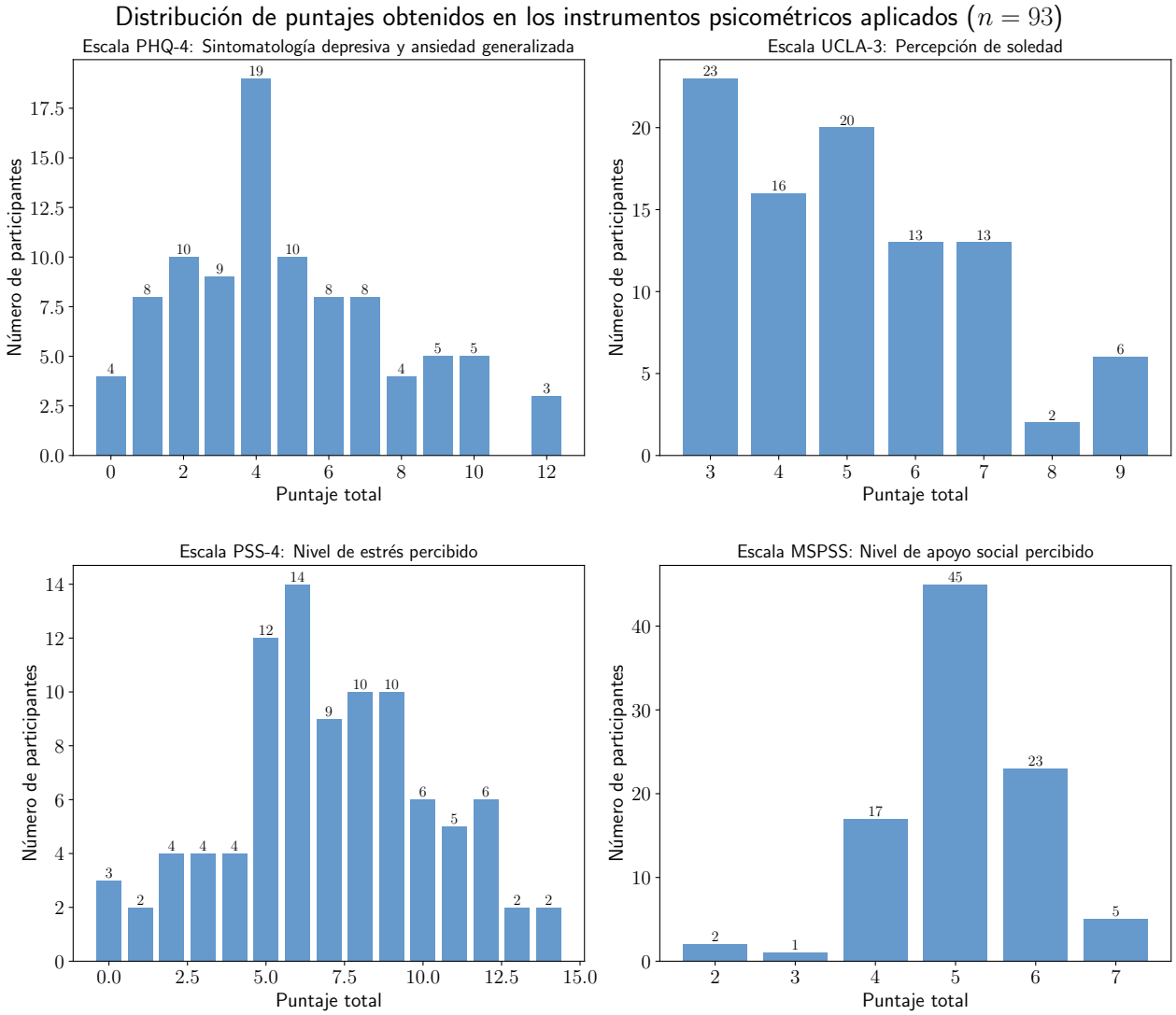


FIGURA 4.4: Distribución de los puntajes obtenidos en los instrumentos psicométricos aplicados en la etapa de caracterización.

Con base en los resultados de los instrumentos psicométricos (Figuras 4.4 y 4.3), se emplea el PHQ-4 como criterio preliminar, dado que permite diferenciar sintomatología ansiosa y depresiva.

Aplicando este y los demás criterios, se seleccionan 36 participantes para las siguientes etapas del estudio, conformando una muestra diversa y metodológicamente sólida en términos emocionales. Las Figuras 4.5 y 4.6 muestran la distribución de puntajes del PHQ-4 y de género en este grupo.

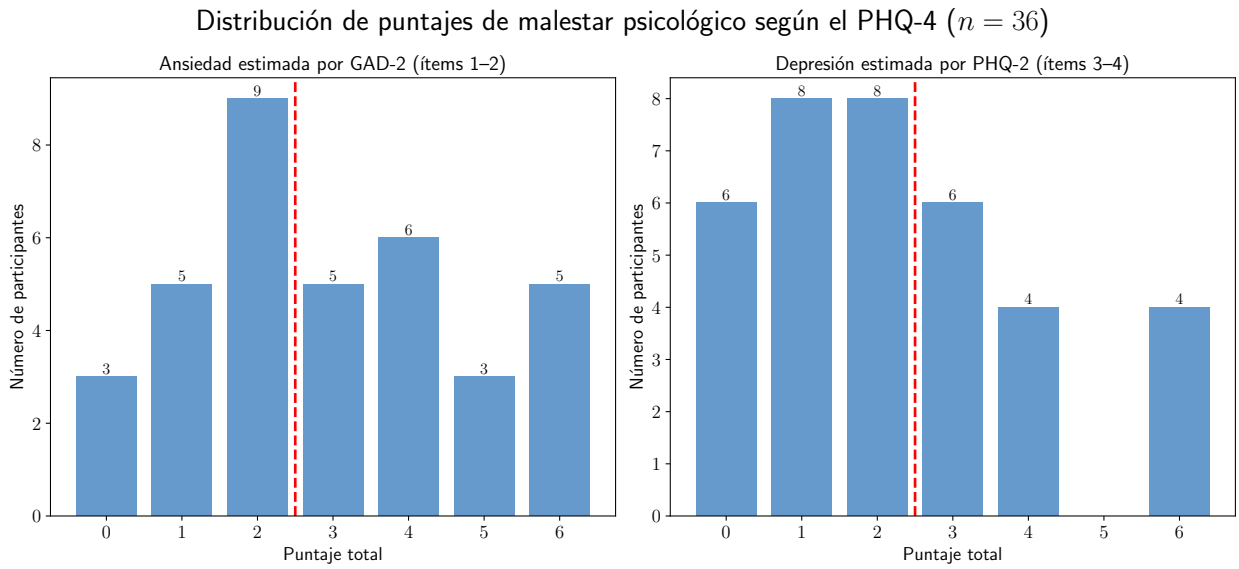


FIGURA 4.5: Distribución de los puntajes obtenidos en el PHQ-4 de los participantes seleccionados.

FIGURA 4.6: Distribución de género de los participantes seleccionados.

4.3. Protocolo experimental

Para los 36 participantes seleccionados, se aplica un protocolo experimental previamente desarrollado, cuya implementación se coordina de manera individual con cada participante para asegurar su disponibilidad presencial. Este protocolo incluyó las siguientes etapas:

1. **Firma del consentimiento informado:** antes de cualquier procedimiento, cada participante asiste presencialmente a una sesión individual en la que se explica en detalle el propósito del estudio, sus implicancias éticas y los derechos asociados a su participación. Posteriormente, firma de manera voluntaria el consentimiento informado (véase Anexo A.1).

2. **Configuración del dispositivo *wearable*:** a cada participante se le entrega un dispositivo de monitoreo biométrico (reloj inteligente), junto con las instrucciones para su uso durante el periodo de recolección (mínimo 48 horas).

3. **Sesión de grabación en entorno controlado:** en la misma instancia presencial, se realiza una grabación audiovisual estructurada en un entorno controlado. El participante responde una breve entrevista en voz alta, siguiendo un guion predefinido. La cámara y el micrófono registran simultáneamente rostro y voz, asegurando calidad técnica para la posterior extracción de características faciales y acústicas.

4. **Almacenamiento y organización de datos:** todos los registros obtenidos —biométricos, de video y audio— fueron anonimizados y almacenados en un sistema de archivos estructurado por identificador de participante. Se mantiene en paralelo una base de contacto para eventuales aclaraciones, garantizando la separación entre identidad personal y datos experimentales.

Para preservar la confidencialidad, todos los registros se almacenan con un identificador único anonimizado, generado mediante el algoritmo `SHA-256` a partir del correo electrónico del participante. Este identificador, criptográficamente irreversible, imposibilita reconstruir la identidad original y asegura la separación entre información personal y datos experimentales, manteniendo la coherencia y consistencia metodológica en la recolección.

4.3.1. Configuración del dispositivo wearable

Para la recolección de datos biométricos, se utiliza el reloj inteligente *Xiaomi Smart Band 9 Active* [35], el cual requiere la instalación de la aplicación móvil *Mi Fitness*, disponible para dispositivos Android e iOS, para su sincronización. Cada dispositivo fue vinculado a una cuenta controlada por el equipo, lo que permite centralizar de forma segura la recolección y almacenamiento de datos.

Se habilita específicamente el monitoreo de datos de salud con una frecuencia de muestreo de un minuto, con el fin de asegurar una resolución temporal adecuada y capturar variaciones fisiológicas relevantes para los análisis posteriores.

No obstante, este proceso de sincronización puede verse afectado por factores externos, tales como:

- Fallas de conectividad entre el reloj y el dispositivo móvil (por ejemplo, pérdida de Bluetooth o batería tanto del reloj como del dispositivo móvil).
- Reinicios o actualizaciones automáticas del dispositivo por parte del sistema.
- Interrupciones en la sincronización diaria debido a falta de acceso a internet o a problemas de permisos en el sistema operativo del teléfono, tal como la ejecución de la aplicación *Mi Fitness* en segundo plano.
- Eliminación accidental o desvinculación de la cuenta de la aplicación.

Estas situaciones pueden provocar lagunas o pérdidas parciales de datos, afectando la continuidad del registro. Para mitigar estos riesgos, se entregan instrucciones específicas a cada participante, solicitando que al menos una vez al día realice una verificación manual del estado de sincronización entre el reloj y la aplicación móvil, así como entre la aplicación y la nube.

4.3.2. Aplicación de la entrevista grabada

La entrevista se divide en dos partes complementarias, con el objetivo de obtener información audiovisual tanto en un entorno controlado como en situaciones más espontáneas. Para todos los participantes se utiliza el mismo formato: una presentación en PowerPoint proyectada frente a

ellos, que incluye una breve introducción, instrucciones específicas para cada actividad y el contenido correspondiente a cada etapa.

Durante la entrevista, cada participante lee en voz alta el contenido que aparece en pantalla, mientras es grabado con la cámara y el micrófono del dispositivo en el que se proyecta la presentación. Las grabaciones se realizaron con un **MacBook Air con procesador Apple M3 (2024)**, que integra una **cámara FaceTime HD 1080p** y un sistema de **tres micrófonos con formación de haz (beamforming)**. Esta técnica de procesamiento de señales permite combinar las entradas de los distintos micrófonos para enfatizar la voz proveniente del hablante y atenuar el ruido de fondo, mejorando la relación señal-ruido de las grabaciones. La cámara permitió registrar video en resolución 1080p a 30 fps, garantizando calidad suficiente para la extracción posterior de características.

Cada sección de la entrevista se almacena en archivos de video independientes y en orden, organizados por etapa. En el caso de las preguntas abiertas, todo el proceso se registra en un único video, con la intención de captar respuestas más naturales y continuas, sin cortes entre una pregunta y otra.

En la primera parte, se solicita a cada participante leer un párrafo estandarizado, diseñado para obtener una muestra neutral de voz, pronunciación y expresiones faciales en condiciones homogéneas para todos. El texto planteado es el siguiente:

> *"El veloz zorro marrón salta sobre el perro perezoso. Hoy es un día hermoso y tengo muchas ganas de aprender algo nuevo."*

A continuación, se formulan cuatro preguntas abiertas, seleccionadas por su capacidad para generar respuestas con una carga emocional significativa, de manera espontánea y con riqueza tanto en expresiones faciales como en entonaciones de voz. Estas preguntas se retoman del estudio de Tlachac et al., donde se respalda su efectividad por ser emocionalmente evocadoras, éticamente seguras y adecuadas para el análisis automático de señales no verbales [22]. Las preguntas aplicadas son las siguientes:

1. Describe el trabajo de tus sueños.
2. Describe un evento que ha marcado una influencia positiva en tu vida.

3. ¿Qué consejo le darías a tu yo más joven?

4. ¿De qué estás más orgulloso en tu vida?

Al final de la entrevista se incorpora una tercera sección en la que cada participante puede expresar cómo se sintió durante el estudio y entregar comentarios o sugerencias sobre la metodología. Esta parte no se considera para análisis emocional, sino que se transcribe con el único propósito de recopilar retroalimentación que permita perfeccionar el protocolo en futuras aplicaciones.

