UNIVERSIDAD EAFIT

# MODELACIÓN PROBABILÍSTICA Y DINÁMICA DE LA ANSIEDAD MEDIANTE TÉCNICAS DE CLUSTERING Y MODELOS OCULTOS DE MÁRKOV

**DIEGO ALEXANDER GIRALDO TIRADO**

**Tesis**

**Asesor**

**Juan Alejandro Peña Palacio**

**UNIVERSIDAD EAFIT**
**ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA**
**MAESTRÍA EN CIENCIAS DE LOS DATOS Y ANALÍTICA**
**MEDELLÍN**
**2026**

# Modelación probabilística y dinámica de la ansiedad mediante técnicas de clustering y Modelos Ocultos de Márkov

[1]Diego A. Giraldo T., [2]Alejandro Peña P.

[1] Escuela de Ciencias aplicadas e ingeniería - Universidad EAFIT, Medellín, Colombia
Email: dagiraldt1@eafit.edu.co

[2] Escuela de Ciencias aplicadas e ingeniería - Universidad EAFIT, Medellín, Colombia
Email: japena@eafit.edu.co

**Abstract.** La ansiedad se ha consolidado como un problema creciente a nivel global, con impactos significativos en el bienestar, la productividad y los costos asociados a la incapacidad. A pesar de los avances recientes en el monitoreo psicológico, persiste la necesidad de enfoques analíticos que capturen su naturaleza fluctuante y dinámica en el tiempo. En respuesta a este desafío, se propone un marco probabilístico para caracterizar el comportamiento dinámico de la ansiedad a partir de variables psicológicas y conductuales. Con el fin de identificar perfiles diferenciados y comprender las transiciones entre estados emocionales, se integran técnicas de aprendizaje no supervisado, modelos de probabilidad no normales y Modelos Ocultos de Márkov (HMM). Indicadores como el nivel de estrés percibido, las horas de sueño, la actividad física y el consumo de estimulantes se emplean para construir agrupaciones conductuales que alimentan el modelo estocástico. Esto permitió modelar la estructura estadística del indicador transformado de ansiedad a través de distribuciones continuas, con el objetivo de derivar una representación parsimoniosa de su comportamiento marginal. Los resultados evidencian una dinámica marcada por la persistencia en estados de riesgo moderado y alto, así como una distribución asimétrica que concentra los episodios más críticos en niveles bajos del indicador de bienestar. El marco resultante ofrece una base técnica para modelar la ansiedad como un proceso dinámico y estocástico, y constituye un insumo potencial para sistemas de monitoreo emocional y para la toma de decisiones en salud mental ocupacional.

**Keywords:** Ansiedad, Modelos Ocultos de Márkov (HMM), Distribuciones no normales, Clustering, Salud mental, Análisis probabilístico, Valor en Riesgo (VaR).

## 1 Introducción

La salud mental se ha consolidado como uno de los principales desafíos del siglo XXI, especialmente en el ámbito laboral. Diversos estudios han mostrado un aumento

sostenido de los trastornos de ansiedad, depresión y estrés en la población trabajadora, afectando el bienestar individual, la productividad y los costos asociados al ausentismo y la incapacidad laboral [1, 2].

La Organización Mundial de la Salud (OMS) estima que las enfermedades mentales representan cerca del 13% de la carga mundial de enfermedad, siendo la ansiedad uno de los trastornos más significativos. En entornos laborales, los síntomas de ansiedad y estrés representan más del 50% de los casos de morbilidad relacionada con el trabajo [3].

En la literatura se pueden encontrar una serie de tendencias de desarrollo entre las que se cuenta los avances tecnológicos en la modelación del estrés y la ansiedad. El creciente interés en el monitoreo continuo del estado psicológico ha promovido la integración de sensores fisiológicos, modelos de aprendizaje automático y enfoques estadísticos avanzados para inferir patrones de estrés y ansiedad en tiempo real [4]

Trabajos pioneros como el de Luis G. Jaimes en 2017 [5] demostraron la viabilidad del uso de Modelos Ocultos de Márkov (HMM) y series temporales con suavizamiento exponencial para predecir episodios de estrés mediante señales fisiológicas recolectadas mediante sensores portátiles, proponiendo un modelo de intervenciones preventivas basado en la detección anticipada del estado emocional.

De forma complementaria, investigaciones recientes en visión por computador y aprendizaje no supervisado han logrado estimar estados de ansiedad, depresión y estrés a partir de imágenes faciales visuales y térmicas, integrando modelos de CNN y HMM en el proceso de clasificación emocional [4].

Una segunda tendencia se centra en la evidencia reciente sobre la ansiedad en el trabajo, en donde se destacan, estudios epidemiológicos y revisiones sistemáticas han evaluado los efectos del estrés y la ansiedad en la productividad laboral, así como la eficacia de estrategias de intervención temprana. En una revisión para la OMS, concluyeron que los programas de tamizaje y monitoreo de salud mental en el trabajo, aunque útiles para la detección, no generan mejoras significativas si no se acompañan de intervenciones terapéuticas preventivas y adecuadas [3].

Por su parte, van Hofffen en 2020 desarrollo modelos predictivos para el ausentismo laboral por trastornos mentales, demostrando que variables psicosociales como la satisfacción laboral, el apoyo social y el nivel de distrés son factores críticos para identificar riesgos de incapacidad prolongada [2].

Otras investigaciones muestran vacíos de investigación y enfoque del presente estudio, los cuales, a pesar de los notables avances en la modelación predictiva de la ansiedad, persiste una significativa brecha metodológica en la comprensión estadística de su comportamiento como fenómeno dinámico [6]. La literatura existente se ha centrado principalmente en tareas de detección y clasificación de la ansiedad, abordando su naturaleza de forma estática. Esta orientación ha dejado de lado la exploración de la distribución probabilística del fenómeno y, crucialmente, la dinámica de transición entre sus posibles estados latentes (o de intensidad), lo cual es fundamental para una caracterización completa de su evolución temporal.

De acuerdo con lo anterior, este trabajo propone abordar esta brecha mediante el uso de Modelos Ocultos de Márkov (HMM) aplicados a variables observables asocia-

das al comportamiento y estilo de vida (sueño, estrés percibido, actividad física, consumo de tabaco, sesiones de terapia, entre otras). El objetivo central de esta modelación es determinar la distribución de probabilidad que mejor describa la evolución temporal de los niveles de ansiedad y, simultáneamente, caracterizar la dinámica de transición entre los estados latentes u ocultos del fenómeno, aportando de forma decidida a los avances tecnológicos en la modelación del estrés y la ansiedad, con el fin de pasar de la detección estática a la comprensión del proceso estocástico asociado a la ansiedad.

Los resultados arrojados por el modelo muestran enfoque probabilístico que trasciende de la modelación descriptiva tradicional, avanzando hacia una comprensión más precisa y explicativa del fenómeno dinámico de la ansiedad. Esta metodología se articula mediante la convergencia de herramientas de análisis de series temporales, métodos estocásticos y aprendizaje estadístico.

Si bien los resultados obtenidos evidencian un avance significativo en la caracterización de la distribución de probabilidad y la dinámica de transición entre estados latentes de ansiedad, el problema no puede considerarse completamente resuelto, dado que persisten desafíos relacionados con la disponibilidad de los datos observacionales, la heterogeneidad individual y la sensibilidad del modelo a las condiciones contextuales. En consecuencia, el trabajo futuro deberá centrarse en la ampliación de la base empírica, la validación cruzada de los modelos en distintas poblaciones laborales y la incorporación de perspectivas multimodales de observación que reconozcan la interacción entre los factores biológicos, psicológicos y sociales que configuran la experiencia de la ansiedad. Esta información permitirá refinar la capacidad predictiva y explicativa de los modelos propuestos. De esta forma, se espera fortalecer los sistemas de vigilancia en salud ocupacional y promover el diseño de intervenciones preventivas más oportunas, personalizadas y rigurosamente basadas en evidencia [2, 3].

El artículo se estructura en cinco secciones. En la primera se presenta la introducción; en la segunda, el estado del arte y la revisión de la literatura; en la tercera, la comprensión del fenómeno, la preparación de los datos y el desarrollo del modelo; la cuarta recoge el análisis y la discusión de los resultados; y, finalmente, la quinta expone las conclusiones, las principales limitaciones del estudio y las líneas de trabajo futuro.

## 2 Estado del arte y revisión de literatura

La ansiedad se ha consolidado como uno de los problemas de salud mental más frecuentes y relevantes en la población general. Estudios recientes en atención primaria destacan su alta prevalencia y el agravamiento de su impacto tras la pandemia de COVID-19[7]. En la actualidad, los trastornos de ansiedad representan un verdadero problema de salud pública debido a su frecuencia y carga asociada para los sistemas sanitario[7]. Según datos epidemiológicos, aproximadamente un tercio de la población experimentará algún trastorno de ansiedad a lo largo de su vida[7], siendo las

fobias específicas y la ansiedad social algunas de las manifestaciones más comunes[7]. Además, existe un gradiente social en la ansiedad: su prevalencia aumenta en grupos socioeconómicamente más vulnerables[7] y afecta casi al doble de mujeres que de hombres[7].

En el contexto laboral, la Organización Mundial de la Salud (OMS) estima que alrededor del 15% de los adultos en edad de trabajar padecían un trastorno mental en 2019, y que cada año se pierden aproximadamente 12.000 millones de días laborales por depresión y ansiedad, con un coste global de 1 billón de dólares en productividad[8] Un entorno de trabajo inadecuado, sobrecarga de trabajo, poco control del empleado o inseguridad laboral son reconocidos como factores de riesgo para la salud mental[8]. La OMS enfatiza que el trabajo digno y un ambiente laboral seguro pueden proteger la salud mental, mientras que condiciones laborales adversas elevan sustancialmente el riesgo de trastornos de ansiedad y depresión[8]Por este motivo, la salud mental en el trabajo ha cobrado protagonismo en la agenda internacional, con la publicación de guías específicas en 2022 que recomiendan intervenciones para prevenir riesgos psicosociales, promover el bienestar en el empleo y apoyar a trabajadores con condiciones mentales[8]

El siglo XXI ha visto emerger cambios sociales, tecnológicos y ambientales que crean un entorno propenso a la ansiedad, un “ambiente ansiógeno” en palabras de Expósito-Duque et al. (2024) [9]. La ansiedad, si bien es una respuesta natural y adaptativa ante amenazas, puede volverse patológica ante estímulos erróneos o patrones de comportamentales [7, 9]. Distintos determinantes sociales se han asociado al aumento de los trastornos de ansiedad en las sociedades contemporáneas. Por un lado, la era digital y la hiperconexión juegan un papel importante: la dependencia de la tecnología y las redes sociales, el ciberacoso y la sobreexposición a información negativa contribuyen significativamente a elevar los niveles de ansiedad en la población [9]. La constante comparación en línea y la preocupación por la imagen corporal, junto con la erosión de la privacidad personal, son factores propios de la vida moderna que también alimentan la ansiedad[9].

Por otro lado, factores macro como las exigencias laborales actuales (competitividad, precariedad, largas jornadas) se han vinculado con mayores niveles de estrés y ansiedad [9]. Además, el cambio climático y las crisis medioambientales han dado lugar al concepto de ecoansiedad: la preocupación crónica por el futuro del planeta, que afecta especialmente a la población más joven [9]. Eventos climáticos extremos y fenómenos asociados, como la migración forzada por desastres, incrementan el riesgo de trastornos mentales en las comunidades afectadas[9]. En resumen, la evidencia apunta a una interacción compleja de factores sociales, tecnológicos, ocupacionales y ambientales detrás del aumento de la ansiedad en la sociedad actual[9]. Este contexto multifactorial exige enfoques integrales para la prevención y manejo de la ansiedad, abordando no solo al individuo sino también su entorno [8].

Dada el alto impacto de la ansiedad, ha aumentado el interés por identificar de manera temprana a las personas en riesgo en contextos como el laboral[3]. Una de las estrategias más exploradas ha sido el tamizaje de salud mental en el lugar de trabajo; sin embargo, la evidencia disponible muestra resultados mixtos respecto a su efectividad. En un análisis reciente de ocho ensayos controlados, con 2.940 empleados apro-

ximadamente evaluó programas de cribado de salud mental laboral y no encontró mejoras significativas en la salud mental de los trabajadores cuando el componente principal consistía únicamente en la aplicación del screening y la oferta de consejería básica, con tamaños de efecto muy pequeños y estadísticamente no significativos[3]. En contraste, aquellos programas en los que el tamizaje se vinculó a un acceso facilitado a intervenciones terapéuticas (por ejemplo, derivación activa a tratamiento psicológico o psiquiátrico) sí evidenciaron una mejoría pequeña pero consistente en los síntomas [3]. Estos resultados sugieren que el cribado, por sí solo, es insuficiente y que su impacto depende de estar articulado con rutas claras de intervención y apoyo clínico [8], en línea con las recomendaciones de la Organización Mundial de la Salud, que aboga por combinar la identificación temprana con mecanismos efectivos de acompañamiento para los trabajadores en riesgo[3].

Otra acercamiento son los modelos de predicción de riesgo basados en factores personales y laborales. Por ejemplo, Van Hoffen et al. (2020) desarrollaron un modelo multivariable en población trabajadora general para predecir el riesgo de ausencia laboral prolongada por trastornos mentales[2]. Este modelo integraba 11 predictores (síntomas de distrés emocional, sexo, estado civil, sector económico, antigüedad, factores psicosociales laborales como claridad de rol, demandas cognitivas, apoyo social, satisfacción laboral, etc.)[2]. En su validación inicial, el modelo logró discriminar moderadamente a los empleados que presentaron ausentismo prolongado por motivos de salud mental (Área Bajo la Curva ROC ≈ 0,71)[2]. Asimismo, los autores desarrollaron un árbol de decisión más simple y fácilmente interpretable para su aplicación en la práctica clínica y ocupacional, aunque con una capacidad de discriminación ligeramente inferior (AUC ≈ 0,70)[2]. En una validación externa posterior, tanto el modelo de predicción original como el árbol de decisión mostraron un desempeño moderado (AUC ≈ 0,70) para identificar trabajadores en riesgo de ausentismo prolongada por trastornos mentales[2]. En conjunto, estos resultados ponen de relieve, por un lado, la dificultad de predecir con alta precisión eventos complejos como las ausencias laborales por ansiedad o depresión[2], y por otro, el valor potencial de las encuestas de salud laboral: la combinación de información sociodemográfica, características del puesto de trabajo y síntomas auto informados permite al menos priorizar a aquellos empleados que podrían beneficiarse de intervenciones preventivas [2, 10].

Los marcos internacionales recomiendan que estas evaluaciones de riesgo en el trabajo se empleen con cautela, como herramienta para informar intervenciones (por ejemplo, derivar al servicio médico ocupacional, implementar ajustes en el puesto) y nunca para discriminar o estigmatizar al trabajador[8, 11]. En línea con ello, los expertos abogan por una aproximación doble: promover entornos laborales saludables (reduciendo las fuentes de estrés organizacionales conocidas) y, a la vez, identificar individuos vulnerables para ofrecerles apoyo personalizado antes de que desarrollen un trastorno establecido[8, 11].

Los avances tecnológicos recientes han abierto nuevas posibilidades para monitorear y predecir el estrés y la ansiedad de forma continua y pasiva. Diversas investigaciones exploran sensores portátiles, visión por computador y otras herramientas digitales para inferir el estado emocional de una persona, incluso sin intervención activa de su parte [12].

En el ámbito de la tecnología vestible, destaca el trabajo de Saito et al. (2022)[12], quienes construyeron un modelo de machine learning para predecir el inicio de trastornos mentales integrando datos de dispositivos wearables con historiales médicos[12]. En concreto, utilizaron datos biométricos recogidos mediante pulseras Fitbit (p. ej., patrones de actividad física, sueño, ritmo cardíaco) junto con información de exámenes médicos periódicos, para anticipar qué individuos podrían desarrollar una condición como depresión o ansiedad en el futuro cercano[13]. Este enfoque demostró la factibilidad de identificar individuos de alto riesgo de manera temprana, aprovechando datos pasivos y continuos del mundo real (p. ej., cambios sutiles en patrones de sueño/actividad que preceden a un episodio mental)[13]. La integración de wearables con datos clínicos representa un paso hacia sistemas de alerta temprana, aunque persisten retos en cuanto a la privacidad, la necesidad de validar estos modelos en poblaciones amplias y asegurar la calidad de los datos recopilados[12, 14].

Otro campo interesante es el de la visión por computador aplicada a la salud mental. Nayak et al. (2021) propusieron un sistema novedoso que analiza secuencias de imágenes faciales en espectro visible y termografía infrarroja para estimar niveles de depresión, ansiedad y estrés en una persona[4]. Mediante técnicas de clustering de los patrones faciales –incluyendo cambios micro expresivos y variaciones de temperatura facial asociadas a respuestas autonómicas– lograron clasificar estados afectivos de forma no invasiva[4]. Si bien el estudio es experimental, demuestra el potencial de las cámaras térmicas y la analítica de imágenes para detectar signos fisiológicos sutiles del estrés que escapan al ojo humano[4].

En cuanto al seguimiento fisiológico, ya es conocida la relación entre ciertas señales corporales (variabilidad cardíaca, actividad electrodermal, niveles de cortisol, etc.) y los estados de estrés [5]. Jaimes et al. (2017) fueron pioneros en aprovechar esta relación para pronosticar la evolución del estrés: introdujeron el concepto de “Future Stress”, donde a partir de series temporales de señales fisiológicas de trabajadores se entrenaron modelos para predecir futuros episodios de estrés [5]. En particular, emplearon una combinación de Modelos Ocultos de Markov (HMM) y técnicas de suavizamiento exponencial para analizar el historial temporal de estrés de cada individuo (obtenido mediante sensores portátiles y evaluaciones en tiempo real)[5]. Los resultados fueron prometedores, logrando anticipar con cierta fiabilidad qué trabajadores experimentarían picos de estrés más adelante[5]. Este enfoque de modelado secuencial es valioso porque reconoce que el estrés (y análogamente la ansiedad) fluctúa en el tiempo y presenta patrones temporales que pueden explotarse predictivamente[5]. De hecho, disponer de un “pronóstico” de estrés permitiría activar intervenciones preventivas just-in-time –por ejemplo, recomendar una pausa o una técnica de relajación al detectar que un empleado probablemente alcanzará un nivel de ansiedad perjudicial en las próximas horas[5]. Si bien el estudio de Jaimes se centró en estrés laboral agudo y utilizó muestras limitadas, sentó las bases para sistemas de alerta temprana basados en wearables y modelos probabilísticos temporales[5].

A medida que la investigación progresa, se reconoce cada vez más que comprender fenómenos como la ansiedad requiere modelar su dinámica subyacente, y no limitarse a observaciones estáticas. En este sentido, están cobrando fuerza los modelos matemáticos y computacionales que tratan a los trastornos mentales como procesos evolu-

tivos[6]. Un ejemplo es el estudio de Göçgün (2024), que planteo formular un problema de distracción mental como un proceso de decisión Markoviano[15], donde se pudo comparar cuantitativamente el efecto de diferentes políticas de manejo sobre el rendimiento de los individuos[15]. Este trabajo, aunque teórico, abre la puerta a utilizar modelos de decisión estocásticos para optimizar intervenciones en salud mental[15]. Sin embargo, según la revisión de De Oliveira et al. (2024), son escasos los ejemplos de micro simulación aplicados a salud mental y muchos adolecen de baja calidad metodológica[16]. Los autores abogan por el desarrollo de modelos más robustos que permitan simular, por ejemplo, cómo evolucionaría la prevalencia de la ansiedad en una cohorte bajo distintas intervenciones de prevención o tratamiento. Este enfoque sería sumamente valioso para fundamentar las políticas públicas[16].

