

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO - CHILE



“CREACIÓN, ANÁLISIS Y OPTIMIZACIÓN DE UN CUESTIONARIO DE SALUD MENTAL MEDIANTE INTELIGENCIA ARTIFICIAL”

YASMINE VICTORIA PÉREZ CERDA

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERA CIVIL EN INFORMÁTICA

Profesor Guía: Nicolás Rojas Morales
Correferente: Nicolás González Rogel

Diciembre - 2025



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título ☐ Tesis de Postgrado

Título del trabajo: Creación, análisis y optimización de un cuestionario de salud mental mediante Inteligencia Artificial

Nombre del candidato(a): Yasmine Victoria Pérez Cerda

Carrera / Grado: Ingeniería Civil Informática

Campus: San Joaquín

Departamento: Departamento de Informática

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Nicolás Rojas Morales, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente

DEJO CONSTANCIA que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución.

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL (marcar una opción)

☒ El trabajo **NO contiene** información que amerite confidencialidad y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (**embargo**) por (**marcar una opción**):

☐ 6 meses ☐ 12 meses ☐ 2 años ☐ 3 años ☐ 5 años ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 14 de octubre del 2026

Firma:

Estudiante o Candidato(a):

Fecha: 14 de octubre del 2026

Firma:

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

DEDICATORIA

Dedicado a mi mamá y a las gemes, que nunca me dejaron caer y siempre confiaron en mí.

AGRADECIMIENTOS

Quiero comenzar agradeciendo a mi mamá, mi hermano, mi cuñada y mi sobrina por acompañarme incondicionalmente desde el primer día, por confiar siempre en mí y por sentirse orgullosos de la persona en la que me he convertido. A mis abuelos, quienes han sido mis pilares a lo largo de toda mi vida y me han transmitido los valores que hoy me definen. Sin ustedes, nada de esto habría sido posible, y agradezco profundamente a la vida poder compartir este momento con quienes han sido mi mayor fortaleza.

Para mis amigas de toda la vida, Fefa, Dani y Andrea, que desde el colegio han elegido cada día quedarse a mi lado y acompañarme en cada etapa de mi vida. Gracias por mostrarme lo bonito de la amistad, por celebrar cada uno de mis logros y por escucharme en mis momentos más vulnerables. Gracias por levantarme cuando sentía que ya no podía más y por recordarme quién soy cuando yo misma lo olvidaba. Son las mejores compañeras de vida, y espero que la sigamos compartiendo por muchos años más.

A mis amigos químicos, en especial a las gemes, Vero, Catita, Javi, Feña, Vivi y Diego Rojas, que se encargaron de que nunca me sintiera sola en este proceso. Sus risas, compañía y preocupación sincera hicieron más llevaderos incluso los momentos más estresantes, y siempre lograron recordarme que era capaz de mucho más de lo que creía. Gracias por hacerme sentir una química más, por acompañarme en cada etapa y por ser parte tan importante de mi vida universitaria y sé que lo seguirán siendo en cada paso que dé en el futuro. No duden que siempre contarán con su amiga informática para todos sus problemas que no tienen nada que ver con Informática, pero que para ustedes soy la más seca.

Por otra parte, quiero agradecer a mis amigos de la carrera, quienes compartieron conmigo las alegrías y tristezas en estos últimos años, estudiando juntos para salvar cada ramo, y que formamos una amistad que ninguno se esperaba.

Finalmente, este trabajo tampoco hubiera sido posible sin el apoyo de mi profesor guía, Nicolás Rojas, y de mi correferente, Nicolás González, quienes desde el primer día confiaron en mí, en mis capacidades y en la visión que tenía para esta memoria. Su orientación, paciencia y disposición constante para resolver dudas y ofrecer consejos fueron esenciales para que este proyecto pudiera realizarse. Más allá de lo académico, agradezco profundamente su calidad humana, el interés genuino por mi bienestar y el compromiso con mi formación profesional. Haber trabajado con ambos en esta etapa final de mi carrera es algo que valoro enormemente y que recordaré con mucho cariño.

RESUMEN

Resumen—La salud mental en estudiantes universitarios es una problemática creciente, especialmente en relación con trastornos como el burnout, la ansiedad y la depresión, que afectan tanto el bienestar emocional como el rendimiento académico. Actualmente, la evaluación de la salud mental en estudiantes se realiza mediante cuestionarios, entendidos como formularios que integran uno o más instrumentos psicométricos validados. Entre los más utilizados se encuentran el PHQ-9, GAD-7, WHOQOL-BREF y la EUBE, los cuales permiten diagnosticar síntomas de depresión, ansiedad y burnout. No obstante, la aplicación conjunta de estos instrumentos genera un cuestionario extenso y complejo, lo que disminuye la cantidad de los datos y la participación estudiantil. Es por ello, que este trabajo tiene como objetivo analizar dichos instrumentos y las respuestas recopiladas de un estudio aplicado en la Universidad Técnica Federico Santa María, con el fin de explorar la creación de un cuestionario unificado y optimizado mediante técnicas de inteligencia artificial. A través del análisis de los ítems de los instrumentos, la identificación de redundancias y correlaciones, y el uso de algoritmos de IA, se busca diseñar una versión más breve y eficiente, capaz de apoyar la detección temprana y facilitar la intervención en estudiantes de riesgo de desarrollar trastornos emocionales.

Palabras Clave—Salud Mental; Inteligencia Artificial; Reducción de dimensionalidad; Selección de características; Análisis semántico.

ABSTRACT

Abstract—Mental health among university students is a growing concern, particularly regarding disorders such as burnout, anxiety, and depression, which negatively affect both emotional well-being and academic performance. Currently, the assessment of student mental health is carried out through questionnaires, understood as forms that integrate one or more validated psychometric instruments. Among the most widely used are PHQ-9, GAD-7, WHOQOL-BREF, and EUBE, which enable the diagnosis of symptoms of depression, anxiety, and burnout. However, the joint application of these instruments results in a lengthy and complex questionnaire, which decreases data quantity and student participation. Therefore, this work aims to analyze these instruments and the responses collected from a study conducted at Universidad Técnica Federico Santa María, with the purpose of exploring the development of a unified and optimized questionnaire through artificial intelligence techniques. By analyzing instrument items, identifying redundancies and correlations, and applying AI algorithms, the goal is to design a shorter and more efficient version, capable of supporting early detection and facilitating intervention in students at risk of developing emotional disorders.

Keywords—Mental Health; Dimensionality Reduction; Artificial Intelligence; Feature Selection; Semantic Analysis.

GLOSARIO

EUBE: Escala Unidimensional de Burnout Estudiantil.

GAD: Generalized Anxiety Disorder (Trastorno de Ansiedad Generalizada).

FA: Feature Agglomeration.

IA: Inteligencia Artificial.

PHQ: Patient Health Questionnaire (Cuestionario de Salud del Paciente).

SBERT: Sentence-Bidirectional Encoder Representations from Transformers.

UMAP: Uniform Manifold Approximation and Projection.

UTFSM: Universidad Técnica Federico Santa María.

VAE: Variational Autoencoder.

WHOQOL: World Health Organization Quality of Life (Organización Mundial de la Salud Calidad de Vida).

ÍNDICE DE CONTENIDOS

RESUMEN	IV
ABSTRACT	IV
GLOSARIO	V
ÍNDICE DE FIGURAS	IX
ÍNDICE DE TABLAS	IX
INTRODUCCIÓN	1
CAPÍTULO 1: DEFINICIÓN DEL PROBLEMA	3
1.1 Objetivos	6
1.1.1 Objetivo General	6
1.1.2 Objetivos Específicos	6
1.2 Resumen del capítulo	6
CAPÍTULO 2: ESTADO DEL ARTE	7
2.1 Salud Mental en el Contexto Universitario	8
2.2 Trastornos emocionales	8
2.2.1 Depresión	8
2.2.2 Ansiedad	8
2.2.3 Burnout	9
2.3 Instrumentos psicométricos en salud mental	9
2.3.1 Patient Health Questionnaire-9 (PHQ-9)	9
2.3.2 Generalized Anxiety Disorder-7 (GAD-7)	10
2.3.3 World Health Organization Quality of Life-BREF (WHOQOL-BREF)	11
2.3.4 Escala Unidimensional de Burnout Estudiantil (EUBE)	11
2.4 Problemas de clasificación	12
2.4.1 Clasificación supervisada	13
2.4.2 Clasificación no supervisada	14
2.4.3 Clasificación semisupervisada	16
2.5 Algoritmos de clasificación	18
2.5.1 Árboles de decisión	18
2.5.2 Random Forest	19
2.5.3 Support Vector Machine (SVM)	21
2.5.4 Redes Neuronales	21
2.5.4.1 Neuronas	21
2.5.4.2 Funciones de activación	23
2.5.4.3 Entrenamiento de redes neuronales	23
2.5.4.4 Funciones de pérdida	24
2.5.4.5 Hiperparámetros	25

2.6	Métricas	26
2.6.1	Matriz de confusión	26
2.6.2	Métricas de evaluación	27
2.7	Selección de características (<i>Feature Selection</i>)	28
2.7.1	Métodos de Selección de Características	29
2.7.1.1	Métodos de filtro (Filter Methods)	29
2.7.1.2	Métodos wrapper (Wrapper Methods)	30
2.7.1.3	Métodos embebidos (Embedded Methods)	32
2.8	Reducción de dimensionalidad	33
2.8.1	Feature Agglomeration	33
2.8.2	Uniform Manifold Approximation and Projection (UMAP)	34
2.8.3	Autoencoders	35
2.8.4	Variational Autoencoders (VAE)	36
2.9	Análisis semántico de ítems	38
2.9.1	Embeddings de texto	38
2.9.1.1	Word2Vec	39
2.9.1.2	BERT	40
2.9.1.3	Sentence-BERT	42
2.10	Aplicaciones previas de IA en salud mental	43
2.11	Resumen del capítulo	47
CAPÍTULO 3: PROPUESTA DE SOLUCIÓN		48
3.1	Preprocesamiento semántico y generación del embedding	50
3.1.1	Reformulación semántica de ítems	50
3.1.2	Prompt Engineering	51
3.1.3	Representación semántica mediante embeddings	52
3.1.4	Fine-tuning del modelo SBERT	53
3.2	Reducción de dimensionalidad con VAE	54
3.2.1	Variantes del modelo VAE	55
3.3	Agrupamiento semántico de ítems	56
3.3.1	Top-Down Clustering	57
3.3.2	K-Means Iterativo	58
3.3.3	Feature Agglomeration	60
3.3.4	Selección del método final	61
3.3.5	Análisis probabilístico de ítems	61
3.4	Screening Dinámico	63
3.4.1	Normalización de Escalas	63
3.4.2	Selección del Ítem Representativo	64
3.4.3	Definición del Umbral de Activación	66
3.5	Diseño del Cuestionario Dinámico Adaptativo	67
3.5.1	Estructura General del Cuestionario	68
3.5.2	Ventajas del Diseño Adaptativo	68
3.6	Resumen del capítulo	69

CAPÍTULO 4: RESULTADOS	70
4.1 Resultados experimentales	70
4.1.1 Comparación de los modelos VAE	71
4.1.2 Proyección UMAP final	72
4.1.3 Selección del modelo de representación semántica	73
4.1.4 Análisis del espacio latente y dimensionalidad	75
4.1.5 Resultados de los métodos de clustering	76
4.1.5.1 Feature Agglomeration	77
4.1.5.2 Método Iterativo	77
4.1.5.3 Método Top-Down	78
4.1.6 Análisis probabilístico	83
4.1.6.1 Estadísticas por instrumento	83
4.1.6.2 Matrices de co-asociación	84
4.2 Diseño del cuestionario optimizado	85
4.2.1 Análisis por instrumento: promedios y discriminación	86
4.2.2 Análisis global: correlación ítem-puntaje total	90
4.2.3 Análisis respuesta-nivel	93
4.2.4 Selección de ítems puerta	94
4.2.5 Definición de umbrales de activación	100
4.3 Estructura Final del Cuestionario Adaptativo	101
4.4 Resumen del capítulo	105
 CAPÍTULO 5: CONCLUSIONES	 106
 ANEXOS	 110
 REFERENCIAS BIBLIOGRÁFICAS	 117

ÍNDICE DE FIGURAS

1	Representación del problema de clasificación, donde a partir de un conjunto de entrada x se asigna una etiqueta de clase y	12
2	Ejemplo de clasificación supervisada.	14
3	Ejemplo de clasificación no supervisada.	15
4	Ejemplo de clasificación semisupervisada mediante propagación de etiquetas.	17
5	Funcionamiento general de un Random Forest.	20
6	Ilustración de una red neuronal artificial.	22
7	Ejemplo de matriz de confusión para un problema de clasificación binario.	27
8	Ejemplo de Forward Selection.	31
9	Ejemplo de Backward Elimination.	31
10	Ejemplo del funcionamiento de UMAP (3D→2D).	35
11	Diagrama de la estructura de un Autoencoder.	36
12	Autoencoder simple vs. Variational Autoencoder.	37
13	Representación de la ubicación de los embeddings de las palabras “triste”, “deprimido” y “alegría” en un espacio vectorial de tres dimensiones.	40
14	Esquema de BERT aplicado a pares de oraciones.	41
15	Esquema de SBERT aplicado a pares de oraciones.	42
16	Flujo general de la propuesta metodológica.	49
17	Procedimiento general del Screening Dinámico.	67
18	Espacios latentes obtenidos por las variantes del VAE.	71
19	Proyección 2D del espacio latente del VAE mediante UMAP.	72
20	Comparación de las proyecciones generadas por las variantes del modelo SBERT.	74
21	Comparación del espacio latente para 8 y 16 dimensiones.	75
22	Resultados FA 10 vs. 20 iteraciones	77
23	Resultados Método Iterativo	78
24	Resultados Método Top Down	79
25	Matrices de co-asociación de cada instrumento psicométrico.	84
26	Árbol de decisión del cuestionario adaptativo.	102
27	Secuencia mínima del cuestionario adaptativo (Parte 1).	103
28	Secuencia mínima del cuestionario adaptativo (Parte 2).	104

ÍNDICE DE TABLAS

1	Ejemplos de ítems que podrían considerarse semánticamente similares	48
2	Ejemplos de la reformulación semántica de ítems mediante Prompt Engineering	52
3	Ejemplos de similitud semántica entre ítems mediante distancia coseno	54
4	Hiperparámetros empleados para el entrenamiento del modelo VAE	55
5	Resumen cualitativo de las variantes evaluadas del modelo VAE.	56
6	Sistema de votación para la matriz de co-asociación.	62
7	Normalización de respuestas según instrumento	64

8	Niveles de severidad definidos para cada instrumento psicométrico	65
9	Clústeres obtenidos en Top-Down.	80
10	Evaluación comparativa del desempeño de los algoritmos de clustering.	82
11	Ejemplos de clústeres estables detectados en la matriz de co-asociación.	83
12	Estadísticas de co-asociación por instrumento.	83
13	Promedios por nivel y discriminación por ítem GAD-7	86
14	Promedios por nivel y discriminación por ítem WHOQOL-BREF	87
15	Promedios por nivel y discriminación por ítem EUBE	88
16	Promedios por nivel y discriminación por ítem PHQ-9	89
17	Top 3 ítems con mayor correlación ítem-puntaje total PHQ-9	90
18	Top 3 ítems con mayor correlación ítem-puntaje total GAD-7	91
19	Top 3 ítems con mayor correlación ítem-puntaje total EUBE	91
20	Top 3 ítems con mayor correlación ítem-puntaje total WHO-FIS	92
21	Top 3 ítems con mayor correlación ítem-puntaje total WHO-PSI	92
22	Top 3 ítems con mayor correlación ítem-puntaje total WHO-SOC	92
23	Top 3 ítems con mayor correlación ítem-puntaje total WHO-AMB	93
24	Selección de ítem representativo para el clúster Inquietud	95
25	Selección de ítem representativo para el clúster Agotamiento	95
26	Selección de ítem representativo para el clúster Relaciones personales	96
27	Selección de ítem representativo para el clúster Condiciones del entorno	96
28	Selección de ítem representativo para el clúster Autoestima	96
29	Selección de ítem representativo para el clúster Preocupación excesiva	96
30	Selección de ítem representativo para el clúster Afecto negativo	97
31	Selección de ítem representativo para el clúster Concentración	97
32	Selección de ítem representativo para el clúster Anhedonia	97
33	Selección de ítem representativo para el clúster Energía y sueño	98
34	Selección de ítem representativo para el clúster Fatiga académica	98
35	Selección de ítem representativo para el clúster Ánimo deprimido	98
36	Selección de ítem representativo para el clúster Frustración e irritabilidad	98
37	Selección de ítem representativo para el clúster Riesgo suicida	99
38	Selección de ítem representativo para el clúster Bienestar general	99
39	Selección de ítem representativo para el clúster Condiciones de vida	100
40	Umbral de activación definidos para cada clúster	101
41	Comparación entre el cuestionario original y el adaptativo.	105

INTRODUCCIÓN

La salud mental en el entorno universitario se ha convertido en una problemática de creciente relevancia a nivel mundial, afectando de manera significativa a estudiantes que enfrentan constantes presiones académicas, sociales y personales. Trastornos como la ansiedad, depresión y burnout, no solo impactan negativamente en el bienestar emocional de los estudiantes, sino que también comprometen su rendimiento académico y calidad de vida.

Actualmente, para identificar de manera oportuna estos trastornos, se utilizan herramientas psicométricas validadas como los instrumentos PHQ-9, GAD-7, WHOQOL-BREF y EUBE. En general, los cuestionarios se diseñan como formularios más amplios que, además de incluir preguntas sociodemográficas, integran uno o más instrumentos psicométricos, cada uno orientado a medir un constructo particular.

Si bien estos instrumentos son efectivos individualmente, su aplicación en un cuestionario para una detección simultánea presenta una limitación crítica: su considerable extensión y complejidad. Esta longitud provoca fatiga y desmotivación en los participantes, lo que reduce la tasa de respuesta y la calidad de los datos obtenidos, dificultando así la detección temprana, lo cual es crucial para una intervención efectiva. En este contexto, cuando se señala que la extensión de los instrumentos “afecta la calidad de los datos”, se hace referencia a que los participantes, al intentar terminar rápidamente, pueden responder de manera apresurada, mecánica o incluso al azar. Esto reduce la validez de la información recopilada, ya que, aunque se espera que las respuestas se entreguen de forma espontánea y sin un análisis profundo, deben ser igualmente conscientes para reflejar con fidelidad el estado real del encuestado.

Frente a esta problemática, la presente memoria propone la creación y optimización de un cuestionario de salud mental unificado empleando técnicas de inteligencia artificial. La metodología se centrará en analizar el conjunto de preguntas de los instrumentos actuales para identificar redundancias, correlaciones y variables prescindibles mediante algoritmos de IA.

Para ello, se utilizará una muestra de 568 estudiantes universitarios, quienes respondieron un cuestionario, en su versión sin optimizar, que contenía todos los ítems de los instrumentos mencionados. Este estudio constituye el caso base de la investigación, sobre el cual se evaluarán redundancias y correlaciones, lo que permitirá fundamentar la propuesta de optimización y utilizarlo referencia frente al cuestionario resultante.

Cabe destacar que no se busca sustituir el rol fundamental de los profesionales en salud mental, sino complementar su labor al proporcionar una herramienta tecnológica eficiente que facilite el proceso diagnóstico temprano y la implementación oportuna de estrategias preventivas.

El presente trabajo se estructura en cinco capítulos: el primero consiste en la definición del problema y los objetivos que guían la investigación. El segundo capítulo establece el estado

del arte, presentando los fundamentos teóricos y metodologías relevantes para la solución. El tercero describe la solución y creación del cuestionario. El cuarto capítulo detalla la validación de la solución. Finalmente, el último capítulo presenta las conclusiones del trabajo, donde se discuten los resultados, contribuciones y trabajo a futuro.

CAPÍTULO 1

DEFINICIÓN DEL PROBLEMA

La salud mental es una de las problemáticas más relevantes en la sociedad actual, especialmente entre los estudiantes universitarios, quienes enfrentan presiones académicas, sociales y personales que aumentan su vulnerabilidad a trastornos emocionales como el *burnout*, la ansiedad y la depresión [1]. En Chile, este problema es particularmente relevante: estudios recientes indican que más del 50 % de los universitarios presenta síntomas de ansiedad o depresión, y un 30 % muestra indicadores de *burnout* académico [2]. Estos trastornos no solo afectan el rendimiento académico, sino que también impactan en la calidad de vida, aumentando el riesgo de deserción académica y problemas de salud a largo plazo [1][3]. La detección temprana de estos trastornos es crucial para implementar estrategias de intervención efectivas y prevenir complicaciones mayores [3], evitando que evolucionen hacia cuadros más graves o crónicos.

Al focalizar el análisis en el contexto chileno, la situación adquiere una dimensión particularmente crítica. Diversos estudios evidencian que el sistema de educación superior del país enfrenta una crisis de salud mental en escalada, ya que previo a la pandemia, ya señalaban una prevalencia alarmante de sintomatología ansiosa y depresiva en la comunidad universitaria, donde se reportaban tasas de prevalencia para la depresión que oscilaban entre 28.3 % y 28.8 %, y para la ansiedad desde un 24.9 % a un 28.3 % [2][4].

Esta situación se enmarca en una problemática más amplia de la juventud chilena, como lo indicaba la Novena Encuesta Nacional de Juventud (2018), donde un 19.1 % de jóvenes reportó haber sentido ganas de terminar con su vida durante el último mes [5], sugiriendo que las presiones universitarias actúan como un catalizador sobre una vulnerabilidad preexistente [6]. Es fundamental comprender que esta crisis no es simplemente la suma de dificultades individuales, sino un fenómeno sistémico, reconocido a nivel institucional y gubernamental. El aumento sostenido en la demanda de servicios de salud mental en la educación superior chilena coincide temporalmente con la implementación de políticas de gratuidad y equidad en el acceso. Si bien estas políticas han democratizado el ingreso a la universidad, también han diversificado el perfil del estudiante, incorporando a jóvenes de estratos socioeconómicos más bajos, quienes a menudo enfrentan mayores estresores sistémicos y tienen un acceso más limitado a servicios de salud mental privados [7].

Esta nueva realidad ha impuesto una presión sin precedentes sobre los sistemas de apoyo estudiantil de las instituciones, que no siempre han contado los recursos para escalar sus servicios de manera proporcional a la demanda. La creación de organismos como la Comisión de Convivencia Universitaria y Salud Mental del Consejo de Rectoras y Rectores de las Universidades Chilenas (CRUCH) en 2022 y la publicación de informes y recomendaciones por parte del Ministerio de Educación (MINEDUC) son una clara evidencia de que el bienestar estudiantil ha trascendido el ámbito institucional para convertirse en un desafío de política

pública nacional [7]. Este contexto sistémico destaca la necesidad imperativa de desarrollar soluciones de detección y apoyo que sean escalables, eficientes y costo-efectivas, ya que los modelos de intervención individual resultan ineficientes para abordar la magnitud del problema.

Las consecuencias de esta crisis se manifiestan de forma directa en los indicadores clave del sistema educativo. Los trastornos de ansiedad y depresión están directamente correlacionados con un deterioro del rendimiento académico, afectando funciones cognitivas esenciales como la atención, la concentración y la memoria, además de disminuir la motivación y la capacidad para gestionar la carga de trabajo, lo que conduce a un ciclo de bajo rendimiento y aumento del malestar emocional. La consecuencia más grave de esta dinámica es el impacto en las tasas de deserción estudiantil. El abandono de los estudios superiores no solo representa un fracaso personal para el estudiante, sino que también implica costos económicos y sociales significativos para sus familias y para el país, que invierte recursos en una formación que no llega a completarse [7]. Por lo tanto, abordar la salud mental de los estudiantes no es solo una responsabilidad ética, sino una necesidad estratégica para garantizar la eficiencia del sistema de educación superior y el desarrollo del capital humano del país.

En este marco, uno de los principales mecanismos utilizados para la detección de los trastornos emocionales mencionados son los instrumentos de evaluación, los cuales han demostrado su efectividad tanto en términos teóricos como empíricos. En el plano teórico, un cuestionario se entiende como un formulario compuesto por uno o más instrumentos psicométricos validados y preguntas sociodemográficas, que permiten evaluar de manera estandarizada el estado de salud mental de una persona. Entre los más utilizados se encuentran el PHQ-9 (*Patient Health Questionnaire-9*), GAD-7 (*Generalized Anxiety Disorder-7*), WHOQOL-BREF (*World Health Organization Quality of Life*), y EUBE (*Escala Universitaria de burnout en Estudiantes*), cada uno diseñado para medir constructos específicos como depresión, ansiedad, calidad de vida y *burnout*.

En términos empíricos, múltiples estudios en contextos clínicos y académicos han confirmado la utilidad de estos instrumentos, mostrando que su aplicación permite identificar de manera confiable, la prevalencia y severidad de síntomas en población universitaria. De esta forma, se utilizan ampliamente como herramientas de tamizaje y diagnóstico temprano, aportando datos fundamentales para medir el impacto de los trastornos en los estudiantes y orientar intervenciones oportunas [8][4][6].