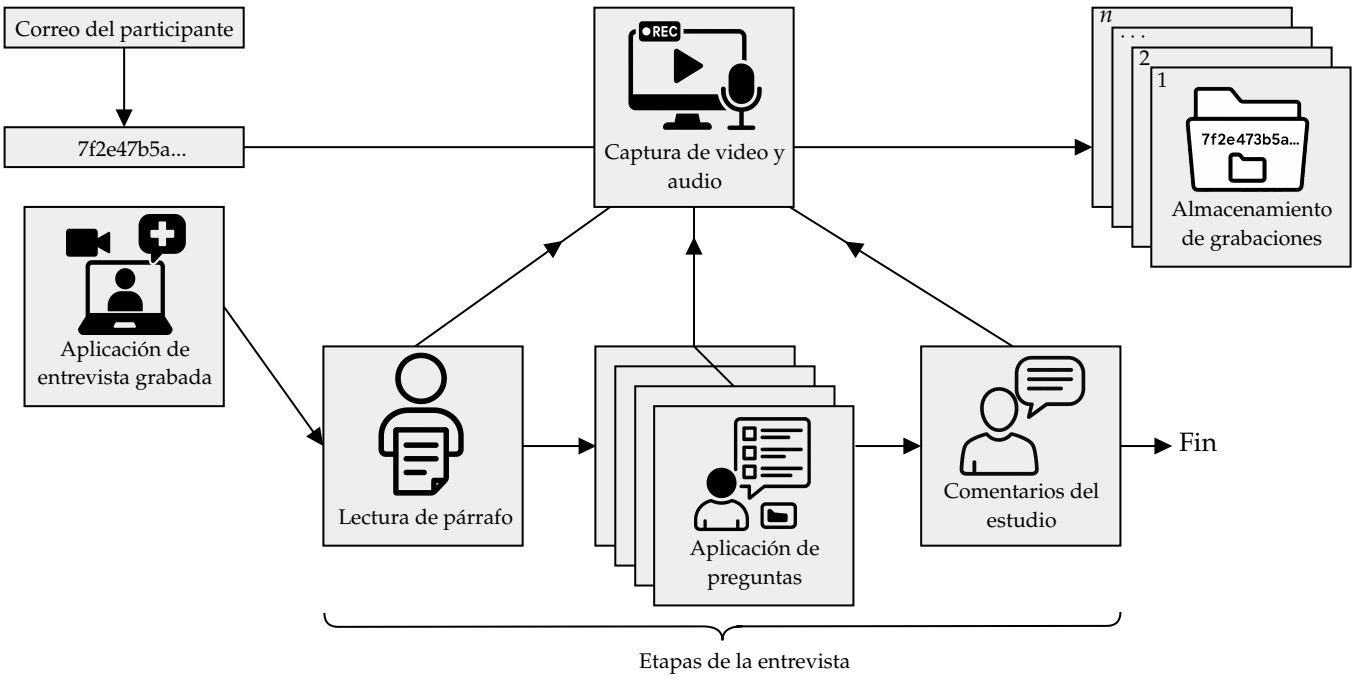


FIGURA 4.7: Diagrama del proceso de entrevista grabada, incluyendo sus tres etapas y el almacenamiento anonimizado de las grabaciones mediante hash SHA-256.

La Figura 4.7 ilustra el flujo completo del proceso de entrevista grabada. En ella se representan las tres etapas descritas anteriormente, así como el proceso de almacenamiento de las grabaciones resultantes. Estas se organizan por participante, utilizando el identificador anonimizado generado mediante el hasheo `SHA-256` del correo electrónico, lo que garantiza la protección de la identidad al tiempo que permite mantener relación interna de los datos.

4.4. Extracción de características

Esta sección describe los procedimientos aplicados para transformar los datos brutos recolectados en representaciones numéricas estructuradas. Estas características, extraídas mediante herramientas especializadas, constituyen la base para el posterior entrenamiento y evaluación de modelos de *machine learning*.

4.4.1. Datos biométricos

Una vez finalizado el periodo mínimo de uso de 48 horas, los participantes entregan el dispositivo en el mismo lugar donde realizaron su primera participación. En ese momento, los datos se recuperan a través de la plataforma en la nube de Xiaomi, dado que cada dispositivo se encuentra vinculado a una cuenta controlada y asociada de forma segura al identificador anónimo del participante.

Este procedimiento permite descargar los datos correspondientes desde la nube, manteniendo la trazabilidad del origen sin comprometer la identidad del participante. Posteriormente, la cuenta se restablece para que el dispositivo pueda ser reutilizado por otra persona en un nuevo ciclo de recolección.

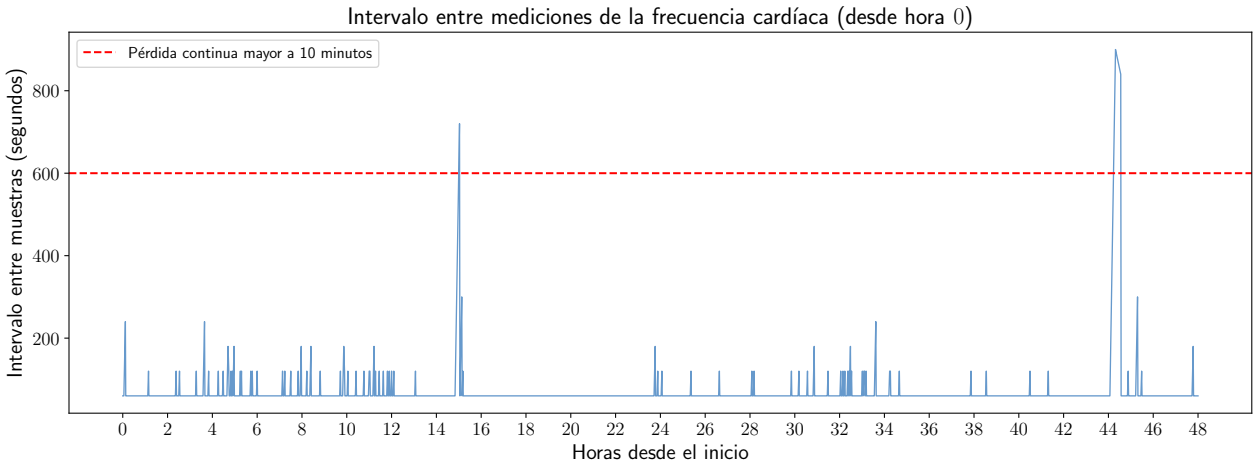


FIGURA 4.8: Ejemplificación de la frecuencia temporal de los registros de frecuencia cardíaca.

En la Figura 4.8 se muestra un ejemplo de la frecuencia temporal de los registros de frecuencia cardíaca. Es importante considerar que la pérdida de datos fue un aspecto contemplado durante el estudio, debido a posibles fallas en la recolección, como se detalla en la Sección 4.3.1, por lo que requiere un tratamiento adicional para garantizar la calidad de los datos descrito en la Sección 4.5.2.

Entre los datos más relevantes obtenidos desde el dispositivo *wearable* se distinguen dos categorías principales: métricas generales de salud y un perfil detallado del sueño. Si bien estos registros corresponden a estimaciones derivadas de los sensores incorporados y no constituyen mediciones clínicas exactas [35], su resolución y registro continuo resultan adecuados para los fines exploratorios de este estudio.

La Tabla 4.2 presenta un resumen de los datos crudos generados por el dispositivo, que sirven como base para el análisis y las interpretaciones posteriores. En ella se incluyen tanto las métricas de salud como aquellas relacionadas con las distintas fases y características del sueño, además de los descriptores de frecuencia cardíaca calculados durante el periodo nocturno.

TABLA 4.2: Resumen de métricas de salud y perfil del sueño relevantes para el estudio capturadas por los dispositivos.

Categoría	**Nombre de la métrica**
Salud	Frecuencia cardíaca [a]
Sueño	Tiempo despierto durante la noche
Sueño	Duración de sueño ligero
Sueño	Duración de sueño REM
Sueño	Duración total del sueño
Sueño	Duración de sueño profundo
Sueño	Descriptores de frecuencia cardíaca[a,b]

[a] La frecuencia cardíaca se mide en latidos por minuto (BPM).
[b] Los descriptores de frecuencia cardíaca son la media, mínima y máxima de la frecuencia cardíaca durante todo el periodo de sueño nocturno.

4.4.2. Datos audiovisuales

Un aspecto clave al trabajar con los registros audiovisuales es la automatización tanto de la segmentación como de la extracción de características, manteniendo la privacidad y la secuencia temporal de las respuestas de cada participante. Para lograrlo, a la presentación utilizada durante la entrevista se incorporan sonidos delimitadores: uno para indicar el inicio y otro para marcar el final de cada actividad o pregunta, los cuales se controlan según el avance del participante en la presentación.

Estos sonidos funcionan como marcadores que permiten delimitar de forma automática los fragmentos relevantes de audio y video. De este modo, es posible identificar con precisión las respuestas de los participantes y analizarlas posteriormente sin intervención manual, lo que además agrega una capa adicional de anonimato y protección a los datos registrados.

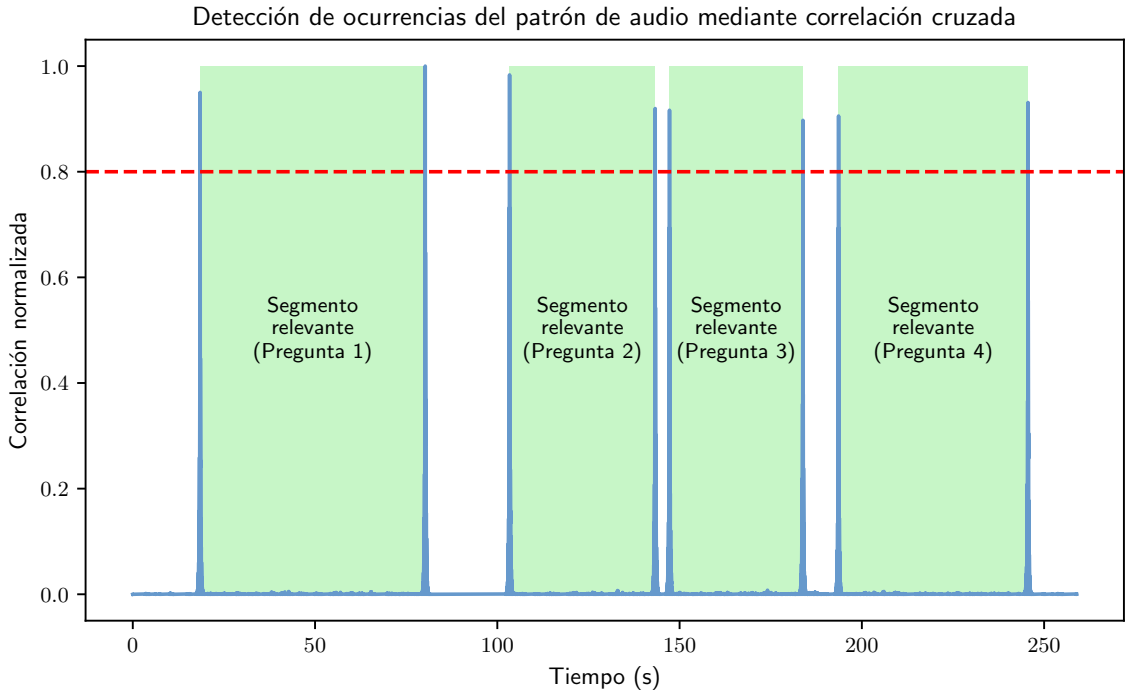


FIGURA 4.9: Correlación entre el audio de la voz del participante y el sonido incluido en la presentación de la entrevista (etapa de preguntas).

A modo de ejemplo, en la Figura 4.9 se muestra la correlación entre el audio capturado en la grabación y el sonido incorporado en la presentación durante la etapa de preguntas. En ella se

observan picos bien definidos que señalan los momentos de inicio y término de cada respuesta, lo que permite segmentar automáticamente los fragmentos relevantes para el análisis. Gracias a este procedimiento, se descartan aquellas secciones del video que no contienen información útil, optimizando el procesamiento posterior.

Extracción de rasgos faciales con OpenFace

Cada segmento de video se procesa de manera independiente, generando un archivo de salida por cada actividad y pregunta de cada participante. El archivo producido por OpenFace [30] contiene una fila por cada fotograma procesado, con un total de 709 características, descritas en la Sección 3.2.4.

Para mejorar la calidad del conjunto de datos, se conservan únicamente los fotogramas con un índice de confianza mayor a 0,85, descartando detecciones poco fiables producidas por condiciones de iluminación adversas, oclusiones parciales del rostro (manos, cabello u objetos) o movimientos bruscos de la cabeza. Posteriormente, se eliminan las columnas correspondientes a *landmarks* faciales y oculares en 2D y 3D, al no ser necesarias para el análisis planteado, dado que se utilizan las características derivadas de la detección de *landmarks*. Este filtrado reduce la dimensionalidad del conjunto de 709 a **49 características** por fotograma, conservando únicamente las categorías de interés.

Extracción de parámetros acústicos con OpenSMILE

Para el análisis de los segmentos de audio se utiliza la herramienta OpenSMILE [31] con la configuración predefinida `eGeMAPSv01a.conf`. Esta opción se selecciona por ofrecer un conjunto reducido y estandarizado de características, específicamente diseñado para el estudio de la voz en contextos afectivos [36]. Su uso permite obtener información relevante evitando la alta dimensionalidad de otros conjuntos de parámetros y manteniendo un bajo costo computacional.

El archivo de salida generado contiene un único vector de características por cada fragmento de audio, compuesto por valores agregados —como medias, desviaciones estándar, percentiles, rangos y pendientes— en lugar de mediciones de una serie temporal. La mayoría de las características principales, como la frecuencia fundamental (F0) y la intensidad, incluyen múltiples descriptores

estadísticos, lo que enriquece la representación acústica. Este formato simplifica la integración con otras modalidades, como características faciales o biométricas, sin necesidad de sincronizar cada instante de la señal.

Entre las principales categorías de características extraídas se encuentran:

- Frecuencia fundamental (F0): relacionada con la entonación y el tono de voz.
- Intensidad y energía: indicadores de proyección vocal y variabilidad expresiva.
- Parámetros espectrales: como formantes, coeficientes cepstrales (MFCCs) y pendientes de espectro, que aportan información sobre timbre y articulación.
- Medidas de perturbación: jitter, shimmer y relación armónicos-ruido (HNR), asociadas a la estabilidad y calidad de la voz.
- Dinámica temporal: métricas sobre la estructura de segmentos sonoros y silencios.