Dentro del campo emergente de la psiquiatría computacional, Zavlis et al. (2025) han compilado por primera vez un estado del arte de cómo se están modelando las dinámicas interpersonales en diversos trastornos [6]. Hallaron 58 estudios donde se aplican desde modelos bayesianos y de aprendizaje por refuerzo, hasta sistemas dinámicos diferenciales para capturar aspectos como la rigidez o flexibilidad de las relaciones sociales en pacientes[6]. Esto es crucial, ya que confirma cuantitativamente nociones clínicas de que las dificultades en las relaciones interpersonales no son solo consecuencias, sino parte integral de la psicopatología[6]. No obstante, Zavlis et al. subrayan que la mayoría de estos modelos aún carecen del rigor esperado en otras disciplinas: pocos reportan métricas de validación sólidas y hay escasa adopción de prácticas abiertas (compartir datos, código)[6]. Por tanto, abogan por elevar los estándares de transparencia y validación en la modelización computacional de la salud mental, para que sus hallazgos ganen confianza y puedan integrarse a la práctica clínica o al diseño de intervenciones[6].

En otras investigaciones generales sobre IA en psiquiatría (p. ej., Lewin et al. 2025) enfatizan tanto las oportunidades como las precauciones necesarias en este campo. Las técnicas de machine learning ya se han aplicado a amplias variantes de problemas: desde predecir respuesta a tratamientos antidepresivos, estimar riesgo de suicidio, hasta detectar recaídas a partir de datos de smartphones[13, 17]. Estas aplicaciones podrían llevar a una psiquiatría más personalizada, donde algoritmos ayuden a tomar decisiones clínicas adaptadas al perfil individual[17]. Sin embargo, los expertos resaltan desafíos clave: es necesario contar con bases de datos amplias y representativas (para que los modelos no queden sesgados a ciertos grupos), y garantizar la explicabilidad de los modelos para que los profesionales de la salud confíen en ellos[13]. Además, desde una perspectiva ética, cualquier sistema de IA debe implementarse asegurando la privacidad de los pacientes y evitando reproducir o agravar desigualdades [17]. En resumen, la IA ofrece herramientas poderosas para la investigación y práctica en salud mental, pero su adopción debe ser cuidadosa y complementaria, nunca reemplazando la evaluación humana sino potenciándola con información adicional[13, 16].

A pesar de los avances descritos, gran parte de la literatura actual aborda la ansiedad de forma estática o transversal[13]: se enfoca en determinar si un individuo está ansioso en un momento dado, o en estimar su riesgo global de padecer ansiedad clínica, basándose en un conjunto de características[6]. Si bien esto es útil, la ansiedad es

un proceso dinámico que fluctúa con el tiempo y las circunstancias[7]. Para alcanzar una comprensión más profunda es necesario modelar explícitamente la evolución temporal de la ansiedad y las transiciones entre sus estados internos [5, 15].

En términos conceptuales, esta propuesta busca dejar de observar la ansiedad como una fotografía clínica aislada y comenzar a tratarla como una película probabilística. Este trabajo se oriental a identificar la distribución de probabilidad que describa la evolución temporal de los niveles de ansiedad y, al mismo tiempo, caracterizar la dinámica de transición entre estados. Para ello se recurre a Modelos Ocultos de Markov (HMM) y sus variantes flexibles, capaces de inferir estados internos a partir de observaciones y de aprender las probabilidades de cambio entre ellos a lo largo del tiempo. En conjunto, este enfoque busca llenar el vacío que deja una literatura predominantemente estática, contribuyendo a pasar de la detección puntual a una comprensión dinámica y estocástica de la ansiedad, y ofreciendo una base formal para sistemas de monitoreo continuo que acompañen el estado mental del individuo en su curso temporal, y no solo en instantáneas aisladas.

## 3 Metodología

El presente estudio se desarrolla bajo el marco de la metodología CRISP-DM[18] (Cross-Industry Standard Process for Data Mining), un estándar ampliamente utilizado en analítica de datos y ciencia aplicada. Este modelo metodológico establece un proceso estructurado e iterativo que permite transformar datos en conocimiento, garantizando la trazabilidad entre el problema de investigación, los métodos empleados y los resultados obtenidos[19].

Sobre esta base, la metodología del estudio adopta un enfoque cuantitativo-estocástico, en el cual la ansiedad se trata como un proceso no determinista, caracterizado por fluctuaciones inherentes al comportamiento humano. Esta aproximación se distingue de los enfoques clínicos tradicionales al concebir la ansiedad como un sistema dinámico de estados interconectados[13], cuya evolución puede describirse mediante procesos de Markov y analizarse a través de distribuciones de probabilidad no normales. Es importante mencionar que el objetivo central del enfoque es capturar la incertidumbre y la variabilidad inherentes a los estados emocionales, proporcionando una representación empírica y probabilística de su dinámica temporal[15].

El proceso metodológico seguido en este estudio se estructuró de manera secuencial e iterativa, tal como se resume en la Figura 1. En primer lugar, se realizó la recolección y consolidación de la base de datos, seguida por una fase de preparación que incluyó limpieza, estandarización y transformación de las variables relevantes. Posteriormente, se llevó a cabo un análisis exploratorio para comprender la distribución inicial de los indicadores psicológicos y conductuales, así como para identificar patrones preliminares en las relaciones entre variables. Con esta base, se aplicaron técnicas de aprendizaje no supervisado para conformar agrupaciones conductuales que sirven como perfiles observables dentro del modelo. Estos perfiles, alimentaron la etapa central del flujo: la estimación de un Modelo Oculto de Márkov, utilizado para

inferir estados latentes y su dinámica de transición. Finalmente, se ajustaron distribuciones continuas sobre el indicador transformado para caracterizar su comportamiento marginal, integrando la información del modelo temporal con el análisis estadístico del riesgo psicológico extremo. Este flujo metodológico permitió articular la estructura observada en los datos con una representación estocástica coherente del fenómeno de la ansiedad.

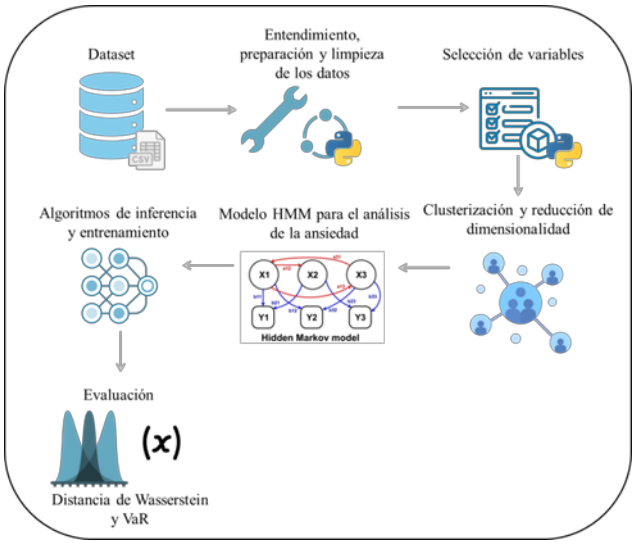


**Fig. 1.** Esquema general del procedimiento metodológico.
El diagrama resume las etapas del estudio, desde la obtención y preparación de los datos hasta el modelamiento temporal mediante HMM y la evaluación del riesgo.

### 3.1 Compresión del fenómeno

Desde una perspectiva dinámica y probabilística, la ansiedad se concibe no como un estado clínico fijo, sino como un proceso estocástico que fluctúa continuamente entre estados de bienestar y malestar. Esta variabilidad es modulada por la interacción de factores psicológicos y conductuales, tales como el estrés percibido, la calidad del sueño, la actividad física o la exposición social, que definen la intensidad del estado de ansiedad [18].

En este sentido, la ansiedad puede concebirse como un sistema estocástico compuesto por estados latentes que evolucionan según probabilidades de transición dependientes del estado previo. Dicho enfoque se alinea con las aproximaciones contemporáneas que reconocen en los fenómenos mentales una estructura no lineal, incierta y sensible al contexto [20, 21]. Por ello, el análisis se apoya en modelos que integran la incertidumbre y permiten estimar la dinámica subyacente a partir de datos observables.

El objetivo de esta fase es la formulación del problema estadístico que orientará las etapas posteriores del estudio. Dicha formulación se centra en dos aspectos clave: determinar si el comportamiento de la ansiedad puede representarse mediante distribuciones de probabilidad no normales y establecer si su evolución temporal responde a un modelo de transición de estados, formalizado a través de Modelos Ocultos de Markov (HMM). Esta aproximación metodológica posibilita una comprensión del fenómeno que trasciende las mediciones descriptivas tradicionales, proporcionando una visión estructural y dinámica del fenómeno de la ansiedad [18].

Finalmente, se establece el alcance no clínico de la investigación: el interés no radica en diagnosticar o predecir trastornos de ansiedad individuales, sino en caracterizar estadísticamente la forma y evolución del fenómeno desde un punto de vista probabilístico.

### 3.2 Comprensión de los datos

En esta fase del proyecto estará centrada en un análisis exploratorio de los datos con el propósito de entender la estructura básica del conjunto de información, verificar la calidad de los registros y caracterizar preliminarmente las variables psicológicas y conductuales relacionadas con la ansiedad.

Se utilizará el Social Anxiety Dataset de Kaggle[22], que contiene información auto-reportada a nivel de individuo sobre variables psicológicas, socioeconómicas y de hábitos, como el nivel de ansiedad, nivel de estrés, calidad del sueño, actividad física, calidad de la dieta y otros factores relacionados con el bienestar emocional. A partir de estas variables se aplicarán se combinan estadísticas descriptivas, visualizaciones gráficas y análisis de dependencias entre variables como insumo para las etapas posteriores de selección de variables, modelamiento y evaluación.

Con el fin de realizar una primera aproximación a la estructura del dataset y apoyar la fase de comprensión de los datos:

- **Gráficos de barras e histogramas:** Se construyeron gráficos de barras e histogramas para las variables ordinales y continuas. Estos gráficos permiten visualizar la distribución de frecuencias en cada escala, identificar categorías poco representadas y reconocer de manera preliminar la forma general de las distribuciones, información que se utiliza posteriormente para la preparación y transformación de las variables[23].
- **Matriz de correlación y correlaciones parciales**: Para las variables numéricas seleccionadas se calculó una matriz de correlación y se representó mediante un mapa de calor. Este tipo de gráfico facilita la inspección visual de la dirección y magnitud de las asociaciones lineales entre pares de variables, y sirve como insumo metodológico para[23]:
    - (i) detectar posibles relaciones fuertes entre variables,
    - (ii) identificar colinealidad que pueda requerir la eliminación o combinación de variables redundantes antes del modelado y
    - (iii) orientar la selección de variables observables que se utilizarán en las etapas de reducción de dimensionalidad y clusterización.

De forma complementaria, se estimaron correlaciones parciales entre las mismas variables, controlando por el efecto del resto de dimensiones del modelo. Este análisis permite distinguir asociaciones directas de aquellas que se explican por la influencia conjunta de otras variables, reduciendo el riesgo de seleccionar indicadores redundantes y aportando un criterio adicional para depurar el conjunto de variables observables antes de aplicar otras técnicas[23].

- **Diagramas de dispersión y boxplots en paneles combinados:** Adicionalmente, se generaron paneles que integran diagramas de dispersión bivariados, histogramas marginales y diagramas de caja para cada variable numérica. Los gráficos de dispersión permiten examinar la distribución conjunta de pares de variables, mientras que los boxplots resumen la tendencia central, la dispersión y la presencia de valores atípicos[23]. Estos paneles se emplean exclusivamente como herramienta descriptiva para verificar la consistencia de los datos y apoyar decisiones posteriores de depuración y estandarización.

En esta sección los gráficos se utilizan únicamente para documentar el procedimiento de exploración y apoyar la toma de decisiones metodológicas; la interpretación detallada de los patrones observados se desarrolla más adelante en el capítulo de resultados.

### 3.3 Preparación de los datos

El conjunto de datos Social Anxiety[22] tiene originalmente una naturaleza transversal: cada registro corresponde a una persona distinta y recoge, en un único momento, su nivel de ansiedad y un conjunto de variables psicológicas y de hábitos (sueño, actividad física, consumo de cafeína, etc.). Este dataset cuenta con 11.000 registros y 19 columnas[22]. En este tipo de bases, las observaciones se asumen independientes entre sí y no existe un eje temporal que vincule repetidamente a un mismo individuo.

Sin embargo, los Modelos Ocultos de Markov (HMM) están diseñados para trabajar con secuencias de observaciones ordenadas en el tiempo, de manera que para cada sujeto $i$ se dispone una trayectoria $(Y_{i1}, Y_{i2}, \dots Y_{iT})$ y el modelo puede estimar las probabilidades de transición entre estados latentes a través de esos pasos temporales[24, 25]. Por esta razón, en la literatura aplicada los HMM se emplean de forma natural sobre datos longitudinales o de panel, donde se registran mediciones repetidas sobre las mismas unidades a lo largo del tiempo [24–26].

En términos de análisis de datos, los datos de panel o longitudinales pueden entenderse como una colección de series de tiempo, una por cada individuo u unidad de análisis, donde cada serie está formada por observaciones repetidas en distintos instantes [26, 27]. Este tipo de estructuras permite estudiar dinámicas de cambio y transiciones entre estados, pero su construcción suele requerir diseños de seguimiento (cohortes, estudios clínicos o encuestas repetidas) que son costosos y no siempre están disponibles. Esto ha motivado enfoques que parten de datos transversales y los reorganizan o combinan con supuestos estructurales para aproximar procesos dinámicos,

lo que en la literatura se conoce como estrategias de modelado a partir de estudios transversales para inferir dinámica temporal o paneles sintéticos[26, 28].

Dado que el conjunto de datos utilizado en este estudio solo proporciona una fotografía transversal, se optó por construir una estructura longitudinal sintética que permitiera ajustar un HMM manteniendo, en la medida de lo posible, la coherencia interna del diseño original[28]. Para ello, se implementó un procedimiento en tres pasos:

1. **Asignación de identificadores de sujeto (ID):** En primer lugar, se definió un número máximo de observaciones por sujeto ($N = 2$) y, a partir del tamaño total de la base de datos, se generó el conjunto de identificadores necesarios. Estos IDs se asignaron a los registros de forma aleatoria y reproducible, utilizando una semilla fija en el generador pseudoaleatorio. De este modo, cada individuo sintético quedó asociado a dos observaciones, lo que permite representar un cambio potencial de estado entre dos cortes temporales consecutivos.
2. **Generación de una variable de fecha simulada:** Para introducir un orden temporal explícito dentro de cada sujeto, se creó una columna de fecha simulada. Se tomó como referencia una fecha base aleatoria y se generaron fechas posteriores espaciadas aproximadamente un mes para cada ID. Sobre estas fechas se añadió una variación aleatoria acotada (entre 0 y 15 días), con el fin de emular la irregularidad típica de los intervalos de medición en contextos clínicos o laborales, y evitar que todas las observaciones se concentraran en instantes idénticos.
3. **Creación de un índice de tiempo discreto:** Finalmente, se construyó una variable entera time que indica la posición temporal de cada observación dentro de cada sujeto (1, 2, …). Este índice se obtuvo ordenando las observaciones por ID y fecha simulada, y aplicando un conteo acumulado por sujeto. La variable time actúa como eje temporal interno para el HMM y garantiza que las secuencias de estados observados se procesen en el orden cronológico correcto durante la estimación de las matrices de transición y emisión.

El resultado de este procedimiento es una base de datos en formato panel, con múltiples sujetos y dos observaciones ordenadas en el tiempo para cada uno, que hace posible el uso de un HMM sobre un conjunto de datos originalmente transversal.

**Detección y tratamiento de valores atípicos**

En esta etapa se llevó a cabo un proceso de depuración de las variables con el propósito de reducir la influencia de valores extremos y mejorar la coherencia interna del conjunto de datos. El foco se centró en las principales variables observables de interés.

A partir de las distribuciones univariadas (histogramas y diagramas de caja) y de los diagramas de dispersión bivariados construidos en la fase de comprensión de los datos, se identificaron valores que se alejaban de manera marcada del rango habitual de la muestra o de los rangos considerados plausibles desde el punto de vista clínico y conductual. Adicionalmente, se revisaron categorías con muy baja frecuencia en las variables discretas, pues su presencia podía generar inestabilidad en las etapas de ajuste de distribuciones y en la estimación de las transiciones estocásticas.

Sobre la base de esta revisión exploratoria se aplicaron reglas de depuración y recodificación, que incluyeron:

(i) eliminación puntual de registros con valores claramente erróneos o incompatibles con la escala original;
(ii) agrupación de categorías poco frecuentes en intervalos más amplios y homogéneos; y
(iii) acotación de valores extremos dentro de rangos operativos definidos para cada variable.

Estas decisiones se documentaron de forma explícita para asegurar la trazabilidad del proceso y garantizar que el conjunto de datos resultante fuera adecuado para las etapas posteriores de modelamiento.

**Normalización y estandarización de variables**

Una vez depurados los valores atípicos y reorganizadas las categorías con baja frecuencia, se procedió a la normalización de las variables con el fin de homogeneizar las escalas de medición y facilitar su uso conjunto en los distintos módulos del modelo.

En primer lugar, se distinguieron las variables numéricas de las categóricas:

- Para las variables numéricas seleccionadas se aplicó una transformación de tipo Min–Max para reescalar sus valores al intervalo $[0,1]$. Esta normalización permite que ninguna variable domine a las demás por diferencia de magnitud y resulta especialmente conveniente para algoritmos basados en distancias, como PCA y K-Means.
- Las variables categóricas binarias se codificaron en formato numérico consistente (0/1), preservando su interpretación original y evitando transformaciones adicionales innecesarias.

En conjunto, el tratamiento de outliers y la posterior normalización conforman una etapa de preparación orientada a obtener un conjunto de datos más estable y comparable entre dimensiones, compatible con los requerimientos del modelamiento estadístico posterior (ajuste de distribuciones continuas, estimación de matrices de transición y simulaciones probabilísticas).

**Selección de variables observables**

La selección de las variables observables se realizó a partir del marco teórico y de la fase de comprensión de los datos, priorizando indicadores psicológicos y de hábitos que capturan dimensiones clave del bienestar y malestar percibido[9, 29]. Se incluyeron únicamente aquellas variables que presentaron coherencia conceptual con el fenómeno de la ansiedad y variabilidad suficiente para el análisis, evitando redundancias innecesarias. Este subconjunto depurado de variables constituye la base para las etapas posteriores de reducción de dimensionalidad y clusterización, a partir de las cuales se derivan los perfiles observables que alimentan el modelo de HMM.

### 3.4 Modelo de clusterización y HMM para el análisis de la ansiedad

El proceso de modelamiento se orienta a capturar la estructura interna del fenómeno de la ansiedad desde una perspectiva estadística y estocástica, integrando técnicas de aprendizaje no supervisado con modelos de dependencia temporal. Dado que la ansiedad se manifiesta como un proceso dinámico y no directamente observable[13], se plantea una estrategia en dos etapas complementarias.

En la primera, se utilizan métodos de aprendizaje no supervisado para identificar patrones latentes y grupos relativamente homogéneos dentro de las variables observacionales, de modo que dichos grupos funcionen como aproximaciones empíricas a posibles estados subyacentes del sistema emocional. La literatura ha mostrado que técnicas como la clusterización tipo K-means, K-Mediods y K-NN resultan especialmente útiles para descubrir estructuras internas en datos multivariados y construir perfiles de comportamiento a partir de medidas psicológicas y de estilo de vida[4].