No obstante, la principal limitación al considerar estos instrumentos en conjunto para detectar los distintos trastornos simultáneamente radica en su extensión y complejidad. En total, la aplicación de estos instrumentos implica responder 57 ítems: 9 del PHQ-9, 7 del GAD-7, 26 del WHOQOL-BREF y 15 de la EUBE. Este volumen de preguntas puede provocar fatiga en los participantes, reduciendo la tasa de respuesta y la calidad de los datos obtenidos [6]. Los estudiantes, al enfrentarse a cuestionarios extensos, pueden desmotivarse o contestar de manera apurada, generando respuestas sesgadas o incompletas. Como consecuencia, se produce una falta de precisión en los resultados, disminuyendo la efectividad de estos instrumentos para la detección temprana de trastornos emocionales. Asimismo, la aplicación

conjunta de estos instrumentos suele carecer de integración tecnológica moderna, ya que a menudo se administran por separado en formatos tradicionales, lo cual dificulta su uso eficiente y la recopilación ágil de resultados unificados para cada estudiante.

Además de las limitaciones técnicas y estructurales de los instrumentos existentes, también es necesario considerar el componente ético asociado a su aplicación, especialmente en entornos remotos. La administración de instrumentos psicométricos que abordan temas sensibles, como la depresión o la ansiedad, puede generar malestar emocional en quienes los responden, particularmente si se encuentran sin acompañamiento profesional al momento de la evaluación. Esta situación plantea desafíos importantes respecto a las condiciones adecuadas para la implementación de estos instrumentos, incluyendo la necesidad de resguardar el bienestar de los participantes, evitar interpretaciones erróneas en contextos no supervisados, y establecer mecanismos de contención en caso de respuestas críticas.

Ante este panorama, surge la necesidad de optimizar un cuestionario que combine adecuadamente los instrumentos mencionados, con el fin de reducir su extensión y mejorar la eficiencia del proceso de evaluación. La solución propuesta radica en analizar los instrumentos actuales, identificar redundancias, posibles correlaciones entre las preguntas y evaluar la factibilidad de construir un cuestionario más conciso y efectivo.

Para abordar esta optimización, la metodología que se empleará contempla tres enfoques complementarios. El primer enfoque corresponde a la selección de características (*feature selection*), donde las “características” corresponden a los ítems o preguntas individuales de los instrumentos. Este proceso consiste en aplicar algoritmos capaces de analizar el conjunto de datos y determinar qué subconjunto de ítems resulta ser más informativo y predictivo respecto de las condiciones de salud mental que se desean detectar, como la presencia de un cuadro depresivo o ansioso. El segundo enfoque considera la aplicación de métodos no supervisados, con técnicas de *clustering* tales como *k-means* o *feature agglomeration*, con el fin de identificar grupos de ítems que tienden a responderse de manera similar. Esto permite detectar redundancias y aplicar técnicas de selección de características sobre dichos clústeres, fortaleciendo el análisis de patrones comunes.

Finalmente, el tercer enfoque aborda la similitud semántica entre ítems, analizando el contenido textual de los ítems a través de representaciones vectoriales (*embeddings*) y espacios latentes. Este análisis permite identificar preguntas que, aunque estén formuladas de diferentes maneras, capturan constructos similares, aportando una perspectiva adicional para la reducción de redundancias.

En conjunto, el objetivo es identificar de manera sistemática las preguntas redundantes, es decir, aquellas que miden el mismo constructo subyacente, o irrelevantes que corresponden a los ítems que aportan poco valor al análisis o no tienen ningún poder predictivo. De esta manera, se busca conservar únicamente aquellas preguntas que resulten informativas y pertinentes, evaluando en qué casos es viable simplificar el cuestionario y en cuáles es necesario mantener la totalidad de los ítems.

Este proceso no solo reduciría la carga sobre los encuestados, fomentando una mayor participación, sino que también mantendría, o incluso mejoraría, la calidad de la información recopilada para facilitar intervenciones tempranas en salud mental.

1.1. Objetivos

1.1.1. Objetivo General

Analizar un conjunto de instrumentos psicométricos existentes que detectan *burnout*, ansiedad y depresión en estudiantes universitarios, con el fin de evaluar la factibilidad de crear un cuestionario integrador optimizado mediante técnicas de inteligencia artificial que reduzca su extensión.

1.1.2. Objetivos Específicos

- Seleccionar métodos de representación semántica y reducción de dimensionalidad documentados en la literatura para su aplicación en el diseño de cuestionarios de salud mental.
- Implementar algoritmos de inteligencia artificial para identificar redundancias, similitudes y agrupamientos semánticos entre los ítems de los instrumentos PHQ-9, GAD-7, EUBE y WHOQOL-BREF.
- Determinar ítems representativos y establecer umbrales empíricos que permitan definir reglas de activación u omisión dentro del cuestionario adaptativo.
- Diseñar una versión preliminar del cuestionario adaptativo que reduzca su extensión ajustando las preguntas según las respuestas previas.

1.2. Resumen del capítulo

En este capítulo se presentó el contexto de la salud mental en estudiantes universitarios, destacando la alta prevalencia de ansiedad, depresión y *burnout* en Chile, así como los efectos negativos que estos trastornos generan en el rendimiento académico y la calidad de vida. Se identificaron las limitaciones de los instrumentos actualmente utilizados, principalmente su extensión y falta de integración, lo que impacta en la calidad de los datos y la participación de los estudiantes. Frente a este escenario, se plantea la necesidad de optimizar un cuestionario integrador mediante técnicas de inteligencia artificial, orientado a facilitar la detección temprana y contribuir en las intervenciones en la comunidad universitaria.

CAPÍTULO 2

ESTADO DEL ARTE

En los últimos años, la salud mental de los estudiantes universitarios ha emergido como una problemática global de creciente preocupación. Existen factores como la presión académica constante, la inestabilidad económica y la transición a la vida adulta, que han contribuido a un aumento sostenido en la prevalencia de trastornos como el burnout, la ansiedad y la depresión en esta población. La detección temprana y la intervención oportuna son cruciales para mitigar el impacto negativo de estos trastornos en el bienestar personal y el rendimiento académico, lo que ha impulsado la necesidad de herramientas de evaluación eficaces y accesibles.

Tradicionalmente, la evaluación de la salud mental se ha realizado mediante cuestionarios psicométricos, que integran instrumentos validados como el PHQ-9, GAD-7, WHOQOL-BREF o la EUBE, cada uno orientado a medir constructos específicos. Si bien la fiabilidad de estos instrumentos está bien documentada en la literatura, la práctica de administrarlos conjuntamente ha dado lugar a cuestionarios de una extensión y complejidad considerable. Esta limitación impacta negativamente en la calidad de los datos recolectados y en la participación de los encuestados, quienes pueden experimentar fatiga de respuesta o abandono. En un escenario donde las instituciones buscan escalar los servicios de apoyo a una población estudiantil cada vez más diversa, la eficiencia de los métodos de tamizaje¹ se vuelve una prioridad.

En este contexto, el avance de la inteligencia artificial ofrece alternativas innovadoras para abordar las limitaciones mencionadas. Herramientas como *feature selection*, la reducción de dimensionalidad y los modelos generativos permiten analizar grandes volúmenes de datos, detectar correlaciones y redundancias entre ítems, e identificar las preguntas más relevantes y predictivas de cada instrumento que son imperceptibles para un análisis humano tradicional. En el ámbito de los cuestionarios psicométricos, estas técnicas posibilitan la creación de versiones optimizadas que permiten reducir la extensión del cuestionario.

El propósito de este capítulo es revisar las investigaciones más relevantes y recientes que han explorado y aplicado técnicas computacionales e inteligencia artificial en la optimización de instrumentos de salud, con énfasis en cuestionarios psicométricos. Se presentarán los conceptos fundamentales, los enfoques metodológicos utilizados en la literatura y los principales resultados reportados, junto con sus limitaciones.

¹El tamizaje es un proceso de evaluación para identificar tempranamente posibles enfermedades, trastornos o condición de salud.

2.1. Salud Mental en el Contexto Universitario

La salud mental de los estudiantes universitarios se ha convertido en una preocupación creciente en las últimas décadas. Diversos estudios epidemiológicos han evidenciado una alta prevalencia de síntomas depresivos, ansiosos y de burnout en la población estudiantil universitaria [9]. Este escenario se atribuye a factores académicos (carga de estudios o presión por el rendimiento), factores sociofamiliares como expectativas familiares, limitaciones económicas y al estrés en general de la vida universitaria. Una comprensión adecuada de los trastornos emocionales más comunes en este contexto es fundamental para diseñar herramientas efectivas de detección y prevención dirigidas a jóvenes universitarios.

2.2. Trastornos emocionales

Los trastornos emocionales engloban afecciones de salud mental caracterizadas por alteraciones del estado de ánimo y respuestas de estrés. En el entorno universitario, los más frecuentes son la depresión, la ansiedad y el síndrome de *burnout*. Estos problemas, de no abordarse, pueden interferir significativamente con el desempeño académico y el bienestar general de los estudiantes.

2.2.1. Depresión

La depresión es un trastorno del estado de ánimo caracterizado por sentimientos persistentes de tristeza, falta de interés en actividades que antes eran placenteras, fatiga y cambios en apetito o sueño. En estudiantes universitarios, se ha reportado que más de la mitad pueden presentar síntomas depresivos clínicamente significativos [10]. Esta alta frecuencia se relaciona con factores como la transición a la vida adulta, la presión académica y la falta de una red de apoyo. Este trastorno no solo afecta la calidad de vida del estudiante, sino que se asocia con un rendimiento académico deficiente y un mayor riesgo de deserción. Es por esto que en muchos campus se utilizan instrumentos de tamizaje como el PHQ-9 para identificar la depresión de forma temprana, permitiendo derivar a los estudiantes a apoyo profesional.

2.2.2. Ansiedad

La ansiedad se manifiesta típicamente como preocupación excesiva, inquietud, tensión muscular y síntomas fisiológicos como palpitaciones o sudoración. Es altamente prevalente en jóvenes universitarios donde estudios con grandes muestras indican que entre el 30 % y 50 % de los estudiantes presentan algún nivel de trastorno de ansiedad durante su formación [7]. Las fuentes de ansiedad en esta etapa incluyen los exámenes, hablar en público, las cargas de trabajo y la incertidumbre sobre el futuro profesional. Si bien cierta ansiedad puede llegar

a ser motivadora, los niveles elevados afectan la concentración y la memoria, repercutiendo en el desempeño académico. Herramientas como el GAD-7 se emplean para cuantificar la severidad de la ansiedad en estudiantes, facilitando su monitoreo y eventual intervención.

2.2.3. Burnout

El *burnout* estudiantil es un síndrome relacionado al estrés crónico académico, análogo al burnout laboral pero en el contexto educacional. Se caracteriza por el agotamiento emocional, la despersonalización (actitudes negativas o distantes hacia el estudio) y baja realización personal. A diferencia de la depresión y la ansiedad, el burnout se origina principalmente en factores situacionales del entorno académico: sobrecarga de tareas, poca retroalimentación, conflictos de rol y altas expectativas [11]. Los estudiantes con *burnout* se sienten exhaustos mental y físicamente, con dificultad para concentrarse y con sensación de ineffectividad en las actividades académicas.

Este síndrome ha sido documentado especialmente en estudiantes de áreas exigentes (medicina, enfermería, ingeniería, etc.), aunque puede presentarse en cualquier carrera. Actualmente, se cuenta con un instrumento específico para medir el *burnout* estudiantil que es la Escala Unidimensional de Burnout Estudiantil (EUBE) de 15 ítems [8]. Detectar *burnout* a tiempo también es crucial, ya que se asocia a mayor riesgo de abandono de los estudios y otros problemas de salud.

2.3. Instrumentos psicométricos en salud mental

Para evaluar de forma estandarizada síntomas de depresión, ansiedad, calidad de vida o burnout en estudiantes, se emplean instrumentos psicométricos validados. Estos instrumentos consisten en ítems (preguntas o afirmaciones) cuyas respuestas suelen darse en escalas Likert, los cuales permiten estimar la severidad de los síntomas y realizar evaluaciones a gran escala de manera eficiente y de bajo costo. A continuación, se describen cuatro instrumentos ampliamente utilizados en salud mental: PHQ-9, GAD-7, WHOQOL-BREF y EUBE, los cuales serán el foco de este estudio por su relevancia en la comunidad universitaria.

2.3.1. Patient Health Questionnaire-9 (PHQ-9)

Instrumento de 9 ítems diseñado para evaluar la presencia y gravedad de síntomas depresivos [12]. Cada ítem corresponde a un criterio de depresión mayor según el DSM-IV², por ejemplo:

²DSM-IV: Manual Diagnóstico y Estadístico de los Trastornos Mentales, cuarta edición, publicado por la American Psychiatric Association

- “Poco interés o placer en hacer las cosas”
- “Se ha sentido decaído, deprimido o sin esperanzas”

Las respuestas van de 0 (“Ningún día”) a 3 (“Casi todos los días”), por lo que la puntuación total oscila entre 0 y 27 puntos. Según el puntaje obtenido, el PHQ-9 permite clasificar el nivel de depresión en diferentes categorías: 0-4 puntos indica depresión mínima, 5-9 leve, 10-14 nivel moderada, 15-19 moderadamente severa y 20-27 depresión severa [12].

El PHQ-9 es muy utilizado en entornos clínicos y de investigación por su brevedad y validez, ya que, además de detectar casos, permite monitorear cambios en síntomas con el tiempo. En el contexto universitario, el PHQ-9 sirve para identificar estudiantes en riesgo que podrían beneficiarse de evaluación y apoyo psicológico. Cabe destacar que existe una versión abreviada PHQ-2, el cual corresponde a las primeras dos preguntas del cuestionario original, empleada para tamizaje rápido, sin embargo, reducen la sensibilidad diagnóstica en comparación al PHQ-9 completo.

2.3.2. Generalized Anxiety Disorder-7 (GAD-7)

Instrumento de 7 ítems orientado a medir síntomas de trastorno de ansiedad generalizada. Indaga con qué frecuencia en las últimas dos semanas la persona ha experimentado síntomas como:

- “Se ha sentido tan inquieto que no ha podido quedarse quieto”
- “Ha tenido miedo de que algo terrible fuera a pasar”

Cada ítem se califica de 0 (“Ningún día”) a 3 (“Casi todos los días”), sumando un puntaje total de 0 a 21 [13]. De acuerdo con la puntuación obtenida, se distinguen cuatro niveles de ansiedad: 0-4 puntos corresponde a nivel mínimo, 5-9 ansiedad leve, 10-14 presenta un nivel moderado de ansiedad y 15-21 equivale a una ansiedad severa.

El GAD-7 ha demostrado alta fiabilidad y validez para detección de trastornos de ansiedad en diversos contextos. En estudiantes universitarios, es útil para cuantificar el nivel de ansiedad académica y general, permitiendo identificar aquellos casos que requieren intervenciones como por ejemplo, técnicas de manejo de estrés o derivación a ayuda psicológica. Similar al PHQ, también existe una versión abreviada GAD-2 de dos ítems, aunque con las limitaciones inherentes de menor sensibilidad.

2.3.3. World Health Organization Quality of Life-BREF (WHOQOL-BREF)

Instrumento diseñado por la Organización mundial de la salud (OMS) en los años 80 basado en WHOQOL-100, siendo WHOQOL-BREF una versión abreviada de este, con tan solo 26 ítems [14]. A diferencia de los instrumentos anteriores centrados en síntomas, el WHOQOL-BREF evalúa el bienestar integral de la persona en cuatro dominios: físico, psicológico, relaciones sociales y entorno. Incluye preguntas sobre satisfacción con la salud, nivel de energía, estado emocional, apoyo social, condiciones de vivienda, entre otros aspectos. Ejemplos de ítems del WHOQOL-BREF incluyen afirmaciones como:

- “¿Cómo de satisfecho/a está de sí mismo?”
- “¿Cómo de satisfecho/a está con su habilidad para realizar sus actividades de la vida diaria?”

Cada ítem se califica en una escala tipo Likert de 1 a 5, donde 1 corresponde a “Muy insatisfecho/a” y 5 a “Muy satisfecho/a”. Los puntajes obtenidos en cada dominio se suman y transforman en escalas comparables, de modo que un mayor puntaje refleja una mejor percepción de calidad de vida en ese ámbito específico. De esta manera, el WHOQOL-BREF no entrega un puntaje único global, sino que permite analizar de forma diferenciada la calidad de vida en cada uno de los cuatro dominios evaluados.

En contextos universitarios, el WHOQOL-BREF permite apreciar cómo factores estresantes impactan la calidad de vida de los estudiantes más allá de síntomas aislados de depresión o ansiedad.

2.3.4. Escala Unidimensional de Burnout Estudiantil (EUBE)

Instrumento desarrollado específicamente para medir el burnout en población estudiantil universitaria. Consta de 15 ítems que se puntúan en una escala Likert, de 0 (“Nunca”) a 3 (“Siempre”), que evalúan el agotamiento y estrés académico del estudiante [3]. Ejemplos de ítems de la EUBE incluyen afirmaciones como:

- “Siento que estudiar me está desgastando físicamente”
- “No creo terminar con éxito mis estudios”

Al ser unidimensional, la escala considera el burnout estudiantil como un único constructo global, centrado principalmente en el agotamiento emocional, en lugar de dividirlo en subdimensiones específicas. Esta herramienta fue propuesta por Barraza en 2008 [8] y validada en población universitaria hispanohablante [3]. Su puntuación total indica el nivel de burnout,

a mayor puntuación, mayor síndrome de desgaste. La EUBE resulta útil para identificar estudiantes en riesgo de burnout que podrían beneficiarse de intervenciones preventivas como talleres de manejo del estrés o asesorías en técnicas de estudio.

2.4. Problemas de clasificación

La clasificación corresponde a una de las tareas fundamentales en el aprendizaje automático y se refiere al proceso de asignar una categoría o etiqueta a una instancia, que corresponde a un conjunto de características que representa un único punto de datos, en función de sus características observadas [15]. En términos generales, este proceso busca encontrar una relación entre un conjunto de variables de entrada y una variable de salida, de modo que un modelo entrenado pueda generalizar y predecir correctamente etiquetas para nuevas observaciones.

Formalmente, un problema de clasificación puede expresarse como la búsqueda de una función:

$$f : X \rightarrow Y \quad (1)$$

donde cada instancia $x = (x_1, x_2, \dots, x_n)$ está definida por un vector de características (x_i) , mientras que $y \in Y$ corresponde a la clase asignada. En el marco de esta memoria, las características representan las respuestas individuales que los estudiantes proporcionan a los ítems de los cuestionarios psicométricos, y la clase denota la condición de salud mental inferida, como por ejemplo “depresión severa” o “ansiedad leve”. Así, una instancia equivale al conjunto completo de respuestas de un estudiante, constituyendo la unidad fundamental de análisis a partir de la cual se aborda el problema de clasificación.

La Figura 1 resume el funcionamiento general de un problema de clasificación desde una perspectiva algorítmica. Dado un vector de atributos x , el modelo de clasificación busca aproximar una función f que mapea cada instancia al conjunto de clases posibles. El esquema permite observar que el conjunto de características constituye la entrada del sistema, mientras que la salida corresponde a una etiqueta discreta y . Este proceso formaliza la idea de que, a partir de variables observables, el modelo genera una decisión categórica, lo que constituye la base para aplicar técnicas de aprendizaje supervisado en contextos como la salud mental estudiantil.

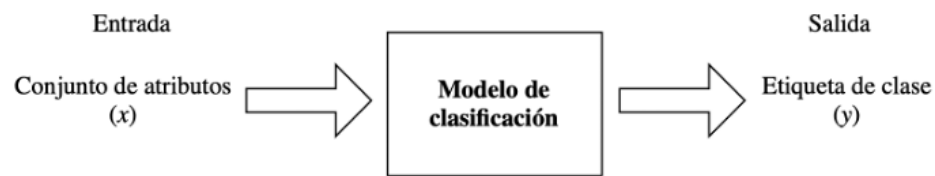


Figura 1: Representación del problema de clasificación, donde a partir de un conjunto de entrada x se asigna una etiqueta de clase y .

Fuente: [15]

Una vez definido el problema de clasificación, es necesario un proceso de entrenamiento mediante el cual un algoritmo ajusta su estructura o sus reglas de decisión, dependiendo del tipo de algoritmo, para reconocer patrones en los datos. Dicho entrenamiento consiste en utilizar ejemplos representativos que contienen tanto las características de entrada como la información asociada a las clases, agrupados en lo que se denomina **conjunto de entrenamiento (training set)**. A partir de este conjunto, el algoritmo aprende a relacionar los atributos con sus etiquetas, generando un modelo capaz de realizar predicciones sobre instancias nuevas. Durante este proceso, uno de los principales desafíos es manejar la presencia de “**ruido**” en los datos, entendido como errores, respuestas inconsistentes o valores atípicos que distorsionan los patrones reales de las observaciones. Este tipo de ruido puede originarse por fallos en la medición o por respuestas poco atentas de los encuestados, afectando la capacidad del modelo para generalizar correctamente y reduciendo su precisión en la predicción de nuevas instancias.

Sin embargo, evaluar el desempeño únicamente sobre los datos utilizados en el entrenamiento llevaría a una estimación sesgada, es decir, una estimación no realista de la capacidad del modelo para clasificar correctamente nuevas instancias, ya que el modelo estaría siendo evaluado con los mismos ejemplos que utilizó para aprender. En consecuencia, los resultados obtenidos no reflejarían su verdadera capacidad para generalizar a datos no observados. Por ello, se reserva una porción distinta de los datos, conocida como **conjunto de validación (validation set)**, utilizado durante el entrenamiento para ajustar los hiperparámetros y seleccionar el mejor modelo, o **conjunto de prueba (test set)**, empleado únicamente al final del proceso para evaluar el desempeño general del modelo sobre datos no vistos. De esta forma, se evita que el modelo sea evaluado con los mismos ejemplos utilizados para aprender, asegurando que la estimación de su rendimiento refleje su verdadera capacidad de generalización.

Dado que el objetivo de esta investigación es procesar y analizar respuestas de cuestionarios psicométricos, resulta pertinente distinguir entre clasificación supervisada, no supervisada y semisupervisada, cuyos enfoques que permiten abordar el problema desde diferentes perspectivas y que se presentan en las siguientes subsecciones.

2.4.1. Clasificación supervisada

La clasificación supervisada consiste en el aprendizaje de una función de decisión $f : X \rightarrow Y$ a partir de un conjunto de instancias previamente etiquetadas [16]. En este enfoque, cada elemento del conjunto de entrenamiento se representa como un par (x, y) , donde $x = (x_1, x_2, \dots, x_n)$ corresponde al vector de características de una instancia e $y \in Y$ es la etiqueta de clase asignada. El objetivo del algoritmo es encontrar una aproximación de f que, a partir de las características observadas en x , logre predecir correctamente la etiqueta asociada a nuevas instancias no vistas durante el entrenamiento.

Esto se puede ilustrar en la Figura 2, donde se representan dos tipos de instancias (círculos

y triángulos) y la función f aprendida por el modelo corresponde a la frontera de decisión que separa ambas clases. De esta forma, al introducir nuevas instancias, el modelo puede asignarles la etiqueta adecuada según el lado de la frontera en el que se ubiquen.

En el contexto de esta memoria, un ejemplo de instancia sería el conjunto completo de respuestas de un estudiante a los ítems de un instrumento psicométrico, mientras que la etiqueta y correspondería al diagnóstico clínico asociado, como “depresión” o “burnout”. El procedimiento de entrenamiento permite que el modelo identifique patrones y relaciones entre las características y las clases correspondientes, de modo que pueda generalizar este conocimiento a casos futuros.

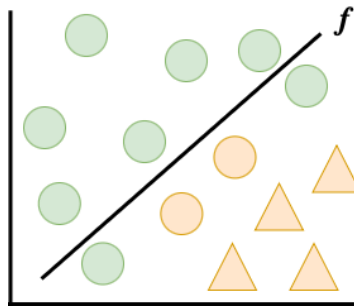


Figura 2: Ejemplo de clasificación supervisada.
Fuente: Elaboración propia.

Existen múltiples algoritmos para clasificación supervisada, entre ellos árboles de decisión, Random Forest, las máquinas de soporte vectorial (SVM) y las redes neuronales artificiales, cada uno con distintas estrategias para aprender la función f . La descripción detallada de estos algoritmos se presenta en la Subsección 2.5. El éxito de este enfoque depende críticamente de la calidad de las etiquetas empleadas como *ground truth*³ y de que el conjunto de entrenamiento sea lo suficientemente representativo de la población objetivo, ya que de lo contrario el modelo puede aprender patrones erróneos y perder capacidad de generalización.

2.4.2. Clasificación no supervisada

La clasificación no supervisada, comúnmente denominada *clustering*, agrupa instancias en función de sus similitudes sin necesidad de contar con etiquetas previamente definidas [17]. Formalmente, dado un conjunto de instancias $x_1, x_2, \dots, x_m$ en un espacio de características X , el objetivo es particionar dicho conjunto en k grupos $C_1, C_2, \dots, C_k$ tales que las instancias dentro de un mismo grupo presenten alta similitud interna y, al mismo tiempo,

³El término *ground truth* se refiere a las etiquetas reales o de referencia que sirven como verdad objetiva para entrenar y evaluar un modelo de aprendizaje automático.

diferencias significativas respecto a las instancias de otros grupos. En este enfoque, el algoritmo no busca aproximar una función $f : X \rightarrow Y$, sino descubrir la estructura latente de los datos, entendida como los patrones o relaciones subyacentes que no son directamente observables en las variables originales pero que explican cómo se organizan naturalmente las instancias.