En total, la configuración `eGeMAPSv01a` produce **88 características** por cada segmento de audio analizado, cada una representada mediante sus respectivos descriptores estadísticos cuando corresponde. Estas características son ampliamente utilizadas en estudios de análisis de voz y han demostrado ser efectivas para capturar aspectos emocionales y psicológicos en la comunicación verbal [36, 37].

Transcripción y embeddings textuales con Whisper y BETO

El procesamiento de los segmentos de audio correspondientes a las respuestas verbales se realiza mediante el modelo Whisper, desarrollado por OpenAI, el cual permite generar transcripciones automáticas de audio a texto con alta precisión y soporte para múltiples idiomas, incluido el español, como se describe en la Sección 3.2.4. Cada archivo de audio se transcribe en su totalidad, generando un texto limpio por cada respuesta. Esta transcripción automática asegura consistencia en el registro verbal y reduce la posibilidad de sesgos humanos en el proceso de anotación.

Con el texto transcrito, se obtienen representaciones semánticas a través de BETO [28], un modelo de lenguaje entrenado específicamente para el español. Cada respuesta se transforma en un vector de *embeddings* con una **dimensión de 768** (véase Sección 3.2.2), lo que permite capturar de

forma numérica el contenido semántico y contextual de las respuestas de los participantes. Estas representaciones vectoriales facilitan la comparación entre respuestas, así como su integración con modalidades no verbales para el entrenamiento de modelos multimodales.

De esta manera, la información textual obtenida complementa a los indicadores biométricos, faciales y acústicos, aportando una capa semántica que enriquece la interpretación global del estado emocional y psicosocial de cada participante.

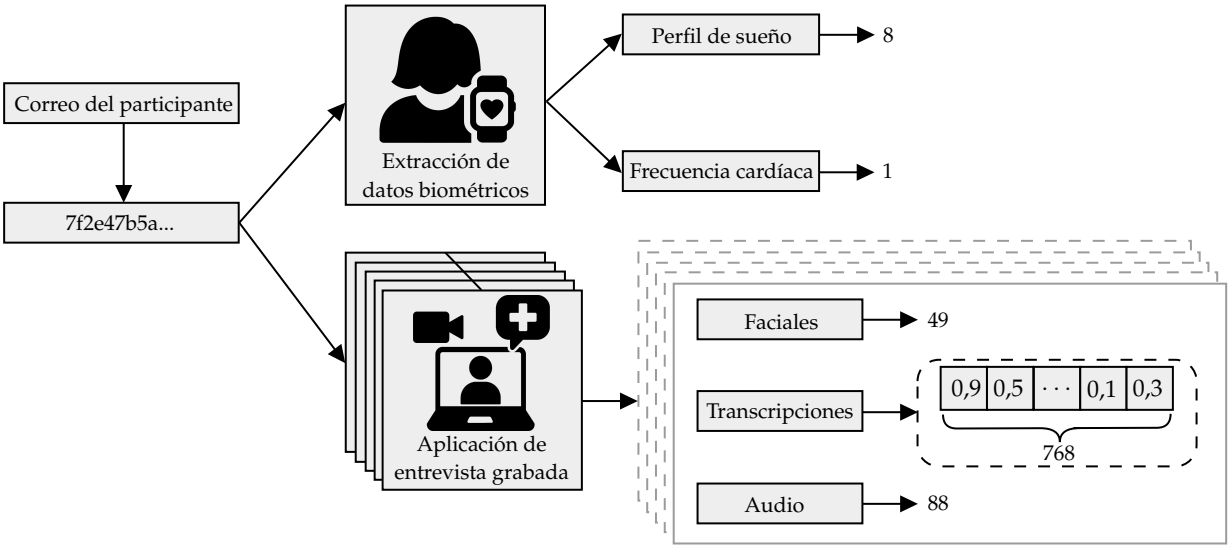


FIGURA 4.10: Flujo de extracción de características biométricas, faciales, acústicas y textuales, con la dimensionalidad resultante en cada modalidad.

En resumen, la Figura 4.10 representa el flujo completo del proceso de extracción de características, que inicia con la captura de las entrevistas grabadas y la recolección de datos biométricos. Posteriormente, cada modalidad se procesa para generar vectores de características según lo descrito en las secciones anteriores, asegurando un marco metodológico unificado que permite la futura integración de la información en un modelo multimodal de análisis.

4.5. Validación y control de calidad de datos

Con el objetivo de garantizar la confiabilidad de las modalidades que conforman el análisis multimodal, se implementa un proceso de validación y control de calidad tanto para las transcripciones de audio como para los datos biométricos obtenidos de los dispositivos *wearables*.

4.5.1. Validación de datos textuales

La calidad de las transcripciones obtenidas mediante el sistema de reconocimiento automático de voz (ASR) se evalúa a partir de la lectura controlada de un párrafo estándar incluida en la **etapa 1** de la entrevista (véase Sección 4.3.2). Esta instancia proporciona un texto de referencia idéntico para todos los participantes, lo que permite cuantificar objetivamente la precisión del sistema de transcripción y estimar su desempeño esperado en las respuestas a preguntas abiertas.

Las transcripciones se normalizan convirtiendo todo el texto a minúsculas, eliminando tildes, suprimiendo signos de puntuación y reduciendo los espacios múltiples a un solo espacio.

Métricas de evaluación

Se calcula la métrica *Word Error Rate* (WER), ampliamente utilizada en la evaluación de sistemas de reconocimiento automático de voz. El WER cuantifica el número de errores cometidos en una transcripción considerando sustituciones, eliminaciones e inserciones de palabras respecto a un texto de referencia.

Formalmente, el WER se define como:

$$WER = \frac{S + D + I}{N} \tag{4.1}$$

donde S corresponde al número de sustituciones, D al de eliminaciones, I al de inserciones y N al número total de palabras en la transcripción de referencia.

El uso del WER se justifica en este estudio porque las transcripciones automáticas constituyen la base para la generación de embeddings textuales. En este contexto, los errores de sustitución, inserción u omisión de palabras pueden alterar el contenido semántico de las representaciones, lo que afecta directamente la calidad del vector multimodal consolidado y, en consecuencia, el desempeño de los modelos de clasificación.

Resultados y criterios de aceptación

Los resultados que se presentan en la Figura 4.11 muestran que la mayoría de los participantes obtiene transcripciones exactas (WER $= 0$). Un grupo menor presenta entre una y cuatro palabras incorrectas, lo que corresponde a un rango de 0,05 $\leq$ WER $\leq$ 0,20, considerando un total de 20 palabras en el párrafo de referencia. No se registran casos con errores superiores a seis palabras (WER $>$ 0,25). En este contexto, se aceptan todas las transcripciones para el análisis, bajo el supuesto de que el desempeño observado se mantiene en las respuestas a las preguntas abiertas.

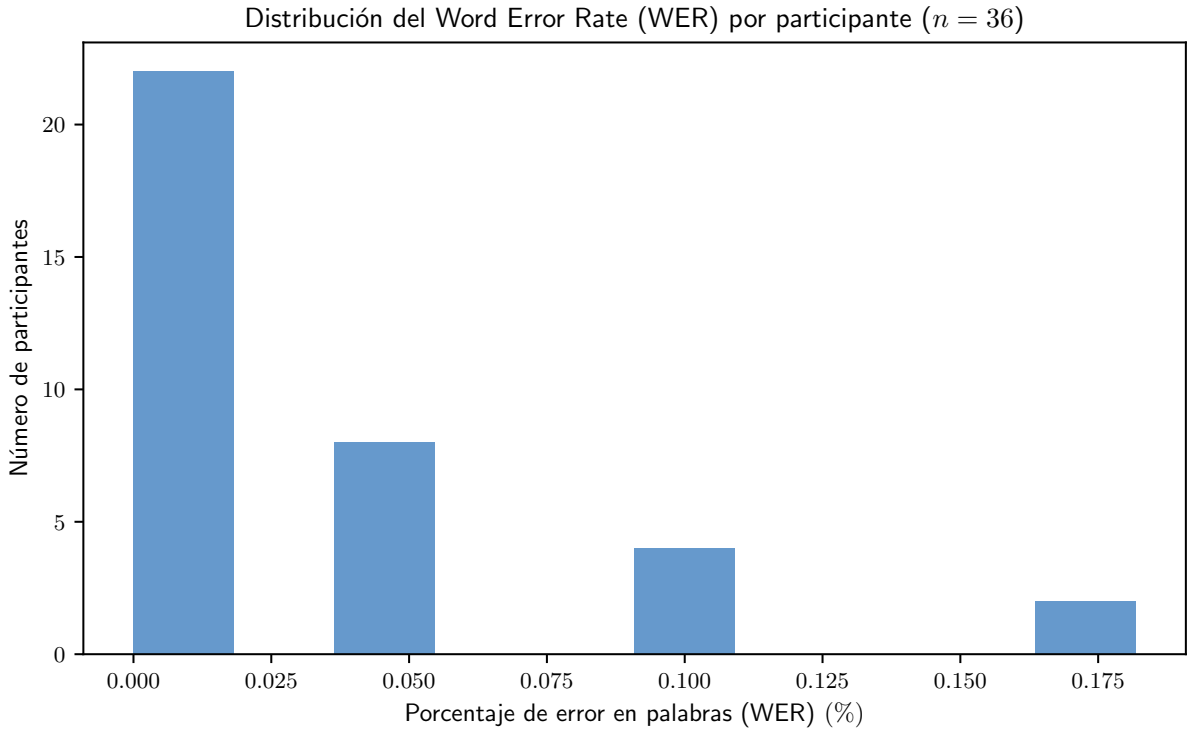


FIGURA 4.11: Distribución del porcentaje de error en palabras (*Word Error Rate*, WER) obtenido en la validación de transcripciones por participante.

4.5.2. Control de los datos biométricos

El análisis se centra en la señal de frecuencia cardíaca registrada por dispositivos *wearable* con una resolución de un minuto. Se identifican intervalos con ausencia de datos, atribuibles principalmente a pérdidas de conectividad o problemas de sincronización. En esta etapa no se evalúa la calidad

de la medición realizada por los sensores, sino que se enfatiza únicamente la cuantificación de la pérdida de datos durante el periodo mínimo de 48 horas de uso del dispositivo.

Cuantificación de datos capturados

Durante el periodo de 48 horas de uso del dispositivo se espera la captura de un registro por minuto, equivalente a 2880 mediciones por participante. No obstante, se anticipa una pérdida mínima de datos ocasionada por factores externos, como el tiempo requerido para la devolución del dispositivo, que puede variar según la disponibilidad del participante. Esta pérdida se estima en un rango de 1 a 2 horas (60 a 120 registros).

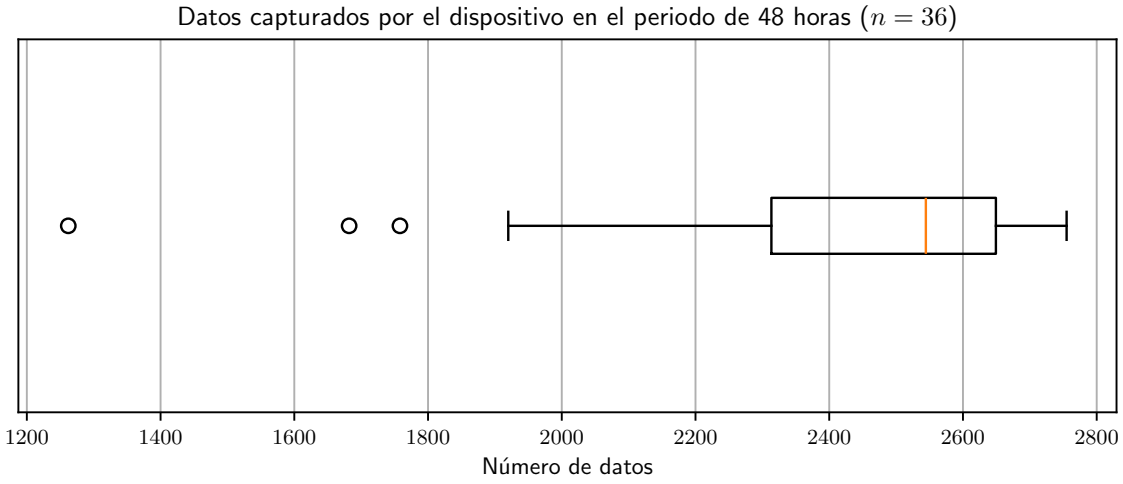


FIGURA 4.12: Distribución de la cantidad de datos capturados por cada participante durante el periodo de 48 horas de medición con el dispositivo.

La Figura 4.12 muestra la distribución de la cantidad de datos capturados por los dispositivos durante el periodo de 48 horas de medición en los 36 participantes. Se observa que, en la mayoría de los casos, la captura de registros se aproxima al valor esperado (2880 registros), con una mediana cercana a los 2500 datos. Sin embargo, también se identifican algunos valores atípicos con una cantidad considerablemente menor de mediciones (alrededor de 1200 a 1700), lo que evidencia pérdidas asociadas a factores externos o problemas de sincronización. El rango intercuartílico se concentra entre aproximadamente 2300 y 2700 registros, reflejando que, pese a las pérdidas ocasionales, la mayoría de los participantes alcanzó una cobertura de datos cercana a lo planificado.