En la segunda etapa, estos estados se incorporan en un marco temporal mediante Modelos Ocultos de Markov (HMM), que permiten representar procesos dinámicos en los que la variable de interés es parcialmente observable y evoluciona en el tiempo de acuerdo con una estructura de estados latentes y probabilidades de transición[25]. Los HMM han demostrado ser una herramienta flexible para el análisis de series temporales y datos longitudinales en contextos donde el estado subyacente no es directamente observable, incluyendo aplicaciones en salud, ciencias del comportamiento y otros dominios[5].

**Métodos de agrupamiento y reducción de dimensionalidad**

En este estudio se emplean técnicas de agrupamiento para identificar perfiles relativamente homogéneos de individuos a partir de sus características psicológicas y de hábitos. El objetivo es organizar las observaciones en grupos internos lo más similares posible y, al mismo tiempo, distintos de otros grupos, siguiendo el enfoque clásico de aprendizaje no supervisado[30, 31]

**K-means**

El algoritmo K-Means particiona el conjunto de datos en $k$ clústeres mediante un procedimiento iterativo que alterna entre: (i) asignar cada observación al centroide más cercano y (ii) recalcular los centroides como el promedio de las observaciones de cada grupo. El criterio de optimización consiste en minimizar la suma de distancias cuadráticas intra–clúster (inercia), lo que lo hace eficiente y ampliamente utilizado cuando se trabaja con variables numéricas escaladas[31]

**K-medoids**

Dado que K-means es sensible a valores extremos, se consideran variantes robustas en las que el centro del clúster se define mediante medidas menos afectadas por outliers, como la mediana o los medoids. Estas formulaciones mantienen la lógica de partición en $k$grupos, pero reemplazan el promedio por un representante robusto, mejorando el comportamiento del algoritmo en presencia de distribuciones asimétricas o con valores atípicos[30].

**K-Nearest Neighbors (KNN)**

Aunque KNN es un método supervisado de clasificación, su regla basada en la vecindad se utiliza de forma complementaria para la asignación de nuevas observaciones a clústeres ya definidos: un nuevo individuo se asigna al grupo predominante entre sus $k$ vecinos más cercanos en el espacio de características. Esta regla de decisión se fundamenta en el trabajo clásico de Cover y Hart, donde se analiza la consistencia y el error de clasificación del esquema de vecinos más cercanos[32].

Para determinar el número adecuado de clústeres $k$ y comparar el desempeño de los distintos métodos de agrupamiento, se recurre tanto al método del codo como a indicadores de validación interna:

- **Método del codo:** Se calcula la inercia total (suma de distancias cuadráticas intra–clúster) para distintos valores de $k$. A medida que $k$ aumenta, la inercia disminuye, pero lo hace con rendimientos decrecientes; el "codo" de la curva —el punto a partir del cual la reducción adicional de inercia es marginal— se utiliza como referencia para seleccionar un número de clústeres que equilibre ajuste y parsimonia.[30]
- **Índice de Silhouette:** Evalúa, para cada observación, qué tan similar es a su propio clúster en comparación con otros clústeres. Su valor promedio oscila entre $-1$ y 1, donde valores cercanos a 1 indican buena separación entre clústeres, valores alrededor de 0 sugieren solapamiento y valores negativos indican asignaciones potencialmente erróneas. En este estudio se espera maximizar el índice de Silhouette para preferir particiones compactas y bien separadas [30]
- **Índice de Calinski–Harabasz:** Mide la relación entre la dispersión entre clústeres y la dispersión dentro de los clústeres. Valores más altos del índice indican particiones con grupos más compactos y mejor separados, por lo que también se busca maximizar este indicador al comparar configuraciones de $k$y diferentes algoritmos de agrupamiento[33].
- **Índice de Davies–Bouldin**: Cuantifica, para cada clúster, la razón entre su dispersión interna y la separación con respecto a los clústeres más cercanos, y promedia estas razones sobre todos los grupos. En este caso, valores más bajos indican una mejor partición, de modo que el objetivo es minimizar el índice de Davies–Bouldin para seleccionar configuraciones con clústeres compactos y bien diferenciados[30].

El conjunto de estas métricas permite comparar de manera sistemática las soluciones generadas por los métodos de agrupación, así como orientar la elección del número de clústeres más adecuado para representar los perfiles conductuales y de riesgo que posteriormente se integrarán al modelo temporal.

**Análisis de Componentes Principales (PCA)**

El Análisis de Componentes Principales (PCA) se utiliza como técnica de reducción de dimensionalidad previa y complementaria al clustering[33]. El PCA transforma el conjunto de variables originales en un nuevo sistema de componentes lineales no correlacionados entre sí, ordenados según la varianza explicada, lo que permite repre-

sentar los datos en un espacio de menor dimensión manteniendo la mayor parte de la información relevante [33].

En el contexto de este estudio, el PCA cumple un rol específico: el dataset integra múltiples indicadores psicológicos y de hábitos (por ejemplo, niveles de estrés, sueño, actividad física, consumo de estimulantes, participación en terapia) que pueden estar fuertemente correlacionados entre sí[4]. Trabajar directamente con todas estas variables puede introducir redundancia, sesgar los algoritmos de agrupamiento hacia dimensiones con mayor escala o variabilidad y dificultar la interpretación de los perfiles resultantes.

De este modo, el uso de PCA no se limita a un propósito puramente técnico de reducción de dimensionalidad, sino que se integra de manera funcional al proyecto: contribuye a definir un espacio latente más limpio y manejable desde el cual se derivan los perfiles observables que, en etapas posteriores, alimentan el Modelo Oculto de Markov para describir la dinámica estocástica de la ansiedad[4, 30].

**Modelos de cadenas de Markov y procesos ocultos (HMM)**

Los procesos estocásticos constituyen una herramienta fundamental para describir sistemas cuyo comportamiento evoluciona en el tiempo bajo condiciones de incertidumbre. En estos modelos, el estado futuro depende únicamente del estado presente, propiedad conocida como supuesto de Markov, expresada formalmente como:

$$P(q_i|q_1 \dots q_{i-1}) = P(q_i|q_{i-1}), \quad (1)\ ,$$

lo cual simplifica la complejidad temporal al asumir una memoria de primer orden [34].

De manera general, una cadena de Markov permite modelar la secuencia de estados observables (por ejemplo, niveles de ansiedad discretizados), mientras que los Modelos Ocultos de Markov (HMM) amplían esta estructura al incluir estados latentes no observables, que representan la dinámica interna del sistema psicológico.

De forma general, un HMM se define por el conjunto de parámetros $\lambda$ [34]:

$$\lambda = (A, B, \pi), \quad (2)$$

Donde $A$ es la matriz de transición entre estados latentes, $B$ es la matriz de emisión que vincula estados y observaciones, y $\pi$ es el vector de probabilidades iniciales de los estados[34].

En primer lugar, la evolución temporal entre los estados se describe a través de la matriz de transiciones $A = [a_{ij}]$[35]. La Matriz transición $A$ captura la dinámica interna del sistema, representando las probabilidades de transición entre estados mentales. Cada elemento de la matriz de transición se define como[35]:

$$a_{ij} = P(q_{t+1} = j \mid q_t = \text{i}) \quad (3)$$

$a_{i,j}$: Indica la probabilidad de pasar del estado $i$ al estado $j$ en el siguiente instante temporal [34].
Por otro lado, la relación entre $n$ los estados mentales y clusters de las variables observables se describe a través de la matriz de emisión $B = [b_j(o_t)]$[34]. Esta matriz especifica, para cada estado oculto, cómo se distribuyen las probabilidades de observar cada tipo de emisión:

$$b_j(o_t) = P(O_t = o_t | q_t = j) \qquad (4)$$

$b_j(o_t)$: Indicada la probabilidad de observar un evento $o_t$ dado el estado oculto $j$[34]

Finalmente, $\pi$ es el vector de probabilidades iniciales de los estados, definido como[34]:

$$\pi = (\pi_1, \pi_2, \ldots, \pi_N) \qquad (5)$$

donde $N$ es el número de estados latentes del modelo. Cada componente $\pi_i$ representa la probabilidad de que la cadena de Markov comience en el estado $i$; es decir, $\pi_i = P(q_1 = i)$[34]. Algunos estados $j$ pueden tener $\pi_j = 0$, lo que implica que no son estados de inicio posibles, y el vector completo debe satisfacer la condición de normalización $\sum_{i=1}^{N} \pi_i = 1$ [34].
Bajo esta parametrización $\lambda = (A, B, \pi)$, la probabilidad de una secuencia de observaciones $O = (o_1, \ldots, o_T)$ puede evaluarse de forma consistente, lo que permite ajustar y comparar diferentes configuraciones del HMM para modelar la dinámica de los estados de ansiedad [5, 35].

**Algoritmos fundamentales para la inferencia y estimación en HMM**

Los Modelos Ocultos de Márkov (HMM) ofrecen un marco probabilístico robusto para representar la evolución temporal de procesos latentes, como los distintos niveles de ansiedad que pueden experimentarse a lo largo del tiempo[5]. La capacidad del modelo para describir la persistencia, progresión o recuperación de estos estados depende de un conjunto de algoritmos esenciales que operan sobre su estructura interna. Entre ellos, los algoritmos Forward–Backward [35]y Viterbi [35] permiten realizar inferencia sobre las probabilidades de los estados ocultos y determinar la trayectoria más probable, mientras que el algoritmo Baum–Welch[36] estima iterativamente los parámetros del modelo a partir de los datos observados.

El algoritmo Forward-Backward calcula la probabilidad total de la secuencia observada $P(O \mid \lambda)$, pero también permite obtener la probabilidad instantánea de estar en cada estado oculto a lo largo del tiempo. Esta probabilidad se representa mediante la variable $\alpha_t(j)$, definida como[35]

$$a_t(j) = P(o_1, o_2, \ldots, o_t, q_t = j \mid \lambda), \qquad (6)$$

Donde: $a_t(j)$ describe la probabilidad acumulada de haber observado la secuencia $(o_1, \ldots, o_t)$ y encontrarse en el estado $j$ en el instante $t$.

El conjunto de valores $\boldsymbol{\alpha}_t = [\alpha_t(1), \alpha_t(2), \ldots, \alpha_t(N)]$ constituye un vector de probabilidades que refleja la distribución de ocupación esperada de cada estado en el tiempo[35]. Esta distribución puede promediarse para obtener el posterior medio por estado ($\bar{\gamma}$), que indica el porcentaje del tiempo que el modelo espera que cada estado esté activo a lo largo de la secuencia[35]:

$$\bar{\gamma}_t = \frac{1}{T}\sum_{t=1}^{T} \bar{\gamma}_t(j)\,, \text{donde } \bar{\gamma}_t(j) = P(q_t = j \mid O, \lambda) \quad (7)$$

Donde $\bar{\gamma}_t$: se interpreta como la ocupación esperada de cada estado oculto.
Esta información proporciona una visión probabilística de la dinámica subyacente de la ansiedad, mostrando cómo la probabilidad de cada estado varía a lo largo del tiempo y permitiendo identificar períodos de mayor vulnerabilidad o estabilidad emocional [34, 35].

El algoritmo Viterbi busca identificar la secuencia de estados más probable $Q^* = (q_1^*, q_2^*, \ldots, q_T^*)$ que explica las observaciones, resolviendo[34, 35]:

$$Q^* = argmax_Q\, P(Q|O, \lambda) \qquad (8)$$

A diferencia del Forward, que suma probabilidades sobre todas las trayectorias, Viterbi aplica el operador máximo para cada paso temporal [34, 35]:

$$\delta_t(j) = max_{1\le i\le N}\left[\delta_{t-1}(i)a_{ij}b_j(o_t)\right] \qquad (9)$$

Donde, $\delta_t(j)$: representa la probabilidad del camino más probable que termina en el estado $j$ tras $t$observaciones
Mientras estima las probabilidades, el algoritmo guarda en cada instante la ruta más probable recorrida hasta ese punto. Gracias a estos punteros de retroceso, es posible seguir el camino en sentido inverso al final del cálculo y obtener la secuencia de estados que mejor explica las observaciones [34]. En este caso el algoritmo de Viterbi identifica la evolución temporal más probable de los niveles de ansiedad, permitiendo caracterizar fases de estabilidad, transición o recaída emocional dentro de una población [35].

El algoritmo Baum-Welch es un caso particular del método de Expectation-Maximization (EM) aplicado al entrenamiento de los parámetros del HMM. Su objetivo es estimar las matrices de transición $A$ y emisión $B$ que maximizan la verosimilitud de las observaciones $P(O \mid \lambda)$ [36].

$$\xi_t(i,j) = \frac{\alpha_t(i)\alpha_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{P(O|\lambda)} \qquad (10) \qquad \gamma_t(j) = \frac{\alpha_t(j)\beta_t(j)}{P(O|\lambda)} \quad (11)$$

En la fase de Maximization (M-step), se actualizan los parámetros como:

$$a_{ij}^{new} = \frac{\sum_{t=1}^{T-1}\xi_t(i,j)}{\sum_{t=1}^{T-1}\gamma_t(i)} \quad (12)\,, \qquad b_{ij}^{new}(v_k) = \frac{\sum_{t:o_t=v_k}\gamma_t(j)}{\sum_{t=1}^{T}\gamma_t(j)} \quad (13)$$

Este procedimiento iterativo continúa hasta que la variación de la verosimilitud converge[36]. En términos psicológicos, el algoritmo ajusta las probabilidades de transi-

ción entre estados mentales y las probabilidades de observación de cada indicador hasta que el modelo describe de forma óptima la evolución temporal de la ansiedad.

**Distribución estacionaria en HMM**

La distribución estacionaria constituye uno de los conceptos fundamentales en la teoría de cadenas de Márkov y, por extensión, en los Modelos Ocultos de Márkov (HMM) [37]. En términos generales, una distribución estacionaria es un vector de probabilidades $\pi^*$que permanece invariante bajo la dinámica del proceso, de modo que[37, 38]:

$$\pi^* = \pi^* A \qquad (14)$$

Donde; $A$: representa la matriz de transición del sistema. Esta condición implica que $\pi^*$es un autovector izquierdo de $A$ asociado al autovalor 1, sujeto a las restricciones $\pi_i^* \geq 0$ y $\sum_i \pi_i^* = 1$[38]. En otras palabras, si la cadena se inicializa con la distribución $\pi^*$, la probabilidad de encontrarse en cada estado se mantiene constante en todos los instantes temporales[38, 39].

Esta condición expresa que, una vez alcanzado el equilibrio, la probabilidad de encontrarse en cada uno de los estados no cambia con el tiempo, independientemente de la distribución inicial del proceso[38] . Desde esta perspectiva, la distribución estacionaria refleja el comportamiento asintótico del modelo, indicando la proporción de tiempo que, en el largo plazo, el sistema tiende a permanecer en cada estado[38, 39].
En síntesis, la distribución estacionaria proporciona un marco conceptual sólido para interpretar el comportamiento de largo plazo de los estados ocultos en un HMM[39]. Su análisis no solo revela la estructura global de persistencia entre estados, sino que también permite comprender cómo se organiza la dinámica interna del fenómeno estudiado, facilitando interpretaciones que trascienden el análisis puntual de transiciones inmediatas[37].

## 3.5 Métricas

En epidemiología y salud mental es cada vez más frecuente representar los fenómenos clínicos como procesos probabilísticos, en los que el interés no recae únicamente en valores promedio, sino en la distribución completa de estados de salud y en la probabilidad de eventos adversos[15, 16]. En este tipo de enfoques se utilizan modelos de Markov, modelos ocultos, microsimulación y técnicas de inferencia bayesiana para describir trayectorias de enfermedad[16], evaluar intervenciones y cuantificar el riesgo asociado a diferentes escenarios.[15]

En particular, los modelos probabilísticos permiten analizar no solo el comportamiento central de una variable sino también el peso de las colas de la distribución, donde se ubican los episodios más graves o menos frecuentes [27, 29]. Esta perspectiva es especialmente relevante en salud mental, donde la ocurrencia de crisis, recaídas o estados de alto malestar suele ser poco frecuente pero clínicamente crítica[27], y donde los estudios recientes de psiquiatría computacional enfatizan la importancia de

describir la dinámica de riesgo a lo largo del tiempo en lugar de limitarse a mediciones puntuales[29].

En este proyecto, la evaluación del modelo se enmarca en esa perspectiva probabilística [6, 13, 15, 16]. A partir de la distribución del indicador de ansiedad, se compara de manera formal la distribución empírica con la distribución teórica ajustada mediante medidas de distancia entre distribuciones y, adicionalmente, se cuantifica el riesgo extremo asociado a los niveles más bajos de bienestar psicológico. Para ello se emplean, respectivamente, la distancia de Wasserstein[40], que resume de manera robusta la discrepancia global entre distribuciones, y el Valor en Riesgo (VaR)[41], que permite caracterizar el comportamiento de la cola crítica de la distribución. Estas herramientas se describen en detalle en las secciones siguientes.

**Evaluación de bondad de ajuste y distancia de Wasserstein**

La evaluación de la bondad de ajuste en modelos probabilísticos requiere métricas capaces de capturar no solo diferencias en parámetros de tendencia central o dispersión, sino también en la forma completa de la distribución. En este sentido, la distancia de Wasserstein[40, 42], también conocida como earth-mover's distance o distancia de transporte óptimo, constituye una herramienta robusta para comparar distribuciones de probabilidad[40, 42].

Según Panaretos y Zemel (2019) [40], la distancia de Wasserstein mide el mínimo esfuerzo necesario para transformar una distribución μ en otra ν, considerando la geometría del espacio donde se definen. Esta propiedad la hace especialmente adecuada para analizar fenómenos complejos, donde las variables subyacentes presentan asimetrías, colas largas y no linealidades que dificultan la comparación bajo métricas tradicionales como las divergencias (Kullback–Leibler y Jensen–Shannon) o por pruebas clásicas de bondad de ajuste como Kolmogorov–Smirnov, Anderson–Darling y Shapiro–Wilk [40, 42].

Matemáticamente, la distancia $W_p(\mu, \nu)$ se define como el valor mínimo esperado del costo $\| X - Y \|^p$para todos los posibles acoplamientos $(X, Y)$ de las medidas μ y ν [40]. Esta formulación permite cuantificar la proximidad entre una distribución empírica observada y una distribución teórica ajustada, lo que resulta esencial en la validación de modelos estadísticos y estocásticos [40, 42].

En los últimos años, la distancia de Wasserstein ha sido incorporada dentro del campo de la inferencia estadística exacta. Le Duy y Takeuchi (2022) [42] demostraron que es posible derivar inferencias precisas para esta métrica sin recurrir a aproximaciones asintóticas, lo cual la convierte en una alternativa confiable para evaluar ajustes de modelos incluso en muestras pequeñas.

De esta manera, el uso de la distancia de Wasserstein en este trabajo se justifica como una estrategia metodológica para evaluar el grado de correspondencia entre las distribuciones empíricas de los niveles de ansiedad observados y las distribuciones teóricas ajustadas, bajo el criterio de que valores más bajos de la distancia indican un mejor ajuste del modelo a los datos.