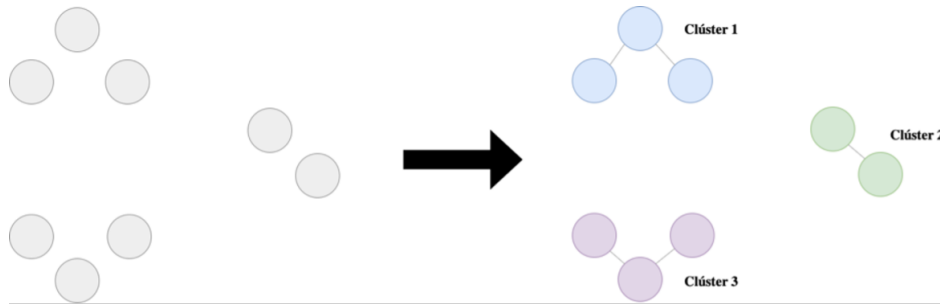


Figura 3: Ejemplo de clasificación no supervisada.

Fuente: Elaboración propia.

Esto se puede demostrar con la Figura 3, donde en una primera etapa todas las instancias aparecen en gris, reflejando la ausencia de etiquetas conocidas. Posteriormente, el algoritmo identifica la estructura latente y organiza las instancias en diferentes clústeres. De esta forma, se observa cómo las instancias con características similares quedan agrupadas, mientras que aquellas con diferencias o de mayor distancia, se separan en distintos conjuntos.

En el contexto de esta memoria, las técnicas no supervisadas permiten, por ejemplo, aplicar *K-Means* para identificar ítems de los cuestionarios que tienden a responderse de forma similar por parte de los estudiantes, lo que puede evidenciar redundancia entre preguntas. Asimismo, los métodos de agrupamiento resultan útiles para detectar patrones emergentes, caracterizar subtipos de estudiantes de acuerdo con sus perfiles de salud mental, o realizar reducción de dimensionalidad al identificar conjuntos de variables altamente correlacionadas.

K-Means

El algoritmo *K-Means* es una técnica de clasificación no supervisada ampliamente utilizada para agrupar observaciones en función de su similitud. Formalmente, dado un conjunto de datos $X = \{x_1, x_2, \dots, x_n\}$, donde cada $x_i \in \mathbb{R}^d$, el objetivo es particionar X en k grupos disjuntos $C = \{C_1, C_2, \dots, C_k\}$ minimizando la suma de las distancias cuadradas entre los puntos y el centroide de su clúster:

$$\operatorname{argmin}_C \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2$$

donde μ_i representa el centroide del clúster C_i , calculado como el promedio de los puntos asignados a dicho grupo.

El algoritmo sigue un proceso iterativo:

1. **Inicialización:** Selección de k centroides iniciales, que pueden ser elegidos aleatoriamente o mediante métodos específicos como k-means++.
2. **Asignación:** Cada punto de datos se asigna al clúster cuyo centroide esté más cercano, según una métrica de distancia (usualmente la distancia euclidiana).
3. **Actualización:** Recálculo de los centroides de cada clúster como el promedio de los puntos asignados a ellos.
4. **Repetición:** Los pasos de asignación y actualización se repiten hasta que las asignaciones de los puntos ya no cambian o se alcanza un criterio de convergencia predefinido.

Entre los principales *hiperparámetros* del algoritmo se encuentran:

- **Número de clústeres (k):** Número de clústeres a generar que determina la granularidad de la partición.
- **Método de inicialización ($init$):** Método para seleccionar los centroides iniciales. El esquema más común es k-means++, que selecciona los centroides iniciales de manera probabilística para mejorar la estabilidad y evitar soluciones subóptimas.
- **Número de inicializaciones independientes (n_{init}):** Cantidad de ejecuciones independientes del algoritmo con distintas inicializaciones. El resultado final corresponde al agrupamiento con menor inercia total, lo que reduce la dependencia de una inicialización específica.
- **Semilla aleatoria ($random_state$):** Semilla del generador de números aleatorios que permite garantizar la reproducibilidad de los resultados.

Aunque K-Means es valorado por su simplicidad, rapidez y buena escalabilidad con grandes volúmenes de datos, presenta ciertas limitaciones: requiere definir a priori el número de clústeres k , es sensible a la inicialización de los centroides y puede verse afectado por la presencia de outliers o por la forma no esférica de los grupos en el espacio de características.

2.4.3. Clasificación semisupervisada

Un problema de clasificación semisupervisada corresponde a un enfoque intermedio entre la clasificación supervisada y la no supervisada. En este caso, el conjunto de datos disponible

contiene tanto instancias etiquetadas como no etiquetadas, siendo estas últimas generalmente la mayoría. Formalmente, se dispone de un conjunto reducido de pares (x_i, y_i) con $y_i \in Y$, junto con un conjunto más amplio de instancias x_i sin etiqueta asociada. El objetivo del algoritmo es aprovechar simultáneamente la información de ambas fuentes para construir una función de decisión $f : X \rightarrow Y$ con mayor capacidad de generalización que la que se obtendría usando únicamente los datos etiquetados [16].

En el ámbito de la salud mental universitaria, este paradigma resulta especialmente relevante, ya que obtener etiquetas confiables suele ser costoso y requiere intervención de especialistas, mientras que recopilar respuestas a cuestionarios es mucho más sencillo y escalable. Así, los métodos semisupervisados permiten explotar la gran cantidad de datos sin etiquetar (respuestas sin diagnóstico clínico confirmado) junto con un subconjunto reducido de casos validados, mejorando el aprendizaje del modelo.

Entre las estrategias más utilizadas se encuentran los enfoques de auto-entrenamiento (*self-training*), donde un modelo inicial asigna etiquetas provisionales a los datos no clasificados y posteriormente se reentrena utilizando este conjunto ampliado. También destacan los métodos basados en grafos, que representan las relaciones de similitud entre instancias etiquetadas y no etiquetadas para propagar la información de la clase a través de la estructura del grafo. En la Figura 4 se ejemplifica este proceso: los nodos coloreados corresponden a instancias con etiquetas conocidas, mientras que los nodos grises representan datos sin etiquetar. Luego, tras aplicar el algoritmo, las etiquetas se extienden a estos últimos en función de su similitud con los datos ya clasificados, generando así las pseudo-etiquetas que incrementan el volumen de información disponible para el entrenamiento.

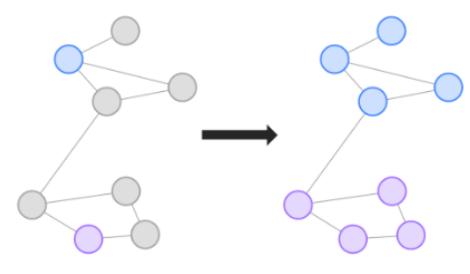


Figura 4: Ejemplo de clasificación semisupervisada mediante propagación de etiquetas.
Fuente: [18]

En el contexto de la presente investigación, la clasificación semi-supervisada ofrece un marco adecuado para escenarios donde no se cuenta con un diagnóstico clínico exhaustivo de todos los estudiantes. Pero sí se dispone de cuestionarios masivos y de un subconjunto acotado de casos validados por especialistas.

2.5. Algoritmos de clasificación

En el ámbito del aprendizaje automático, los algoritmos de clasificación cumplen un rol fundamental en tareas donde se requiere asignar una etiqueta o categoría a partir de un conjunto de características observadas. En el contexto de salud mental, permiten predecir si un estudiante presenta o no síntomas de ansiedad, depresión o burnout en base a sus respuestas a cuestionarios psicométricos. El objetivo central de estos modelos es identificar patrones y dependencias entre las variables de entrada, que corresponden a los ítems de los cuestionarios, y la variable objetivo, que indica la presencia de sintomatología. Además, buscan lograr una adecuada capacidad de generalización hacia nuevas instancias no vistas durante el entrenamiento.

Dentro de la amplia gama de algoritmos disponibles, destacan los Árboles de decisión, Random Forest, Support Vector Machine (SVM) y Redes Neuronales, que han demostrado ser especialmente útiles por su versatilidad para manejar distintos tipos de datos, su capacidad de capturar relaciones complejas y, en algunos casos, su potencial para ofrecer interpretaciones claras sobre los factores que influyen en la clasificación.

2.5.1. Árboles de decisión

Los árboles de decisión son algoritmos de clasificación que representan el proceso de predicción mediante una estructura jerárquica en forma de árbol. Cada nodo interno corresponde a una condición sobre una variable, las ramas representan los posibles resultados de esa condición y cada hoja indica la clase asignada a las instancias que cumplen con la secuencia de reglas trazada desde la raíz hasta la hoja [19].

La construcción del árbol se realiza de manera recursiva, dividiendo el conjunto de datos en particiones cada vez más homogéneas de acuerdo con un criterio de impureza. Este criterio mide el grado de mezcla de clases en un nodo, por lo que mientras menor sea la impureza, más “puro” es el nodo, es decir, más instancias pertenecen a una sola clase. Entre los criterios más utilizados destacan [18]:

- **Índice de Gini:** Mide la probabilidad de clasificar incorrectamente una instancia seleccionada al azar si se le asigna una etiqueta de acuerdo con la distribución de clases del nodo. Se define como:

$$Gini(t) = 1 - \sum_{i=1}^C p_i^2 \quad (2)$$

Donde p_i es la proporción de instancias de la clase i en el nodo t y C es el número total de clases. Un valor cercano a 0 indica menor impureza del nodo.

- **Entropía:** Mide la cantidad de información necesaria para representar la distribución

de clases en un nodo. Se define como:

$$Entropy(t) = - \sum_{i=1}^C p_i \log_2(p_i) \quad (3)$$

Donde, al igual que en el criterio anterior, p_i es la proporción de instancias de la clase i . Un valor bajo de entropía indica mayor homogeneidad en el nodo y un nodo es más homogéneo cuando la mayoría de las instancias que contiene pertenecen a una misma clase.

Ambas métricas cuantifican la impureza del nodo y, por tanto, son comparables. Sin embargo, difieren en la forma en que penalizan la incertidumbre: la entropía tiene una base logarítmica y tiende a asignar un peso mayor a las clases minoritarias, mientras que el índice de Gini es una medida más simple y menos sensible a estas diferencias. En la práctica, ambas suelen producir particiones muy similares, y la elección entre una u otra depende del criterio de implementación o de la estabilidad buscada en el modelo.

Durante el entrenamiento, en cada nodo el algoritmo evalúa las posibles divisiones de acuerdo con estos criterios y selecciona aquella que produce la mayor reducción de impureza, es decir, la que genera nodos hijos más homogéneos que el nodo padre. Este proceso continúa hasta cumplir una condición de parada como alcanzar una profundidad máxima o un número mínimo de instancias en cada nodo. Como resultado de este procedimiento, los árboles de decisión no solo permiten realizar predicciones, sino que también ofrecen una representación gráfica del proceso de clasificación.

Una de sus principales ventajas es la interpretabilidad, ya que permiten visualizar de manera clara qué atributos o umbrales resultan determinantes en la clasificación. Sin embargo, presentan ciertas limitaciones: son sensibles al ruido en los datos y tienden a sobreajustarse si no se aplican técnicas de control de complejidad, como la poda o la regularización de la profundidad del árbol.

2.5.2. Random Forest

Por definición, el Random Forest es un clasificador que consiste en una colección de árboles de decisión, cada uno construido de manera independiente a partir de un muestreo aleatorio de los datos y de las variables disponibles. Cada árbol genera de manera independiente una predicción para una entrada X , y el resultado final se obtiene mediante un proceso de votación mayoritaria entre todos los árboles [20]. El término bosque hace referencia al conjunto de múltiples árboles generados, mientras que el adjetivo aleatorio proviene de la introducción de variabilidad tanto en los datos como en las características utilizadas en la construcción de cada árbol.

Cuando se entrena un número suficientemente grande de árboles, cada uno con un vector de características aleatorio, el algoritmo asigna la clase final en función de la categoría más

votada entre los árboles individuales. Este procedimiento permite reducir la varianza de las predicciones, mejorar la capacidad de generalización y otorgar mayor estabilidad frente al sobreajuste que los árboles de decisión.

Esto se ilustra en la Figura 5, donde un mismo conjunto de datos se utiliza para generar múltiples árboles de decisión independientes, cuyas salidas se combinan en un proceso de votación o promedio para producir el resultado final.

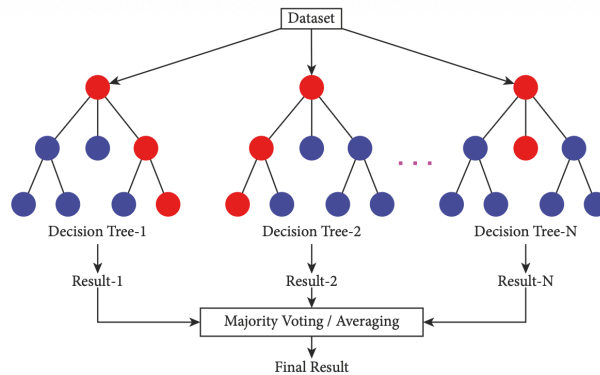


Figura 5: Funcionamiento general de un Random Forest.

Fuente: [21][20]

Al implementar un modelo de Random Forest, resulta fundamental ajustar los hiperparámetros que controlan tanto la construcción de los árboles individuales como el funcionamiento conjunto del modelo, basado en la combinación de múltiples árboles de decisión. Los más relevantes son los siguientes [20]:

- **Criterio:** Función utilizada para evaluar la calidad de las particiones en cada nodo, siendo los más comunes el índice de Gini y la entropía.
- **Cantidad de estimadores:** Número total de árboles a generar para el ensamble final, es decir, para la combinación final de predicciones de todos los árboles. Un mayor valor incrementa la estabilidad del modelo pero aumenta el costo computacional.
- **Profundidad máxima:** Límite de niveles que puede alcanzar cada árbol en forma vertical, lo que permite controlar su complejidad y prevenir sobreajuste.
- **Casos mínimo por rama:** Número mínimo de instancias requeridas en un nodo para permitir la creación de una partición, evitando divisiones irrelevantes en ramas con pocos datos.
- **Cantidad de características:** Número de variables consideradas en cada división, lo que introduce aleatoriedad y favorece la diversidad entre árboles.

2.5.3. Support Vector Machine (SVM)

Son algoritmos de clasificación supervisada que buscan encontrar el hiperplano óptimo que separe las instancias de distintas clases dentro del espacio de características. Este hiperplano se define de tal manera que el margen de separación entre clases, es decir, la distancia a los puntos más cercanos de cada clase, sea máximo, lo que le otorga al modelo una mayor robustez frente a instancias cercanas al límite de decisión [22].

Una característica fundamental de las SVM es la incorporación de funciones kernel, que permiten proyectar los datos originales a un espacio de mayor dimensión, donde resulta posible encontrar separaciones lineales incluso cuando en el espacio original las clases no son linealmente separables. Este enfoque amplía significativamente la capacidad de las SVM para capturar relaciones no lineales entre variables, mejorando así su poder de clasificación en comparación con los modelos estrictamente lineales.

2.5.4. Redes Neuronales

Las redes neuronales artificiales (ANN) son modelos de aprendizaje automático inspirados en la estructura y funcionamiento del cerebro humano. Están compuestas por unidades elementales llamadas neuronas artificiales, organizadas en capas y conectadas mediante pesos sinápticos que determinan la importancia que tiene cada entrada en la activación de la siguiente neurona [23]. El procesamiento de la información ocurre de forma progresiva a través de varias capas, lo que permite transformar las representaciones iniciales de los datos de entrada en características de mayor nivel de abstracción.

2.5.4.1 Neuronas

Componente básico de una red neuronal artificial, que simula el proceso de recibir múltiples entradas, ponderarlas y generar una salida transformada [24]. Matemáticamente, se define como [20]:

$$N = \sigma \left(\sum_{i=1}^n W_i X_i + b \right) \quad (4)$$

Siendo:

- σ : Función de activación, que introduce no linealidad en el modelo y permite capturar relaciones complejas entre variables.
- X : Datos de entrada vectorizados.
- W : Vector de pesos conformado por la información entregada de cada neurona que conforma la capa. Cada neurona tiene un peso específico asociado a la conexión entre ella y otra neurona.

- b : Sesgo o *bias* que representa un valor opcional adicional para favorecer uno de los resultados en las neuronas o la activación de ellas mismas.

Una red neuronal se organiza en capas de neuronas interconectadas. Existen tres tipos [23][20]:

- **Capa de entrada (Input layer):** Encargada de recibir la información, transformarla y enviarla a la siguiente capa.
- **Capas ocultas (Hidden layers):** Realizan transformaciones intermedias de los datos mediante combinaciones lineales de las entradas seguidas de la aplicación de funciones de activación no lineales.
- **Capa de salida (Output layer):** Se encarga de tomar la información de todas las capas anteriores y generar la predicción final, entregando un resultado traducible al lenguaje humano.

En la Figura 6 se ilustra la arquitectura general de una red neuronal artificial. La capa de entrada (color verde) recibe las variables del conjunto de datos, representadas como $x_0, x_1, \dots, x_n$. Estas entradas se combinan linealmente con los pesos w , que representan la importancia de cada característica en la predicción. La información resultante se transmite hacia una o más capas ocultas (color azul), donde cada neurona aplica una función de activación f_h sobre la suma ponderada de las entradas. Esto permite introducir no linealidad en el modelo y aprender representaciones de mayor nivel de abstracción. Finalmente, la capa de salida (color rojo) recibe las transformaciones de las capas ocultas y produce la predicción $\hat{y}$, mediante una función de activación apropiada al tipo de tarea: sigmoide en clasificación binaria, softmax en clasificación multiclase o una activación lineal en regresión.

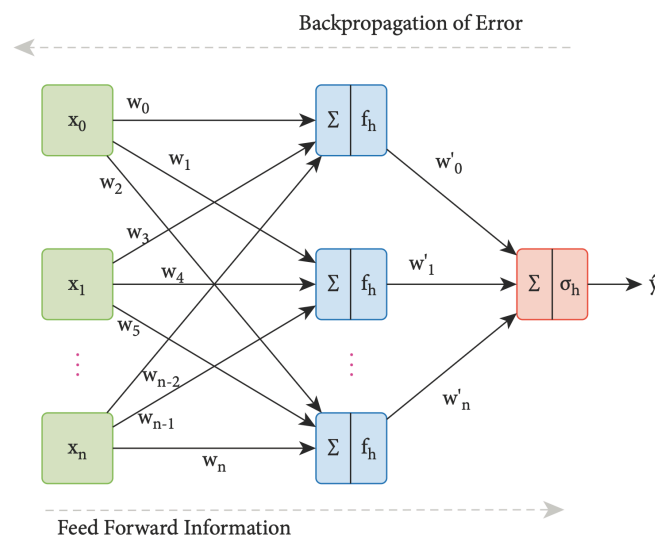


Figura 6: Ilustración de una red neuronal artificial.

Fuente: [21]

2.5.4.2 Funciones de activación

Las funciones de activación permiten que las redes neuronales modelen relaciones no lineales. Entre las más utilizadas destacan [24]:

- **Sigmoide:** Función que mapea cualquier valor real a un rango entre 0 y 1, definida como:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (5)$$

Es especialmente útil en la capa de salida para tareas de clasificación binaria, ya que su salida puede interpretarse como una probabilidad.

- **ReLU (Rectified Linear Unit):** Función que retorna el valor de entrada si es positivo y cero en caso contrario, definida como:

$$ReLU(x) = \max(0, x) \quad (6)$$

Es ampliamente utilizada en capas ocultas debido a su simplicidad computacional y a que ayuda a mitigar el problema del desvanecimiento del gradiente durante el entrenamiento.

- **Softmax:** Función que convierte un vector de valores reales en un vector de probabilidades, donde cada componente está en el rango (0,1) y la suma total es 1. Se define como:

$$Softmax(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \quad (7)$$

Es comúnmente utilizada en la capa de salida para tareas de clasificación multiclase.

2.5.4.3 Entrenamiento de redes neuronales

El aprendizaje de una red neuronal se realiza mediante un proceso iterativo que ajusta los pesos y sesgos para minimizar la discrepancia entre las predicciones y los valores reales. Este proceso se conoce como retropropagación del error (*backpropagation*) y funciona de la siguiente manera [23]:

1. **Propagación hacia adelante (Feedforward):** Los datos de entrada se pasan a través de la red, capa por capa, aplicando las funciones de activación en cada neurona para generar una predicción $\hat{y}$.
2. **Cálculo del error:** Se evalúa la diferencia entre la predicción $\hat{y}$ y el valor real y utilizando una función de pérdida adecuada, como el error cuadrático medio para regresión o la entropía cruzada para clasificación.

3. **Retropropagación (Backpropagation):** El error calculado se propaga hacia atrás a través de la red, computando los gradientes de la función de pérdida con respecto a cada peso y sesgo mediante la regla de la cadena.
4. **Actualización de pesos:** Los pesos y sesgos se ajustan utilizando un algoritmo de optimización, como el descenso del gradiente, que actualiza cada parámetro en la dirección que reduce el error. Este proceso se repite durante múltiples iteraciones (épocas) hasta que el modelo converge a una solución óptima o se alcanza un criterio de parada.

Este ciclo se repite durante varias iteraciones o épocas, hasta alcanzar un nivel de error bajo o cumplir un criterio de parada.

2.5.4.4 Funciones de pérdida

En el aprendizaje automático, el entrenamiento de un modelo requiere un criterio que permita cuantificar la discrepancia entre las predicciones generadas $\hat{y}^{(i)}$ y los valores reales $y^{(i)}$. Este criterio se conoce como función de pérdida (*loss function*), y constituye el elemento central que guía el ajuste de los parámetros del modelo durante la optimización.

Su valor se calcula para cada instancia o lote de instancias, y el algoritmo de optimización modifica los parámetros del modelo en dirección opuesta al gradiente de la pérdida, buscando minimizarla progresivamente.

Formalmente, dada una muestra de entrenamiento de tamaño n , la función de pérdida L se define como:

$$L(\hat{y}^{(i)}, y^{(i)}) = \frac{1}{n} \sum_{i=1}^n \ell(\hat{y}^{(i)}, y^{(i)}) \quad (8)$$

Las funciones de pérdida más utilizadas son:

- **Error cuadrático medio (MSE):** Se define como el promedio del cuadrado de las diferencias entre ambos valores, donde n corresponde al número total de instancias.

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}^{(i)} - y^{(i)})^2 \quad (9)$$

- **Error absoluto medio (MAE):** Se define como el promedio de las diferencias absolutas entre ambos valores.

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}^{(i)} - y^{(i)}| \quad (10)$$

- **Entropía cruzada (Cross-Entropy):** Utilizada en tareas de clasificación, mide la diferencia entre dos distribuciones de probabilidad: la distribución real y la predicha por el

modelo. Para clasificación binaria, se define como:

$$CE = -\frac{1}{n} \sum_{i=1}^n [y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})] \quad (11)$$

Para clasificación multiclase, se generaliza a:

$$CE = -\frac{1}{n} \sum_{i=1}^n \sum_{c=1}^C y_c^{(i)} \log(\hat{y}_c^{(i)}) \quad (12)$$

donde C es el número de clases y $y_c^{(i)}$ es un indicador binario que vale 1 si la clase correcta para la instancia i es c , y 0 en caso contrario.

2.5.4.5 Hiperparámetros

El rendimiento de una red neuronal depende en gran medida de la elección adecuada de sus hiperparámetros. Entre los más relevantes destacan [3]:

- **Función de activación:** Función de activación para una capa del modelo.
- **Tamaño de muestra (Batch size):** Número de instancias que se procesan antes de actualizar los parámetros del modelo.
- **Cantidad de iteraciones (Epochs):** Número de veces que el algoritmo recorrerá todo el conjunto de datos durante el entrenamiento.
- **Tasa de aprendizaje (Learning rate):** Número que define la velocidad con la que el algoritmo aprende, es decir, la velocidad de actualización de los pesos en cada iteración.
- **Cantidad de capas ocultas:** Determina la profundidad del modelo y su capacidad para aprender representaciones complejas.
- **Cantidad de neuronas:** Cantidad de neuronas presentes en cada capa oculta.

En síntesis, las ANN son una de las herramientas más potentes del aprendizaje automático, capaces de modelar relaciones altamente no lineales y complejas entre los datos. No obstante, presentan limitaciones importantes que deben considerarse. Su interpretabilidad es reducida en comparación con modelos más transparentes⁴ como los árboles de decisión o Random Forest. Además, requieren grandes volúmenes de datos y recursos computacionales para alcanzar un buen rendimiento y están expuestas al riesgo de sobreajuste si no se aplican técnicas de regularización como dropout, normalización o ajuste de hiperparámetros. Por ello, su implementación debe equilibrar la capacidad predictiva con la necesidad de garantizar confiabilidad y transparencia, especialmente en ámbitos sensibles como la salud mental.

⁴El término transparencia hace referencia a la facilidad con la que se puede comprender cómo las variables influyen en la predicción final.

2.6. Métricas

La evaluación de un modelo de clasificación requiere el uso de métricas cuantitativas que permitan medir de forma objetiva su desempeño y determinar la validez y confiabilidad de los resultados. Estas métricas ofrecen distintas perspectivas complementarias sobre el comportamiento del modelo. En un contexto supervisado, la matriz de confusión y las medidas derivadas de ella (precisión, exactitud, recall, F1-score) constituyen herramientas fundamentales para analizar la capacidad del modelo de distinguir correctamente entre clases. Adicionalmente, el Error Cuadrático Medio (MSE) permite cuantificar la magnitud de los errores entre las predicciones del modelo y los valores reales.