Tratamiento de datos faltantes

Dado que el porcentaje de datos perdidos dentro del periodo de captura es bajo en la mayoría de los casos, se aplica una estimación de valores faltantes mediante interpolación lineal. Este procedimiento se limita al intervalo comprendido entre el primer y el último registro del dispositivo, con el fin de preservar la variabilidad natural de la señal. De este modo, no se fuerza la serie a alcanzar las 2880 mediciones teóricas.

La Figura 4.13 muestra la distribución de la cantidad de registros perdidos por diversos motivos. En general, las pérdidas se mantienen en valores bajos, aunque se observan algunos casos puntuales con mayor nivel de interrupción.

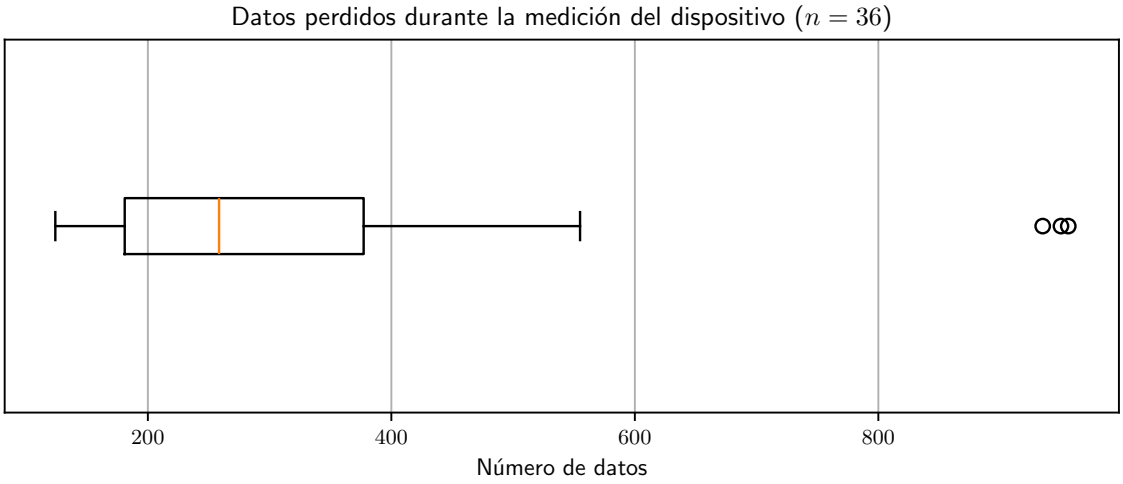


FIGURA 4.13: Distribución de la cantidad de datos perdidos por participante durante el periodo de medición del dispositivo.

La Figura 4.14 ilustra la cantidad de registros disponibles tras la interpolación aplicada, donde la mayoría de los participantes alcanzan valores cercanos al total teórico esperado.

Este ajuste permite recuperar de manera adecuada la continuidad de las series temporales y recalcular métricas en los perfiles de sueño, en particular aquellas relacionadas con los descriptores de la frecuencia cardíaca, que podrían verse afectadas por la pérdida de datos en distintos periodos.

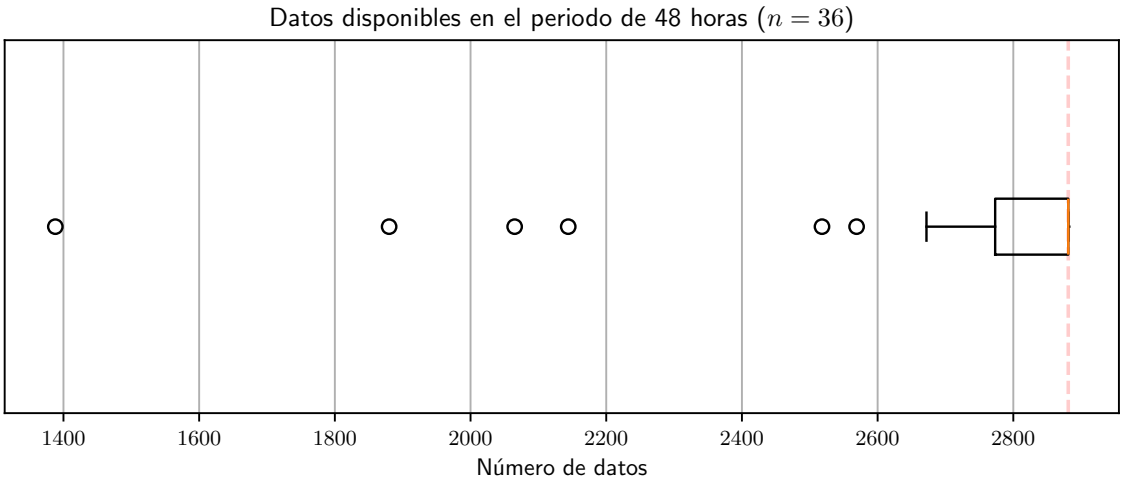


FIGURA 4.14: Distribución de la cantidad de datos disponibles por participante tras la estimación de valores faltantes mediante interpolación lineal, durante el periodo de 48 horas de medición. La línea roja punteada indica el valor teórico esperado de 2880 mediciones.

4.5.3. Aplicación de descriptores estadísticos

Con el objetivo de ampliar el análisis de la señal de frecuencia cardíaca durante el periodo de medición, se aplican los descriptores estadísticos descritos en la Sección 3.2.1 a la serie temporal obtenida tras la interpolación de los datos faltantes. De esta manera, se pasa de contar con una única característica (registro de la frecuencia cardíaca por minuto) a un conjunto de diez métricas que describen de forma más completa su comportamiento en el periodo de estudio.

El mismo procedimiento se aplica a las series temporales correspondientes a las características extraídas mediante OpenFace, abarcando las variables asociadas a la pose de la cabeza, las unidades de acción facial y la dirección de la mirada. En este caso, se calculan todas las métricas excepto la entropía, de modo que cada variable queda representada por un conjunto de descriptores que resumen su comportamiento global durante el periodo analizado y permiten identificar patrones y variaciones relevantes.

Este enriquecimiento de la representación posibilita capturar no solo las propiedades de tendencia central y dispersión, sino también aspectos relacionados con la forma y la complejidad de la distribución de las señales, favoreciendo un análisis más profundo de las variaciones fisiológicas

y conductuales.

4.6. Modelación y evaluación

El conjunto de datos utilizado para el modelamiento se genera y prepara siguiendo la metodología descrita en las secciones anteriores, a partir de cinco instancias por participante —cuatro correspondientes a preguntas abiertas y una a la actividad de lectura controlada—. Cada una de estas instancias se registra en tres modalidades principales: audio (88 características), rasgos faciales con sus descriptores estadísticos ($49 \times 9 = 441$ características) y *embeddings* textuales obtenidos de las transcripciones (vector de 768 dimensiones). De forma complementaria, se incorporan métricas derivadas de la señal de frecuencia cardíaca y características relacionadas con el perfil de sueño del participante, obtenidas a partir del monitoreo continuo.

No obstante, la actividad de lectura controlada no se considera para el modelamiento, ya que su propósito principal es proporcionar un punto de referencia uniforme para la validación de las transcripciones y, de este modo, incrementar la fiabilidad de los datos lingüísticos obtenidos (véase Sección 4.5.1), más que capturar señales emocionales.

En esta etapa se busca probar de manera experimental la viabilidad de integrar estas modalidades heterogéneas en un solo modelo, más que optimizar el rendimiento absoluto. La naturaleza piloto del estudio implica que la muestra es reducida (36 participantes), con una dimensionalidad inicial elevada, lo que incrementa el riesgo de sobreajuste y limita la capacidad de generalización de los resultados.

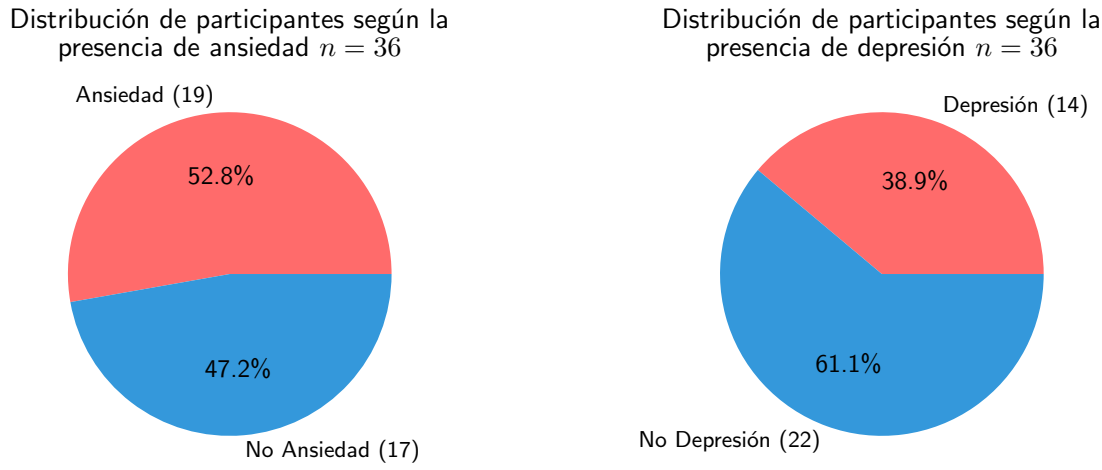
El tamaño reducido aumenta la probabilidad de que el conjunto de entrenamiento y el de prueba no presenten una distribución perfectamente equilibrada de las etiquetas, lo que puede introducir sesgos en la evaluación y dificultar la interpretación del rendimiento. Esta situación resulta especialmente relevante si las clases de interés no están representadas de manera uniforme en la muestra, ya que un desbalance en la proporción de etiquetas entre entrenamiento y prueba puede favorecer la predicción de la clase mayoritaria y subestimar la capacidad real del modelo para discriminar la minoritaria.

4.6.1. Enfoque del problema y construcción de etiquetas

El modelamiento se plantea como un problema de clasificación supervisada, en el cual las etiquetas se asignan a partir de los puntajes obtenidos en el instrumento PHQ-4 aplicado en la etapa de caracterización de participantes (véase Sección 4.2). Dado que este cuestionario evalúa sintomatología de ansiedad y depresión mediante sus subescalas GAD-2 y PHQ-2, se construyen dos clasificadores binarios independientes según las escalas de los instrumentos (véase Tabla 4.1):

- **Clasificador de ansiedad:** etiqueta positiva si el participante presenta síntomas de ansiedad y negativa en caso contrario.
- **Clasificador de depresión:** etiqueta positiva si el participante presenta síntomas de depresión y negativa en caso contrario.

El análisis previo de la distribución de etiquetas en ambos clasificadores permite identificar el grado de balance entre clases, lo que a su vez orienta la selección de métricas de evaluación más adecuadas que la *accuracy* en contextos potencialmente desbalanceados. La Figura 4.15a muestra la proporción de participantes en las clases positiva y negativa para ansiedad, mientras que la Figura 4.15b presenta la distribución equivalente para depresión.



(A) Ansiedad. (B) Depresión.

FIGURA 4.15: Distribución de etiquetas binarias según el instrumento PHQ-4 ($n = 36$), para las condiciones de ansiedad (izquierda) y depresión (derecha). Convención de color: rojo = presencia del síntoma, azul = ausencia.

4.6.2. Descripción y partición del conjunto de datos

Previo al modelamiento, se consolida el conjunto final de características combinando todas las modalidades extraídas: facial, acústica, textual y biométrica. Cada participante queda representado por un único vector resultante de la concatenación de todas las modalidades, generando así una representación multimodal completa.

La Tabla 4.3 resume las modalidades utilizadas, la dimensionalidad resultante de cada bloque de características, el tipo de agregación temporal aplicado y observaciones relevantes sobre su procesamiento.

TABLA 4.3: Resumen de modalidades y características utilizadas en el modelamiento

Modalidad	**Dims.**[a]	**Agregación temporal**	**Observaciones**
Facial	441[c]	Agregación estadística por característica	Filtrado por índice de confianza > 0.85 para descartar detecciones poco fiables.
Acústica	88	Vector agregado por pregunta	Descriptores estadísticos predefinidos para análisis afectivo.
Textual	768	Vector por pregunta	Representación densa del contenido semántico de las respuestas transcritas.
Biométrica	18	Vector único por participante	Incluye descriptores de frecuencia cardíaca y métricas del periodo de sueño[b].

[a] Dimensión del bloque de características.
[b] Descriptores: tiempo despierto, sueño ligero, REM, profundo y total.
[c] Se realizó la aplicación de descriptores estadísticos a cada una de las características obtenidas por cada una de las preguntas de la entrevista realizada al participante.

La concatenación de estas modalidades genera un vector de **5206 características por participante**, para un total de 36 participantes en el conjunto de datos. Cada participante cuenta con dos etiquetas binarias independientes, correspondientes a los clasificadores de ansiedad y depresión, dado que se construyen dos clasificadores independientes. De estas características, 1297 provienen de cada una de las preguntas de la entrevista realizada a los participantes.

Para la evaluación, el conjunto se divide en un **70 % para entrenamiento** y un **30 % para prueba**, lo que corresponde a 25 y 11 participantes, respectivamente. La división se lleva a cabo mediante

muestreo estratificado, con el fin de preservar la proporción original de clases en cada clasificador. La Tabla 4.4 presenta la distribución de participantes y etiquetas en cada subconjunto.

TABLA 4.4: Distribución de participantes y etiquetas en la partición 70/30.

Clasificador	**Conjunto**	**# Participantes**	**Positivas**	**Negativas**
Ansiedad	Entrenamiento	25	13	12
	Prueba	11	6	5
Depresión	Entrenamiento	25	10	15
	Prueba	11	4	7

4.6.3. Integración multimodal

Con el fin de aprovechar de manera conjunta la información proveniente de las distintas fuentes de datos, se diseña un esquema de integración multimodal. Este procedimiento busca combinar de forma coherente las señales faciales, acústicas y textuales, así como las métricas biométricas complementarias, con el objetivo de construir una representación unificada de cada participante, que tras la reducción de dimensionalidad resulta más compacta e interpretable que un vector de características concatenado en bruto.