**Valor en Riesgo (VaR)**

El Valor en Riesgo (Value at Risk, VaR) [41, 43]es una métrica ampliamente utilizada en el análisis de riesgos para cuantificar la magnitud de eventos extremos dentro de una distribución de probabilidad. Formalmente, para un nivel de confianza $\alpha$, el VaR se define como el cuantil inferior de orden $1-\alpha$ de la distribución de una variable aleatoria $X$ [43]:

$$\mathrm{VaR}_{\alpha}(X) = F^{-1}(1-\alpha), \quad (15)$$

Donde, $F^{-1}$: representa la función cuantílica. Esta métrica permite identificar escenarios adversos que no son evidentes mediante estadísticas centrales como la media o la mediana, y constituye una herramienta fundamental para evaluar el comportamiento de la distribución en sus colas [43]. La utilidad del VaR es particularmente destacada en contextos donde las variables presentan distribuciones no normales, asimétricas o con colas pesadas [43]. En tales situaciones, los eventos extremos pueden desviarse de manera sustancial del comportamiento promedio, lo que hace necesario recurrir a métricas basadas en cuantiles para caracterizar adecuadamente el riesgo[41]. Al enfocarse directamente en la probabilidad y magnitud de estos eventos, el VaR proporciona una medida robusta y comparativamente simple para evaluar la severidad de los escenarios menos favorables [41, 43].

Cuando se combina con técnicas complementarias el VaR se integra en un marco analítico más amplio que permite caracterizar de manera precisa el riesgo extremo asociado a un fenómeno de interés[41]. Esta aproximación resulta especialmente pertinente en dominios donde la variabilidad y las colas de la distribución contienen información crítica para la comprensión del proceso subyacente[41, 43].

# 4 Análisis y Discusión de Resultados

## 4.1 Análisis exploratorio de los datos

Partiendo del conjunto de estudio: el Social Anxiety Dataset[22], un conjunto de datos público disponible en Kaggle. Este recurso reúne más de 10.000 registros individuales e integra factores conductuales, de estilo de vida y psicológicos, con el propósito de modelar la ansiedad social a partir de variables como hábitos cotidianos, características demográficas e indicadores autoinformados de malestar emocional[22]. Su estructura, orientada al análisis estadístico y al entrenamiento de modelos de aprendizaje automático, ofrece un marco comparativo útil para contextualizar los resultados del presente estudio y reforzar la validez externa de los patrones identificados muestra analizada.

| Columna | Cantidad de datos no nulos | Tipo de dato |
|---|---|---|
| Age | 11000 | float64 |
| Gender | 11000 | object |
| Occupation | 11000 | object |
| Sleep Hours | 11000 | int32 |
| Physical Activity | 11000 | int32 |
| Caffeine Intake | 11000 | float64 |
| Alcohol Consumption | 11000 | int64 |
| Smoking | 11000 | object |
| Family History of Anxiety | 11000 | object |
| Stress Level (1-10) | 11000 | int64 |

| Columna | Cantidad de datos no nulos | Tipo de dato |
|---|---|---|
| Heart Rate (bpm) | 11000 | float64 |
| Breathing Rate | 11000 | int64 |
| Sweating Level (1-5) | 11000 | int64 |
| Dizziness | 11000 | object |
| Medication | 11000 | object |
| Therapy Sessions | 11000 | int64 |
| Recent Major Life Event | 11000 | object |
| Diet Quality (1-10) | 11000 | int64 |
| Anxiety Level (1-10) | 11000 | float64 |

**Fig. 2.** Esta figura presenta la información descriptiva del dataset, mostrando el tipo de dato y el número de observaciones por variable.

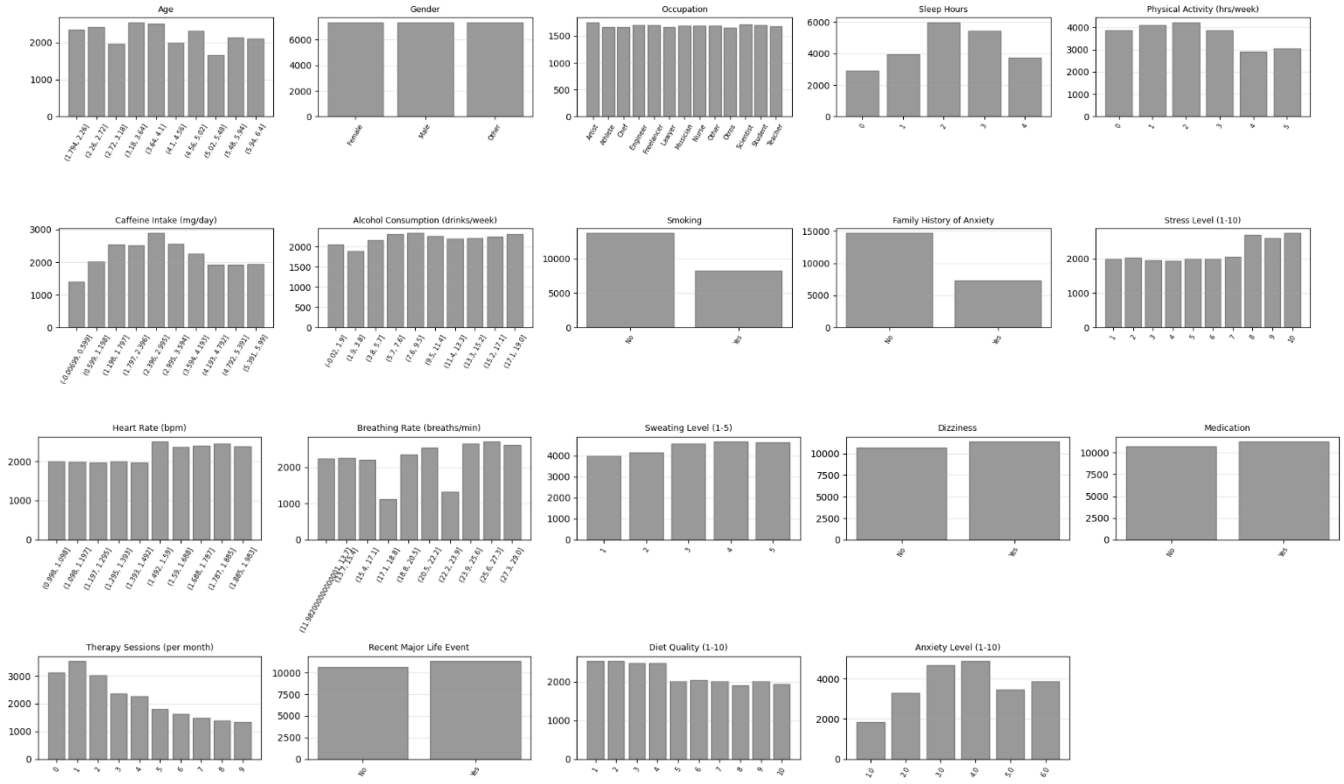


**Fig. 3.** Gráfico de barras e histograma

La figura 4 presenta la distribución de frecuencias para todas las variables del y Dataset. En términos demográficos, la edad se encuentra relativamente bien distribuida a lo largo de los distintos rangos etarios y el género muestra proporciones similares entre las categorías, al igual que los diferentes tipos de ocupación, lo que sugiere una muestra heterogénea. En las variables de estilo de vida se observa que la mayoría de los individuos duerme entre 7 y 8 horas, realiza niveles intermedios de actividad física y se concentra en rangos moderados de consumo de cafeína y alcohol. Predominan los no fumadores, las personas sin eventos vitales estresantes recientes y quienes no están en tratamiento farmacológico ni en terapia frecuente, lo que apunta a una población mayoritariamente no clínica. Los indicadores fisiológicos (frecuencia cardiaca y respiratoria) se concentran en intervalos típicos, sin colas extremas pronunciadas. Finalmente, las escalas de estrés y ansiedad muestran valores repartidos a lo largo de toda la escala, con ligera concentración en niveles intermedios, lo que indica variabilidad suficiente para el análisis posterior de patrones de riesgo psicológico.

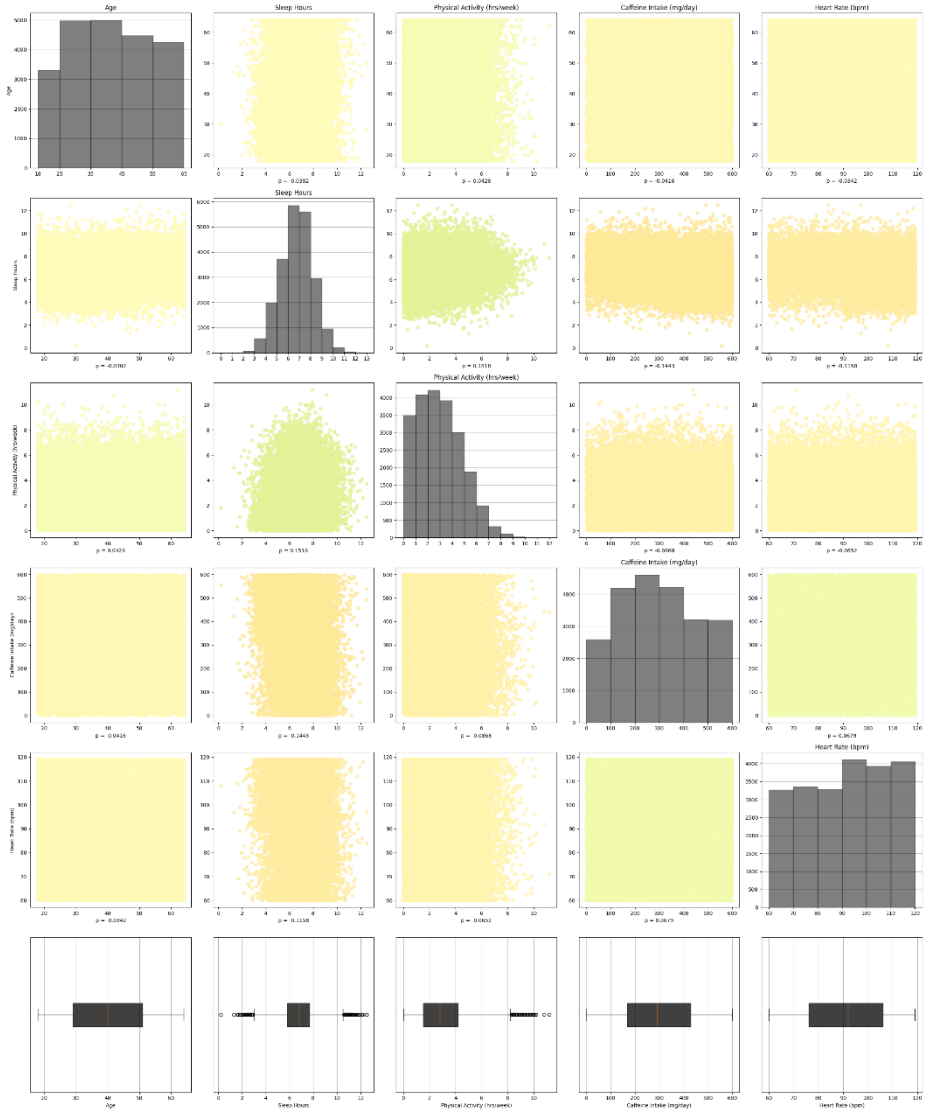


**Fig.** 4. Distribución univariada, relaciones bivariadas y valores atípicos de las variables continuas (Age, Sleep Hours, Physical Activity, Caffeine Intake y Heart Rate)

Esta figura combina histogramas, diagramas de dispersión y boxplots para cinco variables continuas. En la diagonal se aprecia que Sleep Hours sigue una distribución aproximadamente normal centrada alrededor de 7–8 horas, mientras que Physical Activity está claramente sesgada a la derecha (la mayoría realiza entre 2 y 6 horas/semana). El Caffeine Intake se concentra en rangos intermedios, sin colas extremas muy marcadas, y la Heart Rate se distribuye de forma relativamente uniforme

dentro del rango observado; la Age se reparte entre adultos jóvenes y de mediana edad, con ligera concentración en los tramos intermedios.
Los diagramas de dispersión muestran nubes de puntos densas, pero sin patrones definidos, lo que sugiere correlaciones lineales débiles o cercanas a cero entre las variables (por ejemplo, la relación entre edad y sueño o entre sueño y actividad física parece mínima). En la parte inferior, los boxplots confirman la ausencia de asimetrías extremas y permiten identificar algunos valores atípicos puntuales (especialmente en actividad física y consumo de cafeína), pero sin indicios de datos aberrantes masivos. En conjunto, la figura indica que las variables aportan información relativamente complementaria y que no existe una colinealidad fuerte entre estos indicadores de estilo de vida y parámetros fisiológicos.

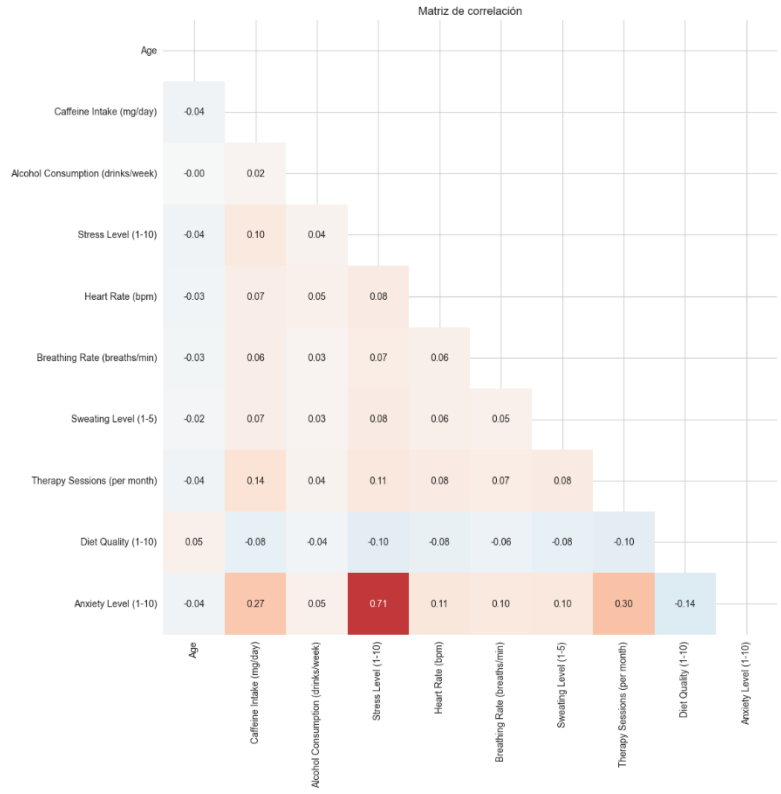


**Fig. 5.** Matriz de correlación de las variables continuas del dataset

La matriz de correlación sobre las variables del dataset muestra, en primer lugar, que la relación más intensa se da entre Stress Level y Anxiety Level ($\rho \approx 0{,}71$), lo que indica que los mayores niveles de ansiedad se asocian de manera consistente con un incremento del estrés percibido. En segundo lugar, la ansiedad presenta correlaciones positivas pero moderadas con el consumo de cafeína ($\rho \approx 0{,}27$) y con la frecuencia de Therapy Sessions ($\rho \approx 0{,}30$), lo que sugiere que tanto ciertos hábitos estimulantes

como la búsqueda de ayuda profesional tienden a concentrarse en los individuos con mayor malestar psicológico. En contraste, Diet Quality muestra una correlación negativa débil con la ansiedad ($\rho \approx -0{,}14$), compatible con la idea de que un peor estado emocional se acompaña de hábitos alimentarios menos saludables, aunque sin una relación lineal marcada.

El resto de las asociaciones entre las variables continuas, así como entre estas y la ansiedad, se mantiene en rangos muy bajos ($|\rho| < 0{,}15$), en particular para los indicadores fisiológicos (Heart Rate, Breathing Rate, Sweating Level) y para Age y Alcohol Consumption. Este patrón sugiere una baja colinealidad entre los predictores y respalda su uso conjunto en etapas posteriores de modelación, dado que cada variable aporta información relativamente complementaria sobre el estado de bienestar/malestar psicológico sin redundancias fuertes.

## 4.2 Preparación de los datos

**Detección y tratamiento de valores atípicos**

Inicialmente, la variable Anxiety Level presentaba una escala amplia (1 a 10). Con el fin de reducir la dispersión y concentrar los valores extremos, se agruparon las categorías superiores (6 a 10) en un solo nivel. Esta transformación permitió obtener una nueva escala de seis niveles (1–6), conservando la estructura ordinal y facilitando la comparación entre grupos de distinta intensidad

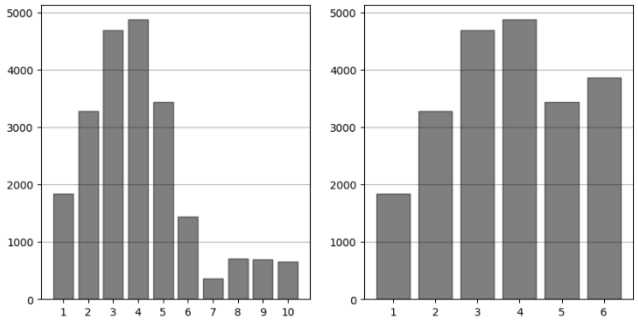


**Fig. 6.** Distribución original (1–10) y transformada (1–6) de la variable Anxiety Level, tras agrupar los niveles superiores (6–10) con el fin de reducir la dispersión y concentrar los valores extremos.

De manera similar, la variable Therapy Sessions se encontraba originalmente en un rango de 0 a 12. Con el fin de controlar el efecto de valores infrecuentes (10 a 12 sesiones), estos se consolidaron en una categoría superior denominada “9 o más”. Esta decisión permitió conservar la representatividad de los casos habituales y reducir la variabilidad, manteniendo la escala ordinal del fenómeno observado.

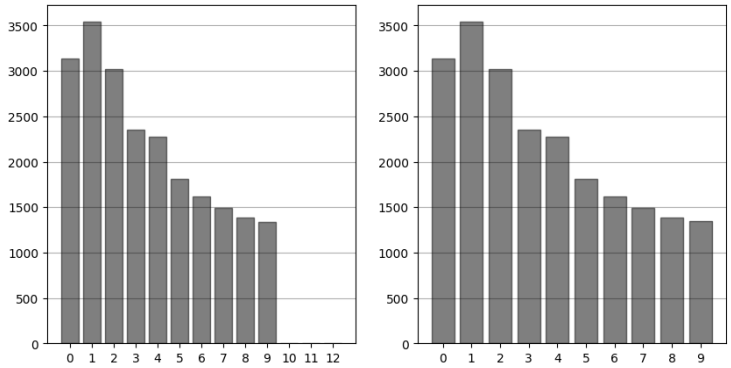


**Fig. 7.** Distribución original y transformada de la variable Therapy session, tras agrupar los niveles superiores con el fin de reducir la dispersión y concentrar los valores extremos.

La variable Sleep Hours fue discretizada en intervalos significativos para la interpretación clínica: menor de 5, entre 5–6, 6–7, 7–8 y más de 8 horas. Este proceso de categorización respondió a la necesidad de capturar patrones de descanso asociados con niveles de ansiedad y bienestar, convirtiendo una variable continua en una discreta con cinco grupos interpretables. Los histogramas resultantes mostraron una concentración central en los rangos de 6–7 y 7–8 horas, coherente con los valores esperados en población adulta.

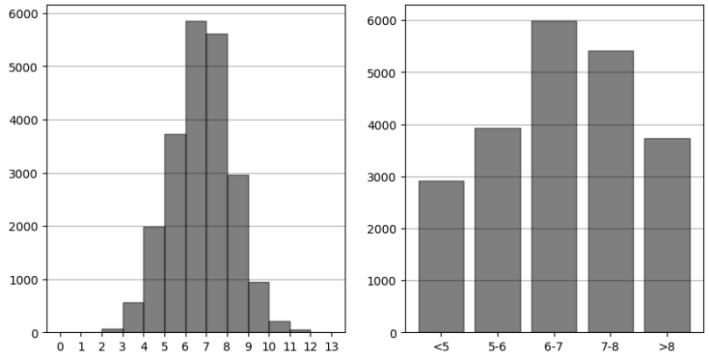


**Fig. 8.** Distribución original y transformada de la variable Sleep Hours, tras agrupar los niveles superiores con el fin de reducir la dispersión y concentrar los valores extremos.