En conjunto, estas métricas proporcionan una visión integral del rendimiento del modelo y son esenciales para validar la efectividad de la solución propuesta, así como para comparar diferentes enfoques de clasificación.

2.6.1. Matriz de confusión

La matriz de confusión constituye una herramienta fundamental para la evaluación de modelos de clasificación, ya que permite representar de manera estructurada la relación entre las clases reales y las clases predichas por el modelo. Se organiza en una tabla cuadrada donde las filas corresponden a las clases reales y las columnas a las predicciones realizadas. De esta forma, cada celda refleja la cantidad de instancias clasificadas en esa combinación, lo que permite identificar cuatro categorías principales: verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN) [3]. Esta representación facilita un análisis detallado tanto de los aciertos como de los errores cometidos por el modelo y constituye la base para el cálculo de métricas derivadas, tales como la exactitud, la precisión, el recall y la medida F1.

Esto se puede ejemplificar con la Figura 7, en la que se presenta una matriz de confusión para un problema de clasificación binaria. En ella, los cuadrantes diagonales indican las predicciones correctas (TP y TN), mientras que los cuadrantes fuera de la diagonal reflejan las clasificaciones erróneas (FP y FN). De esta manera, la matriz permite una interpretación inmediata del desempeño del modelo y sirve de punto de partida para cuantificar su eficacia mediante diferentes métricas de evaluación.

Valor real	Positivo	TP=50	FN=10
	Negativo	FP=5	TN=35
		Positivo	Negativo
		Predicción del modelo	

Figura 7: Ejemplo de matriz de confusión para un problema de clasificación binario.
Fuente: Elaboración propia.

2.6.2. Métricas de evaluación

Estas métricas permiten evaluar de manera cuantitativa el desempeño del modelo en distintos aspectos [25]. Cada una enfatiza una perspectiva particular sobre la calidad de las predicciones:

- **Exactitud (Accuracy):** Corresponde a la proporción de predicciones correctas respecto del total de casos evaluados. Su valor puede ser interpretado como la probabilidad de que una clasificación del modelo sea correcta.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$

- **Precisión (Precision):** Mide la proporción de instancias clasificadas correctamente como positivas en relación con todas las instancias predichas como positivas.

$$Precision = \frac{TP}{TP + FP} \quad (14)$$

- **Recall (Sensibilidad):** Indica la proporción de instancias positivas que fueron correctamente identificadas por el modelo.

$$Recall = \frac{TP}{TP + FN} \quad (15)$$

- **F1-Score:** Corresponde a la media armónica entre precisión y recall, equilibrando ambas métricas y ofreciendo una evaluación más robusta en escenarios con clases desbalanceadas.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (16)$$

- **AUC-ROC:** Mide la capacidad global del modelo para discriminar entre clases positivas y negativas, calculando el área bajo la curva (AUC) que relaciona la tasa de verdaderos positivos con la tasa de falsos positivos. Su valor varía entre 0 y 1, donde valores cercanos a 1 reflejan un buen desempeño.

$$AUC = \int_0^1 TPR(FPR^{-1}(x)) dx \quad (17)$$

- **Silhouette Score:** Evalúa la calidad de una partición generada por un algoritmo de clustering. Su valor oscila entre -1 y 1, donde valores cercanos a 1 indican buena separación y cohesión; valores cercanos a 0 reflejan solapamiento entre clústeres; y valores negativos sugieren una asignación incorrecta.

Para cada muestra i , se define:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

donde $a(i)$ es la distancia promedio entre i y los puntos de su mismo clúster, y $b(i)$ es la distancia promedio entre i y los puntos del clúster más cercano distinto del suyo.

2.7. Selección de características (*Feature Selection*)

Es una técnica de reducción de dimensionalidad cuyo objetivo es identificar, dentro de un conjunto amplio de variables, aquel subconjunto más relevante para la tarea de interés [26]. A diferencia de otros métodos de reducción que generan nuevas representaciones de los datos, feature selection mantiene un subconjunto de las variables originales, lo que facilita la interpretación de los resultados y su aplicación práctica. En nuestro contexto, las características corresponden a los ítems de los cuestionarios psicométricos. Reducir el número de ítems presenta ventajas prácticas, tales como disminuir el tiempo de aplicación, reducir la fatiga de los encuestados y favorecer la escalabilidad en estudios poblacionales. No obstante, la selección debe realizarse cuidadosamente para evitar eliminar variables que aportan información diagnóstica esencial.

Diversos enfoques basados en algoritmos de aprendizaje automático permiten identificar los ítems más informativos, ya sea mediante medidas de importancia de variables como en modelos de Random Forest, o mediante técnicas de selección iterativa.

En términos aplicados, la selección de características puede operar no solo reduciendo ítems redundantes, sino también adaptando dinámicamente la aplicación de un cuestionario. Por ejemplo, ítems con alta correlación pueden agruparse y representarse por una sola pregunta, mientras que otros pueden aplicarse de forma condicional, es decir, presentar preguntas adicionales únicamente cuando las respuestas iniciales sugieren la presencia de un síntoma relevante. Así, si un estudiante responde que no presenta problemas de sueño, no resulta

necesario indagar más sobre ese dominio. En cambio, si indica dificultades para dormir, se pueden habilitar ítems adicionales que permitan caracterizar la severidad y frecuencia del problema.

A continuación se presentan los principales métodos de selección de características, agrupados en tres enfoques: métodos de filtro, métodos wrapper y métodos embebidos.

2.7.1. Métodos de Selección de Características

2.7.1.1 Métodos de filtro (Filter Methods)

Los métodos de filtro constituyen un enfoque de selección de características basado en el análisis de propiedades estadísticas e informativas de los datos, sin involucrar directamente un modelo predictivo. Su propósito es identificar aquellas variables que muestran mayor relevancia en relación con la variable objetivo o que aportan información no redundante al conjunto [26]. Entre los procedimientos más comunes se incluyen:

- **Correlación:** Medida estadística que cuantifica el grado y la dirección de la relación lineal entre dos variables [27]. Se expresa mediante el coeficiente de correlación de Pearson (r), donde $cov(X, Y)$ es la covarianza entre las variables X e Y , y σ_X, σ_Y sus desviaciones estándar. Los valores de r varían entre -1 y +1.

$$r = \frac{cov(X, Y)}{\sigma_X \sigma_Y} \quad (18)$$

- **Chi-cuadrado (χ^2):** Evalúa si existe asociación estadísticamente significativa entre un ítem y la clase. Se define como [27]:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (19)$$

Donde O_{ij} representa la frecuencia observada de la categoría del ítem i en la clase j , mientras que E_{ij} corresponde a la frecuencia esperada bajo la hipótesis de independencia entre ambas variables. Aquí, r es el número de categorías del ítem y c el número de clases o grupos de la variable objetivo. Valores altos de χ^2 indican una relación más fuerte entre la característica y la clase.

- **Información mutua:** Cuantifica la cantidad de información compartida entre un ítem y la clase, es decir, cuánto reduce la incertidumbre de una variable al conocer la otra. Una mayor información mutua implica mayor relevancia de la característica para predecir la clase [27].

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) \quad (20)$$

- **Análisis de varianza (ANOVA):** Evalúa si existen diferencias estadísticamente significativas en las medias de un ítem entre distintos grupos de la variable objetivo. Valores de F más altos indican mejor capacidad discriminativa del ítem. Se calcula como [27]:

$$F = \frac{MS_{between}}{MS_{within}} \quad (21)$$

Donde $MS_{between}$ es la media de los cuadrados entre grupos y MS_{within} la media de los cuadrados dentro de los grupos.

- **Feature Agglomeration:** Técnica de agrupamiento jerárquico aplicada sobre variables en lugar de instancias. Define una medida de similitud $d(x_i, x_j)$ entre características, como la correlación o la distancia euclidiana. Este enfoque agrupa ítems altamente similares en clústeres, seleccionando un representante en cada grupo para reducir la dimensionalidad. La distancia euclidiana se define como [24]:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2} \quad (22)$$

2.7.1.2 Métodos wrapper (Wrapper Methods)

Los métodos wrapper son técnicas de selección de características que evalúan directamente el desempeño de un modelo de predicción entrenado sobre distintos subconjuntos de variables, entendidos como combinaciones específicas de características extraídas del subconjunto total [26]. Por ejemplo, si un cuestionario contiene 9 ítems, un subconjunto podría estar formado por 3 de ellos, otro por 4 distintos, y así sucesivamente. El algoritmo compara el rendimiento alcanzado por cada subconjunto y selecciona aquellos que maximizan el desempeño del modelo. Entre las estrategias más utilizadas se encuentran [28]:

- **Forward Selection:** Comienza con un conjunto vacío y agrega en cada iteración la característica que produce la mayor mejora en el rendimiento. En cada paso, el algoritmo evalúa todas las características aún no seleccionadas, incorporando temporalmente cada una al conjunto actual y calculando una métrica de desempeño mediante validación cruzada. La variable que genera la mayor mejora en la métrica elegida se incorpora de manera definitiva, repitiéndose el proceso hasta que no se observe un incremento significativo en el rendimiento o se alcance un número máximo de características.

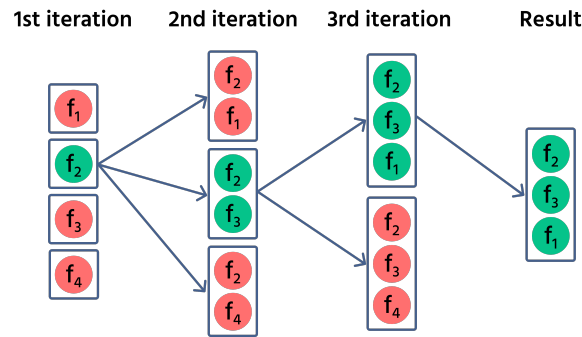


Figura 8: Ejemplo de Forward Selection.
Fuente: [29].

- **Backward Elimination:** El procedimiento inicia con todas las características disponibles y elimina en cada paso la menos relevante, es decir, la que menos aporta a la mejora del rendimiento.

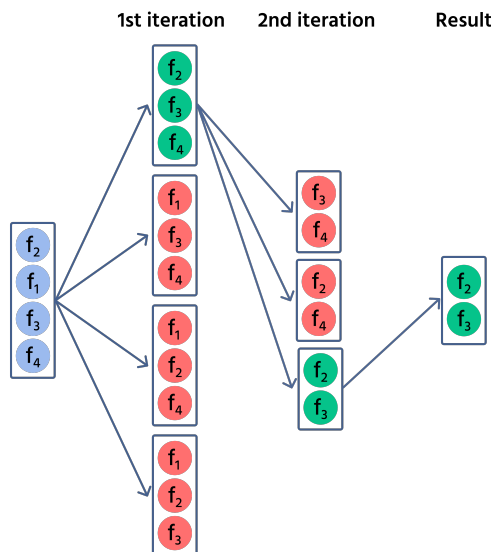


Figura 9: Ejemplo de Backward Elimination.
Fuente: [29].

- **Recursive Feature Elimination (RFE):** Consiste en entrenar un modelo con todas las variables, eliminar la de menor importancia y repetir el procedimiento de manera recursiva hasta obtener el subconjunto de características deseado. En la práctica, el número final de características puede fijarse de antemano como un hiperparámetro o determinarse automáticamente mediante validación cruzada, seleccionando el punto en que el rendimiento del modelo deja de mejorar significativamente.

2.7.1.3 Métodos embebidos (Embedded Methods)

Los métodos embebidos realizan la selección de características de manera simultánea al entrenamiento del modelo predictivo. A diferencia de los métodos de filtro o wrapper, la selección se integra en el propio proceso de aprendizaje: el modelo ajusta sus parámetros no solo para optimizar la función de pérdida, sino también para controlar su complejidad y resaltar la relevancia de las variables [26]. En este enfoque destacan dos grandes mecanismos:

1. **Regularización:** Consiste en añadir términos de penalización a la función de pérdida con el objetivo de limitar la complejidad del modelo y evitar el sobreajuste. Formalmente, para un modelo lineal, donde se tiene que cada variable de entrada x_j se asocia con un coeficiente β_j que representa el peso o contribución de esa característica en la predicción, existen las siguientes variantes:

- **Regularización L_1 (Lasso):** Algunos coeficientes se reducen exactamente a cero, eliminando implícitamente las variables irrelevantes [15].

$$L(\beta) = \sum_{i=1}^n l(\beta^{(i)}, y^{(i)}) + \lambda \sum_{j=1}^p |\beta_j| \quad (23)$$

En esta expresión, n representa el número de observaciones del conjunto de datos y p el número total de características o variables consideradas en el modelo. El término de penalización $\lambda \sum_{j=1}^p |\beta_j|$ se aplica a todos los coeficientes β_j , forzando a algunos a ser exactamente cero cuando λ es suficientemente grande. De este modo, la regularización actúa sobre todas las variables, pero sólo algunas son retenidas al minimizar la función de pérdida.

2. **Importancia de variables:** Técnica que permite estimar el impacto de cada variable en el desempeño del modelo.

- **Valores de Shapley:** Basados en la teoría de juegos cooperativos, miden la contribución de cada característica en la predicción final. Dada una predicción $f(x)$, un conjunto de características N y un subconjunto S , el valor de Shapley para una característica j se define como [30]:

$$\phi_j(f) = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|! (|N| - |S| - 1)!}{|N|!} [f(S \cup \{j\}) - f(S)] \quad (24)$$

donde N representa el conjunto total de características, S un subconjunto de características sin incluir j , y $f(S)$ la predicción del modelo cuando solo se consideran las variables en S . El término $f(S \cup \{j\}) - f(S)$ expresa el cambio en la predicción al incorporar la característica j , ponderado por el número de posibles combinaciones en las que j podría ser añadida. De esta forma, $\phi_j(f)$ cuantifica la contribución promedio de la variable j al resultado final del modelo.

2.8. Reducción de dimensionalidad

La reducción de dimensionalidad es un proceso cuyo objetivo es representar un conjunto de datos en un espacio de menor número de dimensiones, preservando la mayor cantidad posible de la información contenida en las variables originales. Formalmente, consiste en encontrar una función de transformación $f : \mathbb{R}^n \rightarrow \mathbb{R}^k, k < n$, que proyecte los datos desde un espacio de alta dimensión $\mathbb{R}^n$ a un subespacio de dimensión reducida $\mathbb{R}^k$, de manera que la estructura esencial se mantenga [31].

A diferencia de la selección de características, que conserva un subconjunto de las variables originales, la reducción dimensional generalmente genera nuevas características transformadas o combinaciones de las originales. Entre los enfoques más representativos destacan técnicas clásicas como PCA (Análisis de Componentes Principales), que construye variables ortogonales, denominadas componentes, que explican la mayor varianza de los datos, y métodos modernos basados en redes neuronales, como los autoencoders, que aprenden representaciones comprimidas en su espacio latente. El espacio latente se refiere a la representación de los datos en un número reducido de dimensiones que captura sus factores subyacentes más relevantes. Este proceso puede ayudar tanto a eliminar ruido y redundancia como a facilitar la visualización de datos complejos en pocas dimensiones.

En el contexto de cuestionarios psicométricos, la reducción de dimensionalidad puede utilizarse para identificar variables latentes que capturen patrones comunes entre ítems altamente correlacionados. Por ejemplo, los nueve ítems del PHQ-9 podrían sintetizarse en dos o tres factores subyacentes que representen constructos como “Estado de ánimo depresivo”, facilitan así tanto el análisis como la predicción.

2.8.1. Feature Agglomeration

La técnica de *Feature Agglomeration* corresponde a un método jerárquico de reducción de dimensionalidad que agrupa variables o características similares en función de su correlación o distancia semántica, reduciendo la redundancia informativa del conjunto de datos.

Formalmente, el método aplica un procedimiento de *agglomerative clustering* sobre las columnas de la matriz de características $X \in \mathbb{R}^{n \times d}$, en lugar de sobre las filas (instancias), con el fin de construir una jerarquía de fusiones de variables hasta obtener un número reducido de componentes $k < d$. Cada componente resultante puede representarse como el promedio o combinación lineal de las variables originales pertenecientes a su clúster.

En el contexto de los cuestionarios psicométricos, esta técnica permite agrupar ítems con alto grado de correlación semántica o estadística, reduciendo el número total de preguntas sin perder representatividad del constructo evaluado. Por ejemplo, ítems como “duermo mal” y “me cuesta dormir” podrían fusionarse en una sola dimensión latente asociada a la “calidad del sueño”.

2.8.2. Uniform Manifold Approximation and Projection (UMAP)

El algoritmo UMAP es una técnica de reducción de dimensionalidad no lineal basada en la teoría de variedades y en la construcción de grafos de similitud. Su propósito es preservar la estructura topológica del espacio original en una representación de menor dimensión, manteniendo las relaciones de vecindad entre las observaciones. Formalmente, UMAP modela los datos como un grafo ponderado $G = (V, E)$, donde los vértices V representan las instancias y las aristas E se definen a partir de una función de conectividad local. Posteriormente, busca una proyección $Y \in \mathbb{R}^{n \times d'}$, con $d' \ll d$, que minimice la divergencia de Kullback-Leibler entre las distribuciones de conectividad de los espacios original y reducido, obteniendo así una representación topológicamente coherente y de menor dimensión.

Los principales *hiperparámetros* utilizados en UMAP son los siguientes:

- **n_neighbors**: Número de vecinos considerados para construir el grafo de similitud.
- **min_dist**: Distancia mínima entre puntos en el espacio reducido.
- **metric**: Métrica de distancia utilizada para calcular la similitud entre puntos.
- **n_components**: Dimensión del espacio resultante.
- **random_state**: Semilla para el generador aleatorio, que permite reproducibilidad.

UMAP destaca por su capacidad para preservar tanto la estructura local como la global de los datos, siendo computacionalmente eficiente y escalable a grandes volúmenes de información. Además, su combinación con técnicas de agrupamiento como *K-Means* o *Feature Agglomeration* permite detectar estructuras semánticas latentes con alta coherencia conceptual. Sin embargo, su principal limitación radica en que la interpretación de las dimensiones reducidas no siempre es directa y depende de la configuración de sus hiperparámetros, los cuales determinan el equilibrio entre la preservación topológica del espacio original y la claridad visual de la proyección obtenida.

A continuación, la Figura 10 ilustra un ejemplo del funcionamiento de UMAP, mostrando cómo un conjunto de datos en tres dimensiones (3D) puede ser proyectado a un espacio bidimensional (2D) manteniendo la estructura de los datos, con los diferentes clusters, identificados por color, separados y definidos en la nueva proyección.

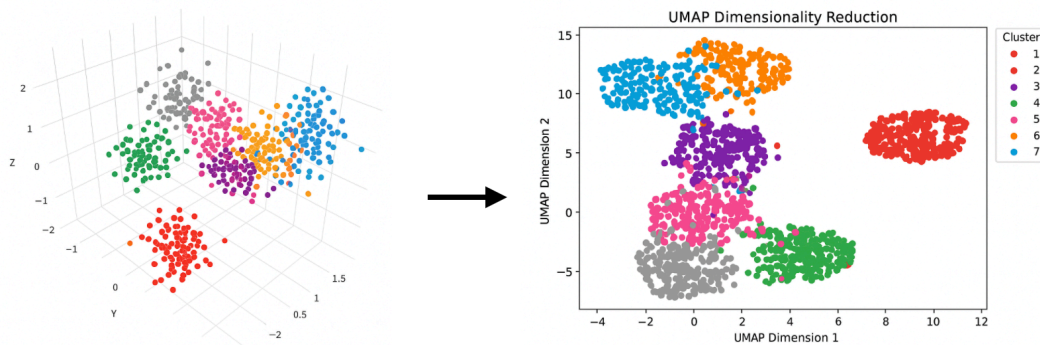


Figura 10: Ejemplo del funcionamiento de UMAP (3D→2D).
Fuente: [32].

2.8.3. Autoencoders

Un autoencoder es un tipo de red neuronal artificial no supervisada, diseñada con el objetivo fundamental de aprender una representación eficiente y comprimida de los datos de entrada, a la cual se le denomina código o espacio latente. El espacio latente es un espacio vectorial de menor dimensión que captura las características más relevantes o informativamente significativas y la estructura subyacente de los datos de entrada, descartando la información redundante. La distinción entre información esencial y redundante se define de manera implícita durante el entrenamiento: el modelo aprende a conservar las propiedades que permiten reconstruir los datos originales con la menor pérdida posible, mientras que los patrones o variaciones que no contribuyen significativamente a esa reconstrucción tienden a ser descartados. Este proceso se guía por una función de pérdida que mide la diferencia entre los datos de entrada y los datos reconstruidos, ajustando los pesos de la red para minimizar dicho error.

Se puede entender como una versión resumida de los datos originales, donde cada punto en este espacio representa una combinación significativa de las características principales de los datos [33]. Su principal característica radica en que el objetivo de entrenamiento es la reconstrucción de sus propios datos de entrada, minimizando la pérdida de información en el proceso de comprensión. A diferencia de otras arquitecturas, un autoencoder no tiene una tarea de predicción de etiquetas, sino que se enfoca en la reducción de dimensionalidad y la extracción de características intrínsecas de los datos.

La arquitectura de un autoencoder se compone de dos partes funcionales interconectadas: el encoder y el decoder. El **encoder** es responsable de mapear los datos de entrada, $x \in \mathbb{R}^n$, a una representación de menor dimensión en el espacio latente, $h \in \mathbb{R}^d$, donde $d < n$. Formalmente, esta transformación es como una función de codificación $h = f(W_1x + b_1)$, donde W_1 es la matriz de pesos y b_1 es el vector de sesgo del encoder. A continuación, el **decoder** toma esta representación latente o codificada, h , y la utiliza para reconstruir la entrada

original, produciendo una salida $x' \in \mathbb{R}^n$. Esta fase de decodificación se representa como $x' = g(W_2h + b_2)$, donde W_2 y b_2 son los parámetros del decoder. Los parámetros son los valores numéricos internos de la red que el modelo aprende y ajusta durante el entrenamiento para realizar la transformación de datos de manera óptima [34].

El entrenamiento del autoencoder se lleva a cabo ajustando sus parámetros W_1 , b_1 , W_2 , b_2 , con el fin de minimizar la función de pérdida de reconstrucción, $L(x, x')$. Esta función mide la diferencia entre los datos de entrada y su versión reconstruida. Una métrica comúnmente utilizada es el error cuadrático medio, definida anteriormente en la Sección 2.5.4.4. Al minimizar esta pérdida, el autoencoder captura las características esenciales de los datos, eliminando la redundancia y el ruido, y almacenándolas en el código de baja dimensión h . Cuando la reconstrucción es satisfactoria, el modelo ha logrado una representación latente que encapsula la variabilidad más significativa de los datos [33][34].

La Figura 11 ilustra de manera esquemática la arquitectura básica de esta red. En la capa de entrada, se observa el vector de datos originales X . A través del encoder, estos datos se comprimen en la capa oculta, representada por el vector h . Posteriormente, el decoder utiliza esta representación latente para generar la salida reconstruida X' en la capa de salida. Las conexiones entre capas están mediadas por las matrices de pesos W_1 , para el encoder, y W_2 para el decoder.

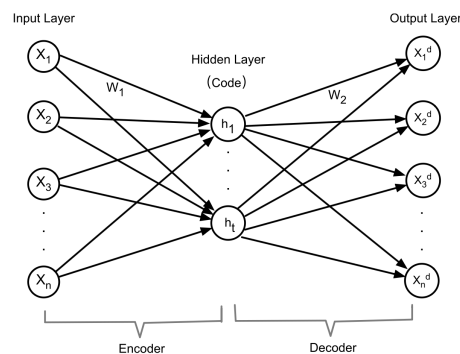


Figura 11: Diagrama de la estructura de un Autoencoder.
Fuente: [34].

2.8.4. Variational Autoencoders (VAE)

Los Autoencoders Variacionales (VAE) son una extensión de los autoencoders tradicionales que incorporan principios de inferencia variacional y modelos generativos dentro de su arquitectura [33]. A diferencia de un autoencoder convencional, cuyo objetivo es aprender una representación comprimida y reconstruir los datos de entrada, el VAE introduce una formulación estadística que permite modelar explícitamente la distribución latente de los datos [35].

En un VAE, el encoder no asigna un punto fijo en el espacio latente para cada instancia, sino una distribución probabilística, generalmente Gaussiana, caracterizada por parámetros de media $\mu(x)$ y varianza $\sigma^2(x)$ [33]. Durante el muestreo, cada dato se codifica como un vector latente $z \sim N(\mu, \sigma^2)$, lo que introduce continuidad y permite generar nuevas instancias plausibles a partir del espacio latente.

El entrenamiento de un VAE se basa en optimizar la llamada Evidence Lower Bound [24]:

$$\mathcal{L}(x) = \mathbb{E}_{q(z|x)}[\log p(x|z)] - D_{KL}(q(z|x)||p(z)) \quad (25)$$

donde el primer término $\mathbb{E}_{q(z|x)}[\log p(x|z)]$ corresponde al error de reconstrucción [24], que mide qué tan bien el decodificador reproduce la entrada original, y el segundo término D_{KL} es la divergencia Kullback-Leibler [24], que regulariza las distribuciones latentes $q(z|x)$ para aproximarlas a una distribución gaussiana estándar $p(z) = N(0, 1)$.