Cada entrevista se divide en cuatro instancias correspondientes a las preguntas abiertas, tratadas como bloques independientes de características. Cada bloque integra tres modalidades principales: facial, acústica y textual. En esta última, las transcripciones se representan mediante *embeddings*, generando vectores de 768 dimensiones por cada respuesta. La elevada dimensionalidad resultante, combinada con el reducido número de participantes, incrementa de manera considerable el riesgo de sobreajuste y limita la posibilidad de interpretar directamente los resultados.

Para mitigar esta problemática, se plantea de forma experimental un primer enfoque de integración directa, consistente en la concatenación de todas las modalidades junto con las métricas biométricas. Posteriormente, se aplica una normalización del conjunto seguida de una reducción de dimensionalidad mediante PCA, preservando el 90 % de la varianza.

La Figura 4.16 ilustra el procedimiento de este enfoque de integración directa, considerando los datos en su estado original.

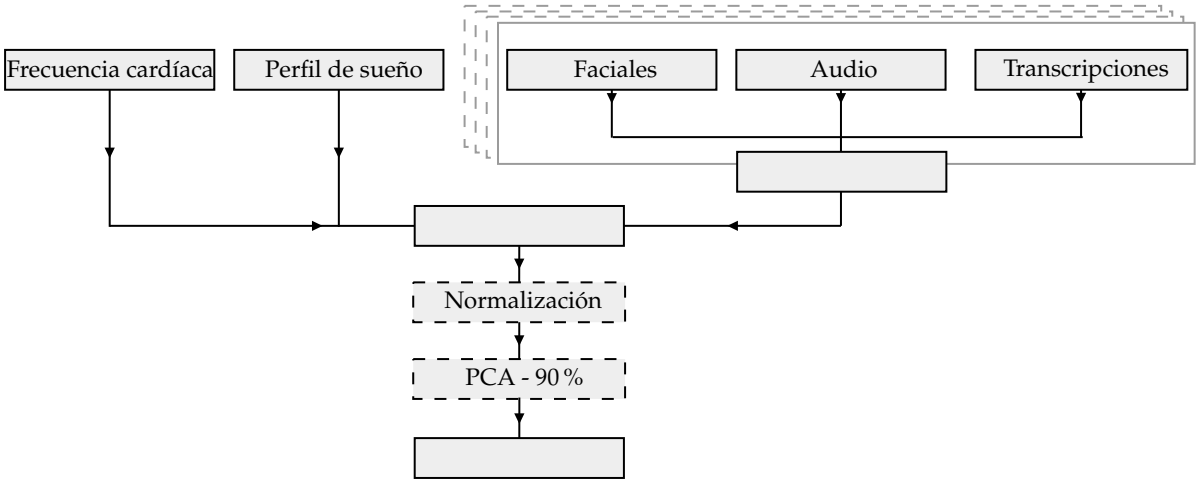


FIGURA 4.16: Integración multimodal de datos faciales, acústicos y textuales, junto con métricas biométricas (enfoque directo).

4.6.4. Selección de características relevantes

Con el fin de complementar el enfoque anterior, se explora una segunda estrategia que incorpora un proceso de selección de características previo a la integración multimodal. Para ello, se aplica un análisis de correlación de *Pearson* entre cada variable y la etiqueta de interés, considerando únicamente el conjunto de entrenamiento para evitar sesgos en la evaluación. Este coeficiente se emplea por su uso en la literatura para la selección de características extraídas con herramientas como OpenFace, como es el caso de este estudio, replicando parcialmente la metodología de trabajos previos [23].

El análisis se realiza de forma independiente para los dos problemas de clasificación (véase Sección 4.6.1), aplicando un umbral mínimo sobre el valor absoluto de la correlación para retener únicamente las variables más informativas, según el coeficiente de *Pearson* (véase Sección 3.2.3). Además, para capturar posibles variaciones contextuales, el procedimiento se replica de manera separada en cada una de las preguntas abiertas, lo que permite identificar tanto atributos consistentes a lo largo de los distintos estímulos como otros cuya relación con la etiqueta depende del contenido y la carga emocional de la pregunta.

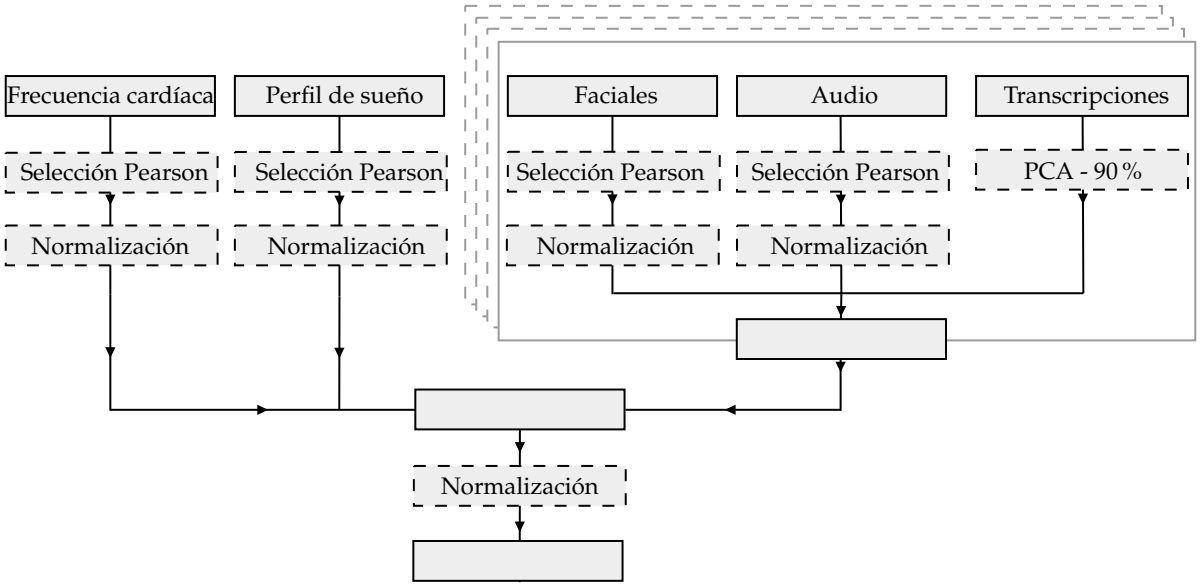


FIGURA 4.17: Pipeline de integración multimodal con selección de características basada en correlación de *Pearson*.

El pipeline de este enfoque se presenta en la Figura 4.17, donde se muestra la secuencia de pasos seguida: en cada bloque de pregunta se realiza la selección de características por modalidad junto con su respectiva normalización.

En el caso particular de las transcripciones, se aplica además una reducción de dimensionalidad mediante PCA. Finalmente, los vectores resultantes se integran en una representación multimodal unificada, la cual se normaliza nuevamente antes de utilizarse en el entrenamiento del modelo.

De manera complementaria, en las Figuras 4.18 y 4.19 se presentan los resultados del análisis exploratorio, destacando las cinco características con mayor correlación respecto de cada etiqueta, desglosadas por pregunta abierta. En términos generales, en ambas condiciones se observa un predominio de variables faciales, lo cual resulta esperable dada su elevada dimensionalidad (441 frente a 88 acústicas).

No obstante, también se identifican variables acústicas con niveles de correlación comparables a las faciales, lo que sugiere que la señal de audio constituye una fuente complementaria de información relevante.

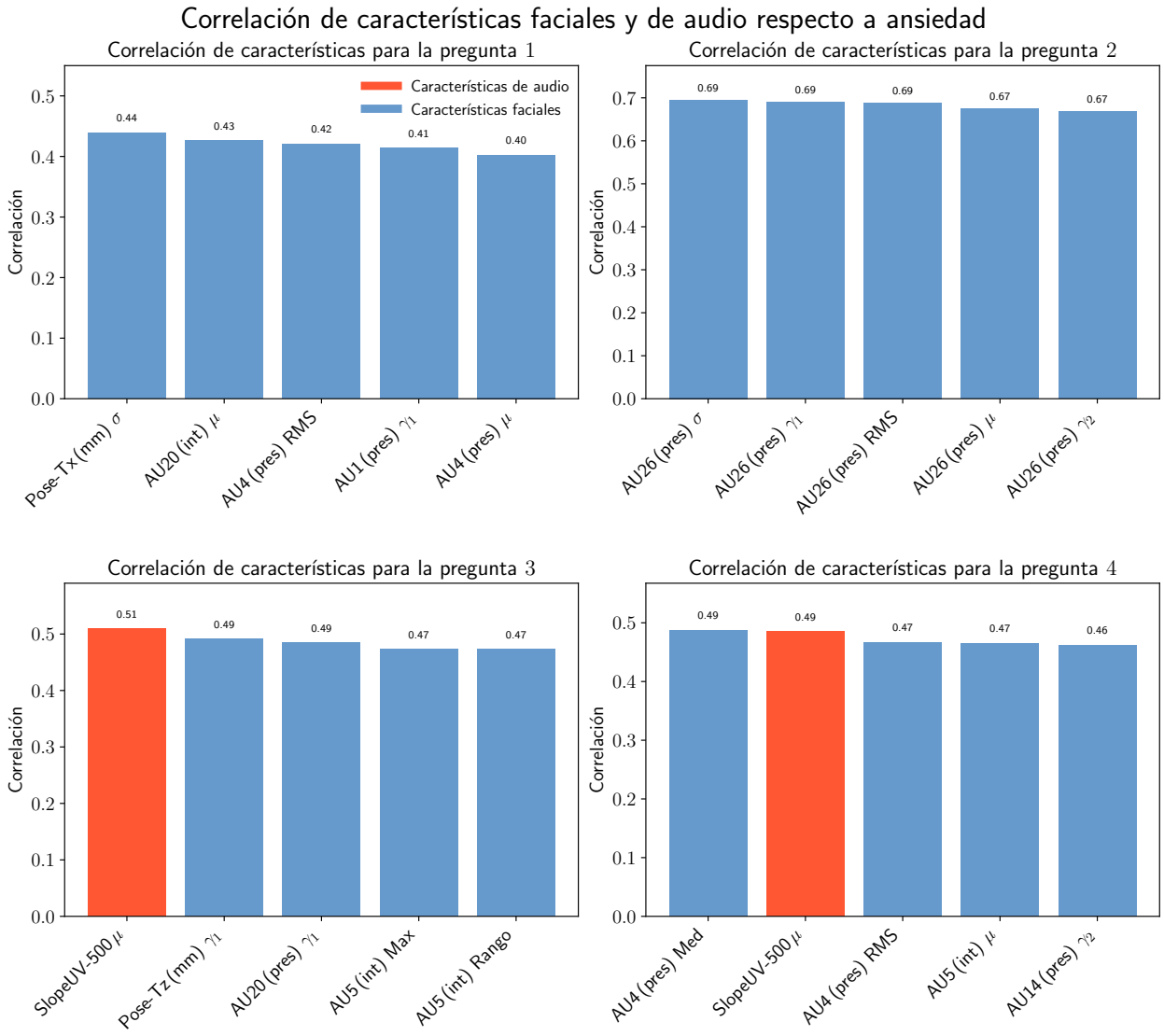


FIGURA 4.18: Top 5 características más correlacionadas con la etiqueta de ansiedad, evaluadas por pregunta abierta.

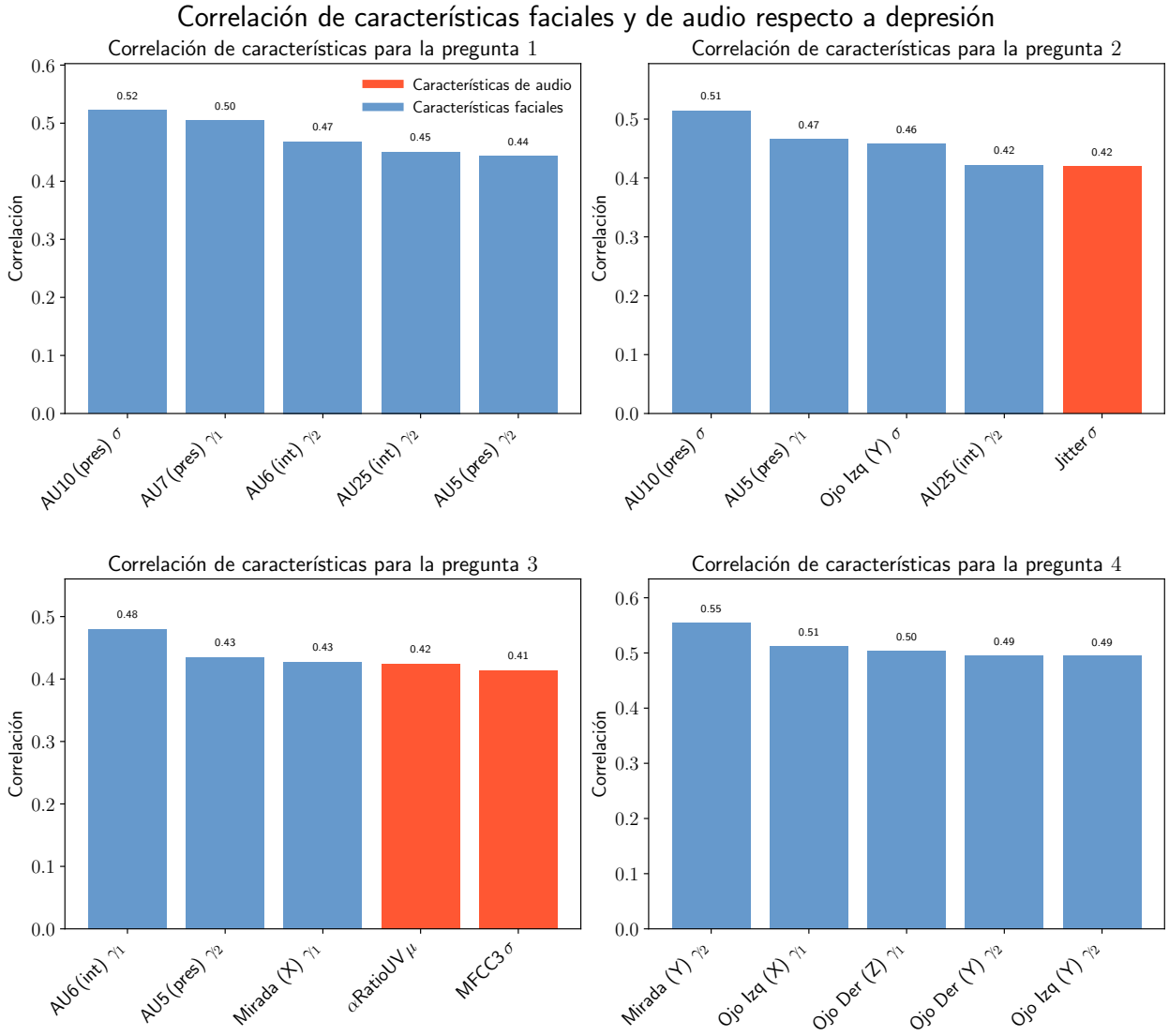


FIGURA 4.19: Top 5 características más correlacionadas con la etiqueta de depresión, evaluadas por pregunta abierta.