Finalmente, la variable Physical Activity fue reagrupada en intervalos de una hora (0–1, 1–2, 2–3, 3–4, 4–5 y más de 5 horas). Esta discretización permitió eliminar sesgos derivados de registros atípicos de alta frecuencia y facilitar comparaciones posteriores en los análisis de correlación y agrupamiento.

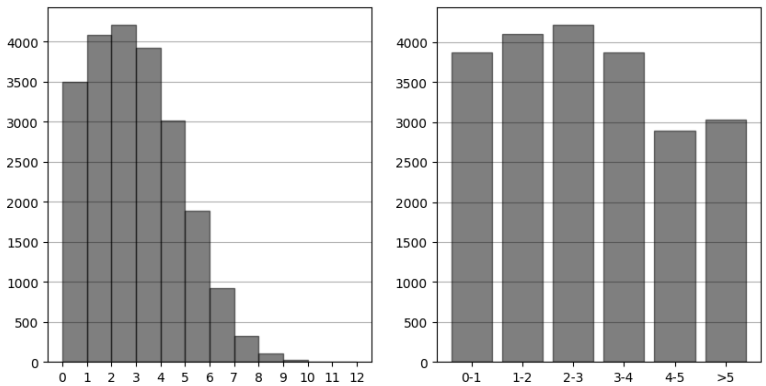

**Fig. 9.** Distribución original y transformada de la variable Physical Activity, tras agrupar los niveles superiores con el fin de reducir la dispersión y concentrar los valores extremos.

### Normalización y estandarización de variables

Luego de la normalización min–max, todas las variables continuas y ordinales quedaron reescaladas al rango [0,1], manteniendo su forma original pero en una escala común. En las variables categóricas ordinales (Stress Level, Anxiety Level, Sweating Level, Diet Quality, Therapy Sessions, Sleep Hours, Physical Activity) se observan picos discretos bien definidos, mientras que en las variables más continuas (Age, Caffeine Intake, Alcohol Consumption, Breathing Rate y Heart Rate) las distribuciones se tornan más suaves y extendidas. Este reescalamiento facilita la comparación entre indicadores heterogéneos y evita que las diferencias de unidades o magnitudes distorsionen las etapas posteriores de modelación estadística y de aprendizaje automático.

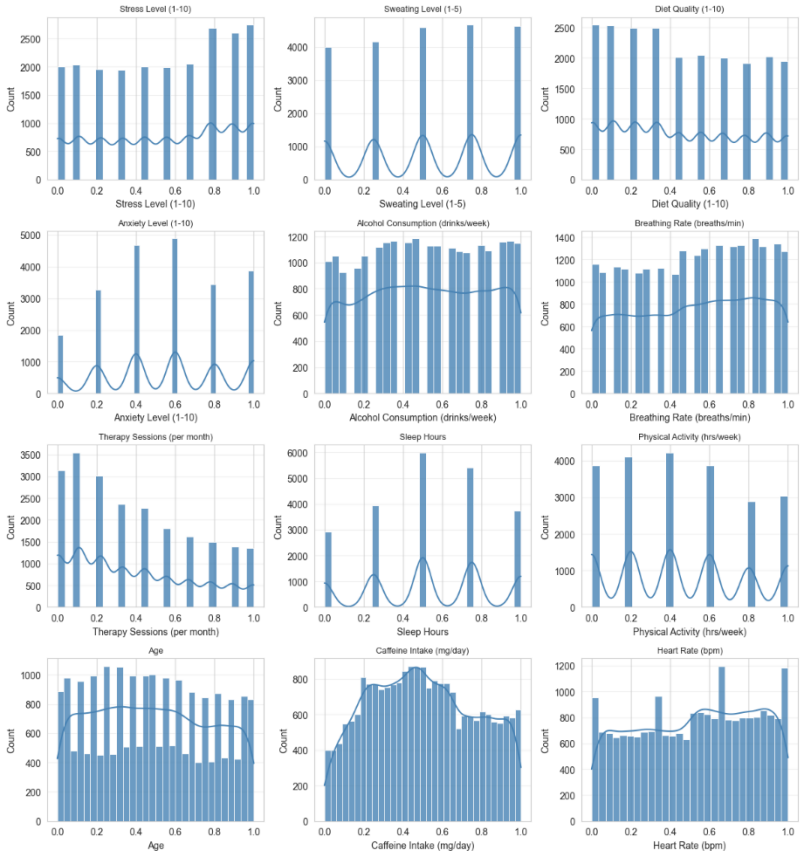


**Fig. 10.** Distribuciones de las variables continuas y ordinales del Dataset después de la normalización min–max al rango [0,1]

En todas las variables dicotómicas se observa un claro predominio de la categoría 0 (ausencia del factor), lo que indica que la mayoría de los participantes no fuma, no presenta mareos frecuentes, no se encuentra en tratamiento farmacológico ni ha experimentado recientemente un evento vital mayor; aun así, la proporción no despreciable de casos en la categoría 1 aporta variabilidad suficiente para explorar su relación con los niveles de ansiedad.

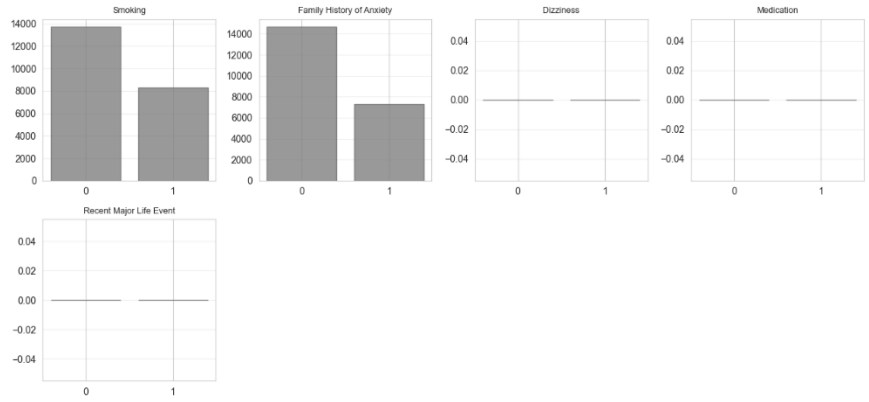


**Fig. 11.** Distribución de frecuencias de las variables binarias del dataset

### 4.3 Selección de variables

En esta etapa del análisis se decidió trabajar únicamente con variables continuas de carácter psicológico y de hábitos de comportamiento (por ejemplo, Stress Level, Sleep Hours, Physical Activity, Caffeine Intake, Therapy Sessions), excluyendo variables sociodemográficas como género, edad, profesión u otras características estructurales. Esta decisión responde a dos criterios principales: en primer lugar, estas variables psicológicas y conductuales representan indicadores más proximales y potencialmente modificables del estado de ansiedad, y por tanto resultan más coherentes con el objetivo de modelar su dinámica temporal y de diseñar intervenciones sobre el estilo de vida[8, 9]. En segundo lugar, las variables sociodemográficas se utilizaron de forma complementaria para describir y caracterizar los perfiles obtenidos, pero no se incorporaron como entradas del modelo a fin de evitar que los patrones latentes quedaran dominados por diferencias estructurales (por ejemplo, entre grupos de edad o tipos de ocupación) más que por la expresión conductual y psicológica del fenómeno de interés.

A partir de la matriz de correlación (Fig. 12), se priorizaron como variables observables con una asociación al menos ligera con el nivel de ansiedad ($|r| \geq 0{,}15$) y con relevancia teórica para el bienestar psicológico[9]. Bajo este criterio se conservaron Stress Level, Sleep Hours, Therapy Sessions, Physical Activity y Caffeine Intake, mientras que Diet Quality y Heart Rate fueron excluidas del conjunto final de variables continuas, dado que mostraron las correlaciones más bajas con Anxiety Level (|r|

≤ 0,14) y aportaban una capacidad explicativa limitada respecto al constructo de interés.

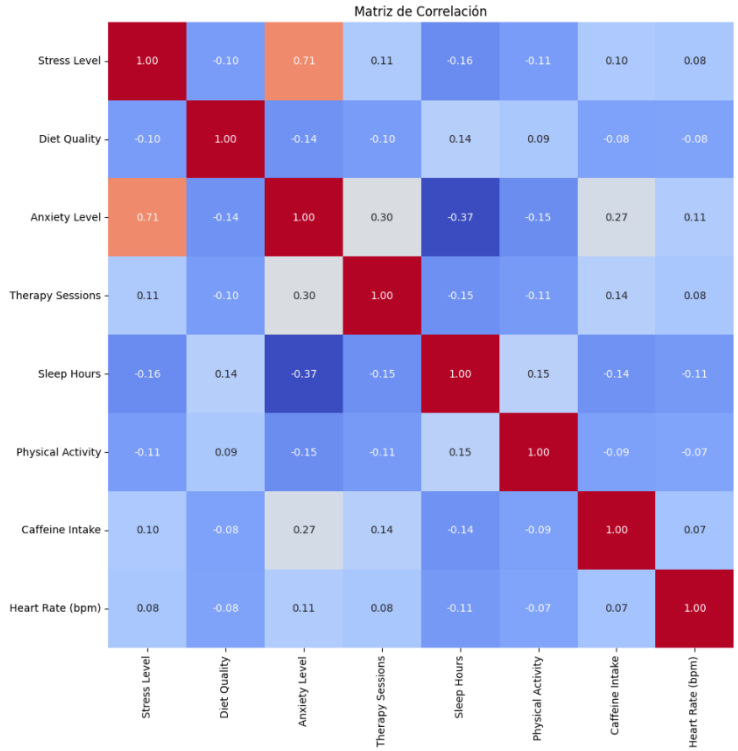


**Fig. 12.** Matriz de correlación de las variables continuas

Dentro del conjunto de variables binarias consideradas (Smoking, Family History of Anxiety, Dizziness, Medication y Recent Major Life Event), se realizó un análisis comparativo de su variabilidad y relación con las dimensiones emocionales. De este proceso se identificó que la variable Smoking mostraba una mayor coherencia teórica y estadística con el fenómeno en estudio, por lo que se mantuvo como factor complementario en el modelo.

El hábito de fumar se reconoce en la literatura psicológica y médica como un comportamiento compensatorio frente al estrés o la ansiedad, frecuentemente utilizado como mecanismo de regulación emocional a corto plazo. Numerosos estudios han documentado una asociación bidireccional entre el consumo de tabaco y la presencia de síntomas ansiosos, en la cual la nicotina puede generar una sensación temporal de alivio, pero también contribuir al incremento de la activación fisiológica y la dependencia emocional[29].

La inclusión de Smoking en el modelo responde a su valor explicativo como indicador de un hábito conductual y su potencial para reflejar patrones compensatorios dentro del conjunto de observables[29]. Su permanencia aporta una dimensión adicional al análisis, permitiendo interpretar la ansiedad no solo desde parámetros psicoló-

gicos y fisiológicos, sino también a través de comportamientos de afrontamiento asociados al estilo de vida[29].

Con el fin de confirmar la independencia relativa entre las variables seleccionadas y evitar redundancias estadísticas, se realizó un análisis de correlaciones parciales. Este permitió estimar la relación específica entre cada par de variables controlando por el efecto de las demás. Los resultados mostraron coeficientes bajos o moderados, lo cual sugiere una estructura de covariación controlada y una baja multicolinealidad entre los indicadores.

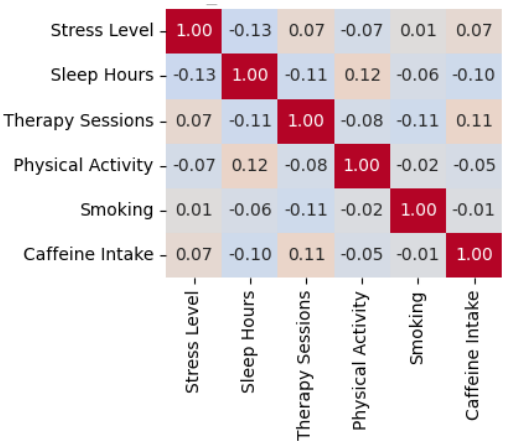


**Fig. 13.** Matriz de correlaciones parciales de las variables seleccionadas

En los resultados del análisis de correlaciones parciales se observó que Stress Level mantiene una relación leve y negativa con Sleep Hours (r = –0.13), en línea con la evidencia que vincula la falta de descanso con mayores niveles de tensión emocional.

Por su parte, Caffeine Intake mostró asociaciones débiles con el resto de las dimensiones, lo que refuerza su utilidad como indicador independiente de hábitos estimulantes que pueden influir en la activación fisiológica. Finalmente, Physical Activity presentó correlaciones cercanas a cero con las demás variables, lo que evidencia que aporta información complementaria y no redundante dentro del conjunto seleccionado.

Este análisis permitió consolidar el subconjunto de variables observables con suficiente independencia estadística, reforzando su idoneidad para ser modeladas conjuntamente bajo la metodología de Modelos Ocultos de Markov (HMM).

## 4.4 Clusterización y reducción de dimensionalidad

**Variables observables**

Con el fin de identificar la estructura latente de las variables observables y obtener una segmentación conductual robusta, se evaluaron tres metodologías de clusteriza-

ción ampliamente utilizadas en análisis multivariado: K-Means, K-Medoids y K-Nearest Neighbours.

La evaluación se realizó sobre el espacio reducido mediante Análisis de Componentes Principales, manteniendo tres componentes que preservaron más del 80% de la varianza total del sistema. Para comparar el desempeño de cada algoritmo se utilizaron métricas internas de validación no supervisada: el Silhouette Score, el índice de Calinski–Harabasz y el índice de Davies–Bouldin. Estas métricas cuantifican, respectivamente, la coherencia interna de cada grupo, la separación entre clústeres y el grado de solapamiento entre ellos.

| Método | *Silhouette* | *Calinski-Harabasz* | *Davies-Bouldin* |
|---|---|---|---|
| K-Means | 0.3027 | 9765.7 | 1.196 |
| K-Medoids | 0.3011 | 9723.6 | 1.202 |
| K-NN | 0.2798 | 7872.4 | 1.205 |

**Fig. 14.** Tabla comparativa de las métricas de los algoritmos de clusterización.

Los resultados muestran diferencias relevantes entre los métodos. K-Means obtuvo las puntuaciones más altas en Silhouette (0.3027) y Calinski–Harabasz (9765.7), además del valor más bajo en Davies–Bouldin (1.196), lo que indica una estructura más nítida, compacta y mejor diferenciada en el espacio PCA. K-Medianas presentó un comportamiento muy cercano, con valores ligeramente inferiores en las tres métricas, reflejando una segmentación similar pero más conservadora. Por su parte, el método basado en proximidad local KNN obtuvo los valores menos favorables, evidenciando clústeres menos definidos y una mayor dispersión interna respecto a las alternativas centradas en centroides. Desde una perspectiva metodológica y empírica, la consistencia de los resultados respalda la selección de K-Means como el algoritmo principal de clusterización para esta etapa. Su capacidad para maximizar simultáneamente la cohesión y la separación entre grupos lo convierte en una herramienta adecuada para sintetizar patrones conductuales relevantes y para generar estados observables estables que posteriormente alimentan el Modelo Oculto de Markov.

El número de clústeres se evaluó inicialmente mediante el método del codo, utilizando la inercia como criterio de segmentación. La gráfica obtenida muestra una disminución marcada entre k = 2, k = 3 y k = 4, seguida de una reducción significativa en valores superiores. Este patrón sugiere que k = 3 o k = 4 son candidatos razonables para capturar la estructura subyacente.

| *K* | *Silhouette* | *Calinski-Harabasz* | *Davies-Bouldin* |
|---|---|---|---|
| *3* | 0,3027 | 9.765,7 | 11.963 |
| *4* | 0,3012 | 9.998,94 | 10.845 |

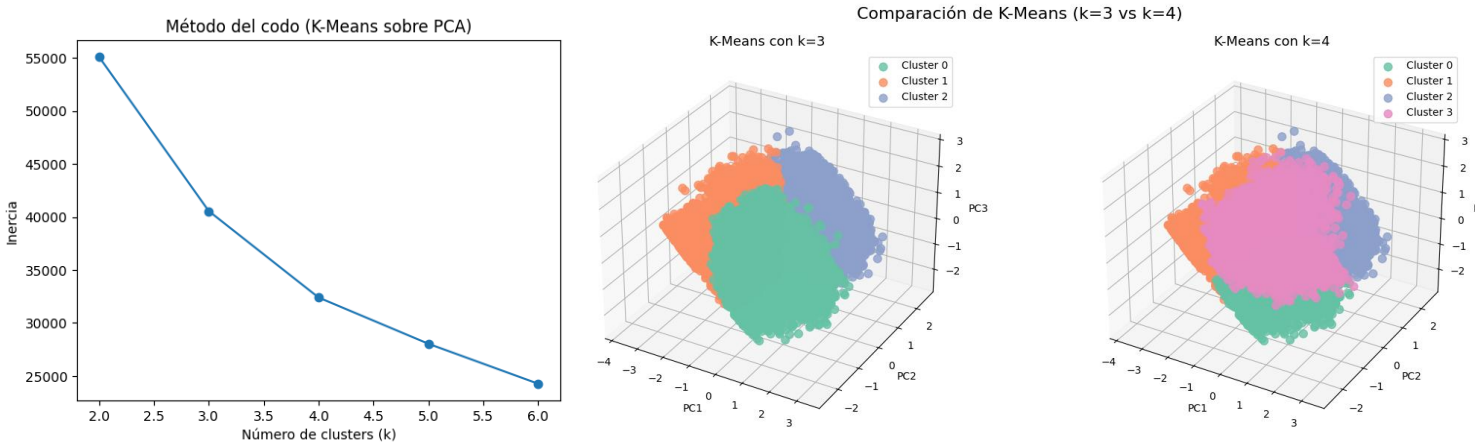


**Fig. 15.** Tabla y grafica comparativa de los clusters, métricas y método del codo con k=3 y k=4

Interpretando los resultados, la métrica de Silhouette es prácticamente equivalente para ambos valores, lo que sugiere cohesión interna similar, en cuanto a los indicadores de Calinski–Harabasz y Davies–Bouldin, el k=4 tiene mejor desempeño en ambos, indicando una mejor separación entre los clústers y a su vez reflejando clusters más definidos. Considerando de manera conjunta el método del codo, la visualización de los clúster y las métricas evaluadas, la evidencia empírica favorece una solución de k = 4, dado que permite capturar mayor heterogeneidad en los perfiles conductuales.

A pesar de que las métricas internas muestran un desempeño comparable entre k = 3 y k = 4, la elección final se orientó hacia k = 3, dado que ofrece un nivel de granularidad más coherente con los propósitos del modelo. En el contexto de los Modelos Ocultos de Markov, un número excesivo de estados puede dificultar la interpretación de las trayectorias, generar transiciones poco estables y fragmentar perfiles que conceptualmente pertenecen a una misma categoría de riesgo. La solución de tres clústeres, por el contrario, produce grupos bien definidos, suficientemente diferenciados y alineados con una estructura latente interpretable.