Este doble objetivo permite que el VAE no solo aprenda representaciones comprimidas, sino que también construya un espacio latente estructurado, continuo y regularizado, donde instancias similares en los datos originales se ubican cercanas entre sí.

Entre las ventajas de los VAE destaca su capacidad de capturar relaciones no lineales complejas en los datos, lo que los convierte en una herramienta poderosa para la reducción de dimensionalidad. En el ámbito de cuestionarios psicométricos, esto permite representar ítems correlacionados como combinaciones de dimensiones latentes, identificando redundancias y agrupando preguntas que miden constructos similares.

La Figura 12 muestra la diferencia entre un autoencoder simple, que mapea instancias a puntos fijos en el espacio latente, y un VAE, que los representa como distribuciones, garantizando continuidad y capacidad generativa.

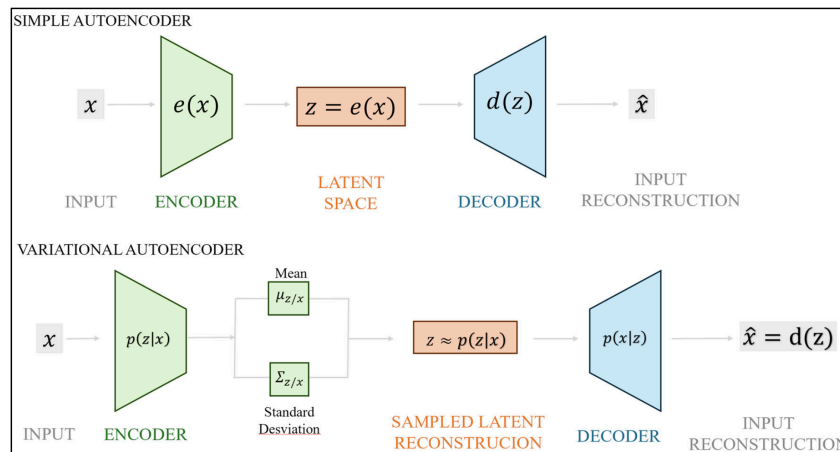


Figura 12: Autoencoder simple vs. Variational Autoencoder.

Fuente: [35].

2.9. Análisis semántico de ítems

Además de evaluar redundancias estadísticas entre variables, resulta fundamental analizar los ítems desde una perspectiva semántica, es decir, considerar el significado del texto de cada pregunta.

En los instrumentos psicométricos, diferentes ítems pueden expresar constructos⁵ conceptualmente cercanos, generando solapamientos en el contenido. Por ejemplo, el ítem “¿Te sientes triste y sin esperanzas?” y “¿Has perdido el interés en actividades que antes disfrutabas?”, abordan dimensiones similares del estado de ánimo depresivo, aunque con formulaciones distintas.

El propósito del análisis semántico es cuantificar estas similitudes de significado entre ítems, lo que permite identificar redundancias conceptuales y aportar criterios adicionales para la optimización de cuestionarios. Para ello se emplean técnicas de representación vectorial del lenguaje, conocidas como *embeddings* de texto, que proyectan las palabras o frases en un espacio matemático donde la proximidad refleja la cercanía semántica.

2.9.1. Embeddings de texto

Los *embeddings* de texto son representaciones vectoriales que permiten codificar palabras, frases o documentos en un espacio matemático de dimensión reducida, de modo que la proximidad geométrica entre vectores refleje la similitud semántica entre los textos [24][36]. Formalmente, un *embedding* es una función de mapeo:

$$f : T \rightarrow \mathbb{R}^d \quad (26)$$

donde T es el conjunto de secuencias de texto y $\mathbb{R}^d$ es el espacio latente de dimensión d , con $d \ll |V|$ y $|V|$ el tamaño del vocabulario.

A diferencia de las representaciones tradicionales basadas en conteos (como *bag-of-words* o TF-IDF), los *embeddings* capturan información contextual y semántica, permitiendo que expresiones con significados cercanos se ubiquen próximas en el espacio vectorial. Por ejemplo, los vectores de “triste” y “deprimido” estarán más cercanos entre sí que al vector de “alegría”. De manera ilustrativa, una palabra como “triste” podría representarse en un espacio de 5 dimensiones mediante un vector como:

$$\text{triste} \rightarrow [0.2, -0.1, 0.4, 0.7, -0.3] \quad (27)$$

A continuación se revisan diferentes enfoques para generar *embeddings*.

⁵Un constructo es una variable abstracta o concepto teórico, como depresión o ansiedad, que no se observa directamente, pero que se mide a través de ítems o indicadores en un instrumento psicométrico.

2.9.1.1 Word2Vec

Es un método de representación distribuida de palabras desarrollado por Mikolov et al. (2013) [36], que permite codificar términos del lenguaje en vectores numéricos dentro de un espacio de baja dimensionalidad. Su objetivo es ubicar palabras con significados similares en posiciones cercanas dentro del espacio vectorial.

El modelo se entrena mediante dos arquitecturas principales:

- **Continuous Bag of Words (CBOW):** Predice una palabra a partir de su contexto, es decir, del conjunto de palabras que la rodean.
- **Skip-Gram:** Predice el contexto de una palabra a partir de la palabra central.

En ambos casos, el modelo ajusta los pesos de manera que la similitud entre vectores refleje la probabilidad de coocurrencia en el lenguaje natural. Para cuantificar dicha similitud, se utiliza la similitud coseno, definida como:

$$\text{sim}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \quad (28)$$

donde $A \cdot B$ corresponde al producto punto entre los vectores A y B , y $\|A\|$, $\|B\|$ son sus magnitudes euclidianas⁶. Esta medida toma valores entre -1 y 1, siendo cercanos a 1 cuando los vectores apuntan en direcciones similares (alta similitud semántica) y próximos a 0 cuando son ortogonales (sin relación).

En la Figura 13 se muestra un ejemplo de cómo Word2Vec organiza palabras en un espacio vectorial de tres dimensiones. En este caso, “triste” y “deprimido” aparecen cercanas, reflejando su similitud semántica, mientras que “feliz” se ubica con mayor distancia, reflejando su menor similitud semántica, pero dentro del mismo espacio que “triste”, ya que ambos corresponden a un estado de ánimo.

En el contexto de instrumentos psicométricos, Word2Vec permite representar cada palabra como un vector que captura su significado en función de las palabras que lo componen, posibilitando detectar similitudes semánticas entre preguntas con distinta redacción pero que evalúan el mismo constructo.

⁶Las magnitudes euclidianas corresponden a la longitud del vector en el espacio, calculada como la raíz cuadrada de la suma de los cuadrados de sus componentes.

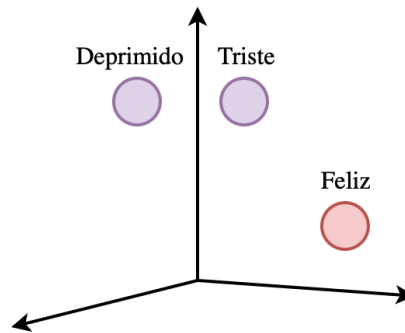


Figura 13: Representación de la ubicación de los embeddings de las palabras “triste”, “deprimido” y “alegría” en un espacio vectorial de tres dimensiones.

Fuente: Elaboración propia.

2.9.1.2 BERT

BERT (*Bidirectional Encoder Representations from Transformers*), desarrollado por Devlin et al. (2019) [36], es un modelo de lenguaje basado en la arquitectura Transformer que genera representaciones contextuales de palabras y frases. A diferencia de Word2Vec, que asigna un único vector fijo a cada palabra independientemente de su contexto, BERT produce *embeddings* distintos según el entorno en el que aparece la palabra, lo que permite capturar matices semánticos más complejos.

Para comprender su funcionamiento, es necesario introducir algunos conceptos claves [36]:

- **Tokenización:** El texto se divide en unidades mínimas llamadas tokens. En BERT, además de palabras y subpalabras, se incluyen tokens especiales como *[CLS]*, utilizado para representar la secuencia completa, y *[SEP]* para separar oraciones cuando se analizan en pares.
- **Embeddings de entrada:** Cada *token* se convierte en un vector que combina tres tipos de información: *embedding* de palabra, *embedding* posicional, el cual indica la posición del *token* en la secuencia, y, *embedding* de segmento, que distingue si un *token* pertenece a la *Sentence A* o *Sentence B*.
- **Mecanismo de autoatención:** Núcleo de la arquitectura Transformer que permite que cada token se relacione con todos los demás de la secuencia, asignando pesos según su relevancia. De esta manera, BERT captura dependencias a largo plazo y relaciones bidireccionales entre palabras.
- **Representaciones contextuales:** El resultado del proceso anterior es que cada palabra obtiene un *embedding* distinto según el contexto en el que aparece. Por ejemplo, la palabra “banco” tendrá una representación diferente si se refiere a una institución financiera o a un asiento.

El entrenamiento de BERT se realiza mediante dos tareas principales [36]:

- **Masked Language Modeling (MLM):** Predicción de tokens ocultos en una oración, utilizando el contexto bidireccional.
- **Next Sentence Prediction (NSP):** Determina si una oración sigue lógicamente a otra, lo que permite a BERT modelar relaciones entre frases.

Esto se refleja en la Figura 14, donde se muestra el proceso aplicado a un par de oraciones (*Sentence A* y *Sentence B*). Tras la tokenización, ambas oraciones son procesadas por la arquitectura Transformer de BERT, generando *embeddings* contextuales. Posteriormente, se obtiene una representación conjunta (*embedding z*) que puede emplearse en tareas como el cálculo de similitud semántica entre frases, usualmente a través de una función de activación como la sigmoide y una capa de comparación.

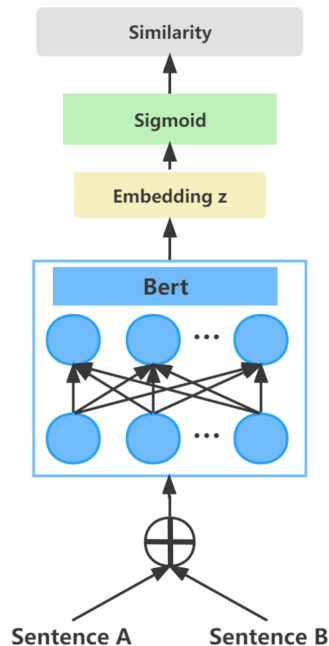


Figura 14: Esquema de BERT aplicado a pares de oraciones.

Fuente: [37].

De esta forma, BERT aprende relaciones profundas entre palabras y frases, lo que lo hace adecuado para tareas de comprensión semántica compleja. En el ámbito de cuestionarios, BERT puede capturar matices en la formulación de ítems que Word2Vec no logra distinguir, como diferencias entre “ánimo bajo” y “desmotivación persistente” en contextos distintos.

2.9.1.3 Sentence-BERT

Sentence-BERT (SBERT) es una variante de BERT propuesta por Reimers y Gurevych (2019) [36], específicamente diseñada para generar representaciones vectoriales de oraciones completas que puedan compararse de manera eficiente en términos de similitud semántica. Mientras que BERT produce *embeddings* a nivel de tokens que requieren cálculos de similitud costosos, SBERT emplea una arquitectura siamés/triplete que permite mapear directamente oraciones completas a un espacio vectorial denso y de dimensión fija, optimizado para cálculos de similitud.

Una arquitectura siamés consiste en dos redes neuronales idénticas, con parámetros compartidos, que procesan entradas distintas en paralelo. Su objetivo es aprender un espacio de representación donde ejemplos similares queden cercanos y los diferentes se ubiquen más alejados. Cuando el diseño se amplía a tres redes, se denomina arquitectura triplete, y se entrena de modo que la oración de referencia y la oración con significado similar estén más próximos que con el de diferente significado.

En el esquema de la Figura 15, dos redes BERT paralelas procesan oraciones diferentes, y a sus salidas se aplica un mecanismo de agregación como *mean pooling*, que consiste en promediar los vectores de todos los tokens de la oración, para obtener un único *embedding* por cada oración. Posteriormente, se mide la similitud entre los vectores generados mediante funciones como la similitud coseno o la distancia euclidiana. Durante el entrenamiento, el modelo minimiza la distancia entre frases semánticamente equivalentes y maximiza la separación entre aquellas que no guardan relación, logrando así un espacio vectorial en el que la proximidad refleja la cercanía semántica.

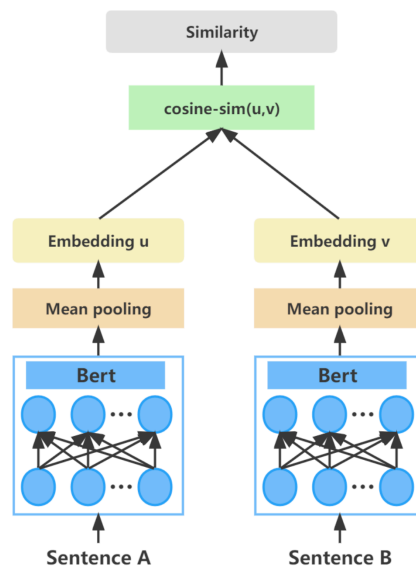


Figura 15: Esquema de SBERT aplicado a pares de oraciones.

Fuente: [37].

En el contexto de cuestionarios psicométricos, SBERT permite representar cada ítem como un vector de frase que captura su significado global. Esto posibilita cuantificar con mayor precisión la redundancia semántica entre ítems completos, más allá de las palabras individuales, mejorando la detección de preguntas conceptualmente equivalentes.

2.10. Aplicaciones previas de IA en salud mental

En la actualidad, la salud mental se encuentra en un punto de inflexión, impulsado por la creciente integración de la inteligencia artificial y el aprendizaje automático. Esta sinergia no solo ofrece nuevas herramientas para el diagnóstico y la evaluación, sino que también está redefiniendo el tratamiento y la intervención. Diversos estudios recientes demuestran la eficacia de estas tecnologías, abordando desafíos que van desde la diferenciación precisa de trastornos complejos hasta la optimización de cuestionarios mediante técnicas de IA, con el objetivo de reducir el número de ítems o integrar varios instrumentos en versiones más breves que los originales.

La aplicación de la IA en la salud mental ha permitido optimizar la evaluación y el diagnóstico de diversas maneras. Un ejemplo destacado es el uso de técnicas avanzadas de selección de características. Erguzel *et al.* (2015) emplearon registros de coherencia en medidas que reflejan el grado de sincronización entre las señales eléctricas captadas en distintas regiones cerebrales, de pacientes con Trastorno Depresivo Mayor (MDD) y Trastorno Bipolar (BD), inicialmente conformados por 48 variables [38]. Para reducir la dimensionalidad, aplicaron un algoritmo de optimización por colonia de hormigas mejorado (IACO), donde cada solución candidata representaba un subconjunto de características. La calidad de cada subconjunto se evaluaba mediante el desempeño de un clasificador, de modo que las variables con mayor aporte discriminativo acumulaban mayor probabilidad de ser seleccionadas en las siguientes iteraciones. Este proceso permitió identificar un conjunto óptimo de 22 características relevantes alcanzando una precisión del 80,19 %, superando a métodos alternativos como ACO estándar y algoritmos genéticos.

Otra aproximación relevante es el uso de clasificadores híbridos para mejorar la predicción de la salud mental. Yadav y Bokhari (2024) plantearon el desarrollo de modelos híbridos de clasificación con el objetivo de mejorar la predicción de problemas de salud mental a partir de datos clínicos y de instrumentos psicométricos [30]. El algoritmo comienza con una fase de preprocesamiento e ingeniería de características, realizando una limpieza de datos y según las variables derivadas generadas, se seleccionaron aquellas que aportaban mayor información relevante. Posteriormente, entrenaron modelos de referencia individuales: Árboles de Decisión, k-Vecinos Más Cercanos (kNN), Redes Neuronales y Random Forest, evaluando su rendimiento de manera independiente a través de dos enfoques.

En el primer enfoque, el modelo DT+kNN se organizó de forma secuencial: el Árbol de decisión se utilizó para realizar una partición inicial del espacio de características, simplificando la estructura de los datos, y sobre cada subconjunto resultante se aplicó kNN para identi-

ficar similitudes locales entre instancias. De este modo, el árbol aportó interpretabilidad y reducción de la complejidad, mientras que kNN se encargó de refinar la clasificación dentro de cada grupo. Por otro lado, en el segundo enfoque, el modelo RF+NN integró un Random Forest como primera capa de selección y ponderación de variables, y sobre sus salidas se entrenó una Red Neuronal, capaz de capturar relaciones no lineales más complejas y de alto nivel. Así, el Random Forest funcionó como un generador de representaciones enriquecidas y robustas, que la red neuronal aprovechó para optimizar la clasificación final.

Al comparar los resultados, los modelos híbridos superaron de manera consistente a sus contrapartes individuales. El enfoque DT+kNN alcanzó una precisión del 86,69 %, mientras que el modelo RF+NN llegó al 93,54 %, confirmando que la combinación estratégica de algoritmos complementarios permite capturar patrones tanto lineales como no lineales presentes en los datos y ofrecer un camino intermedio entre la interpretabilidad y la alta precisión.

En paralelo, se han explorado estrategias para optimizar directamente cuestionarios psicométricos. Glavin *et al.* (2023) exploraron el uso de algoritmos de machine learning para optimizar el cuestionario PHQ-9 y generar versiones ultra breves destinadas al tamizaje de depresión [39]. El objetivo fue comparar combinaciones de solo dos ítems del cuestionario completo con el tradicional PHQ-2, que incluye únicamente las preguntas sobre “ánimo deprimido” y “pérdida de interés o placer”. Para ello, los autores trabajaron con bases de datos clínicas en las que el PHQ-9 había sido previamente administrado a pacientes, y realizaron un análisis sistemático de todas las posibles combinaciones de pares de ítems. Cada combinación se evaluó entrenando clasificadores de aprendizaje automático que estimaban la presencia de depresión clínica según criterios diagnósticos estandarizados. Los modelos se validaron mediante procedimientos de validación cruzada y pruebas en conjuntos externos de datos, lo que permitió medir de manera robusta el desempeño de cada pareja de ítems. Los resultados mostraron que varias combinaciones alternativas superaban ligeramente el rendimiento del PHQ-2. La más destacada fue la formada por los ítems de “ánimo deprimido” y “baja energía”, que alcanzó un AUC $\approx 0,954$, mientras que el PHQ-2 tradicional obtuvo un AUC $\approx 0,942$. Este hallazgo indica que es posible acortar instrumentos como el PHQ-9 manteniendo un nivel aceptable de sensibilidad, siempre que los ítems sean seleccionados estratégicamente.

De manera similar, Wang *et al.* (2024) desarrollaron versiones abreviadas del cuestionario GAD-7 utilizando el algoritmo RiskSlim, con el propósito de obtener instrumentos más breves pero igualmente válidos para la detección de ansiedad en poblaciones multiculturales [40]. El estudio inicia a partir de bases de datos clínicas internacionales que incluían respuestas completas al GAD-7, lo que permitió evaluar el desempeño de modelos en distintos contextos culturales y lingüísticos. El procedimiento consistió primero en entrenar el modelo RiskSlim, un método de aprendizaje supervisado que construye reglas de puntuación escasas y lineales, optimizadas para obtener clasificadores simples y fácilmente interpretables. El algoritmo fue aplicado de forma iterativa para generar y evaluar múltiples versiones reducidas del cuestionario, que incluían entre dos y seis ítems seleccionados en función de su capacidad para discriminar la presencia de ansiedad clínica. Cada versión abreviada fue validada

internamente mediante validación cruzada y externamente en bases de datos independientes. Los indicadores de desempeño (AUC, sensibilidad y especificidad) se compararon con los obtenidos al aplicar GAD-7 completo y con la escala ultrabreve GAD-2, recomendándose finalmente la versión de tres ítems (GAD-7R3) por su equilibrio entre brevedad y exactitud diagnóstica, mostrando una generalización robusta en todos los países evaluados y superando al GAD-2.

Complementariamente, otros estudios han utilizado algoritmos evolutivos para abreviar escalas específicas. En el caso del Inventario de Mindfulness para el Deporte, se implementaron algoritmos genéticos (GA) como estrategia evolutiva para reducir la longitud de la escala original de 95 ítems [41]. La propuesta inicia con la generación de una población inicial de posibles subconjuntos de ítems, representados como cadenas binarias donde un “1” indicaba que un ítem era retenido y un “0” que era descartado. Cada subconjunto fue evaluado mediante una función de aptitud que combinaba indicadores de validez factorial, consistencia interna (alfa de Cronbach) y capacidad de explicar la varianza total del constructo. Los subconjuntos con mayor puntuación se seleccionaban para reproducirse mediante operadores de cruce y mutación, generando nuevas combinaciones de ítems en las siguientes generaciones. Iteración tras iteración, el algoritmo aprende a conservar aquellos ítems que aportaban mayor información psicométrica y eliminar aquellos redundantes o con baja carga factorial. Tras converger, el GA identificó una versión reducida de 37 ítems, que mantuvo niveles equivalentes de fiabilidad y validez en comparación con la escala completa, confirmando que los algoritmos evolutivos permiten optimizar cuestionarios extensos preservando su rigor científico.

Más allá de la reducción de instrumentos estáticos, la administración adaptativa y la combinación de cuestionarios apoyadas en inteligencia artificial ofrecen un gran potencial. Colledani *et al.* (2025) desarrollaron cuestionarios adaptativos para la evaluación de depresión y ansiedad, con el propósito de reducir la carga de aplicación manteniendo la validez diagnóstica [42]. El enfoque se inicia a partir de las versiones completas de los cuestionarios y utiliza árboles de decisión como motor de selección dinámica de ítems. En este enfoque, cada nodo del árbol correspondía a una pregunta, y las ramas definían las posibles rutas de acuerdo con la respuesta del participante. El algoritmo fue entrenado con datos clínicos para identificar cuáles ítems tenían mayor capacidad predictiva de la severidad, de manera que fueran presentados primero. Si la respuesta indicaba ausencia de síntomas en un dominio específico, el árbol “podaba” automáticamente las preguntas subsecuentes relacionadas con ese dominio. En cambio, si la respuesta revelaba la presencia de un síntoma, se habilitaban ítems adicionales que permitían caracterizar su intensidad y frecuencia. La lógica adaptativa se evaluó comparando los resultados obtenidos con la aplicación completa de los cuestionarios. Los autores comprobaron que, en promedio, los cuestionarios adaptativos requerían menos de la mitad de los ítems originales para llegar a estimaciones de severidad equivalentes, lo que demuestra que el uso de árboles de decisión no solo reduce el número de preguntas administradas, sino que también personaliza la experiencia de cada encuestado. Este enfoque ofrece así una vía metodológica eficaz para diseñar cuestionarios inteligentes que equilibran eficiencia y precisión en la evaluación clínica.

En la misma línea, Kuleindiren *et al.* (2022) diseñaron una aplicación móvil que integró ítems de los cuestionarios PHQ-9 y GAD-7 en un solo instrumento adaptado para tamizaje digital [43]. El objetivo fue evaluar si una combinación de preguntas seleccionadas estratégicamente podía mantener la capacidad diagnóstica de ambos instrumentos, reduciendo la redundancia y mejorando la experiencia de uso en entornos móviles. Para ello, los autores realizaron un análisis de los ítems más representativos de cada escala, considerando tanto su peso estadístico en la predicción de los diagnósticos como su relevancia clínica según expertos en salud mental. El sistema desarrollado presentaba primero un conjunto base de preguntas comunes a ambos cuestionarios, y, de acuerdo con las respuestas, habilitaba ítems adicionales específicos de depresión o ansiedad. Esta estructura permitió integrar dominios de los dos instrumentos en un flujo único, evitando la aplicación duplicada de preguntas con contenido similar. Posteriormente, el rendimiento del cuestionario combinado fue evaluado con modelos de machine learning entrenados sobre las respuestas de muestras clínicas y comparado con los resultados obtenidos al aplicar PHQ-9 y GAD-7 de manera independiente, así como con las versiones ultrabreves PHQ-2 y GAD-2. Los resultados mostraron que el cuestionario combinado alcanzaba un nivel de precisión equivalente al de los instrumentos completos aplicados por separado, pero con una administración más rápida y una mejora significativa frente a las escalas ultrabreves. De esta manera, el estudio evidencia que la integración de ítems mediante IA y diseño adaptativo puede generar cuestionarios digitales más eficientes y con un desempeño clínico robusto.

Este enfoque adaptativo ha sido explorado en mayor profundidad por Tallroth y Ålander (2016) con un enfoque híbrido para la reducción de ítems en cuestionarios de autoinforme de salud mental, combinando Random Forest (RF) con Algoritmos Genéticos (GA) [44]. El estudio considera inicialmente un cuestionario de 160 ítems administrado a población general, con el objetivo de identificar un subconjunto reducido que mantuviera un poder predictivo elevado. En primer lugar, Random Forest se utilizó para calcular la importancia de cada ítem, empleando medidas como la disminución de la impureza y la reducción del error de clasificación cuando la variable era incluida en los árboles. Estas puntuaciones de importancia se incorporaron como criterios dentro del algoritmo genético, que operaba sobre representaciones binarias de los ítems (1 = conservar, 0 = eliminar). A través de iteraciones sucesivas, el GA generó poblaciones de posibles subconjuntos de preguntas, los evaluó en función de su capacidad predictiva y psicométrica, y aplicó operadores de cruce y mutación para refinar las soluciones en cada generación. Como resultado, el cuestionario fue reducido de 160 a alrededor de 12 ítems, conservando una capacidad de predicción elevada y comparable al instrumento completo.