4.6.5. Entrenamiento preliminar

Con el objetivo de evaluar de forma preliminar el rendimiento de los vectores obtenidos tras la integración multimodal —considerando los enfoques descritos en las secciones anteriores—, se implementa un esquema de validación cruzada estratificada. Este procedimiento permite estimar la capacidad de generalización de los modelos a partir de un conjunto reducido de participantes,

manteniendo en cada partición la proporción original de clases.

Se emplea `StratifiedKFold` con cinco particiones (*folds*), aleatorizando la asignación de las muestras antes de cada división. En cada iteración, cuatro particiones se destinan al entrenamiento y la restante a prueba, rotando estos roles hasta completar las cinco evaluaciones. Este enfoque resulta especialmente adecuado en escenarios con pocos datos —como en este caso, con 36 participantes—, ya que asegura que cada instancia participe tanto en el entrenamiento como en la validación, reduciendo la dependencia de una única división del conjunto. Cuando corresponde (véase Sección 4.6.2), la validación cruzada se aplica exclusivamente sobre el subconjunto de entrenamiento, resguardando que la selección de modelos e hiperparámetros no incorpore información del conjunto de prueba.

En cada modelo se realiza un ajuste de hiperparámetros mediante búsqueda en malla (*Grid Search*), explorando múltiples combinaciones para *Random Forest*, *Support Vector Machine* y *Logistic Regression*. Tanto la selección de características basada en correlación de *Pearson* como la reducción de dimensionalidad (PCA) se incluyen dentro del *pipeline* de aprendizaje, por lo que se ajustan exclusivamente con los datos de entrenamiento de cada *fold*, evitando fuga de información. La métrica empleada para seleccionar el mejor modelo en la búsqueda es *F1-score* (ver detalles en los resultados), manteniéndose el mismo espacio de búsqueda y los mismos clasificadores en ambos esquemas de integración, de modo que la única diferencia entre evaluaciones corresponde a la estrategia de integración multimodal empleada.

4.6.6. Evaluación comparativa del modelo multimodal

Con el objetivo de analizar el efecto de las dos estrategias de integración propuestas, se evalúan de manera independiente dos configuraciones experimentales para cada problema de clasificación (ansiedad y depresión). La primera corresponde al enfoque de integración directa descrito en la Sección 4.6.3, en el cual se concatenan todas las modalidades junto con las métricas biométricas y se aplica únicamente una reducción de dimensionalidad mediante PCA. La segunda incorpora el procedimiento de selección de características descrito en la Sección 4.6.4, donde se retienen únicamente las variables que superan el umbral de correlación definido por modalidad y pregunta, integrándose posteriormente en un vector consolidado por participante.

En ambos casos, el rendimiento se calculó exclusivamente sobre el conjunto de prueba definido previamente (véase Sección 4.6.2), utilizando los mismos clasificadores —*Random Forest*, *Support Vector Machine* y *Logistic Regression*— y los hiperparámetros resultantes del mismo espacio de búsqueda, de manera que la única diferencia entre configuraciones corresponde a la estrategia de integración empleada. Las métricas consideradas fueron *F1-score*, *ROC-AUC* y *Accuracy*, seleccionadas por su complementariedad en contextos con posibles desbalances de clase.

TABLA 4.5: Comparación de métricas para ansiedad entre modelo con una integración directa y con selección de características.

Enfoque	**Clasificador**	**F1-score**	**ROC-AUC**	**Accuracy**	**Recall**	**Precision**
ID	SVM	0,706	0,500	0,545	**1,000**	0,545
ID	Logistic Regression	0,615	0,533	0,545	0,667	0,571
ID	Random Forest	0,706	0,333	0,545	**1,000**	0,545
SC	SVM	0,500	0,500	0,455	0,500	0,500
SC	**Logistic Regression**	**0,800**	**0,800**	**0,818**	0,667	**1,000**
SC	Random Forest	0,600	0,633	0,636	0,500	0,750

ID: Integración Directa.

SC: Selección de Características.

Los resultados obtenidos se presentan en la Tabla 4.5 para el caso de **ansiedad** y en la Tabla 4.6 para **depresión**. En el primer caso, la estrategia con selección de características permitió un incremento claro en el desempeño, particularmente con el clasificador de *Logistic Regression*, que alcanzó valores superiores en todas las métricas evaluadas. Esto sugiere que, para este escenario, el filtrado previo contribuye a reducir el ruido derivado de la alta dimensionalidad y a facilitar la identificación de patrones relevantes.

TABLA 4.6: Comparación de métricas para depresión entre modelo con una integración directa y con selección de características.

Enfoque	Clasificador	F1-score	ROC-AUC	Accuracy	Recall	Precision
ID	SVM	0,533	0,143	0,364	**1,000**	0,363
ID	Logistic Regression	0,571	0,821	0,727	0,500	**0,667**
ID	**Random Forest**	**0,727**	**0,857**	**0,727**	**1,000**	0,571
SC	SVM	0,533	0,357	0,364	**1,000**	0,364
SC	Logistic Regression	0,444	0,643	0,545	0,500	0,400
SC	Random Forest	0,571	0,750	0,727	0,500	0,667

ID: Integración Directa.

SC: Selección de Características.

En contraste, para **depresión**, los mejores resultados se observaron en el modelo con integración directa, destacando el desempeño de *Random Forest*, que alcanzó valores superiores de F1-score, ROC-AUC y *Accuracy*. Esto podría indicar que, en este caso, la eliminación de variables mediante el criterio de selección basado en el coeficiente de correlación de Pearson redujo información potencialmente valiosa o que los patrones asociados a la etiqueta se encuentran distribuidos en un conjunto más amplio de características, lo cual dificulta su captura mediante un análisis univariado de correlación.

En conjunto, los resultados muestran que la utilidad de la selección de características no es uniforme entre las etiquetas consideradas. Mientras que para ansiedad la estrategia de filtrado previo favoreció un desempeño más robusto, en depresión el modelo con integración directa preservó más información relevante. Estos hallazgos deben entenderse en el marco del carácter exploratorio del estudio y la reducida cantidad de participantes en relación con la dimensionalidad de los datos, lo que obliga a interpretar los resultados con cautela y refuerza la necesidad de seguir evaluando estrategias multimodales más sofisticadas.

Mejor modelo para ansiedad

En el caso de **ansiedad**, el mejor desempeño se obtuvo con el clasificador de *Logistic Regression* bajo la estrategia de selección de características. Este modelo alcanzó un *F1-score* y un *ROC-AUC* superiores a los de *Random Forest* y *SVM*, lo que confirma la utilidad del filtrado previo en la reducción de ruido y la identificación de variables relevantes.

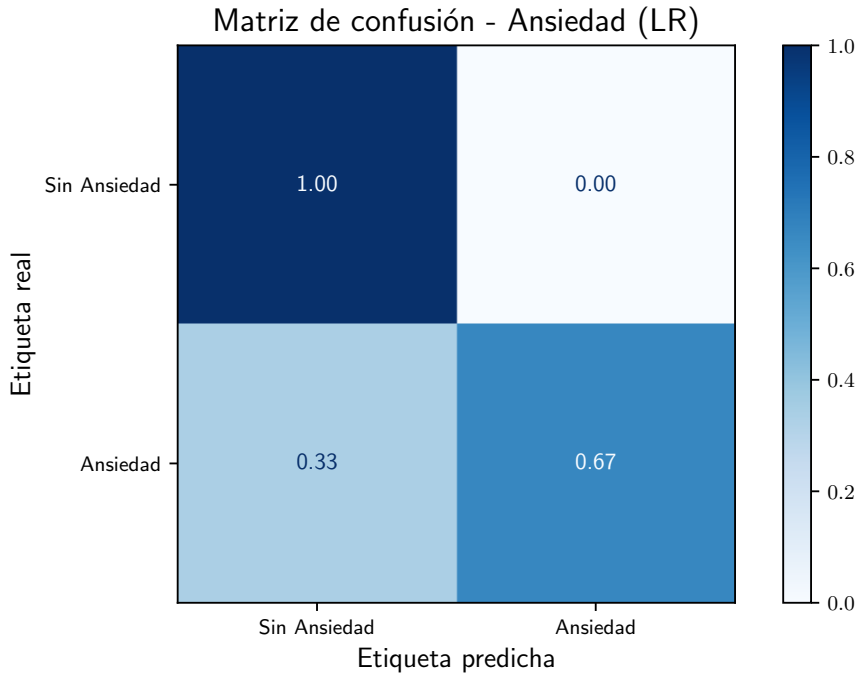


FIGURA 4.20: Matriz de confusión del mejor modelo para ansiedad (Logistic Regression, selección de características).

La Figura 4.20 presenta la matriz de confusión correspondiente, evaluada sobre el conjunto de prueba. Se observa que el modelo alcanzó una tasa de verdaderos negativos de 100 %, clasificando correctamente a todos los participantes sin ansiedad, y una tasa de verdaderos positivos de 67 %, lo que indica que identificó dos tercios de los casos con ansiedad. Estos resultados reflejan un equilibrio adecuado entre ambas medidas, aunque persisten errores en la detección de participantes con sintomatología ansiosa.

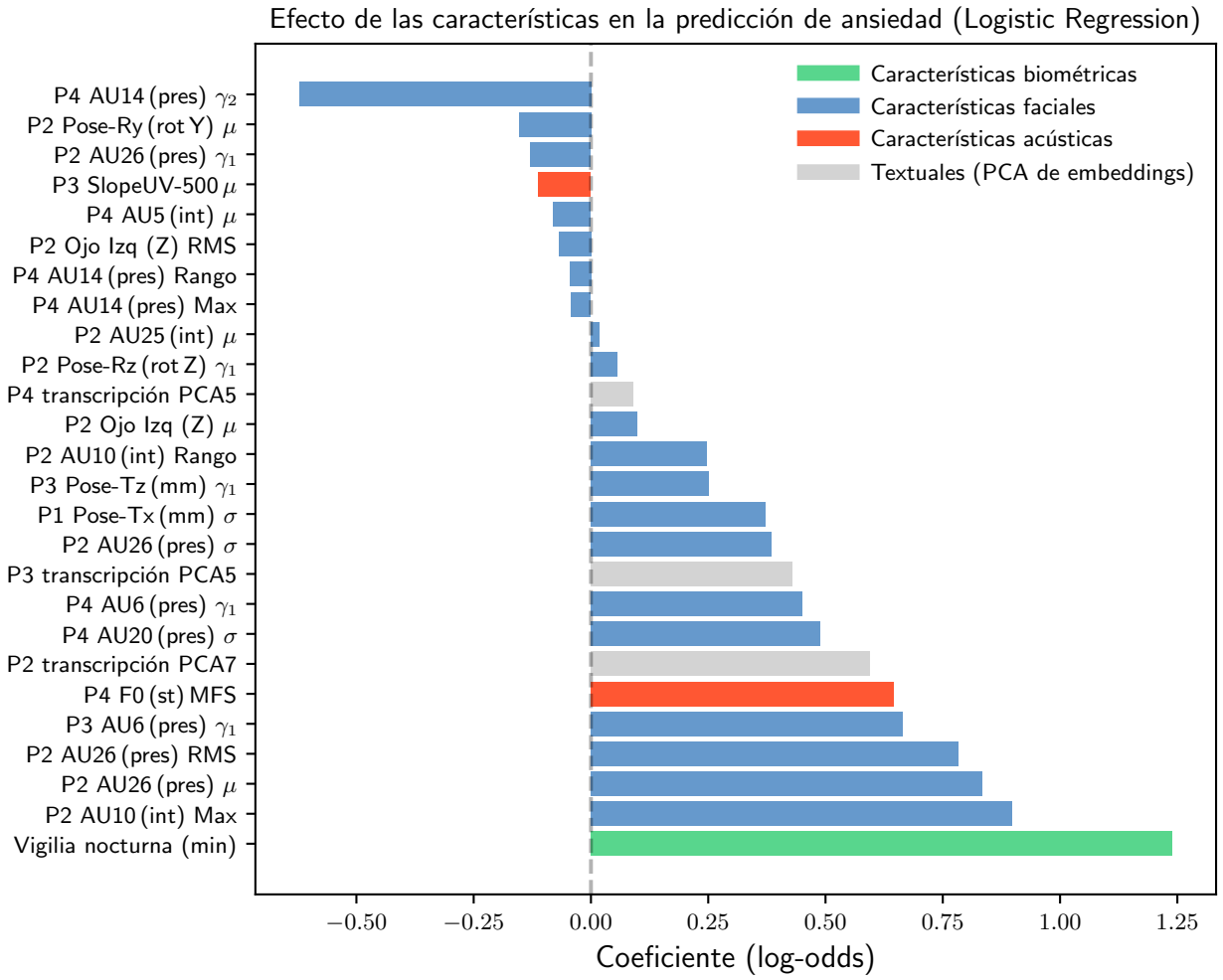


FIGURA 4.21: Efecto de las características para el modelo de Logistic Regression en ansiedad.