Además, la configuración con k = 3 mantiene una estabilidad empírica sólida, con métricas de cohesión y separación prácticamente equivalentes a las obtenidas con k = 4. Bajo esta consideración, la solución con tres grupos maximiza la claridad conceptual sin sacrificar capacidad descriptiva, y proporciona una base más parsimoniosa y manejable para la estimación posterior del HMM. Por estas razones, k = 3 se adopta como la estructura óptima para representar los estados latentes del fenómeno.

La aplicación del algoritmo de segmentación K-Means con k = 3 permitió identificar tres configuraciones diferenciadas de comportamiento y hábitos asociados al bienestar psicológico. Cada grupo agrupa patrones multivariados que reflejan distintos niveles de riesgo y exposición a factores vinculados a la ansiedad.

Los promedios de las variables observables dentro de cada clúster revelan diferencias significativas en estrés percibido, calidad del sueño, actividad física, consumo de estimulantes y uso de servicios terapéuticos. A continuación, se presenta un resumen de estos patrones, facilitando la interpretación comparativa.

| Variable | Clúster 0 - Alto malestar | Clúster 1 - Riesgo moderado | Clúster 2 - Bienestar relativo |
|---|---|---|---|
| **Stress Level** | 0,83 | 0,58 | 0,38 |
| **Sleep Hours** | 0,16 | 0,52 | 0,70 |
| **Therapy Sessions** | 0,71 | 0,43 | 0,24 |
| **Physical Activity** | 0,19 | 0,41 | 0,62 |
| **Smoking** | 0,37 | 0,34 | 0,41 |
| **Caffeine Intake** | 0,73 | 0,54 | 0,37 |

**Fig. 16.** Resumen comparativo de promedios de las variables observables por clúster

El clúster 0 representa el perfil de mayor malestar psicológico. Este grupo se caracteriza por registrar los niveles más altos de estrés, junto con patrones de comportamiento marcadamente disfuncionales: un número significativamente reducido de horas de sueño, baja actividad física y un consumo elevado de cafeína. Adicionalmente, es el clúster con mayor frecuencia de participación en sesiones terapéuticas, lo cual puede interpretarse como una manifestación de búsqueda de apoyo ante el malestar acumulado. Este conjunto de indicadores manifiesta signos claros de deterioro, es coherente con un estado latente de ansiedad alta.

El clúster 1 refleja un perfil intermedio o de riesgo moderado. Los individuos agrupados en este clúster presentan niveles de estrés elevados, pero no extremos y un patrón mixto de hábitos: horas de sueño moderadas, actividad física intermedia y una presencia de estimulantes menor a la observada en el clúster de mayor malestar. La frecuencia en sesiones terapéuticas se ubica igualmente en un punto medio, lo que sugiere que estos individuos experimentan dificultades perceptibles, pero mantienen ciertos elementos de regulación conductual. Este clúster puede considerarse un estado de ansiedad moderada, que ocupa un lugar transicional dentro del continuo de riesgo psicológico.

Por su parte, el clúster 2 constituye el perfil de mayor bienestar relativo dentro de la muestra. Este grupo exhibe los niveles más bajos de estrés, acompañado de patrones de sueño considerablemente más favorables y una mayor actividad física. La participación en sesiones terapéuticas es reducida, lo que coincide con un menor nivel de malestar percibido. Si bien presenta el nivel más alto de tabaquismo, este comportamiento no altera la estructura general de bienestar observada en el conjunto de variables, que es sustancialmente más favorable que en los demás clústeres. En términos latentes, este grupo se corresponde con un estado de ansiedad baja.

En síntesis, los tres clústeres conforman un gradiente progresivo de riesgo psicológico, que va desde el alto malestar (Clúster 0), pasando por un estado intermedio (Clúster 1), hasta un nivel de bienestar relativo (Clúster 2). Esta estructura es consistente tanto con la literatura sobre estratificación del riesgo en salud mental como con la lógica del modelo dinámico propuesto [27]. Además, la diferenciación empírica entre

los clústeres proporciona una base sólida para su utilización como estados latentes dentro del Modelo Oculto de Markov (HMM), permitiendo analizar posteriormente la evolución temporal y las transiciones entre estos perfiles.

**Variables no observables**

La variable de Anxiety Level utilizada en este estudio corresponde a un indicador continuo que sintetiza el grado de malestar psicológico reportado por cada individuo. Para su utilización dentro del Modelo Oculto de Markov (HMM), resulta necesario transformar esta variable en un conjunto reducido de estados discretos, preservando la estructura subyacente del riesgo psicológico[5].

Con este propósito, se aplicó el algoritmo K-Means en una dimensión sobre los valores continuos de ansiedad. La decisión del número de clústeres se evaluó mediante los criterios habituales:

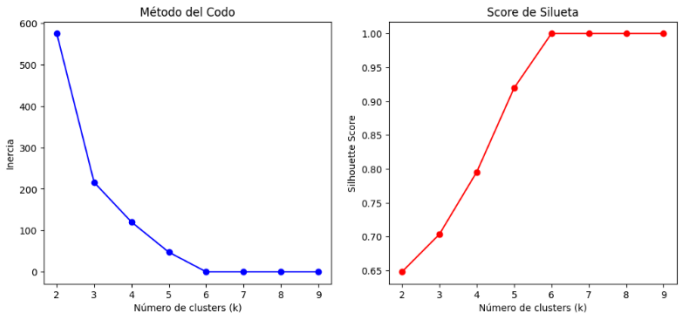


**Fig. 17.** Grafica del método del codo y score de silueta para clusters de la variable no observable

El método del codo basado en la inercia muestra una reducción marcada hasta k = 3 y una ganancia marginal para valores mayores, en cuanto al indicador de silueta, evidencia que alcanza su primer incremento significativo en k = 3, indicando una separación adecuada entre grupos.

Aunque el análisis preliminar mediante el puntaje de silueta mostró aumentos progresivos para valores superiores de k, particularmente a partir de k = 5 y k = 6, la elección del número óptimo de clústeres no puede basarse únicamente en el comportamiento de una métrica. En el caso de la variable continua de ansiedad, la selección debe equilibrar criterios estadísticos con consideraciones de parsimonia, interpretabilidad y coherencia teórica, especialmente debido a su función posterior dentro del modelo temporal.

Desde esta perspectiva, la solución con k = 3 constituye la alternativa metodológicamente más adecuada. En primer lugar, la configuración de tres clústeres ofrece una estructura conceptualmente interpretable, alineada con la representación habitual del riesgo psicológico en niveles de ansiedad baja, moderada y alta. Esta estratificación se encuentra ampliamente respaldada en la literatura y facilita la identificación de esta-

dos latentes clínicamente significativos[9]. En comparación, segmentaciones más finas como las generadas por k = 6 producen categorías demasiado específicas para ser diferenciadas conceptualmente y tienden a fragmentar un continuo sin aportar una mejora sustantiva en la comprensión del fenómeno.
En segundo lugar, el método del codo muestra una disminución significativa hasta k = 3, seguida de reducciones marginales para valores mayores. Este comportamiento sugiere que las transformaciones más significativas en la estructura del continuo de ansiedad se concentran antes del valor k = 4. Los incrementos posteriores al k=4 parecen corresponder principalmente a subdivisiones que son más redundantes del mismo eje, en lugar de identificar grupos verdaderamente diferenciados. En consecuencia, la evidencia respalda que k = 3 representa un punto de equilibrio apropiado entre complejidad analítica y capacidad discriminativa, evitando una sobre segmentación que no aporta valor interpretativo adicional.

Un elemento adicional que respalda la elección de k = 3 es la robustez en el tamaño de los clústeres: la solución final produce tres grupos de 5.125, 9.568 y 7.307 individuos, respectivamente. Esta distribución equilibrada evita la aparición de clústeres marginales o sobredimensionados y garantiza la estabilidad de las estimaciones en el Modelo Oculto de Markov (HMM), donde tamaños muy pequeños comprometerían la fiabilidad de las probabilidades de transición.

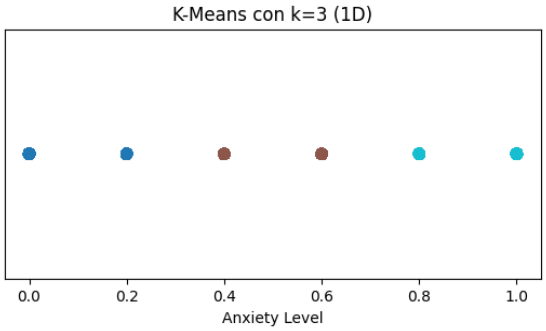


**Fig. 18.** Segmentación del indicador continuo de ansiedad mediante K-Means con k = 3 en una dimensión. Cada punto representa una observación en la escala normalizada de Anxiety Level, coloreada según el clúster asignado (bajo, moderado y alto nivel de ansiedad).

La aplicación final del algoritmo K-Means con k = 3 permitió la identificación de tres agrupaciones robustas y bien definidas, distribuidas de manera ordenada a lo largo del eje del indicador de ansiedad. Estas categorías representan de manera natural tres estados latentes de ansiedad:

- Estado 0 (Baja Ansiedad): Valores bajos del indicador, asociados a un nivel reducido de malestar psicológico.
- Estado 1 (Riesgo Moderado): Valores intermedios, interpretados como un riesgo moderado de desregulación.
- Estado 2 (Alta Ansiedad): Valores elevados del continuo, correspondientes a un nivel alto de ansiedad.

La visualización unidimensional de la distribución de los datos confirma la coherencia y solidez de la estructura obtenida, evidenciando centros de grupo bien separados y consistentes con la progresión teórica esperada del fenómeno. Esta discretización es esencial, ya que los tres estados discretos derivados (0, 1 y 2) constituyen la base empírica para la etapa posterior de modelado temporal. Dichos estados serán incorporados como nodos latentes dentro de la estructura de un Modelo Oculto de Markov (HMM). Esto permitirá la estimación de las probabilidades de transición entre los estados y el análisis de las trayectorias dinámicas del riesgo psicológico a lo largo del tiempo.

## 4.5 Modelo HMM para el análisis de la ansiedad

**Matriz de transición**

La matriz de transición $A$ representa la probabilidad de pasar del estado latente $i$ en el tiempo $t$ al estado $j$ en el tiempo $t + 1$.

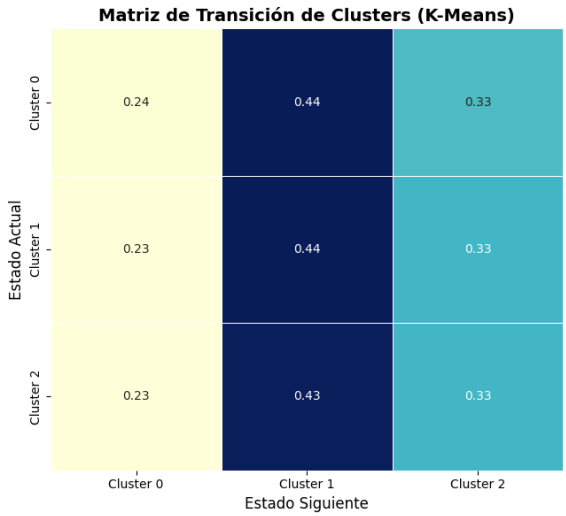


**Fig. 19.** Matriz de transición entre estados ocultos.

La matriz de transición muestra que el Estado 1 (riesgo moderado) funciona como el punto de mayor probabilidad en la evolución temporal del fenómeno, concentrando alrededor del 44% de las transiciones desde cualquier estado. Las probabilidades hacia los niveles extremos (baja y alta ansiedad) son similares (≈0.23–0.33), lo que indica que los cambios ocurren principalmente de manera gradual y rara vez implican saltos directos entre los extremos. Esta estructura sugiere una dinámica estable y homogénea, coherente con la naturaleza continua del malestar psicológico.

**Matriz de emisión**

La matriz de emisión *B* describe cómo se relacionan los estados observables de ansiedad con los estados latentes del HMM, es decir, la probabilidad de pertenecer a cada estado oculto dada la categoría observada de ansiedad.

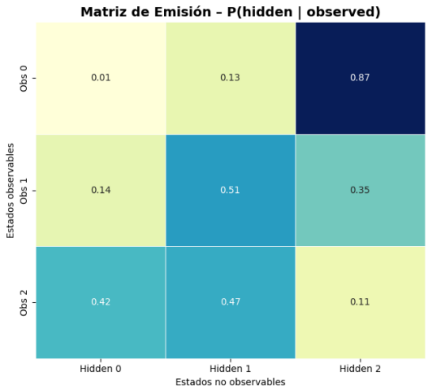


**Fig. 20.** Matriz de emisión

Esta matriz revela una correspondencia robusta entre los patrones conductuales observables y los estados latentes de ansiedad modelados por el HMM. El clúster caracterizado por alto malestar presenta una probabilidad del 87% de asociarse al estado oculto de alta ansiedad, lo que confirma la coherencia estructural entre las señales conductuales y el proceso emocional subyacente. En contraste, el clúster de riesgo moderado exhibe un comportamiento más heterogéneo, distribuyéndose entre los tres estados latentes, aunque conservando su mayor probabilidad en el estado intermedio (51%), lo que refleja la naturaleza gradual y transitoria del fenómeno.

Finalmente, el clúster de bienestar relativo muestra un desajuste relevante entre lo observable y lo latente: solo el 42% se asocia a baja ansiedad, mientras que un 47% se ubica en un estado de riesgo moderado, evidenciando que las manifestaciones externas de bienestar no necesariamente capturan la dinámica interna del malestar psicológico. En conjunto, estos resultados justifican el uso de modelos estocásticos con estados ocultos y aportan evidencia empírica sobre la complejidad del proceso ansioso en contextos laborales.

**Vector de probabilidades iniciales de estados**

El vector de distribución inicial $\pi$ describe la probabilidad de que un individuo se encuentre en cada estado latente en el primer instante observado. En este estudio, la estimación empírica muestra que el 43% de los registros se ubican inicialmente en el Estado 1 (riesgo moderado), seguido por un 33% en el Estado 2 (alta ansiedad) y un 23% en el Estado 0 (baja ansiedad).

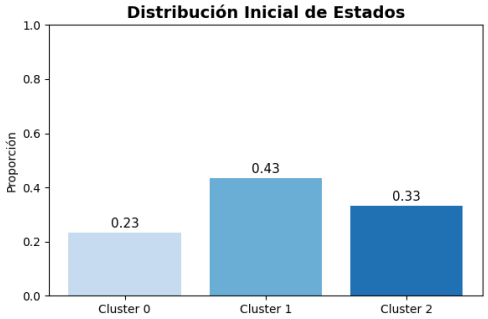


**Fig. 21.** Vector de probabilidades iniciales de estados.

Este patrón sugiere que, en el punto de partida de la serie temporal, la mayoría de los sujetos no se encuentran ni en un nivel de baja ansiedad ni en un extremo crítico, sino en una situación intermedia de vulnerabilidad psicológica. Al mismo tiempo, la proporción relativamente alta en el Estado 2 indica la presencia de un grupo no despreciable con niveles elevados de ansiedad desde el inicio, mientras que el Estado 0 concentra la menor fracción de la muestra. En conjunto, la distribución inicial refuerza la idea de un contexto con predominio de malestar moderado a alto, sobre el cual operan las transiciones dinámicas capturadas por el HMM. En síntesis, la distribución inicial sugiere que la muestra parte de un escenario en el que predomina un nivel de ansiedad entre moderado y alto, lo cual constituye el punto de partida sobre el que se desarrollan las transiciones modeladas por el HMM.

## 4.6 Algoritmos de inferencia y entrenamiento

El algoritmo Forward-Backward permite calcular la probabilidad total de la secuencia observada $P(O \mid \lambda)$, la cual resulta de multiplicar numerosas probabilidades individuales menores que uno. En secuencias extensas, este producto conduce a valores extremadamente pequeños, cercanos al límite de representación numérica. Por esta razón, es habitual expresar dicha probabilidad en su forma logarítmica —el log-likelihood— que preserva la información estadística del modelo y garantiza estabilidad computacional.

En este estudio, el modelo obtuvo un log-likelihood de –22335.94, valor que corresponde a una probabilidad muy baja en escala estándar, pero completamente interpretable en escala logarítmica. La transformación logarítmica no modifica el significado probabilístico del resultado, sino que reescala la verosimilitud para permitir su análisis y comparación[37].

Desde una perspectiva inferencial, log-likelihood menos negativos indican una mayor compatibilidad entre la dinámica propuesta por el HMM y las observaciones. Así, el valor obtenido evidencia que el modelo asigna una verosimilitud coherente y no degenerada a la secuencia analizada, cumpliendo con la función central del algoritmo Forward y proporcionando un criterio sólido para la comparación con configuraciones alternativas del modelo. la distribución de ocupación esperada de cada estado en el tiempo[39].

Complementariamente, el análisis de la distribución de ocupación esperada en el tiempo de los estados latentes, derivada de los valores $\gamma_t(j)$, aporta una caracterización más fina de la estructura interna del modelo. La distribución obtenida:

$$\bar{\gamma} = [0.233,\ 0.435,\ 0.332],$$

Donde $\bar{\gamma}$: representa la proporción de tiempo en que el proceso se ubica en cada estado a lo largo de la secuencia. Los resultados muestran que el estado 1 presenta la mayor ocupación (43.5%), seguido de los estados 2 (33.2%) y 0 (23.3%). Este patrón indica que el modelo utiliza de manera efectiva los tres estados disponibles y evita tanto la concentración excesiva en un único estado (propia de modelos degenerados) como la dispersión uniforme, característica de modelos poco informativos.

La mayor ocupación del estado 1 sugiere la existencia de un estado latente con mayor estabilidad temporal, mientras que las proporciones asignadas a los estados 0 y 2 evidencian una participación significativa en la dinámica secuencial. En conjunto, esta distribución posterior complementa la información provista por el log-likelihood y confirma que el modelo captura de manera coherente y diferenciada la estructura latente subyacente.

El algoritmo Viterbi se empleó para obtener la secuencia más probable de estados ocultos que dio origen a la serie de observaciones. Viterbi estima una trayectoria determinista que maximiza la verosimilitud conjunta de toda la secuencia [35].

$$Q^* = [2, 0, 1, 1, 2, 2, 2, 0, 1, 1, 2, 0, 1, 2, 2, 2, 2, 0, 1, 2, 0, 1, 2, 2, 0, 1, 2, 0, 1, 2]$$

Los resultados mostraron una ruta caracterizada por una alta presencia del Estado 2, asociado a niveles elevados de ansiedad, seguida de transiciones frecuentes hacia el Estado 1 (riesgo moderado) y, en menor medida, episodios breves del Estado 0 (baja ansiedad).

Este patrón es consistente con la estructura del modelo, en la que las probabilidades de transición favorecen la progresión $0 \rightarrow 1 \rightarrow 2$ y limitan la permanencia prolongada en el estado de bienestar. En conjunto, la trayectoria estimada refleja una dinámica latente dominada por episodios de malestar alto, intercalados con periodos de transición moderada, lo cual coincide con los patrones inferidos a partir de la distribución estacionaria y de las probabilidades de emisión del HMM.

El algoritmo de Baum-Welch ayuda a comprender la dinámica interna del fenómeno de la ansiedad y su relación con los patrones observables, se estimó un Modelo Oculto de Márkov cuyos parámetros fueron optimizados mediante el algoritmo Baum–Welch. A partir de este procedimiento iterativo, el modelo ajustó de manera conjunta las probabilidades de transición entre estados latentes y las probabilidades de emisión hacia los clusters identificados en las variables observables.