En el contexto universitario, Tabares *et al.* (2024) aplicaron modelos de machine learning para el tamizaje simultáneo de múltiples trastornos en población juvenil, integrando datos de los cuestionarios PHQ-9 y GAD-7 junto con variables psicosociales y contextuales [45]. La muestra incluyó jóvenes en edad escolar y universitaria, lo que permitió evaluar tanto síntomas clínicos como factores de riesgo asociados a su entorno social y familiar. El procedimiento inicia con la recopilación y preprocesamiento de los datos, que incluyó la normalización de escalas y la codificación de variables categóricas. Posteriormente, se entrenaron distintos

clasificadores supervisados para detectar síntomas de depresión, ansiedad y otros trastornos, siendo Random Forest el algoritmo que alcanzó el mejor desempeño. Este modelo construyó múltiples árboles de decisión a partir de subconjuntos aleatorios de características y observaciones, y combinó sus predicciones para mejorar la robustez y reducir el sobreajuste. Random Forest logró una precisión del 91 % en la detección de ansiedad, además de un rendimiento elevado en la identificación de depresión y otros problemas psicológicos. El análisis de importancia de variables reveló que, junto con los ítems de los cuestionarios clínicos, tenían un peso significativo factores psicosociales como el nivel educativo de los padres y el consumo de alcohol, mostrando que la combinación de instrumentos psicométricos estandarizados con variables contextuales permite construir modelos predictivos más completos y precisos.

Finalmente, Magalhães *et al.* (2025) evaluaron a estudiantes de psicología en Portugal mediante cuestionarios de burnout académico, con el objetivo de predecir el riesgo de este síndrome a partir de modelos de machine learning [46]. El estudio considera múltiples dimensiones del burnout, incluyendo dificultad en la regulación emocional, fatiga, baja calidad del sueño y carga académica percibida. Los datos fueron preprocesados para estandarizar escalas y manejar valores faltantes, y posteriormente se entrenaron clasificadores supervisados como SVM, Random Forest y XGBoost. El procedimiento consistió en dividir la muestra en conjuntos de entrenamiento y prueba, optimizando los hiperparámetros de cada modelo mediante validación cruzada. Los tres algoritmos alcanzaron desempeños muy altos, con precisiones entre 94 % y 98 %, y coincidieron en señalar que la regulación emocional deficiente y la mala calidad del sueño eran los predictores más influyentes. Estos resultados muestran que el uso de algoritmos de clasificación avanzados permite identificar de forma temprana a estudiantes en riesgo de burnout, aportando evidencia para diseñar estrategias preventivas y de apoyo en el contexto universitario.

2.11. Resumen del capítulo

En este capítulo se revisaron las bases conceptuales y técnicas para la optimización de instrumentos psicométricos. El análisis se inició estableciendo el contexto universitario como un escenario de alta prevalencia de burnout, ansiedad y depresión, lo que subraya la necesidad urgente de herramientas de tamizaje eficientes. Para ello, se revisaron los fundamentos del machine learning aplicados a la detección de patrones en datos de salud mental, junto con técnicas de optimización clave, como la selección de características y la reducción de dimensionalidad. Por último, se exploró el aporte crucial de la IA, incluyendo el uso de embeddings y modelos avanzados, confirmando la viabilidad de reducir y adaptar estos instrumentos de manera efectiva.

CAPÍTULO 3

PROPUESTA DE SOLUCIÓN

En los instrumentos psicométricos aplicados a salud mental universitaria es común encontrar ítems que, aunque estén formulados de manera diferente, expresan un contenido semántico similar. Esta redundancia introduce una sobrecarga de ítems en la construcción de un cuestionario unificado al integrar múltiples ítems que finalmente apuntan a medir el mismo constructo, lo que puede provocar fatiga en los encuestados y afectar negativamente en la tasa de participación y la calidad de las respuestas recolectadas.

A continuación se presentan algunos ejemplos de ítems que podrían considerarse semánticamente similares y que miden el mismo constructo:

Tabla 1: Ejemplos de ítems que podrían considerarse semánticamente similares

Constructo	Ítem	Enunciado
Concentración	EUBE-5	“Cada vez me es más difícil concentrarme en clases”
	EUBE-9	“Cada vez me cuesta más ponerle atención al maestro”
	PHQ-7	“Ha tenido dificultad para concentrarse en ciertas actividades, tales como leer el periódico o ver la televisión”
	WHO-5	“¿Cuál es su capacidad de concentración?”
Nivel de energía	WHO-8	“¿Tienes energía suficiente para la vida diaria?”
	PHQ-4	“Se ha sentido cansado(a) o con poca energía”
Desmotivación	PHQ-2	“Se ha sentido decaído(a), deprimido(a) o sin esperanzas”
	EUBE-2	“Mis problemas escolares me deprimen fácilmente”
	EUBE-6	“Me desilusionan mis estudios”

El objetivo de esta propuesta metodológica es verificar si efectivamente tales ítems son semánticamente similares y, además, identificar otros que no resultan evidentes a simple vista.

La identificación de posibles redundancias en los ítems evidencia la necesidad de contar con un enfoque sistemático que permita evaluar estas similitudes de manera objetiva. La detección manual, basada únicamente en la percepción del investigador, resulta ineficiente, ya que existen relaciones semánticas menos evidentes que podrían pasar desapercibidas y que, sin embargo, contribuyen a la sobrecarga del cuestionario.

Por esta razón, el presente capítulo propone una solución metodológica basada en técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático (ML) que permitan identificar, agrupar y seleccionar los ítems más representativos, con el objetivo de reducir,

cuando sea posible, la cantidad de preguntas del cuestionario optimizado.

La propuesta metodológica se estructura en una secuencia de etapas que van desde la preparación de los datos hasta el diseño del cuestionario optimizado. En primer lugar, se realiza un preprocesamiento de los ítems y respuestas, transformando cada pregunta en un vector semántico mediante técnicas de *embeddings*, lo que permite capturar relaciones de similitud más allá de la coincidencia léxica. Posteriormente, estos vectores se ingresan a un *Variational Autoencoder* (VAE), que proyecta la información en un espacio latente donde es posible identificar patrones subyacentes.

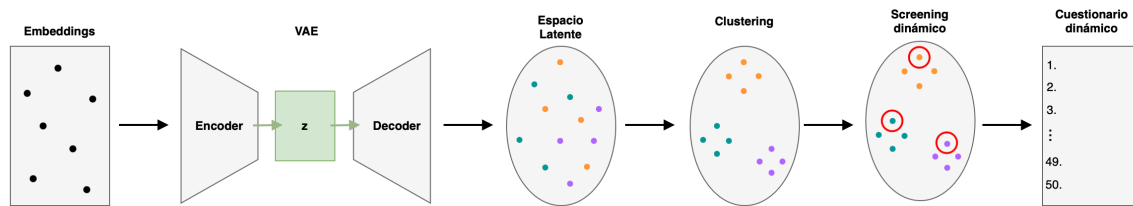


Figura 16: Flujo general de la propuesta metodológica.

Fuente: Elaboración propia.

A partir del espacio latente generado por el modelo *Variational Autoencoder* (VAE), se implementaron tres variantes de análisis de agrupamiento orientadas a evaluar distintas estrategias de estructuración semántica de los ítems. Cada variante explora una forma diferente de descomponer el espacio de representaciones latentes y permite comparar la coherencia de las agrupaciones obtenidas.

- **Top-Down Clustering:** Estrategia jerárquica descendente que inicia con un número reducido de clústeres amplios y los divide recursivamente según su disimilitud semántica. En cada iteración, se aplican particiones mediante K-Means para identificar subconjuntos más homogéneos y fusiones de ítems con el cluster más similar a ellos, hasta alcanzar agrupaciones con coherencia semántica interna.
- **Clustering Iterativo:** Estrategia ascendente de refinamiento iterativo que inicia su ejecución con un número elevado de clústeres pequeños generados por K-Means. En cada iteración, se evalúa la similitud semántica entre grupos y se fusionan aquellos más cercanos hasta alcanzar un equilibrio entre compacidad y coherencia.
- **Feature Agglomeration:** Método jerárquico ascendente que agrupa los ítems de forma progresiva, fusionando aquellos que presentan mayor similitud en su representación latente. Se utilizó principalmente como técnica de reducción de dimensionalidad y como punto de comparación para evaluar la coherencia de los resultados obtenidos con los métodos anteriores.

De la comparación entre estas tres variantes se selecciona el resultado más consistente para proceder con una etapa de *screening* dinámico. En esta fase, se identifica dentro de cada grupo el ítem más representativo (aquel con mayor correlación ítem-total), que actuará como *puerta de entrada*.

La lógica del *screening* consiste en adaptar la aplicación del cuestionario: si la respuesta del estudiante a este ítem supera un umbral definido, se despliegan las preguntas adicionales asociadas al mismo constructo. En caso contrario, se omiten y se continúa con el siguiente constructo.

Finalmente, el objetivo central de esta propuesta es el diseño de un **cuestionario dinámico adaptativo**, orientado a reducir la cantidad total de ítems presentados a cada participante, de modo que sólo se contesten aquellas preguntas estrictamente necesarias según sus respuestas previas. Este enfoque busca optimizar el proceso de evaluación mediante un sistema que adapta el recorrido del cuestionario en tiempo real según la activación de cada constructo.

Para alcanzar este objetivo, el presente capítulo describe la secuencia metodológica implementada: se inicia con el preprocesamiento y reformulación de los ítems (Sección 3.1.1), seguido por la generación de representaciones vectoriales semánticas (Sección 3.1.3), la reducción de dimensionalidad mediante modelos latentes (Sección 3.2), el agrupamiento semántico de ítems (Sección 3.3) y, finalmente, el diseño del cuestionario optimizado a través del proceso de *screening* dinámico (Sección 3.4).

3.1. Preprocesamiento semántico y generación del embedding

3.1.1. Reformulación semántica de ítems

Previo a la generación de los embeddings, se llevó a cabo un proceso de reformulación semántica de los ítems mediante técnicas de *prompt engineering*, con el objetivo de estandarizar la estructura lingüística de las preguntas y facilitar su procesamiento por modelos de lenguaje. Esta etapa fue necesaria debido a que los instrumentos incluidos no fueron redactados bajo un mismo propósito ni con un mismo estilo lingüístico. Mientras que PHQ-9, GAD-7 y WHOQOL-BREF fueron diseñados para evaluar síntomas clínicos, la EUBE fue formulada desde un enfoque educacional, orientado a describir experiencias académicas como el desgaste por estudio o la percepción de rendimiento. Estas diferencias de contexto implican que, sin una estandarización previa, los modelos de lenguaje podrían interpretar los ítems atendiendo a claves superficiales de su redacción, atribuyéndoles significados propios de dominios distintos al constructo que realmente evalúan. La reformulación semántica permitió, por tanto, alinear el estilo lingüístico de todos los instrumentos, manteniendo su contenido psicológico pero evitando que diferencias de contexto, propósito o formato influyeran en la representación semántica capturada por el modelo.

Los instrumentos psicométricos originales —*PHQ-9*, *GAD-7*, *EUBE* y *WHOQOL-BREF*— presentan diferencias significativas en la redacción de sus ítems, en el uso de tiempos verbales (“Ha tenido.” vs. “Me he sentido.”), el tipo de sujeto gramatical (primera, segunda o tercera persona) y tienen formas interrogativas o declarativas. Estas variaciones introducen ruido semántico, interfiriendo en la capacidad del modelo para capturar relaciones semánticas entre preguntas que, pese a estar formuladas de manera diferente, miden un mismo constructo psicológico (por ejemplo, ansiedad, energía, ánimo o calidad del sueño). Por esta razón, se reformularon los ítems procurando que los cambios en la redacción no alteraran el constructo que se busca evaluar, sino que permitieran expresar la misma idea con una forma más coherente y comparable entre instrumentos. Algunos ítems se mantuvieron en su forma original, otros se adaptaron solo al uso de primera persona, y varios pasaron por múltiples reformulaciones hasta lograr que el modelo pudiera identificar correctamente su sentido semántico y agruparlos junto a los ítems con significado similar.

3.1.2. Prompt Engineering

El *prompt engineering* se define como una metodología de diseño y ajuste de instrucciones (*prompts*) dirigidas a modelos de lenguaje con el propósito de guiar la generación o reformulación del texto según objetivos específicos [47]. En este contexto, se utilizó para generar versiones lingüísticamente equivalentes de los ítems originales, de modo que la similitud semántica reflejara el constructo subyacente y no dependiera de la forma gramatical o del estilo del enunciado.

El proceso no fue un simple reemplazo automático, sino un diseño analítico e iterativo. Se observó en las primeras iteraciones que las reformulaciones sutiles, como solo cambiar a primera persona, no eran suficientes para que el modelo capturara las similitudes semánticas deseadas. El modelo SBERT, si bien es semánticamente robusto, resultaba ser excesivamente literal, ya que podía agrupar correctamente “desilusión” con “deprimido”, pero fallaba al separar “aburrido” de “falta de interés”, o “problemas de sueño” de “insatisfacción con el sueño”, tratando los matices como constructos completamente diferentes.

El resultado de este proceso se puede observar en la Tabla 2, la cual presenta ejemplos representativos de ítems reformulados provenientes de los distintos instrumentos. Este procedimiento permitió obtener un conjunto de ítems con una redacción más uniforme y clara, donde las diferencias entre ellos reflejan principalmente el significado psicológico que buscan medir, y no solo la manera en que estaban redactados. Esta versión reformulada de los ítems constituye la base para el siguiente paso del proceso: su transformación en representaciones numéricas o *embeddings*, las cuales permiten analizar de forma computacional la similitud semántica entre las preguntas.

Tabla 2: Ejemplos de la reformulación semántica de ítems mediante Prompt Engineering

Instrumento	Ítem original	Ítem reformulado
PHQ-9	“Ha tenido dificultad para concentrarse en ciertas actividades, tales como leer el periódico o ver la televisión.”	“He tenido una baja capacidad para mantener la atención y la concentración en tareas cotidianas, como leer o ver TV.”
GAD-7	“No ha sido capaz de parar o controlar su preocupación”	“He tenido una preocupación excesiva y persistente que me resulta difícil de controlar, pensando demasiado en diferentes motivos.”
EUBE	“Creo que estudiar hace que me sienta agotado”	“He sentido agotamiento y desgaste por mis estudios: sintiéndome completamente agotado(a).”
WHOQOL-BREF	“¿Con qué frecuencia tiene sentimientos negativos, tales como tristeza, desesperanza o depresión?”	“He notado que tengo con frecuencia sensaciones de malestar emocional, como tristeza, ansiedad o una tensión constante que no desaparece.”

3.1.3. Representación semántica mediante embeddings

Una vez obtenidos los ítems reformulados, se procedió a representar cada uno de ellos en un espacio vectorial continuo mediante técnicas de embeddings semánticos. Estos modelos permiten transformar el significado de un texto en un vector numérico de alta dimensionalidad, de modo que ítems con contenidos conceptualmente similares se ubiquen próximos entre sí dentro de ese espacio.

A diferencia de enfoques tradicionales como *bag of words* o *TF-IDF*, los embeddings capturan relaciones contextuales y semánticas profundas, lo que resulta especialmente útil para detectar redundancias entre preguntas redactadas de forma distinta pero que evalúan un mismo constructo psicológico.

En este trabajo se utilizó como base el modelo preentrenado **paraphrase-multilingual-MiniLM-L12-v2**, perteneciente a la familia Sentence-BERT (SBERT) [36]. Este modelo fue elegido no sólo por su compatibilidad multilingüe, sino también por su capacidad para representar enunciados completos con matices semánticos complejos, preservando las relaciones de contexto que suelen encontrarse en ítems psicométricos.

Durante las etapas iniciales se probaron distintos modelos preentrenados como *all-MiniLM-L6-v2* y *bert-base-spanish-wwm-cased*, pero su desempeño dependía fuertemente de la lon-

gitud y la formulación gramatical de los ítems. Tras realizar múltiples iteraciones de ajuste y comparación, se seleccionó el modelo que mejor mantenía la coherencia entre ítems de un mismo constructo y, a la vez, separaba con mayor claridad los distintos dominios.

3.1.4. Fine-tuning del modelo SBERT

El *fine-tuning* (o ajuste fino) consiste en reentrenar parcialmente un modelo de lenguaje pre-entrenado para adaptarlo a un dominio o tarea específica [48]. En este caso, el objetivo fue ajustar los pesos del modelo SBERT a las características lingüísticas y semánticas propias de los ítems psicométricos reformulados. El proceso se llevó a cabo utilizando un conjunto de pares de ítems con etiquetas de similitud semántica, permitiendo que el modelo aprendiera qué enunciados son más cercanos entre sí en términos conceptuales como por ejemplo los ítems “Poco interés o placer en hacer cosas” (PHQ9-1) y “El asistir a clases se me hace aburrido” (EUBE-10).

El entrenamiento se realizó empleando el script mostrado a continuación, especificando los hiperparámetros utilizados en el proceso:

```
python3 scripts/fine_tuning.py \
  --train_path data/finetuning_v5.xlsx \
  --model_name paraphrase-multilingual-MiniLM-L12-v2 \
  --output_dir models/sbert_finetuned_v5 \
  --epochs 3 \
  --batch_size 32 \
  --lr 2e-5
```

Estos parámetros corresponden a un proceso de ajuste de un modelo en tres épocas (*epochs*) con un tamaño de lote (*batch_size*) de 32 y una tasa de aprendizaje (*learning rate*) de 2×10^{-5} . Bajo estas condiciones, el modelo SBERT preentrenado se expone a pares de ítems previamente emparejados según su nivel de similitud semántica, es decir, duplas de preguntas que fueron previamente definidas como conceptualmente cercanas o distantes en función del constructo psicológico que evalúan. Esto permite que el modelo internalice de manera más fina las relaciones lingüísticas entre los distintos enunciados, reconociendo qué preguntas expresan un mismo constructo psicológico y cuáles pertenecen a dimensiones diferentes. De este modo, el modelo ajusta sus representaciones internas al dominio específico de los cuestionarios de salud mental.

Como resultado del proceso de *fine-tuning* se obtiene una versión del modelo especializada en el dominio psicológico, cuya configuración interna ha sido optimizada para capturar de forma más precisa las similitudes y diferencias entre los ítems.

La Tabla 3 muestra ejemplos de pares de ítems junto con sus respectivas similitudes coseno calculadas a partir de los embeddings generados por el modelo fine-tuneado, demostrando su capacidad de relacionar ítems que son similares entre sí.

Tabla 3: Ejemplos de similitud semántica entre ítems mediante distancia coseno

Ítem 1	Ítem 2	Similitud coseno
He tenido dificultad para concentrarme en ciertas actividades, tales como leer el periódico o ver la televisión.	Cada vez me es más difícil concentrarme en clases.	0.81
No ha sido capaz de parar o controlar su preocupación.	Se ha preocupado demasiado por motivos diferentes.	0.78
Ha tenido miedo de que algo terrible fuera a pasar.	Ha tenido dificultad para quedarse o permanecer dormido, o ha dormido demasiado.	0.25

Finalmente, los embeddings obtenidos se integraron con las respuestas normalizadas de los participantes, conformando vectores extendidos de 385 componentes (384 correspondientes a la representación semántica de cada ítem y uno al valor de respuesta). Estos vectores constituyen la entrada del modelo VAE, descrito en la siguiente sección, cuyo propósito es explorar y visualizar la estructura latente de los ítems en función de sus relaciones semánticas.

3.2. Reducción de dimensionalidad con VAE

Dado que los *embeddings* generados por el modelo *Sentence-BERT* son vectores de alta dimensionalidad, se empleó un Autoencoder Variacional (VAE) con el propósito de obtener una representación latente compacta y estructurada que permitiera analizar de manera más clara las relaciones semánticas entre los ítems. Los VAE ofrecen una aproximación probabilística que resulta especialmente adecuada para capturar variaciones sutiles en el significado psicológico de los enunciados, representándolos dentro de un espacio continuo que preserva relaciones conceptuales relevantes.

El VAE está compuesto por dos redes neuronales: un *encoder*, encargado de comprimir la entrada en un conjunto reducido de variables latentes, y un *decoder*, cuya función es reconstruir la entrada original a partir de dichas variables. A diferencia de un autoencoder clásico, el VAE aprende distribuciones probabilísticas para cada dimensión del espacio latente, lo que confiere mayor flexibilidad y estabilidad al modelo al representar relaciones subyacentes de los datos.

En este caso, el modelo recibió como entrada vectores extendidos de 385 componentes. Sin embargo, solo los 384 componentes semánticos fueron utilizados en el proceso de reconstrucción, ya que el componente correspondiente al valor de respuesta se mantuvo separado para su uso posterior en la etapa de *screening* dinámico (Sección 3.4). De este modo, el espa-

cio latente aprendido por el VAE refleja exclusivamente la estructura semántica de los ítems, sin influencias derivadas de las respuestas de los participantes.

El entrenamiento del modelo se realizó utilizando los hiperparámetros resumidos en la Tabla 4, seleccionados tras un proceso exploratorio orientado a equilibrar capacidad de reconstrucción, estabilidad numérica e interpretabilidad del espacio latente.

Tabla 4: Hiperparámetros empleados para el entrenamiento del modelo VAE

Parámetro	Valor
Dimensión del espacio latente (z)	16
Dimensión de la capa oculta	96
Tasa de aprendizaje (lr)	1×10^{-3}
Tamaño de lote (<i>batch size</i>)	256
Número de épocas	60
<i>Dropout</i>	0.1
β (peso de regularización KL)	10.0
Capacidad máxima C_{max}	5.0
Época de saturación C_{stop}	30
Semilla aleatoria	42

3.2.1. Variantes del modelo VAE

Para obtener el modelo VAE que se utilizó en los pasos posteriores, se entrenaron diversas variantes del VAE modificando parámetros como la arquitectura interna, el tamaño de las capas ocultas, la fuerza de regularización y las condiciones de entrenamiento. Todas estas variantes utilizaron exactamente el mismo conjunto de embeddings de entrada, de manera que cualquier diferencia observada en las representaciones latentes se atribuye exclusivamente a la configuración interna del modelo.

Cada variante generó una estructura latente distinta, lo que permitió evaluar cómo las decisiones arquitectónicas influyen en la organización semántica de los ítems. Para ello, las medias latentes (μ_z) producidas por cada modelo fueron posteriormente proyectadas mediante UMAP, con el objetivo de visualizar y comparar las distintas estructuras generadas.

Para poder comparar las distintas variantes del VAE, todas las proyecciones fueron generadas utilizando una proyección bidimensional con $n_neighbors = 12$ y $min_dist = 0.5$, empleando la métrica de distancia coseno y una semilla aleatoria fija ($random_state = 42$) para garantizar la reproducibilidad y capturar relaciones locales entre ítems sin forzar una separación artificial entre grupos, equilibrando la preservación de la estructura global y la interpretabilidad visual del espacio proyectado. La Tabla 5 resume los patrones observados en cada varian-

te del modelo VAE y la evaluación cualitativa correspondiente, incluyendo la configuración seleccionada como modelo final (Variante F).

Tabla 5: Resumen cualitativo de las variantes evaluadas del modelo VAE.

Variante	Patrón observado en el espacio latente	Evaluación y decisión
A	Agrupamientos excesivamente compactos, con varios ítems concentrados en regiones muy reducidas del espacio proyectado.	Indica una regularización demasiado fuerte y pérdida de variabilidad entre ítems. Variante descartada.
B	Estructura similar a la Variante A, con conglomerados muy densos y poca separación interna entre ítems.	Presenta problemas de sobrecompactación; no permite analizar fines detalles semánticos. Variante descartada.
C	Distribución más extendida, con ítems repartidos a lo largo del espacio pero sin formación clara de grupos coherentes.	Genera un patrón disperso y de difícil interpretación. Variante descartada.
D	Estructura también dispersa, con clústeres poco definidos y solapamiento entre dominios.	La falta de agrupamientos nítidos dificulta la identificación de clústeres estables. Variante descartada.
E	Organización más coherente, con agrupamientos visibles y cierta separación entre dominios, aunque aún con zonas de solapamiento.	Se considera una configuración intermedia; mejora la estructura, pero no es la más clara. Variante exploratoria.
F	Estructura latente claramente organizada, con clústeres definidos y buena separación entre los instrumentos y dominios psicológicos.	Presenta la organización más estable y coherente; se selecciona como configuración final del VAE.

3.3. Agrupamiento semántico de ítems

Una vez obtenido el espacio latente del modelo VAE, se procedió con el agrupamiento de los ítems para identificar patrones de similitud semántica entre preguntas pertenecientes a distintos instrumentos psicométricos. Este proceso busca reconocer ítems que evalúan constructos conceptualmente equivalentes, aun cuando difieran en su redacción, y servir como base para la selección de los ítems representativos que se utilizarán posteriormente en el *screening* dinámico (Sección 3.4).

El agrupamiento semántico tiene tres objetivos principales:

1. Distinguir preguntas que miden el mismo constructo psicológico.
2. Identificar subdimensiones dentro de los constructos principales (ansiedad, depresión, burnout, calidad de vida).
3. Establecer grupos semánticamente coherentes que faciliten el diseño de un cuestionario dinámico más eficiente.