La Figura 4.21 presenta los coeficientes estimados por el modelo de *Logistic Regression* en la predicción de ansiedad. La vigilia nocturna (minutos despierto durante la noche) se identifica como la característica con mayor coeficiente positivo, lo que indica que un incremento en el tiempo de vigilia se asocia con una mayor probabilidad de ansiedad. En contraste, la AU14 (Dimpler, sonrisa social) muestra el coeficiente negativo de mayor magnitud, señalando que su mayor presencia se relaciona con una menor probabilidad de ansiedad. Estos resultados ponen de relieve la contribución conjunta de los biomarcadores fisiológicos del sueño y de las expresiones faciales como variables clave para la predicción de ansiedad.

Mejor modelo para depresión

Para la clasificación de **depresión**, el rendimiento más alto se obtuvo con el modelo de *Random Forest* bajo el esquema de integración directa de modalidades. A diferencia de lo observado en ansiedad, en este caso la selección de características redujo información relevante, lo que se tradujo en un peor desempeño de los modelos entrenados con esa estrategia.

La matriz de confusión del modelo ganador, mostrada en la Figura 4.22, refleja una capacidad razonable para discriminar entre participantes con y sin síntomas de depresión, aunque con cierta tendencia a favorecer la clase negativa.

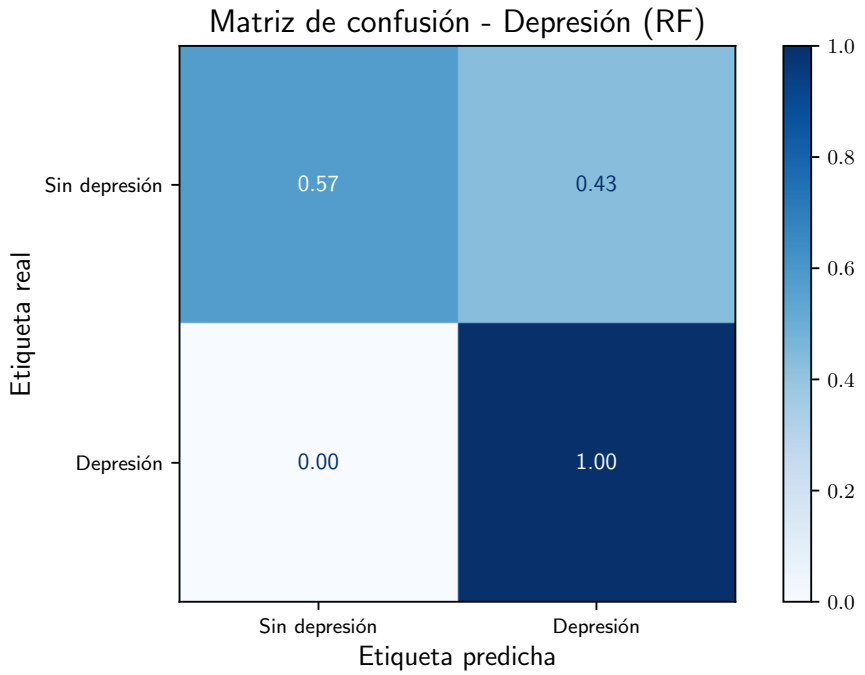


FIGURA 4.22: Matriz de confusión del mejor modelo para depresión (Random Forest, integración directa).

Conclusiones

El desarrollo del estudio demuestra que la metodología planteada es factible y coherente con los objetivos definidos. La fase de caracterización inicial de los participantes, sustentada en un instrumento sociodemográfico y psicométrico diseñado específicamente para este fin, proporciona una base sólida para el análisis. La incorporación de escalas validadas internacionalmente, como el PHQ-4, no solo entrega indicadores confiables del bienestar, sino que también asegura la comparabilidad metodológica con trabajos previos. Asimismo, la consideración de aspectos de representatividad, como el balance de género, contribuye a reducir sesgos y a mejorar la diversidad de condiciones evaluadas, reforzando la solidez del diseño experimental.

La implementación del protocolo experimental confirma la viabilidad técnica de realizar una captura sincronizada y no invasiva de datos biométricos, faciales y acústicos en un contexto universitario. Este proceso se respalda en procedimientos de validación y control de calidad, como el uso del WER para evaluar las transcripciones y la cuantificación de pérdidas en los registros generados por los dispositivos *wearables*. La metodología aplicada para la obtención y filtrado de datos permite preservar la integridad de la información, aunque se reconocen oportunidades para incorporar estrategias de validación más avanzadas en futuros estudios con el fin de incrementar la confiabilidad del sistema.

En la etapa de modelación se desarrolla un primer prototipo multimodal que integra de manera estructurada las diferentes fuentes de información —biométrica, facial, acústica y textual— en un único vector de características por participante. Este modelo, entrenado y evaluado de forma preliminar, aborda la clasificación binaria de síntomas de ansiedad y depresión utilizando el PHQ-4 como referencia, y explora estrategias de selección de características basadas en correlación de

Pearson para optimizar su rendimiento. Aunque los resultados deben interpretarse con cautela debido al tamaño reducido de la muestra, se evidencia que la integración de modalidades heterogéneas y la selección de atributos relevantes pueden mejorar de manera consistente la capacidad predictiva.

De cara al futuro, se recomienda ampliar tanto el tamaño como la diversidad de la muestra con el propósito de mejorar la capacidad de generalización del modelo. Asimismo, resulta pertinente profundizar en el análisis del aporte específico de cada modalidad —biométrica, facial, acústica y textual— y en cómo las distintas preguntas planteadas durante la entrevista inciden en el desempeño del sistema. Este tipo de evaluación permitiría identificar combinaciones de variables y estímulos más informativas, optimizando la estructura del modelo. Paralelamente, la exploración de técnicas avanzadas de integración multimodal y de métodos de aprendizaje profundo podría facilitar la detección de relaciones complejas y no lineales entre las fuentes de datos.

Bibliografía

[1] Jinnan Liu et al. "Estimation of the Global Disease Burden of Depression and Anxiety between 1990 and 2044: An Analysis of the Global Burden of Disease Study 2019". *Healthcare* 12.17 (29 de ago. de 2024), 1721. ISSN: 2227-9032.

[2] Soeun Kim et al. "Psychosocial Alterations during the COVID-19 Pandemic and the Global Burden of Anxiety and Major Depressive Disorders in Adolescents, 1990–2021: Challenges in Mental Health amid Socioeconomic Disparities". *World Journal of Pediatrics* 20.10 (oct. de 2024), 1003-1016. ISSN: 1708-8569, 1867-0687.

[3] *¿Qué es el bienestar? Definición, tipos y habilidades para el bienestar. | Psychology Today en español*. https://www.psychologytoday.com/es/blog/que-es-el-bienestar-definicion-tipos-y-habilidades-para-el-bienestar.

[4] Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado. *Depresión y ansiedad: una amenaza para tu bienestar*. http://www.gob.mx/issste/articulos/depresion-y-ansiedad-una-amenaza-para-tu-bienestar-292262.

[5] Verónica López et al. *Salud mental en estudiantes de educación superior: un desafío post-pandemia*. Policy Brief Propuestas para políticas Inclusivas. Centro de Investigación para la Educación Inclusiva, 2024.

[6] Consejo de Rectores de las Universidades Chilenas. *Comisión de Convivencia Universitaria y Salud Mental: Prevención del suicidio es un desafío de toda la comunidad universitaria*. 2023.

[7] Tianqi Jia. "A Study of the Causes and Effects of Anxiety among Students in Higher Education". *Journal of Education, Humanities and Social Sciences* 45 (26 de dic. de 2024), 111-115. ISSN: 2771-2907.

[8] Nan Gao et al. *Investigating the Reliability of Self-report Data in the Wild: The Quest for Ground Truth*. 2021. arXiv: 2107.00389 [cs].

[9] Bernd Löwe et al. "A 4-Item Measure of Depression and Anxiety: Validation and Standardization of the Patient Health Questionnaire-4 (PHQ-4) in the General Population". *Journal of Affective Disorders* 122.1-2 (abr. de 2010), 86-95. ISSN: 01650327.

[10] Xiaoqian Liu y Xiaoyang Wang. "Automatic Identification of a Depressive State in Primary Care". *Healthcare* 10.12 (22 de nov. de 2022), 2347. ISSN: 2227-9032. PMID: 36553871.

[11] Deepika Sharma et al. "Demystifying Mental Health by Decoding Facial Action Unit Sequences". *Big Data and Cognitive Computing* 8.7 (9 de jul. de 2024), 78. ISSN: 2504-2289.

[12] Louise V. Coutts et al. "Deep Learning with Wearable Based Heart Rate Variability for Prediction of Mental and General Health". *Journal of Biomedical Informatics* 112 (dic. de 2020), 103610. ISSN: 15320464.

[13] Shuroug A. Alowais et al. "Revolutionizing Healthcare: The Role of Artificial Intelligence in Clinical Practice". *BMC Medical Education* 23.1 (sep. de 2023), 689. ISSN: 1472-6920.

[14] Md Manjurul Ahsan y Zahed Siddique. *Machine Learning-Based Heart Disease Diagnosis: A Systematic Literature Review*. Dic. de 2021. arXiv: 2112.06459 [cs].

[15] Jie Mei, Christian Desrosiers y Johannes Frasnelli. *Machine Learning for the Diagnosis of Parkinson's Disease: A Systematic Review*. Oct. de 2020. arXiv: 2010.06101 [cs].

[16] Emmanuel G. Pintelas et al. "A Review of Machine Learning Prediction Methods for Anxiety Disorders". *Proceedings of the 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion*. DSAI 2018: 8th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion. Thessaloniki Greece: ACM, 20 de jun. de 2018, 8-15. ISBN: 978-1-4503-6467-6.

[17] Kevin Hilbert et al. "Separating Generalized Anxiety Disorder from Major Depression Using Clinical, Hormonal, and Structural MRI Data: A Multimodal Machine Learning Study". *Brain and Behavior* 7.3 (mar. de 2017), e00633. ISSN: 2162-3279, 2162-3279.

[18] Daniel W. Russell. "UCLA Loneliness Scale (Version 3): Reliability, Validity, and Factor Structure". *Journal of Personality Assessment* 66.1 (1996), 20-40.

[19] Gregory D. Zimet et al. "The Multidimensional Scale of Perceived Social Support". *Journal of Personality Assessment* 52.1 (1988), 30-41.

[20] Sheldon Cohen, Tom Kamarck y Robin Mermelstein. "A Global Measure of Perceived Stress". *Journal of Health and Social Behavior* 24.4 (1983), 385-396. ISSN: 00221465, 21506000.

[21] Alaa Abd-alrazaq et al. "Wearable Artificial Intelligence for Anxiety and Depression: Scoping Review". *Journal of Medical Internet Research* 25 (19 de ene. de 2023), e42672. ISSN: 1438-8871.

[22] M. L. Tlachac et al. "Voice Recordings from Short Mobile Sessions versus Clinical Interviews for Mental Illness Screening: A Comparative Study with Deep Transfer Learning". *ACM Transactions on Computing for Healthcare* 6.3 (jul. de 2025), 1-30. ISSN: 2637-8051.

[23] Ravi Prasad Thati et al. "A Novel Multi-Modal Depression Detection Approach Based on Mobile Crowd Sensing and Task-Based Mechanisms". *Multimedia Tools and Applications* 82.4 (feb. de 2023), 4787-4820. ISSN: 1380-7501, 1573-7721.

[24] Kuang Chua Chua et al. "Application of Higher Order Statistics/Spectra in Biomedical Signals—A Review". *Medical Engineering & Physics* 32.7 (2010), 679-689. ISSN: 1350-4533.

[25] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York: Springer, 2006. ISBN: 978-0-387-31073-2.

[26] *What Is Classification Threshold*. Iguazio. URL: `https://www.iguazio.com/glossary/classification-threshold/`.

[27] *What Are Vector Embeddings? | A Comprehensive Vector Embeddings Guide*. URL: `https://www.elastic.co/what-is/vector-embedding`.

[28] José Cañete et al. "Spanish Pre-Trained BERT Model and Evaluation Data". *PML4DC at ICLR 2020*. 2020.

[29] Felipe L. Gewers et al. "Principal Component Analysis: A Natural Approach to Data Exploration". *ACM Computing Surveys* 54.4 (mayo de 2022), 1-34. ISSN: 0360-0300, 1557-7341. arXiv: `1804.02502 [cs]`.

[30] Tadas Baltrušaitis et al. "OpenFace 2.0: Facial Behavior Analysis Toolkit". *13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. 2018, 59-66.

[31] Florian Eyben, Martin Wöllmer y Björn Schuller. "Opensmile: the munich versatile and fast open-source audio feature extractor". *Proceedings of the 18th ACM International Conference on Multimedia*. MM '10. Firenze, Italy: Association for Computing Machinery, 2010, 1459–1462. ISBN: 9781605589336.