En cuanto a la dinámica temporal, la matriz de transición reveló un comportamiento marcadamente no estacionario. El Estado 0 presenta una baja permanencia y una fuerte tendencia a transitar hacia el Estado 1, lo que sugiere que las condiciones de bienestar psicológico tienden a ser transitorias. A su vez, el Estado 1 opera como un

punto de inestabilidad, con una probabilidad elevada de desplazarse hacia el Estado 2, lo cual refleja la progresión natural hacia periodos de mayor malestar. Finalmente, el Estado 2 alterna entre mantenerse y retornar al Estado 0, evidenciando que los episodios de ansiedad alta pueden remitir, aunque estos retornos no garantizan estabilidad prolongada.

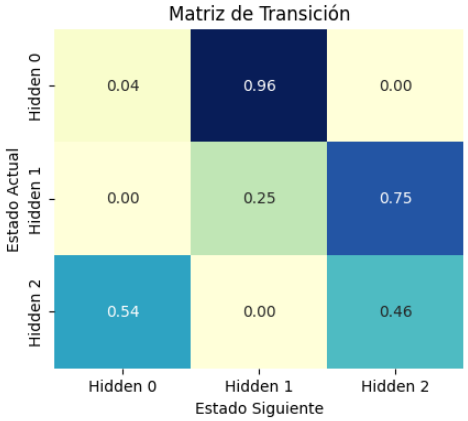


**Fig. 22.** Matriz de transición optimizada con Baum-Welch

La matriz de emisión evidenció un patrón claramente diferenciado entre los tres estados ocultos. El Estado 0, asociado a baja ansiedad, mostró una mayor probabilidad de generar observaciones correspondientes a los clusters de bienestar relativo y riesgo intermedio, reflejando un perfil de estrés reducido y menor demanda de apoyo terapéutico. Por su parte, el Estado 1, interpretado como riesgo moderado, se vinculó predominantemente con el clúster intermedio, lo que sugiere que este estado captura periodos de variabilidad emocional sin llegar a niveles extremos de malestar. Finalmente, el Estado 2, correspondiente a alta ansiedad, emitió de manera determinística hacia el clúster de alto malestar, indicando una correspondencia directa entre este estado latente y los valores extremos de las variables observables.

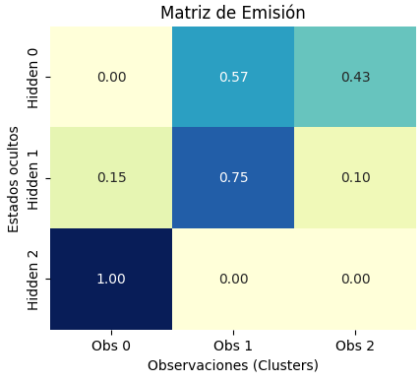


**Fig. 23.** Matriz de emisión optimizada con Baum-Welch

En conjunto, estos resultados indican que el fenómeno de la ansiedad en la población analizada presenta una estructura cíclica, caracterizada por oscilaciones entre estados de bienestar, riesgo y malestar severo. La convergencia del modelo bajo Baum–Welch demuestra que los patrones observables contienen suficiente información para discriminar con claridad los estados latentes y que las transiciones entre ellos siguen una lógica consistente con la literatura sobre fluctuaciones emocionales y procesos de desregulación psicológica.

La distribución estacionaria constituye un elemento central en el análisis dinámico de los Modelos Ocultos de Márkov, pues describe la proporción de tiempo que el sistema tiende a permanecer en cada estado latente cuando evoluciona durante un horizonte temporal suficientemente largo.

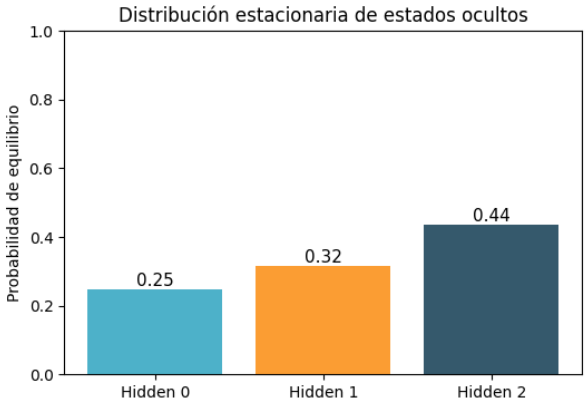


**Fig. 24.** Grafica de la distribución estacionaria de los estados ocultos

En este estudio, la distribución estacionaria obtenida fue $\pi^* = (0.246, 0.317, 0.436)$ para los estados de baja ansiedad, riesgo moderado y alta ansiedad, respectivamente. Este resultado refleja una estructura de equilibrio en la cual el proceso emocional tiende a concentrarse predominantemente en los estados de mayor malestar psicológico, mientras que la permanencia esperada en el estado de bienestar es relativamente baja. La importancia de esta distribución radica en que sintetiza la dinámica global del sistema, independientemente del estado inicial o de la secuencia específica observada, proporcionando una caracterización estable del comportamiento latente asociado al fenómeno de la ansiedad[38].

Para evaluar la estabilidad temporal del modelo y su velocidad de convergencia hacia el equilibrio, se analizaron las distribuciones de probabilidad proyectadas para los pasos futuros $(t+1, t+2\ y\ t+3)$ a partir de la distribución posterior del último instante observado.

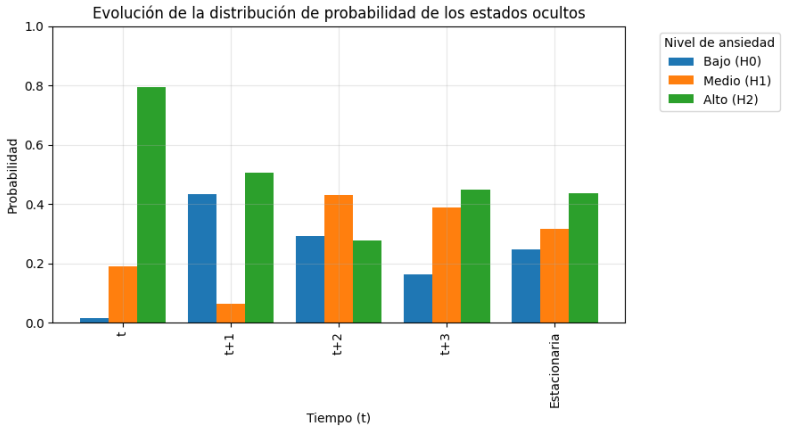


**Fig. 25.** Grafica de la evolución de la distribución de probabilidad de los estados ocultos

Inicialmente, el sistema se encontraba fuertemente concentrado en el estado de alta ansiedad (0.796), pero en los pasos siguientes evolucionó hacia configuraciones más equilibradas: en t+1 la probabilidad se redistribuyó hacia baja ansiedad (0.433) y alta ansiedad (0.505), mientras que en t+2 y t+3 emergió una mezcla dominada por los estados moderado y alto. Estas trayectorias muestran que, aun partiendo de un extremo, la cadena converge rápidamente hacia un patrón estable que se aproxima a la distribución estacionaria.

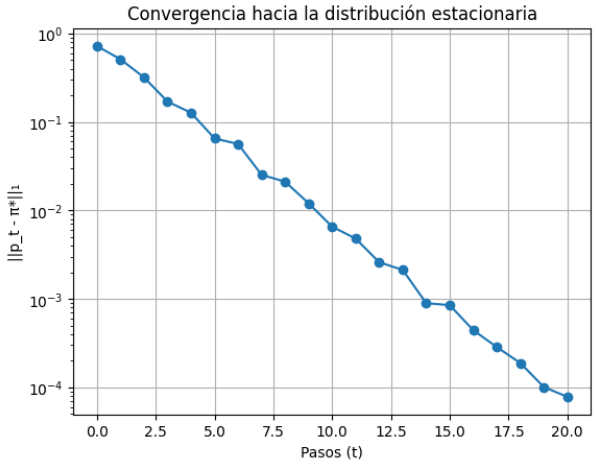


**Fig. 26.** Gráfico de convergencia hacia la distribución estacionaria. Elaboración propia

El análisis de la distancia $\| p_t - \pi^* \|_1$confirmó esta tendencia, evidenciando una caída exponencial en menos de veinte iteraciones. En conjunto, estos resultados indican que la dinámica latente del modelo es altamente estable y presenta un mecanismo de mezcla fuerte, de modo que la estructura de transiciones conduce de manera consistente hacia un equilibrio dominado por estados de ansiedad moderada y alta, con baja persistencia del estado de bienestar.

La distribución estacionaria obtenida ofrece una lectura consistente con la teoría psicológica de los procesos emocionales recurrentes. En términos de dinámica afecti-

va, este equilibrio sugiere que el sistema tiende a estabilizarse en configuraciones donde el malestar psicológico es predominante, mientras que los estados de baja ansiedad, aunque posibles, son relativamente inestables y de menor duración [44]. Esta estructura refleja un patrón de vulnerabilidad crónica o de sensibilización emocional, donde las transiciones hacia estados de mayor activación ansiosa ocurren con mayor probabilidad que los retornos sostenidos hacia el bienestar [18].

Desde una perspectiva clínica, este equilibrio puede interpretarse como un indicador de que la población analizada presenta ciclos prolongados de activación emocional que tienden a reforzarse en el tiempo, coherentes con modelos de rumiación, hipervigilancia o reactividad ante estresores [18, 44]. La escasa permanencia del estado de baja ansiedad sugiere que, aun cuando ocurren episodios de alivio, estos no representan el punto de estabilidad del sistema, sino estados transitorios precedidos o seguidos por niveles más elevados de desregulación. En conjunto, el equilibrio estacionario del HMM permite caracterizar la ansiedad no como un fenómeno episódico, sino como un proceso dinámico donde los estados de mayor malestar son estructuralmente más persistentes, aportando evidencia empírica sobre la naturaleza cíclica del malestar psicológico en esta población [44].

## 4.7 Evaluación

**Transformación del indicador**

Para la etapa de evaluación se implementó una transformación fundamental del indicador principal: trabajar con $1 - nivel\ de\ ansiedad$ en lugar del nivel de ansiedad original. Esta decisión metodológica responde a razones estadísticas, topológicas y de interpretación del riesgo.

En primer lugar, la transformación invierte la escala de medición y permite interpretar valores altos como mayor bienestar psicológico, lo cual resulta más intuitivo cuando se analizan variables continuas acotadas en $[0,1]$. Además, este ajuste genera una distribución con cola derecha más pronunciada, característica que facilita el uso de distribuciones teóricas positivas y asimétricas —como la log-normal— para aproximar el comportamiento de la variable. Al trabajar con $1 - nivel\ de\ ansiedad$, la distribución resultante presenta coeficientes de asimetría positivos y una topología más coherente con modelos que asumen positividad, monotonicidad y sesgo a la derecha.

Desde la perspectiva del riesgo, esta transformación es especialmente relevante. Los valores críticos ya no se encuentran en la cola derecha, sino en la cola inferior de $1 - nivel\ de\ ansiedad$, que corresponde a episodios de bienestar extremadamente bajo, es decir, a los momentos de mayor ansiedad. De esta manera, el análisis de riesgo extremo se expresa naturalmente en términos de cuantiles bajos (p1, p5) o métricas de tipo Valor en Riesgo (VaR)[41], donde un valor cuantitativo pequeño refleja un escenario psicológico severo. Esto enriquece el análisis aplicado, pues permite utilizar herramientas de la matemática del riesgo para describir probabilísticamente los estados psicológicos más críticos.

Finalmente, trabajar con esta variable transformada también contribuye a mejorar la compatibilidad entre la distribución marginal observada y el ajuste continuo posterior.

Esto permite comparar con mayor precisión la distribución empírica versus la log-normal ajustada y evaluar su proximidad mediante métricas como la distancia de Wasserstein. Así, la transformación constituye un paso metodológico clave antes de analizar la distribución teórica, la bondad de ajuste y la estimación de riesgos extremos.

**Construcción de la distribución discreta empírica**

Como punto de partida para la comparación entre la distribución empírica y la distribución teórica ajustada, se construyó una distribución discreta basada en la dinámica de largo plazo del modelo oculto. Para ello se utilizó la distribución estacionaria obtenida del HMM, la cual describe la proporción de tiempo que el proceso latente permanece en cada uno de los estados de ansiedad en equilibrio[39]. Con base en estas proporciones, se generó un conjunto de datos de tamaño $n = 1000$, asignando a cada estado un número de observaciones proporcional a su peso estacionario. Este procedimiento permitió extraer muestras aleatorias de los registros originales pertenecientes a cada cluster, preservando tanto la variabilidad real del indicador como la estructura de equilibrio del modelo.

A partir de esta muestra balanceada se extrajeron los valores del indicador transformado $1 - ansiedad$, los cuales fueron depurados para garantizar su validez numérica dentro del soporte teórico $(0,1]$. Sobre estos valores se estimó la función de masa de probabilidad (PMF), definida en cinco niveles discretos del indicador. Esta representación empírica resume el comportamiento marginal del bienestar psicológico y constituye la referencia inicial para evaluar el ajuste de un modelo continuo.

De manera paralela, se ajustó una distribución log-normal mediante máxima verosimilitud, obteniendo su función de densidad (PDF) y una muestra simulada con el mismo tamaño que la distribución empírica. La distribución log-normal ajustada permite contrastar la PMF empírica con un modelo teórico y verificar si este reproduce correctamente la forma del indicador. Esto es especialmente importante en los valores bajos de $1 - ansiedad$, donde se concentran los escenarios críticos de ansiedad elevada.

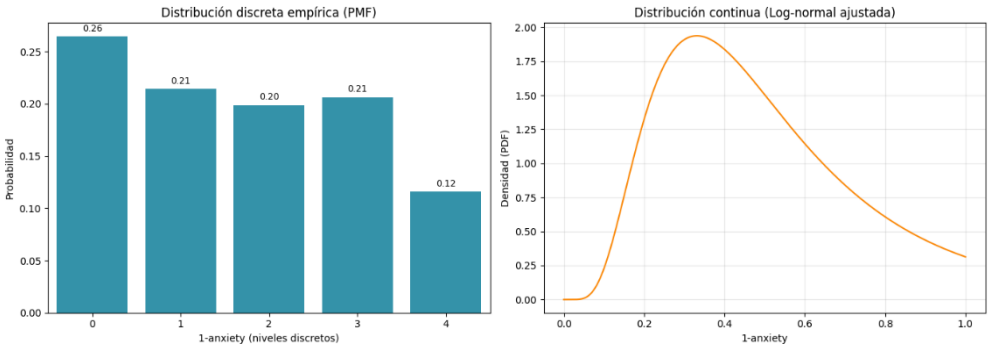


**Fig. 27.** Comparación entre la distribución empírica discreta y la distribución log-normal ajustada

| Parámetros de la distribución discreta empírica | |
|---|---|
| **Niveles** | [0.2, 0.4, 0.6, 0.8, 1.0] |
| **PMF** | [0.264, 0.214, 0.199, 0.207, 0.116] |
| **Media** | 0.539 |
| **Varianza** | 0.074 |
| **Entropía** | 1.578 |
| $\boldsymbol{q_{0.05}}$ / $\boldsymbol{q_{0.5}}$/$\boldsymbol{q_{0.95}}$ | 0.2 / 0.6 / 1.0 |

| Parámetros de la distribución Log-normal simulada | |
|---|---|
| $\boldsymbol{\sigma}$ | 0.579 |
| $\boldsymbol{\mu}$ | - 0.769 |
| **Media** | 0.548 |
| **Varianza** | 0.119 |
| **Mediana** | 0.463 |
| $\boldsymbol{q_{0.05}}$ / $\boldsymbol{q_{0.5}}$/$\boldsymbol{q_{0.95}}$ | 0.178 / 0.463 / 1.0 |

**Fig. 28.** Parámetros de la distribución discreta empírica y distribución Log-normal ajustada

La distribución discreta empírica derivada del indicador transformado $1 - ansiedad$ mostró una estructura definida en cinco niveles ordinales, con probabilidades que oscilan entre 11.6% y 26.4%. A partir de esta PMF se obtuvo un valor medio de $0.539$, una varianza de $0.074$ y una entropía de $1.578$, lo que indica una distribución relativamente equilibrada entre los niveles y con una dispersión moderada. Los cuantiles relevantes fueron $q_{0.05} = 0.20$, $q_{0.50} = 0.60$ y $q_{0.95} = 1.0$, evidenciando que el 5% de los casos más críticos se concentra en el nivel más bajo de bienestar.

En contraste, la distribución log-normal ajustada presentó parámetros $\sigma = 0.579$, $\mu = -0.769$, consistentes con una forma asimétrica hacia la derecha. Sus momentos mostraron una media de $0.548$, muy cercana a la empírica, y una varianza mayor $(0.119)$, reflejando una dispersión más acentuada. Los cuantiles teóricos $q_{0.05}$ y $q_{0.95}$ se aproximaron a los valores empíricos una vez truncados al intervalo $[0,1]$, lo que confirma que el modelo continuo captura adecuadamente los extremos relevantes del indicador.

En conjunto, estos resultados muestran que la log-normal constituye una aproximación razonable a la estructura empírica del bienestar psicológico, preservando tanto los momentos globales como el comportamiento en las colas, lo cual es esencial para el análisis de riesgo asociado a los niveles bajos de $1 - ansiedad$.

Con el propósito de evaluar rigurosamente la similitud entre la distribución discreta empírica del indicador $1 - ansiedad$ y la distribución log-normal ajustada, se empleó la distancia de Wasserstein de primer orden. Esta métrica cuantifica el costo mínimo necesario para transformar una distribución en otra y permite comparar representaciones con soportes distintos, por lo que resulta especialmente adecuada en contextos donde coexisten distribuciones discretas y continuas.

Los resultados mostraron un valor de $W_p(\mu, \nu) = 0.06045$ lo cual indica una discrepancia muy baja entre ambas distribuciones.

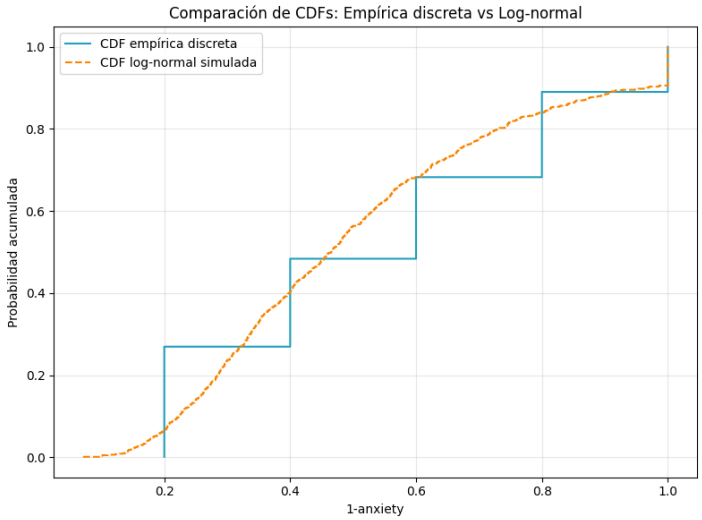


**Fig. 29.** Grafica comparativa de las funciones de distribución acumulada de la distribución empírica y la log normal ajustada

En términos prácticos, este valor sugiere que la log-normal reproduce de manera consistente la estructura global del indicador, incluyendo sus niveles bajos y la forma general de acumulación de probabilidad. La coincidencia observada tanto en las curvas de distribución acumulada (CDF) como en las densidades suavizadas refuerza esta conclusión, mostrando que ambas distribuciones presentan un solapamiento elevado y una forma muy similar.