Normalización y métrica de similitud

Previo a aplicar cualquier algoritmo de agrupamiento, es necesario normalizar los vectores latentes generados por el VAE mediante normalización L2. Esta transformación garantiza que todos los embeddings tengan una longitud unitaria ($\|x\|_2 = 1$), lo que permite que el producto punto entre dos vectores corresponda directamente a su similitud coseno:

$$\cos(x_i, x_j) = \frac{x_i \cdot x_j}{\|x_i\|_2 \|x_j\|_2} = x_i \cdot x_j \quad (29)$$

De este modo, el valor resultante varía entre -1 y 1, representando el grado de similitud semántica entre dos ítems en función de su posición en el espacio latente. Esta métrica es especialmente adecuada para embeddings textuales, ya que la dirección del vector, más que su magnitud, refleja el significado semántico del texto. En otras palabras, dos ítems pueden tener longitudes (normas) distintas por variaciones en la complejidad gramatical o el número de palabras, pero seguir apuntando en la misma dirección en el espacio vectorial si expresan una idea similar. Sin esta normalización, los modelos podrían interpretar erróneamente diferencias de escala como diferencias de significado.

A continuación, se detallan los tres métodos implementados para lograr el agrupamiento de ítems.

3.3.1. Top-Down Clustering

El método *Top-Down* corresponde a una estrategia jerárquica descendente que comienza su ejecución considerando un número reducido de clústeres amplios y los divide recursivamente según su diferencia semántica, refinando las divisiones obtenidas en la iteración anterior. El propósito de este enfoque es identificar inicialmente la estructura global del espacio latente y, a partir de ella, refinar de manera progresiva las divisiones para obtener agrupamientos cada vez más compactos y coherentes.

El algoritmo comienza con una partición inicial del conjunto completo de ítems mediante *K-Means*, y en cada iteración el clúster se subdivide en k_{sub} nuevos subclústeres y se repite el proceso hasta alcanzar una estructura estable. En los casos en que se genera un clúster

con un solo ítem (singleton), este se fusiona con el clúster más semánticamente cercano, determinado por la mayor similitud coseno entre sus centroides. Si no existen clústeres disponibles para fusionar, el algoritmo procede con la subdivisión estándar en k_{sub} particiones.

El procedimiento completo se resume en el pseudocódigo siguiente:

Algorithm 1 Algoritmo de *Top-Down Clustering*

Require: X : matriz de embeddings normalizados (norma L2)

```

1:  $k_{inicial} \leftarrow 4$ 
2:  $k_{sub} \leftarrow 2$ 
3:  $threshold \leftarrow 0.40$ 
4:  $clusters \leftarrow \text{KMeans}(X, n_{clusters} = k_{inicial}, n_{init} = 20, \text{init}=\text{"k-means++"}, \text{random\_state}=42)$ 
5: while True do
6:    $split \leftarrow \text{False}$ 
7:   for cada clúster  $c \in clusters$  do
8:      $size \leftarrow$  número de ítems en  $c$ 
9:      $cohesion \leftarrow \text{mean\_cosine\_similarity}(c)$ 
10:    if  $cohesion < threshold$  then
11:       $subclusters \leftarrow \text{KMeans}$ 
12:      reemplazar  $c$  por  $subclusters$  en  $clusters$ 
13:       $split \leftarrow \text{True}$ 
14:    else if  $size = 1$  then
15:      fusionar ítem con el clúster más semánticamente cercano
16:    end if
17:  end for
18:  if not  $split$  then
19:    break ▷ Se detiene cuando ya no se generan nuevas subdivisiones
20:  end if
21: end while
22: return  $clusters$ 

```

El resultado de este proceso es una estructura jerárquica descendente que refleja cómo los ítems se agrupan de manera natural según su proximidad semántica en el espacio latente. Cada nivel de división permite identificar subdimensiones progresivamente más específicas, las cuales servirán de base para el análisis comparativo y el proceso de *screening* dinámico descrito en la Sección 3.4.

3.3.2. K-Means Iterativo

El método *Iterativo* corresponde a una estrategia ascendente de refinamiento progresivo. A diferencia del enfoque anterior, este comienza con un número elevado de clústeres pequeños y los consolida sucesivamente en función de su similitud semántica.

El algoritmo comienza con una partición inicial del conjunto completo de ítems mediante *K-Means*, generando múltiples clústeres de tamaño reducido. A continuación, en cada iteración, los clústeres con baja cohesión interna ($threshold < 0.40$) se subdividen en k_{sub} nuevos subclústeres, permitiendo separar grupos heterogéneos en unidades más consistentes. El valor del *threshold* se seleccionó de manera empírica, tras evaluar diferentes valores y observar que este punto ofrecía un equilibrio adecuado entre evitar fragmentaciones excesivas y asegurar que los clústeres mantuvieran una coherencia semántica suficiente. En particular, valores más bajos tendían a unir ítems que no compartían un significado psicológico claro, mientras que valores más altos impedían que los ítems con afinidad semántica se reagruparan. El umbral de 0.40 representó así el punto de estabilidad en el cual los ítems se organizaban de manera coherente dentro de cada clúster sin perder la separación conceptual entre dominios.

Cuando se generen clústeres con un solo ítem (singletons), estos se fusionarán con el clúster más semánticamente cercano, determinado por la mayor similitud coseno entre sus centroides. El proceso continúa hasta alcanzar una estructura estable y coherente.

El procedimiento completo de este método se detalla en el siguiente pseudocódigo:

Algorithm 2 Algoritmo de *Clustering Iterativo*

Require: X : matriz de embeddings normalizados (norma L2)

```

1:  $k_{inicial} \leftarrow 12$ 
2:  $threshold \leftarrow 0.40$ 
3:  $clusters \leftarrow \text{KMeans}(X, n_{clusters} = k_{inicial}, n_{init} = 20, random\_state=42)$ 
4: while True do
5:   calcular  $centroides \leftarrow$  promedio de  $X$  en cada clúster
6:   normalizar  $centroides$  con norma L2
7:    $split \leftarrow \text{False}$ 
8:   for cada clúster  $c$  do
9:      $cohesion \leftarrow \text{mean\_cosine\_similarity}(c)$ 
10:    if  $cohesion < threshold$  then
11:      subdividir  $c$  aplicando  $\text{KMeans}(c, n_{clusters} = k_{sub})$ 
12:      reemplazar  $c$  por sus subclústeres
13:       $split \leftarrow \text{True}$ 
14:    else if  $size(c) = 1$  then
15:      fusionar ítem con el clúster más semánticamente cercano
16:    end if
17:  end for
18:  if not  $split$  then
19:    break ▷ Se detiene cuando ya no se generan nuevas subdivisiones
20:  end if
21: end while
22: return  $clusters$ 

```

El enfoque iterativo permite refinar los agrupamientos hasta obtener una estructura más precisa y estable, priorizando la compacidad y la consistencia conceptual de los ítems.

3.3.3. Feature Agglomeration

El método *Feature Agglomeration* se incorporó como técnica complementaria de reducción y comparación, con el propósito de evaluar cómo la estructura latente influye en la coherencia de los agrupamientos obtenidos. A diferencia de los métodos previos que operan directamente sobre los ítems y buscan agrupar preguntas según su similitud semántica, este enfoque actúa sobre las dimensiones del espacio latente generado por el VAE.

En este contexto, cada dimensión latente puede interpretarse como una característica abstracta aprendida por el modelo, por ejemplo, una dirección del espacio que capture cuán orientado está un ítem hacia síntomas de *ánimo bajo*, otra asociada a *agotamiento físico*, u otra que refleje *preocupación excesiva*. Estas variables no existen explícitamente en los cuestionarios originales, pero emergen del VAE como representaciones numéricas que combinan patrones semánticos presentes en los distintos instrumentos.

De esta manera, *Feature Agglomeration* agrupa aquellas dimensiones latentes que muestran comportamientos o correlaciones similares, reduciendo la complejidad del espacio latente y generando una representación más compacta y estable de las características.

El procedimiento consistió en reducir de 16 a 8 dimensiones el espacio latente original mediante un proceso jerárquico ascendente, donde cada fusión entre características preserva la varianza promedio del conjunto. Posteriormente, se aplicó *K-Means* sobre este nuevo espacio reducido y se siguieron las mismas etapas de evaluación empleadas en los enfoques *Top-Down* e *Iterativo*: subdivisión por k_{sub} , fusión de ítems aislados y proyección de resultados con UMAP para poder comparar los resultados obtenidos en cada método.

El proceso descrito se implementó mediante una secuencia de pasos definidos que permiten operacionalizar la aplicación de *Feature Agglomeration* y la evaluación posterior de los clústeres obtenidos. Dichos pasos se detallan a continuación:

1. Normalización L2 de las medias latentes por ítem.
2. Aplicación de *Feature Agglomeration* para reducir la dimensionalidad ($16 \rightarrow 8$).
3. Ejecución de K-Means sobre el nuevo espacio reducido.
4. Evaluación de cohesión y tamaño de clústeres obtenidos.
5. Subdivisión de clústeres dispersos mediante k_{sub} .
6. Fusión de ítems aislados con el clúster más cercano.
7. Repetir los pasos 3–6 hasta alcanzar clústeres con mayor coherencia.

3.3.4. Selección del método final

Finalmente, tras la aplicación y análisis de los tres enfoques de agrupamiento, se seleccionó el método **Top-Down Clustering** como estrategia final de agrupamiento semántico. Esta elección se fundamentó en que dicho enfoque logró una mayor coherencia conceptual entre los ítems dentro de cada grupo, reflejando con mayor fidelidad los constructos psicológicos subyacentes. Los clústeres obtenidos mediante este método presentaron una estructura más interpretable, en la que los ítems compartían un significado semántico común y mostraban una separación más clara respecto de otros dominios.

Cabe señalar que la estructura de clústeres generada por cada uno de los tres métodos evaluados, se presenta y analiza en detalle en el Capítulo 4, como parte de la comparación empírica entre enfoques.

3.3.5. Análisis probabilístico de ítems

El análisis probabilístico tuvo como propósito cuantificar el grado de coincidencia entre los métodos de agrupamiento utilizados y evaluar la consistencia estructural de los clústeres formados. Para ello, se comparó la asignación común de cada par de ítems en los mismos grupos, a lo largo de los tres algoritmos aplicados.

Este procedimiento permitió identificar redundancias, subgrupos semánticos robustos y, especialmente, conjuntos de ítems cuya organización latente se mantiene estable sin depender del método empleado. Estos patrones constituyen la base para seleccionar ítems representativos y definir las reglas de activación del cuestionario adaptativo.

Con este fin, se estableció un procedimiento sistemático compuesto por cuatro etapas principales:

1. **Extracción de asignaciones de clúster:** Para cada uno de los tres métodos de agrupamiento (Top-Down, Iterativo y Feature Agglomeration), se registró el clúster asignado a cada ítem, generando tres particiones independientes del conjunto.
2. **Cálculo de co-ocurrencias:** Se contabilizó cuántas veces cada par de ítems (i, j) coincidió en el mismo clúster. Esta cantidad se denota como v_{ij} y corresponde a la *co-ocurrencia* observada entre métodos.
3. **Construcción de la matriz de co-asociación A :** Definiendo $M = 3$ como el número total de métodos evaluados, la co-asociación se obtuvo como el promedio de co-ocurrencias:

$$A_{ij} = \frac{v_{ij}}{M} \quad (30)$$

Valores cercanos a 1 indican alta estabilidad entre métodos; valores cercanos a 0 reflejan ausencia de coincidencias.

4. **Sistema de votación del nivel de co-asociación:** Con el fin de estandarizar la interpretación de estas coincidencias, se definió un sistema discreto de votación que asigna un valor fijo de co-asociación según la cantidad de métodos que agruparon conjuntamente a los ítems. Este esquema se resume en la Tabla 6.

Tabla 6: Sistema de votación para la matriz de co-asociación.

Cantidad de votos	Nivel de co-ocurrencia	A_{ij}
3 votos	Fuerte co-ocurrencia	1.0
2 votos	Moderada co-ocurrencia	0.66
1 voto	Baja co-ocurrencia	0.33
0 votos	Sin co-ocurrencia	0.0

Este esquema permitió uniformar los niveles de acuerdo entre métodos y facilitar la identificación de relaciones robustas entre ítems.

5. **Identificación de clústeres estables:** Con la matriz A estandarizada, se definió como **clúster estable** a cualquier conjunto de ítems donde cada par cumple:

$$A_{ij} \geq 0.66 \quad (31)$$

Este umbral corresponde a que al menos dos de los tres métodos coincidieron en agruparlos, reflejando estructuras semánticas consistentes dentro del espacio latente.

A partir de la matriz A , se calcularon luego métricas de cohesión por instrumento:

- **mean_A:** promedio de co-asociación entre pares de ítems del mismo instrumento.
- **std_A:** variabilidad de esa co-asociación, que indica si el instrumento se comporta como un conjunto homogéneo o como un grupo diversificado.

Este análisis no busca determinar los clústeres finales, sino evaluar la estabilidad y consistencia de las relaciones entre ítems a través de métodos distintos. Así, la matriz A permite detectar:

- Pares de ítems altamente redundantes.
- Núcleos semánticos robustos dentro de cada instrumento.
- Instrumentos más compactos o más heterogéneos.
- Clústeres estables que podrán ser utilizados posteriormente en el diseño del cuestionario adaptativo.

En síntesis, el análisis probabilístico permitió identificar qué relaciones entre ítems se mantienen estables a través de los distintos métodos de agrupamiento, así como cuantificar la cohesión interna de cada instrumento psicométrico. Estos resultados entregan una base objetiva para seleccionar los ítems más representativos de cada clúster y, especialmente, para definir los puntos de activación utilizados en la etapa adaptativa del cuestionario.

Con esta información, es posible avanzar desde la caracterización estructural de los ítems hacia la construcción de un mecanismo de decisión que determine cuándo profundizar en un constructo y cuándo omitirlo. Este paso corresponde al **screening dinámico**, el cual organiza la secuencia del cuestionario utilizando los ítems representativos y los umbrales derivados del análisis probabilístico.

3.4. Screening Dinámico

Con los clústeres semánticos definidos, el siguiente paso consiste en establecer un mecanismo de *screening* dinámico que organice la secuencia de preguntas de manera más eficiente. La idea central es utilizar la estructura obtenida en el análisis previo para determinar qué grupos de ítems se activan en función de las primeras respuestas entregadas por cada participante. De este modo, el cuestionario adapta su recorrido de forma progresiva, priorizando los clústeres que resultan más pertinentes según el perfil de respuesta inicial.

Cada clúster representa un constructo psicológico, y su activación depende de un **ítem puerta**: un ítem representativo capaz de detectar el inicio del cambio en dicho constructo, el cual determina si se deben presentar o no los otros ítems del clúster al encuestado.

3.4.1. Normalización de Escalas

Para garantizar la comparabilidad entre ítems de distintos instrumentos, todas las respuestas fueron transformadas a un rango común $[0, 1]$. Se mantuvo la trazabilidad de cada ítem respecto a su instrumento de origen, aplicando las fórmulas de normalización específicas detalladas en la Tabla 7. Esta unificación permite analizar conjuntamente ítems provenientes de instrumentos heterogéneos, facilitando la interpretación y la comparación entre constructos en el espacio latente.

Tabla 7: Normalización de respuestas según instrumento

Instrumento	Escala	Fórmula	Respuesta	Valor normalizado
PHQ-9	0-3	$\frac{x}{3}$	0	0.00
			1	0.33
			2	0.67
			3	1.00
GAD-7	0-3	$\frac{x}{3}$	0	0.00
			1	0.33
			2	0.67
			3	1.00
EUBE	1-4	$\frac{x-1}{3}$	1	0.00
			2	0.33
			3	0.67
			4	1.00
WHOQOL-BREF	1-5	$\frac{x-1}{4}$	1	0.00
			2	0.25
			3	0.50
			4	0.75
			5	1.00

3.4.2. Selección del Ítem Representativo

La selección del ítem puerta no se basó únicamente en criterios geométricos (como la centralidad del clúster), sino en una metodología empírica apoyada en los datos del estudio de salud mental aplicado a 390 estudiantes de la UTFSM en 2023. Para este análisis, los niveles de severidad oficiales de los instrumentos PHQ-9, GAD-7, EUBE y WHOQOL-BREF se utilizaron como *criterio externo* de referencia (*ground truth*), entendidos como una clasificación establecida por cada instrumento sobre el grado de afectación del participante. En términos prácticos, estos niveles permiten contrastar cómo varía la distribución de las respuestas de cada ítem frente a grupos de estudiantes con severidades diferentes, facilitando identificar qué preguntas muestran patrones más consistentes con dichas categorías y, por lo tanto, resultan más adecuadas como representantes del clúster.

Para formalizar este procedimiento, sea N_k^i el nivel i de severidad del instrumento k , con $i = 1, \dots, M_k$, donde M_k corresponde al número total de niveles de severidad definidos por el instrumento k . Cada nivel corresponde a una categoría clínicamente definida en los

instrumentos PHQ-9, GAD-7, EUBE y WHOQOL-BREF.

Cada uno de estos niveles puede representarse numéricamente como un intervalo $[d_k^i, u_k^i]$, donde d_k^i y u_k^i corresponden, respectivamente, a los límites inferior y superior establecidos por el instrumento para ese nivel de severidad. A partir de estos intervalos se analizó cómo varía la distribución de las respuestas de cada ítem dentro de los distintos N_k^i , permitiendo evaluar su comportamiento frente a los niveles definidos por cada instrumento.

La Tabla 8 resume los niveles de severidad definidos por cada instrumento, los cuales se utilizan como referencia clínica para el análisis del comportamiento de las respuestas de los ítems y la evaluación de su capacidad de discriminación entre niveles adyacentes. En el caso del WHOQOL-BREF, los niveles se definieron de manera operacional a partir del puntaje total del instrumento, utilizando los cuántiles teóricos de la escala. En particular, se emplearon los puntos de corte correspondientes al primer y tercer cuántil para definir tres niveles de calidad de vida percibida (baja, media y alta), con fines analíticos y comparativos.

Tabla 8: Niveles de severidad definidos para cada instrumento psicométrico

Instrumento	Nivel i	Descripción	Rango $[d_k^i, u_k^i]$
PHQ-9	1	Mínima depresión	[0, 4]
PHQ-9	2	Depresión leve	[5, 9]
PHQ-9	3	Depresión moderada	[10, 14]
PHQ-9	4	Depresión mod. severa	[15, 19]
PHQ-9	5	Depresión severa	[20, 27]
GAD-7	1	Ansiedad mínima	[0, 4]
GAD-7	2	Ansiedad leve	[5, 9]
GAD-7	3	Ansiedad moderada	[10, 14]
GAD-7	4	Ansiedad severa	[15, 21]
EUBE	1	Bajo nivel de burnout	[0, 15]
EUBE	2	Burnout leve	[16, 30]
EUBE	3	Burnout moderado	[31, 45]
EUBE	4	Burnout profundo	[46, 60]
WHOQOL-BREF	1	Baja calidad de vida	[26, 52]
WHOQOL-BREF	2	Calidad de vida media	[53, 104]
WHOQOL-BREF	3	Alta calidad de vida	[105, 130]

A partir de estas mediciones se aplicaron los siguientes criterios:

1. **Correlación con la severidad:** Se evaluó qué ítems mostraban mayor correlación con los niveles de severidad del instrumento principal asociado al constructo. Los ítems

con mayor correlación poseen mayor capacidad de reflejar variaciones reales en el nivel del participante.

2. **Comportamiento por niveles (media y varianza):** El criterio central fue analizar si la respuesta media del ítem aumentaba de manera sistemática a medida que avanzan los niveles de severidad, es decir, si $\mu(N_k^1) < \mu(N_k^2) < \mu(N_k^3)$ para el instrumento correspondiente. Los ítems que muestren este tipo de progresión son considerados candidatos a representar el clúster.
3. **Discriminación entre niveles:** Se evaluó la magnitud del cambio en las respuestas del ítem entre niveles consecutivos de severidad, especialmente entre el primer y segundo nivel, que suelen distinguir entre la ausencia del síntoma y su presencia mínima. Este punto corresponde al primer cambio relevante en el patrón de respuesta, por lo que el umbral se estableció en el límite donde las distribuciones de respuesta dejan de ser homogéneas, lo que permite identificar el punto inicial en el cual un ítem comienza a diferenciar entre grupos.
4. **Coherencia semántica:** Entre los ítems con buen rendimiento estadístico, se selecciona aquel ítem cuya formulación represente de manera más amplia y general el constructo evaluado.

3.4.3. Definición del Umbral de Activación

El umbral de activación del clúster se obtuvo a partir del punto medio entre las medias del ítem puerta en los niveles N1 y N2:

$$\tau = \left\lfloor \frac{\mu_{N1} + \mu_{N2}}{2} \right\rfloor \quad (32)$$

Este valor marca el punto en que las respuestas del ítem dejan de concentrarse en el nivel más bajo (N1) y comienzan a mostrar señales propias del siguiente nivel de severidad (N2). En los modelos de respuesta ordenada, este punto corresponde al *threshold* τ donde aumenta la probabilidad de elegir una categoría superior, por lo que se interpreta como el primer cambio significativo en el comportamiento del ítem.

Posteriormente, el umbral se ajustó a la escala discreta original del instrumento (0-3, 1-4 o 1-5) mediante truncamiento, garantizando consistencia con las opciones reales de respuesta. Esta lógica es coherente con la evidencia empírica utilizada en la validación de instrumentos breves como el PHQ-2 y GAD-2, donde los puntos de corte se seleccionan en función de su capacidad para separar casos y no casos optimizando sensibilidad y especificidad [49].

Así, un clúster se activa cuando la respuesta del ítem puerta supera o cae por debajo del umbral definido para ese constructo, dependiendo de la dirección de la escala del ítem. En

ítems donde puntajes altos indican mayor presencia del constructo como “Creo que estudiar me está desgastando emocionalmente” (EUBE-15), el criterio corresponde a valores mayores o iguales al umbral. En cambio, en ítems formulados en sentido inverso, donde puntajes bajos reflejan menor bienestar o satisfacción, como “¿Cómo de satisfecho está con sus relaciones personales?” (WHO-18) el criterio se aplica utilizando valores menores o iguales al umbral.

El flujo general del procedimiento de *screening* dinámico se ilustra en la Figura 17, donde se muestran las etapas principales para la activación u omisión de cada clúster en función de la respuesta al ítem representativo.

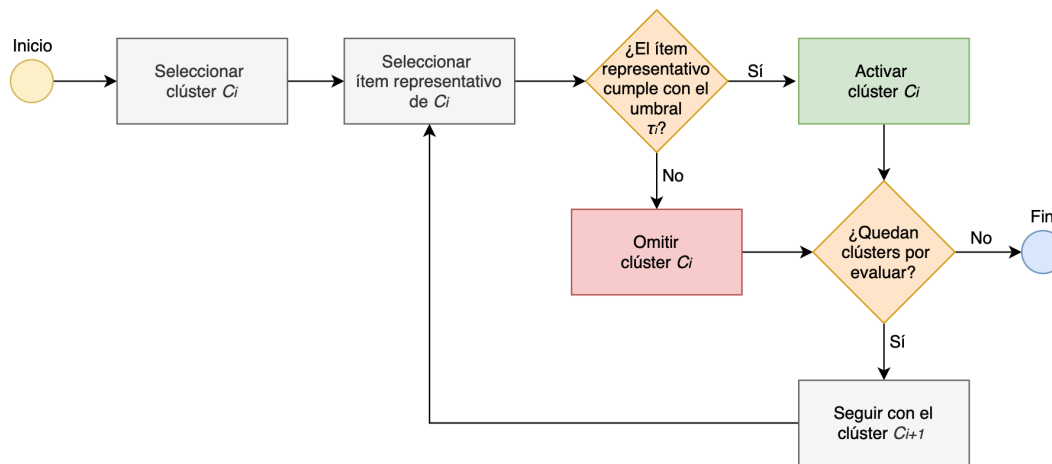


Figura 17: Procedimiento general del Screening Dinámico.
Fuente: Elaboración propia.

3.5. Diseño del Cuestionario Dinámico Adaptativo

El cuestionario final se diseñó bajo un enfoque adaptativo cuyo propósito es reducir la carga de respuesta y, simultáneamente, organizar la secuencia de ítems de manera más eficiente. A diferencia de los instrumentos psicométricos tradicionales, donde todas las preguntas se administran de forma fija, el cuestionario propuesto incorpora una lógica condicional basada en *ítems puerta* y umbrales definidos en las etapas previas del análisis. El detalle completo de los ítems originales de cada instrumento se presentan en los Anexos.

3.5.1. Estructura General del Cuestionario

El cuestionario se organiza en torno a los clústeres semánticos definidos previamente. Cada clúster contiene:

- Un ítem puerta, que actúa como punto de entrada y decisión.
- Uno o más ítems que miden el mismo constructo, que se muestran sólo si el clúster es activado.
- Un umbral de activación τ_i , que determina cuándo el clúster debe ser evaluado con mayor profundidad.

Este mecanismo es similar a los *skip patterns*⁷ usados en encuestas epidemiológicas, pero aplicado siempre con la misma lógica en cada constructo identificado en el análisis de clústeres. Aunque internamente opera como un sistema de decisión adaptativo, desde la perspectiva del usuario, la experiencia de respuesta se mantiene lineal, breve y clara. Esto contribuye a reducir la fatiga cognitiva, aspecto especialmente relevante en mediciones de salud mental.