[32] Alec Radford et al. *Robust Speech Recognition via Large-Scale Weak Supervision*. Dic. de 2022. arXiv: 2212.04356 [eess].

[33] *Introducing Whisper*. https://openai.com/index/whisper/. Abr. de 2022.

[34] David Adzrago, Timothy J. Walker y Faustine Williams. "Reliability and Validity of the Patient Health Questionnaire-4 Scale and Its Subscales of Depression and Anxiety among US Adults Based on Nativity". *BMC Psychiatry* 24.1 (mar. de 2024), 213. ISSN: 1471-244X.

[35] *Xiaomi Smart Band 9 Active - Xiaomi Chile*. https://www.mi.com/cl/product/xiaomi-smart-band-9-active/.

[36] Florian Eyben et al. "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing". *IEEE Transactions on Affective Computing* 7.2 (2016), 190-202. ISSN: 1949-3045.

[37] Fasih Haider et al. *Emotion Recognition in Low-Resource Settings: An Evaluation of Automatic Feature Selection Methods*. 2020. arXiv: 1908.10623 [cs].

A. Anexos

A continuación se presenta la información complementaria que refuerza y amplía el contenido del estudio

A.1. Anexo 1: Formato de consentimiento informado

Estimado (a) participante:

Ud. ha sido invitado/a participar en el estudio modelo multimodal no invasivo para la estimación preliminar del bienestar estudiantil, a cargo de Ricardo Flores Huenchullanca PhD., profesor asistente del Departamento de Ingeniería Informática y Ciencias de la Computación (DIICC) de la Universidad de Concepción.

El objetivo del proyecto es validar un modelo multimodal de caracterización de patrones, a partir de la recolección no invasiva de datos biométricos mediante dispositivos wearables, asociados al bienestar de estudiantes de ingeniería de la Universidad de Concepción.

Si usted decide participar del estudio, se le pedirá que acepte su participación a través del presente consentimiento informado. Su participación consistirá en completar un formulario de información sobre sus características sociodemográficas y screening para estrés, soledad, depresión y percepción de soledad, una breve entrevista grabada (video y audio) y la utilización de un reloj inteligente que mide sus biomarcadores (ritmo cardíaco, saturación de oxígeno, números de pasos, patrones de sueño) durante 2 días.

Un eventual riesgo podría ser que se sintiera incómodo/a mientras contesta este cuestionario. No obstante, usted es libre de dejar el estudio en cualquier momento, sin necesidad de dar ningún

tipo de explicación. La información que usted aporte será de gran valor para la investigación respecto a la relación existente entre los mecanismos de aprendizaje automático para el monitoreo de bienestar estudiantil.

Toda la información derivada de su participación será manejada con estricta confidencialidad a través de un sistema de información anonimizado. Sólo el equipo de investigación tendrá acceso a los datos por usted proporcionados. La información será resguardada según todos los requerimientos que las leyes chilenas explicitan (ley 20.120). Asimismo, tanto en el análisis como en la publicación y difusión científica de los resultados, no se identificará la identidad de ninguno de los/as participantes ni su respectiva organización, para así resguardar el anonimato. La información que entregue mediante su participación sólo será utilizada con fines científicos y relativos a esta investigación y no será usada con fines ajenos a los explícitamente expresados en este documento.

La participación en esta investigación es absolutamente voluntaria y usted puede retirarse en cualquier momento del estudio, sin que ello tenga ninguna consecuencia.

Cualquier pregunta que yo quisiera hacer con relación a mi participación en este estudio será contestada por el/la investigador/a responsable en el correo electrónico riflores@udec.cl. Para cualquier duda que no me haya sido satisfactoriamente resuelta por el investigador responsable me podré dirigir a la Dra. Sandra Saldivia Presidenta del Comité de Ética, Bioética y Bioseguridad de la Universidad de Concepción. Correo: secrevrid@udec.cl.

Después de haber recibido y comprendido la información de este documento y de haber podido aclarar todas mis dudas, otorgó el consentimiento para participar en el estudio: **Modelo multimodal no invasivo para la estimación preliminar del bienestar estudiantil**.

Comprendo y acepto la información que se entregó anteriormente y declaro conocer los objetivos del estudio.

En atención a estas consideraciones, libremente marque la que corresponda.

Yo Acepto ☐ Yo No Acepto ☐

_____

Nombre y firma del participante

A.2. Anexo 2: Descripción de características extraídas con OpenFace y OpenSMILE

Con el fin de complementar la descripción metodológica presentada en el Capítulo 4, en este anexo se detallan las principales características extraídas a partir de las modalidades facial y acústica mediante las herramientas OpenFace y OpenSMILE, respectivamente.

La inclusión de estas modalidades responde a la necesidad de contar con indicadores objetivos y cuantificables que reflejen tanto la expresividad no verbal (movimientos faciales, mirada y micro-expresiones) como la prosodia vocal (entonación, energía y calidad de la voz), dimensiones que la literatura científica ha asociado estrechamente con la manifestación de estados emocionales y alteraciones del bienestar.

En las siguientes secciones se presenta una síntesis de los tipos de variables generadas por cada herramienta, la forma en que fueron representadas y los descriptores estadísticos aplicados para consolidar un vector de características robusto y comparable entre participantes.

A.2.1. Características extraídas con OpenFace

Complementando lo descrito en el marco teórico acerca de OpenFace, en este anexo se presenta un mayor nivel de detalle respecto de las variables generadas por la herramienta. OpenFace constituye una de las implementaciones más utilizadas para la extracción automática de indicadores faciales, incluyendo medidas relacionadas con la pose de la cabeza, la dirección de la mirada y las Unidades de Acción Facial (Action Units, AUs).

Facial Action Coding System (FACS)

El *Facial Action Coding System* (FACS) define un conjunto amplio de Unidades de Acción (AUs), cada una correspondiente a la contracción de un músculo o grupo muscular específico del rostro. Estas unidades permiten descomponer expresiones faciales complejas en componentes musculares básicos y constituyen un estándar ampliamente utilizado en el análisis del comportamiento no verbal.

OpenFace implementa un subconjunto de estas AUs, seleccionadas por su relevancia y frecuencia en estudios de reconocimiento automático de emociones. Dichas AUs se presentan en la Tabla A.1.

TABLA A.1: Unidades de Acción Facial (AUs) reconocidas por OpenFace.

AU	Descripción
AU1	Levantamiento de ceja interna
AU2	Levantamiento de ceja externa
AU4	Fruncimiento de cejas
AU5	Apertura de párpados superiores
AU6	Elevación de mejillas (sonrisa genuina)
AU7	Tensión de párpados inferiores
AU9	Arrugamiento de la nariz (disgusto)
AU10	Elevación del labio superior
AU12	Sonrisa (comisuras elevadas)
AU14	Asimetría bucal (desdén)
AU15	Descenso de comisuras labiales
AU17	Elevación del mentón
AU20	Estiramiento horizontal de labios
AU23	Compresión de labios
AU25	Separación de labios
AU26	Apertura de la mandíbula
AU28	Protrusión de labios
AU45	Parpadeo

Para cada AU, OpenFace entrega dos tipos de medidas:

- **Presencia binaria**, que indica si la unidad de acción está activa en un fotograma (0 = ausente, 1 = presente).
- **Intensidad escalar**, representada en una escala continua que generalmente varía entre 0 y 5, donde valores más altos reflejan contracciones musculares más marcadas.

La combinación de presencia e intensidad permite diferenciar entre microexpresiones sutiles y gestos faciales de mayor magnitud. Por ejemplo, una sonrisa (AU12) puede detectarse con intensidades bajas como una expresión leve o social, mientras que intensidades más altas suelen asociarse a emociones positivas genuinas.

Pose de la cabeza y dirección de la mirada

OpenFace proporciona medidas de la pose de la cabeza y de la dirección de la mirada, las cuales son fundamentales para entender la orientación visual del sujeto y su interacción con el entorno.

Ambas variables se representan mediante vectores tridimensionales que describen la rotación alrededor de los ejes cartesianos x, y, z, lo que permite obtener una estimación precisa de la postura y el foco atencional.

En el caso de la **pose de la cabeza**, los valores corresponden a tres ángulos de rotación:

- **Yaw** (rotación sobre el eje vertical, y): indica la desviación horizontal de la cabeza. Valores positivos representan un giro hacia la derecha del observador, mientras que valores negativos reflejan un giro hacia la izquierda.
- **Pitch** (rotación sobre el eje horizontal, x): describe el movimiento de inclinación vertical. Valores positivos indican que el rostro se orienta hacia arriba, mientras que valores negativos implican una inclinación hacia abajo.
- **Roll** (rotación sobre el eje longitudinal, z): corresponde a la inclinación lateral del rostro. Valores positivos representan una inclinación hacia la derecha y negativos hacia la izquierda.

Por su parte, la **dirección de la mirada** se estima a partir de la posición de las pupilas y la geometría del globo ocular, generando vectores tridimensionales que describen la orientación relativa de ambos ojos respecto del rostro. Estos valores permiten identificar si la mirada está dirigida hacia el frente (valores cercanos a cero) o desviada hacia los lados, arriba o abajo. Un incremento positivo en el eje x refleja una mirada hacia la derecha, mientras que un valor negativo indica una desviación hacia la izquierda. De manera análoga, valores positivos en el eje y representan una mirada hacia arriba y valores negativos, hacia abajo.

La interpretación conjunta de la pose y de la mirada resulta relevante en contextos experimentales, ya que permite diferenciar, por ejemplo, entre un sujeto que desvía los ojos manteniendo la cabeza estable (mirada periférica) y otro que gira toda la cabeza para orientar su atención. Estas distinciones aportan información complementaria sobre la implicación atencional, el contacto visual y la expresión de estados emocionales.

A.2.2. Características extraídas con OpenSMILE

La configuración `eGeMAPSv01a.conf`, utilizada en este estudio, genera un total de 88 características acústicas por segmento de audio. Estas variables se agrupan en categorías diseñadas para capturar dimensiones prosódicas, espectrales y de calidad vocal de la señal, ampliamente utilizadas en investigaciones sobre voz y emociones [36]. La Tabla A.2 resume estas categorías.

TABLA A.2: Resumen de las 88 características acústicas extraídas con la configuración `eGeMAPSv01a` de OpenSMILE.

Categoría	Características
Frecuencia fundamental (F0)	Variaciones básicas del tono de voz: valores medios, dispersión, percentiles y pendientes de ascenso y descenso.
Intensidad y energía	Medidas de sonoridad que reflejan la fuerza y variabilidad del volumen de la voz, incluyendo tendencias crecientes y decrecientes.
Espectro	Distribución de energía en distintas frecuencias: flujo espectral, *alpha ratio*, *Hammarberg index* y pendientes espectrales.
Coeficientes cepstrales (MFCCs)	Primeros cuatro coeficientes en segmentos vocalizados y no vocalizados, que condensan la información del timbre y la resonancia de la voz.
Calidad vocal y perturbación	Indicadores de estabilidad como jitter, shimmer, relación armónicos-ruido (HNR) y medidas logarítmicas relativas al espectro de armónicos.
Formantes	Frecuencia, ancho de banda y amplitud relativa de F1, F2 y F3, relacionados con la articulación del habla y la configuración del tracto vocal.
Dinámica temporal	Organización del discurso a lo largo del tiempo: número de segmentos vocalizados, duración media de vocalizaciones y pausas.
Nivel sonoro equivalente	Potencia acústica global expresada en dB SPL, que resume la intensidad promedio de la voz en un segmento.

Frecuencia fundamental (F0)

La frecuencia fundamental refleja la vibración periódica de las cuerdas vocales y se percibe como el tono de la voz. Sus variaciones están asociadas a la entonación y a la modulación afectiva: aumentos abruptos pueden indicar sorpresa o excitación, mientras que descensos sostenidos suelen relacionarse con tristeza o fatiga emocional.

Intensidad y energía

La intensidad mide el volumen de la voz, mientras que la energía refleja la potencia acústica global. Estas dimensiones aportan información sobre el nivel de activación emocional: voces más fuertes suelen asociarse a estados de excitación o estrés, mientras que niveles bajos pueden reflejar abatimiento o retraimiento.

Características espectrales

Las variables espectrales describen cómo se distribuye la energía a lo largo de las frecuencias. Permiten caracterizar el timbre y la claridad de la articulación, aspectos que cambian con la tensión muscular y el estado emocional del hablante.

Coeficientes cepstrales (MFCCs)

Los MFCCs son una representación compacta de la envolvente espectral. Capturan la estructura armónica y resonante de la voz, y han demostrado ser especialmente útiles para diferenciar entre patrones lingüísticos y variaciones emocionales.

Calidad vocal y perturbación

Este grupo incluye medidas de estabilidad de la señal, como jitter (variaciones en la frecuencia), shimmer (variaciones en la amplitud) y la relación armónicos-ruido. Estos indicadores son sensibles a la tensión fisiológica y psicológica que afecta el control de la voz.

Formantes

Los formantes (F1, F2 y F3) corresponden a las principales resonancias del tracto vocal. Su posición y amplitud aportan información sobre la articulación y claridad del habla, además de reflejar adaptaciones fisiológicas ligadas al estado emocional.

Dinámica temporal

La dinámica temporal captura la organización del habla en el tiempo, incluyendo la duración de pausas y segmentos vocalizados. Alteraciones en este ritmo pueden reflejar cambios en la fluidez y espontaneidad del discurso, característicos en situaciones de ansiedad, estrés o depresión.

Nivel sonoro equivalente

El nivel sonoro equivalente resume la potencia media de la voz durante un intervalo. Valores más altos reflejan un mayor esfuerzo vocal, mientras que niveles bajos pueden asociarse a estados emocionales de menor activación.