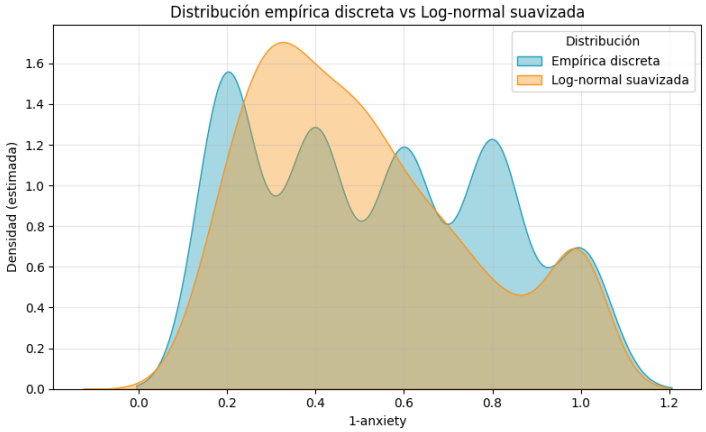


**Fig. 30.** Esta figura presenta la superposición entre la densidad suavizada de la distribución empírica discreta y la densidad de la log-normal ajustada para el indicador $1 -$ ansiedad. Elaboración propia

En conjunto, la distancia de Wasserstein confirma que la distribución log-normal constituye una aproximación paramétrica adecuada para el comportamiento marginal de $1 - ansiedad$, preservando tanto sus momentos principales como el perfil de probabilidad en las colas. Esto habilita su utilización en la estimación de métricas de

riesgo extremo, como percentiles bajos y el Valor en Riesgo (VaR), dentro del análisis de bienestar psicológico.

La estimación del Valor en Riesgo (VaR) aplicado al indicador transformado $X = 1 - ansiedad$ permite identificar de manera rigurosa los episodios correspondientes a los niveles más críticos de bienestar psicológico. Dado que valores bajos de $X$ representan estados de mayor ansiedad, el VaR calculado sobre la cola izquierda de la distribución adquiere un significado directo como umbral de riesgo. El $VaR_{0.05}$ obtenido a partir de la distribución log-normal ajustada y truncada en el intervalo $[0,1]$ fue de 0.179, lo cual indica que el 5 % de las observaciones con mayor ansiedad se sitúan por debajo de este valor de bienestar.

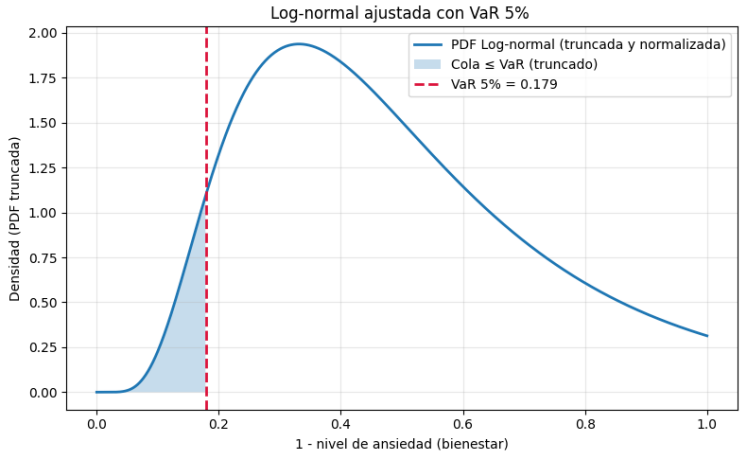


**Fig. 31.** Distribución Log-normal del indicador 1-ansiedad y umbral del VaR del 5%

Este resultado ofrece un punto de referencia cuantitativo para caracterizar los episodios de afectación psicológica más severa dentro del proceso modelado. En términos prácticos, el VaR delimitó un umbral crítico a partir del cual es posible identificar los estados latentes asociados a mayor vulnerabilidad emocional, permitiendo una lectura más fina de la estructura de riesgo extremo. Al trabajar con una distribución truncada coherente con la escala del indicador, el VaR calculado mantiene su interpretación probabilística estándar y se integra de manera consistente en el marco probabilístico propuesto, aportando evidencia sólida para la comprensión de los puntos más extremos del comportamiento del fenómeno.

El análisis integrado de los resultados permite articular la estructura observable del fenómeno, la dinámica latente inferida mediante el HMM y la caracterización probabilística de los niveles de ansiedad. En conjunto, estas evidencias conforman una lectura coherente de la ansiedad como proceso dinámico, heterogéneo y no lineal,

cuyos patrones no pueden capturarse mediante aproximaciones estáticas o exclusivamente descriptivas[6].

La exploración de las variables psicológicas y conductuales reveló una organización interna marcada por heterogeneidad moderada, en la cual coexisten perfiles diferenciados por nivel de estrés, sueño, actividad física, consumo de estimulantes y uso de servicios terapéuticos. La segmentación mediante K-Means permitió sintetizar estos patrones en tres grupos robustos que conforman un gradiente de riesgo psicológicamente interpretable[4, 30]. Estos clústeres —bajo malestar, riesgo moderado y bienestar relativo— capturan configuraciones completas del comportamiento, no solo niveles aislados de cada variable, lo cual refuerza su validez como indicadores compuestos de regulación emocional[6].

La modelación temporal mediante HMM permitió caracterizar la evolución interna del nivel de ansiedad más allá de lo que sugieren las observaciones directas. Los resultados del algoritmo Baum–Welch[36] y las distribuciones estimadas evidencian una dinámica dominada por transiciones graduales, donde el estado moderado funciona como punto central del proceso[35]. La distribución estacionaria confirma esta tendencia[39]: aun cuando los estados de bienestar son posibles, el equilibrio del sistema se inclina hacia configuraciones de malestar moderado o alto.

La ruta inferida por Viterbi[35] muestra que estos estados no se presentan de manera aislada, sino como secuencias que alternan fases de mejoría transitoria con episodios recurrentes de desregulación emocional. El modelo evita concentrarse en un solo estado, lo que indica que la variabilidad interna del fenómeno está bien representada por la estructura de tres niveles adoptada. Los resultados revelan que la ansiedad funciona como un proceso emocional con memoria: una vez que alcanza niveles moderados o altos, tiende a mantenerse allí durante más tiempo, mientras que los retornos al bienestar suelen ser breves y poco estables, patrón compatible con la evidencia reciente de modelos computacionales que describen trayectorias persistentes de vulnerabilidad en psicopatología[6, 13].

La aproximación probabilística del indicador transformado $1 - ansiedad$ muestra que su distribución empírica presenta asimetría positiva y mayor concentración en niveles intermedios, coherente con fenómenos psicológicos que resultan del efecto acumulativo de múltiples factores[45, 46]. El ajuste log-normal reproduce satisfactoriamente esta estructura, como lo evidencia la proximidad de los momentos estimados y la baja distancia de Wasserstein entre ambas distribuciones, en línea con propuestas que utilizan esta métrica como criterio de bondad de ajuste entre distribuciones empíricas y modelos paramétricos[47]. Este resultado sugiere que el fenómeno puede modelarse con un comportamiento multiplicativo, donde el bienestar —y su contraparte, la ansiedad— emerge de la interacción proporcional de determinantes conductuales y emocionales[9].

La estimación del VaR permitió identificar un umbral cuantitativo para los episodios más críticos[41, 43]. El valor del 5% se ubicó en torno a 0.18 en la escala del indicador transformado, lo que implica que los estados de mayor ansiedad corresponden a una fracción pequeña pero significativa de la distribución. Este umbral constituye una medida precisa para delimitar escenarios de alto riesgo psicológico, integrando

en un único parámetro la información derivada del modelo continuo y la distribución empírica del indicador[41, 43].

La concordancia entre los patrones identificados mediante técnicas no supervisadas, la inferencia temporal del HMM y el ajuste paramétrico de la distribución log-normal refuerza la coherencia interna del modelo y valida la solidez de la aproximación adoptada, en línea con propuestas recientes que abogan por combinar análisis estructurales, dinámicos y probabilísticos en el estudio de la psicopatología[5, 6, 15, 16].

En conjunto, estos resultados constituyen una base técnica sólida para la consolidación del comportamiento estadístico del fenómeno y para el desarrollo futuro de modelos orientados al monitoreo y gestión del riesgo psicológico, en consonancia con el creciente interés por herramientas cuantitativas de apoyo a la toma de decisiones en salud mental[8, 13].

# 5 Conclusiones y trabajo futuro

- Los resultados indican que el problema de investigación fue abordado de manera adecuada mediante la combinación de técnicas no supervisadas, modelos estocásticos y análisis probabilístico. La integración entre la segmentación conductual con K-Means, la inferencia temporal mediante Modelos Ocultos de Márkov (HMM) y el ajuste de distribuciones no normales permitió representar de forma consistente el comportamiento dinámico de la ansiedad. En particular, los clústeres obtenidos configuraron un gradiente de riesgo psicológicamente interpretable; el algoritmo Baum–Welch convergió hacia una matriz de transición y una distribución estacionaria no degeneradas, coherentes con la presencia de estados latentes diferenciados; y la baja distancia de Wasserstein entre la distribución empírica del indicador transformado y el modelo log-normal, junto con la estimación del VaR en la cola de mayor riesgo, respaldan la validez del enfoque probabilístico adoptado. En conjunto, estas evidencias sugieren que la metodología propuesta cumple razonablemente el objetivo de caracterizar la dinámica de la ansiedad como un proceso estocástico, aportando una representación coherente, reproducible y cuantitativamente sólida del fenómeno dentro de los alcances del estudio.
- Desde la perspectiva de una aseguradora, el modelo propuesto aporta una capa de información que no ofrecen los enfoques tradicionales basados solo en variables demográficas o historiales de siniestros. Al representar la ansiedad como un proceso dinámico —mediante clústeres conductuales, HMM y métricas de riesgo— el modelo permite estratificar mejor el riesgo, distinguir perfiles con mayor probabilidad de permanecer en estados de malestar moderado o alto y priorizar intervenciones preventivas sobre los poblaciones más vulnerables. Esto se traduce en una herramienta que complementa la tarificación y la gestión técnica del portafolio, ayudando a reducir ausentismo, invalidez y costos asociados a la salud mental, al tiempo que brinda una base cuantitativa para diseñar programas de bienestar más focalizados y eficientes.

- El estudio presenta limitaciones asociadas a la disponibilidad de datos longitudinales reales y a la generalización del modelo más allá del conjunto analizado. Investigaciones futuras deberían incorporar trayectorias más extensas y validación en diferentes poblaciones, con el fin de fortalecer la robustez del modelo y ampliar su aplicabilidad en sistemas de monitoreo del riesgo psicológico.
- Los resultados obtenidos muestran que la metodología implementada puede servir como base para el desarrollo de sistemas automatizados de monitoreo emocional. La capacidad del modelo para identificar transiciones, estados críticos y umbrales de riesgo sugiere su utilidad como componente analítico en plataformas de bienestar laboral, programas de prevención o sistemas de alerta temprana basados en datos conductuales y emocionales.
- En conjunto, el estudio ofrece una caracterización sólida y replicable de la ansiedad como un proceso dinámico y estocástico. La integración de clustering, HMM y análisis probabilístico proporciona una base técnica consistente para investigaciones futuras y aplicaciones orientadas al monitoreo y gestión del riesgo emocional en entornos laborales.

## 6 Declaración de disponibilidad de datos y código

Los conjuntos de datos y el código de análisis que respaldan los resultados de este estudio están disponibles públicamente en el repositorio de GitHub: https://github.com/D2Giraldot/Anxiety_HMM_TdG. El repositorio incluye los scripts de preprocesamiento, las rutinas de modelación y material complementario.

## 7 Bibliografía

1. Chakrabarti Satyajit, Saha HNath (2017) IEEE CCWC-2017 : 2017 IEEE 7th Annual Computing and Communication Workshop and Conference : 09-11 January, 2017, Las Vegas, USA. IEEE
2. van Hoffen MFA, Norder G, Twisk JWR, Roelen CAM (2020) Development of Prediction Models for Sickness Absence Due to Mental Disorders in the General Working Population. J Occup Rehabil 30:308–317. https://doi.org/10.1007/s10926-019-09852-3
3. Strudwick J, Gayed A, Deady M, et al (2023) Workplace mental health screening: A systematic review and meta-analysis. Occup Environ Med 80:469–484
4. Nayak S, Nagesh B, Routray A, et al (2021) Estimation of Depression Anxieties and Stress through Clustering of Sequences of Visual and Thermal Face Images. In: Proceedings of the 2021 IEEE 18th India Council International Conference, INDICON 2021. Institute of Electrical and Electronics Engineers Inc.
5. Jaimes LG, Idalides JV-Laurens, Mustafa IAkbas, Kanwalinderjit Gagneja (2017) Future Stress, Forecasting Physiological Signals. IEEE

6. Zavlis O, Story G, Friedrich C, et al (2025) A systematic review of computational modeling of interpersonal dynamics in psychopathology. Nature Mental Health 3:932–942. https://doi.org/10.1038/s44220-025-00465-9
7. Domínguez Domínguez JA, Expósito Duque V, Torres Tejera E (2024) Epidemiology of anxiety and its context in Primary Care. Atencion Primaria Practica 6:. https://doi.org/10.1016/j.appr.2024.100194
8. World Health Organization (2022) Mental health at work. https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work. Accessed 31 Oct 2025
9. Expósito-Duque V, Torres-Tejera ME, Domínguez Domínguez JA (2024) Social determinants of anxiety in the 21st century. Atencion Primaria Practica 6:. https://doi.org/10.1016/j.appr.2024.100192
10. Madhurima Paul, Swapan Das (2023) Mental health in tech workplace: An analysis. International Journal of Science and Research Archive 10:221–233. https://doi.org/10.30574/ijsra.2023.10.1.0743
11. World Health Organization (2022) WHO guidelines on mental health at work
12. Saito T, Suzuki H, Kishi A (2022) Predictive Modeling of Mental Illness Onset Using Wearable Devices and Medical Examination Data: Machine Learning Approach. Front Digit Health 4:. https://doi.org/10.3389/fdgth.2022.861808
13. Lewin G, Abakasanga E, Titcombe I, et al (2025) Artificial intelligence-enabled predictive modelling in psychiatry: overview of machine learning applications in mental health research. BJPsych Adv 1–7. https://doi.org/DOI: 10.1192/bja.2025.10133
14. Shen S, Qi W, Zeng J, et al (2025) Passive Sensing for Mental Health Monitoring Using Machine Learning With Wearables and Smartphones: Scoping Review (Preprint)
15. Göçgün Y (2024) A PROBABILISTIC ANALYSIS FOR MENTAL HEALTH PROBLEMS: EVALUATION OF POLICIES FOR MANAGING DISTRACTED INDIVIDUALS. Mühendislik Bilimleri ve Tasarım Dergisi 12:643–652. https://doi.org/10.21923/jesd.1418675
16. De Oliveira C, Matias MA, Jacobs R (2024) Systematic Literature Review Microsimulation Models on Mental Health: A Critical Review of the Literature. Value in Health 27:226–246. https://doi.org/10.1016/j
17. Prégent J, Chung VHA, El Adib I, et al (2025) Applications of Artificial Intelligence in Psychiatry and Psychology Education: Scoping Review. JMIR Med Educ 11
18. Barlow DH (2003) Anxiety and Its Disorders: The Nature and Treatment of Anxiety and Panic. Cogn Behav Pract 10:188–189. https://doi.org/10.1016/S1077-7229(03)80027-5
19. Schröer C, Kruse F, Gómez JM (2021) A systematic literature review on applying CRISP-DM process model. In: Procedia Computer Science. Elsevier B.V., pp 526–534

20. Friston KJ, Parr T, de Vries B (2017) The graphical brain: Belief propagation and active inference. Network Neuroscience 1:381–414. https://doi.org/10.1162/netn_a_00018
21. Montague PR, Dolan RJ, Friston KJ, Dayan P (2012) Computational psychiatry. Trends Cogn Sci 16:72–80
22. Zhang C (2023) Social Anxiety Dataset. In: https://www.kaggle.com/datasets/natezhang123/social-anxiety-dataset
23. Turkey J (1977) Exploratory Data Analysis
24. Benoit JS, Chan W, Luo S, et al (2016) A hidden Markov model approach to analyze longitudinal ternary outcomes when some observed states are possibly misclassified. Stat Med 35:1549–1557. https://doi.org/10.1002/sim.6861
25. Zhou J, Song X, Sun L (2020) Continuous time hidden Markov model for longitudinal data. J Multivar Anal 179:. https://doi.org/10.1016/j.jmva.2020.104646
26. Bretó C, Ionides EL, King AA (2020) Panel Data Analysis via Mechanistic Models. J Am Stat Assoc 115:1178–1188. https://doi.org/10.1080/01621459.2019.1604367
27. Collins S, Hoare E, Allender S, et al (2023) A longitudinal study of lifestyle behaviours in emerging adulthood and risk for symptoms of depression, anxiety, and stress. J Affect Disord 327:244–253. https://doi.org/10.1016/j.jad.2023.02.010
28. Tucker A, Li Y, Garway-Heath D (2017) Updating Markov models to integrate cross-sectional and longitudinal studies. Artif Intell Med 77:23–30. https://doi.org/10.1016/j.artmed.2017.03.005
29. Fluharty M, Taylor AE, Grabski M, Munafò MR (2017) The association of cigarette smoking with depression and anxiety: A systematic review. Nicotine and Tobacco Research 19:3–13
30. Oti EU, Olusola MO, Eze FC, Enogwe SU (2021) Comprehensive Review of K-Means Clustering Algorithms. International Journal of Advances in Scientific Research and Engineering 07:64–69. https://doi.org/10.31695/ijasre.2021.34050
31. Morissette L, Chartier S (2013) The k-means clustering technique: General considerations and implementation in Mathematica. Tutor Quant Methods Psychol 9:15–24. https://doi.org/10.20982/tqmp.09.1.p015
32. Cover TM, Hart PE (1952) Nearest Neighbot Pattern Clasification
33. Jollife IT, Cadima J (2016) Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374
34. Jurafsky D, James H M (2024) Speech and Language Processing A.1 Markov Chains
35. Rabiner L (1989) A tutorial on Hidden Markov Models and selected aplications in speech recognition. IEEE
36. Yang F, Balakrishnan S, Wainwright MJ (2017) Statistical and Computational Guarantees for the Baum-Welch Algorithm
37. Bishop C (2006) Pattern Recognition and Machine Learning

38. Levin DA, Peres Y, Wilmer EL (2009) Markov Chains and Mixing Times, second edition
39. Kenemy JG, Snell L (1976) Finite Markov Chains
40. Panaretos VM, Zemel Y (2019) Statistical Aspects of Wasserstein Distances. Annual Review ofStatistics and Its Application. https://doi.org/10.1146/annurev-statistics
41. Johnson CA (2001) Value at risk: Teoría y aplicaciones
42. Duy VN Le, Takeuchi I (2023) Exact statistical inference for the Wasserstein distance by selective inference: Selective Inference for the Wasserstein Distance. Ann Inst Stat Math 75:127–157. https://doi.org/10.1007/s10463-022-00837-3
43. Wainwright JT (2005) Quantitative Risk Management
44. Nolen-Hoeksema S (2000) The Role of Rumination in Depressive Disorders and Mixed Anxiety/Depressive Symptoms
45. Moral De La Rubia J International Journal of Psychology and Counselling Lognormal distribution for social researchers: A probability classic. 16:10–25. https://doi.org/10.5897/IJPC2024.0702
46. Bono R, Blanca MJ, Arnau J, Gómez-Benito J (2017) Non-normal distributions commonly used in health, education, and social sciences: A systematic review. Front Psychol
47. Comas N, Oro O, Catalá B (2021) Prueba de bondad de ajuste para la distribución de distancias en secuencias de datos categóricos