3.5.2. Ventajas del Diseño Adaptativo

El diseño del cuestionario presenta varias ventajas prácticas:

- **Reducción de carga:** Los participantes responden sólo los ítems estrictamente necesarios.
- **Prevención de fatiga:** Se evita la exposición a preguntas redundantes o poco relevantes para el caso individual.
- **Potencial aumento de participación:** Al disminuir el tiempo de respuesta y la carga percibida, aumenta la probabilidad de completar el cuestionario.

En conclusión, la propuesta presentada define cómo se organiza el cuestionario adaptativo y cómo se seleccionan los ítems según las necesidades de cada participante. Con esta base, el siguiente capítulo muestra los resultados obtenidos al aplicar este diseño y cómo se construyó la versión final del cuestionario optimizado.

⁷Los *skip patterns* corresponden a reglas de salto utilizadas en encuestas para omitir bloques de preguntas que no aplican según la respuesta inicial del participante, reduciendo así la carga y el tiempo de respuesta.

3.6. Resumen del capítulo

Este capítulo presentó la metodología utilizada para diseñar el cuestionario optimizado. Se describieron las etapas de reformulación semántica, generación de embeddings y reducción de dimensionalidad mediante un VAE, así como los distintos métodos de agrupamiento evaluados y la selección del enfoque final. Además, se expuso la metodología del análisis probabilístico y del *screening* dinámico, empleados para identificar ítems representativos y definir el flujo adaptativo del cuestionario. Adicionalmente, se compararon distintas variantes del modelo VAE, cada una entrenada con configuraciones internas diferentes, con el propósito de evaluar cómo estas afectaban la estructura latente de los ítems. La comparación cualitativa entre las variantes permitió identificar la configuración más estable y coherente, seleccionada como modelo final. Los resultados derivados de esta metodología se desarrollan en el capítulo siguiente.

CAPÍTULO 4

RESULTADOS

En este capítulo se presentan los resultados obtenidos a partir del proceso metodológico descrito previamente. Se exponen, en primer lugar, los hallazgos experimentales asociados al análisis semántico, incluyendo la estructura del espacio latente generada por los modelos de representación, la organización interna de los ítems en clústeres y el comportamiento probabilístico observado entre niveles de severidad. Posteriormente, se presentan los resultados aplicados correspondientes al diseño final del cuestionario adaptativo, detallando la configuración definitiva de los constructos, la selección de ítems puerta basada en el análisis probabilístico y la lógica de ramificación establecida.

Los resultados presentados en este capítulo se basan en datos recolectados a partir de una aplicación transversal de cuestionarios a estudiantes de Ingeniería Civil Informática de la UTFSM. La recolección de datos se realizó en tres instancias temporales distintas a lo largo del semestre académico, con el objetivo de capturar respuestas bajo contextos de exigencia académica diferenciados. La primera instancia tuvo lugar entre el 26 de septiembre y el 17 de octubre, durante un período de menor carga académica asociado a la semana de Fiestas Patrias. La segunda se desarrolló entre el 2 y el 17 de noviembre, coincidiendo con el proceso de rebaja académica voluntaria, etapa caracterizada por un aumento del estrés percibido y la toma de decisiones académicas relevantes. Finalmente, la tercera instancia se llevó a cabo entre el 23 de noviembre y el 7 de diciembre, hacia el cierre del semestre, período típicamente asociado a una alta demanda evaluativa.

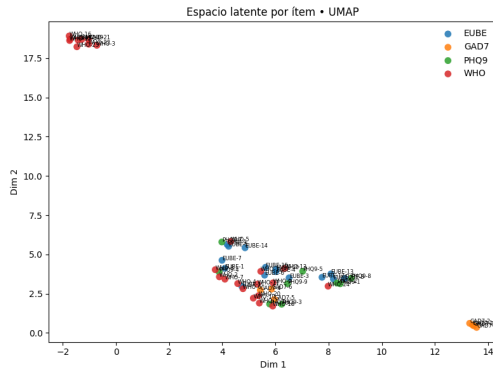
La convocatoria se realizó mediante correos institucionales, difusión en plataformas académicas, redes sociales y actividades presenciales en salas de clases, laboratorios y talleres, complementadas con instancias de seguimiento. La recolección se efectuó a través de formularios administrados mediante Google Forms, incorporando consentimiento informado previo, tanto en modalidad presencial como virtual mediante códigos QR. En total, se obtuvieron 568 respuestas, lo que permitió construir un conjunto de datos amplio y heterogéneo, representativo de distintos momentos del ciclo académico.

4.1. Resultados experimentales

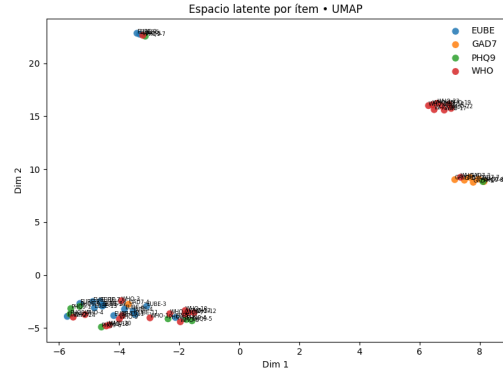
Esta sección reúne los resultados derivados del análisis computacional orientado a identificar redundancias, dependencias y patrones semánticos entre los ítems de los instrumentos seleccionados. Se presentan los hallazgos asociados a la comparación de modelos de embeddings, la evaluación del espacio latente generado mediante un VAE, los distintos enfoques de agrupamiento aplicados y el análisis probabilístico que fundamentó la selección de umbrales para el mecanismo de activación del cuestionario.

4.1.1. Comparación de los modelos VAE

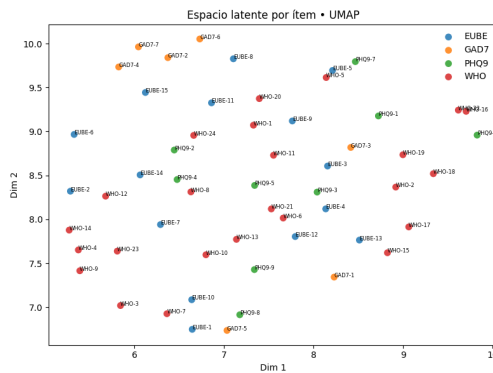
La Figura 18 muestra las proyecciones UMAP de cada variante del modelo, permitiendo comparar la cohesión interna, la separación entre dominios psicológicos y la presencia de ruido estructural.



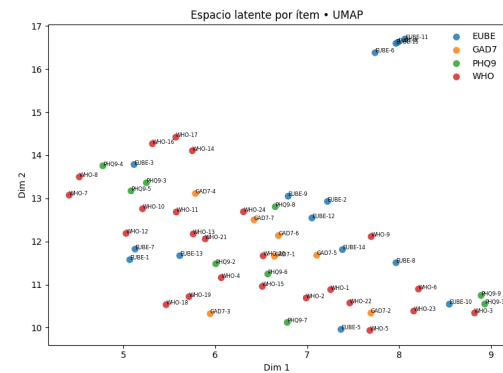
(a) Variante A



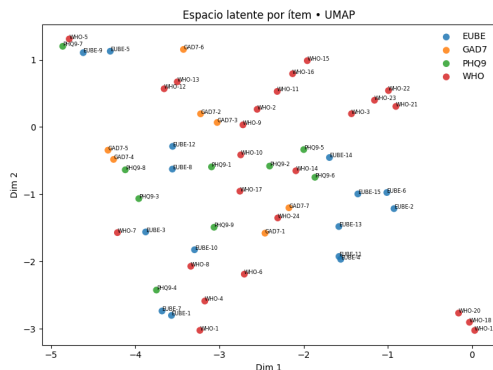
(b) Variante B



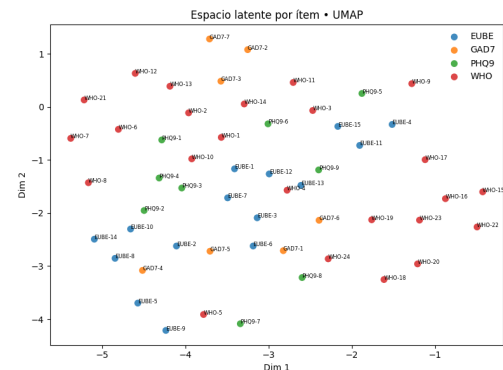
(c) Variante C



(d) Variante D



(e) Variante E



(f) Variante F

Figura 18: Espacios latentes obtenidos por las variantes del VAE.

Fuente: Elaboración propia.

Durante la comparación cualitativa se observaron diferencias claras entre las distintas configuraciones evaluadas. Por ejemplo, las variantes A y B generaron espacios latentes excesivamente compactos, lo que sugiere una regularización demasiado fuerte y una pérdida de variabilidad entre los ítems. En contraste, las variantes C y D produjeron distribuciones más dispersas e inconsistentes, dificultando la interpretación y la identificación de patrones semánticos estables. En cambio, las variantes E y F mostraron una organización más coherente, con agrupamientos definidos y una separación adecuada entre dominios psicológicos. De ellas, la Variante F fue la que presentó la estructura más clara y consistente, reflejando de mejor manera la organización semántica esperada, motivo por el cual fue seleccionada como modelo final.

4.1.2. Proyección UMAP final

Una vez seleccionado el modelo VAE final, se generó la proyección 2D mostrada en la Figura 19. Esta representación corresponde al espacio latente definitivo utilizado en las siguientes etapas del estudio, incluyendo el agrupamiento de ítems y la construcción del sistema de *screening* dinámico.

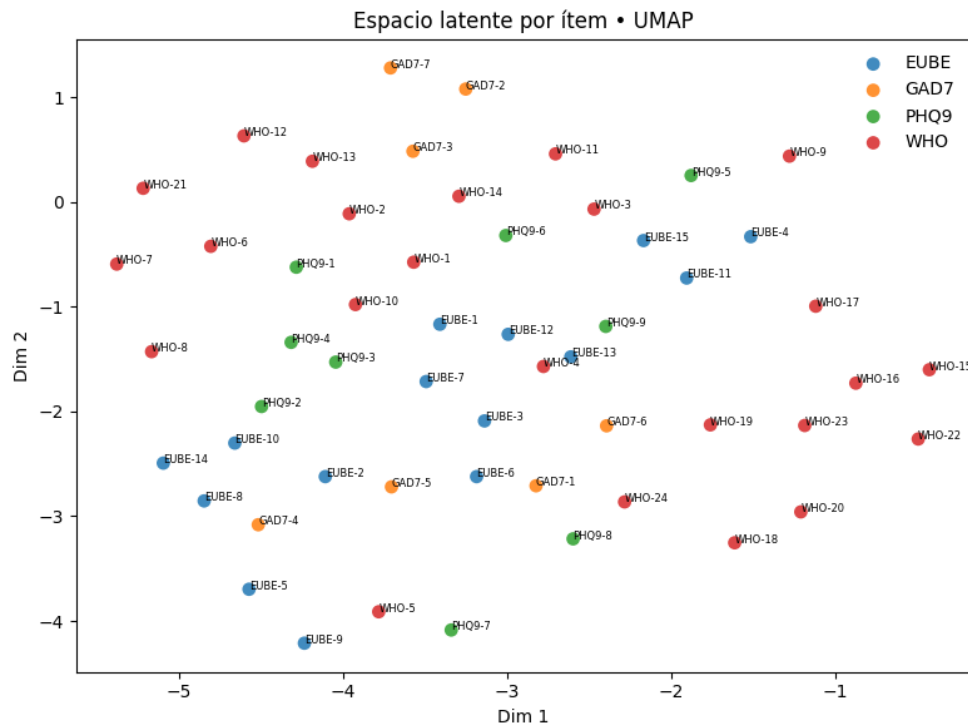


Figura 19: Proyección 2D del espacio latente del VAE mediante UMAP.

Fuente: Elaboración propia.

4.1.3. Selección del modelo de representación semántica

El primer paso realizado, antes del entrenamiento del VAE, fue la evaluación de distintas variantes del modelo SBERT utilizadas para generar los embeddings iniciales de los ítems. Estas representaciones son fundamentales, ya que determinan la calidad semántica de las relaciones aprendidas posteriormente por el VAE y los algoritmos de agrupamiento. Por ello, se compararon los diferentes resultados de una primera clusterización con K-Means para observar su comportamiento en una primera ejecución.

Para evaluar las distintas variantes, se realizó un proceso de validación cruzada cualitativa: se generaron los embeddings, se aplicó un agrupamiento preliminar mediante K-Means y se proyectó el resultado en 2D utilizando UMAP. El objetivo era identificar qué modelo lograba agrupar espontáneamente los ítems conceptualmente similares.

La Figura 20 muestra de manera comparativa la estructura semántica obtenida para las seis variantes del modelo SBERT. En las primeras variantes A y B, se aprecia una distribución más difusa e inestable, con solapamientos considerables entre instrumentos y una organización poco definida. Los clústeres (representados por colores) se superponen significativamente, indicando que el modelo no logra discriminar con nitidez entre matices semánticos distintos como por ejemplo, confundiendo síntomas físicos con emocionales. Esto sugiere que dichas configuraciones capturan relaciones semánticas superficiales, dificultando la identificación de patrones consistentes.

A partir de las variantes intermedias se aprecia una mejora progresiva en la organización espacial: los ítems tienden a agruparse con mayor coherencia y comienzan a emerger estructuras diferenciadas entre constructos como ansiedad, depresión, burnout y calidad de vida.

Finalmente, el modelo seleccionado presenta la organización más clara y estable. Los ítems relacionados a cada instrumento forman agrupamientos definidos, se reducen los solapamientos indeseados y las distancias relativas entre grupos reflejan mejor sus relaciones conceptuales. Esta mayor coherencia fue determinante para seleccionarlo como modelo base para el entrenamiento del VAE, asegurando que el espacio latente final se construya sobre señales semánticas robustas y no sobre ruido vectorial.

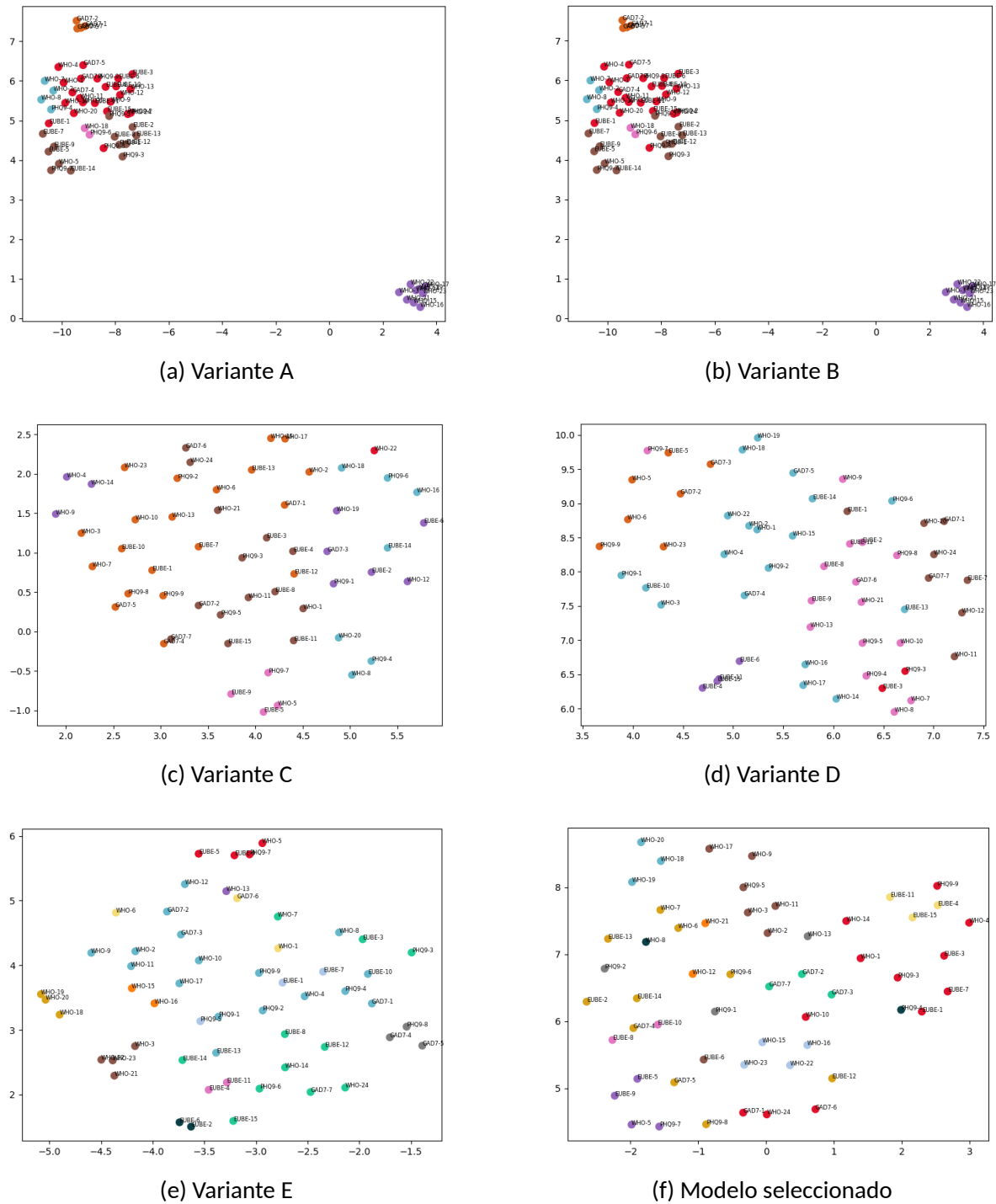


Figura 20: Comparación de las proyecciones generadas por las variantes del modelo SBERT.
Fuente: Elaboración propia.

4.1.4. Análisis del espacio latente y dimensionalidad

El segundo paso del análisis experimental consistió en construir un espacio latente que capturara las relaciones semánticas entre los ítems de los instrumentos PHQ-9, GAD-7, EUBE y WHOQOL-BREF. Para ello se entrenó un *Variational Autoencoder* (VAE), cuyo objetivo fue aprender una representación comprimida y continua capaz de reflejar similitudes conceptuales entre enunciados reformulados.

Con el fin de determinar la dimensionalidad más adecuada (d) para este espacio que equilibrara la compresión de información con la capacidad de representación, se evaluaron distintas configuraciones del VAE variando el número de dimensiones latentes ($d = \{6, 8, 12, 16\}$). Cada configuración fue proyectada posteriormente a dos dimensiones mediante UMAP para analizar visualmente la estabilidad, cohesión y organización semántica de los ítems. Asimismo, se ajustaron hiperparámetros de regularización, específicamente el peso de la divergencia KL β , la tasa de *dropout* y el parámetro de capacidad $C_{\text{máx}}$, los cuales son valores fundamentales para evitar el sobreajuste y garantizar la suavidad del espacio latente. Los resultados de este proceso de ajuste se incluyen en el Anexo para su revisión detallada.

La Figura 21 muestra una comparación directa entre la representación obtenida con 8 dimensiones y la obtenida con 16 dimensiones, esta última correspondiente al modelo final utilizado en el estudio.

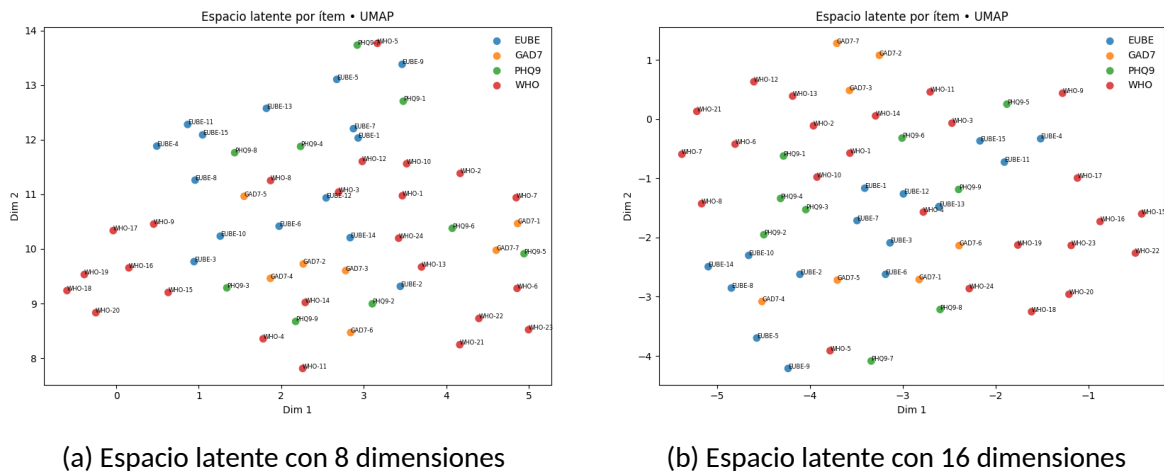


Figura 21: Comparación del espacio latente para 8 y 16 dimensiones.

Fuente: Elaboración propia.

El análisis de la configuración de 8 dimensiones (Figura 21a) revela un espacio latente más disperso y ruidoso, donde los ítems carecen de una estructura consistente entre instrumentos y muestran agrupaciones poco diferenciadas. Por ejemplo, los ítems WHOQOL-BREF (puntos rojos) aparecen mezclados con ítems del EUBE (puntos azules) sin fronteras claras. Este comportamiento indica que la representación es insuficiente para capturar la complejidad

dad semántica de los enunciados, lo cual dificulta la detección posterior de redundancias y patrones de similitud.

Por el contrario, el modelo de 16 dimensiones (Figura 21b) genera un espacio notablemente más estable y coherente. En esta configuración, los ítems muestran una distribución más estructurada, con relaciones semánticas claras y agrupamientos más consistentes dentro de cada instrumento. Es decir, se puede notar cómo los ítems de EUBE forman un clúster compacto (reflejando el contexto puramente académico), mientras que los ítems de ansiedad (GAD-7) y depresión (PHQ-9) ocupan regiones adyacentes con zonas de transición fluida, lo cual es consistente con la superposición sintomática conocida en psicometría. Esta organización favorece tanto la interpretación del espacio latente como la aplicación de métodos de agrupamiento jerárquico, permitiendo una separación más nítida entre constructos psicológicos.

Debido a esta mayor estabilidad y coherencia interna, la configuración de 16 dimensiones (con $dropout = 0.1$ y $\beta = 10$) fue seleccionada como representación final para el resto del análisis. Las proyecciones intermedias (6D y 12D), junto con las variantes de hiperparámetros evaluadas, se presentan en el Anexo para respaldar la elección del modelo final.

Además, la incorporación de $dropout = 0.1$ y $\beta = 10$ en esta configuración final contribuyó a suavizar el espacio, evitando que los ítems colapsaran en puntos singulares y permitiendo una distribución más continua. Debido a esta capacidad para segregar constructos manteniendo la coherencia semántica interna, la configuración de 16 dimensiones fue seleccionada como la base vectorial para las etapas subsiguientes de agrupamiento (*clustering*). Las proyecciones intermedias (6D y 12D), que ilustran la transición gradual de este comportamiento, se detallan en el Anexo.

4.1.5. Resultados de los métodos de clustering

En esta sección se presentan los resultados obtenidos al aplicar los tres enfoques de agrupamiento evaluados: *Feature Agglomeration*, *Top-Down* y el método *Iterativo*. El objetivo del análisis fue identificar estructuras semánticas estables dentro del espacio latente, evaluar la capacidad de cada método para separar adecuadamente los contenidos conceptuales y observar su comportamiento a lo largo del proceso iterativo.

Para cada método se generaron visualizaciones correspondientes a las iteraciones 10 y 20, con el fin de analizar la convergencia del agrupamiento. Estas figuras permiten observar la estabilidad progresiva de los centroides, el grado de solapamiento entre grupos y la coherencia semántica de los ítems asignados en cada etapa del proceso.

A continuación se resumen los principales hallazgos de cada método:

4.1.5.1 Feature Agglomeration

El análisis del método de *Feature Agglomeration* revela un comportamiento de convergencia moderada con limitaciones en la definición de fronteras semánticas.

En la etapa intermedia (Iteración 10), se puede observar en la Figura 22a una fragmentación excesiva en la asignación de clústeres. Regiones del espacio latente que visualmente sugieren cercanía (como el cuadrante inferior izquierdo, asociado a síntomas de ansiedad y estrés académico) aparecen divididas en múltiples subgrupos heterogéneos (tonos rojos, naranjas y rosados). Esto indica que, en fases tempranas, el algoritmo es sensible al ruido local y no logra capturar la estructura global del constructo.

Para la etapa final (Iteración 20), se aprecia una notable consolidación en la Figura 22b. Los grupos pequeños tienden a unificarse, logrando una mayor coherencia en zonas periféricas (como los ítems de bienestar físico en el extremo derecho). Sin embargo, el método persiste en mostrar fronteras difusas en las zonas de transición. Específicamente, se mantiene un solapamiento entre ítems de fatiga académica (EUBE) y síntomas afectivos (PHQ-9/GAD-7), lo que sugiere que *Feature Agglomeration* prioriza la proximidad geométrica inmediata sobre la separación jerárquica de temas, resultando en categorías menos discriminativas para el propósito del cuestionario.

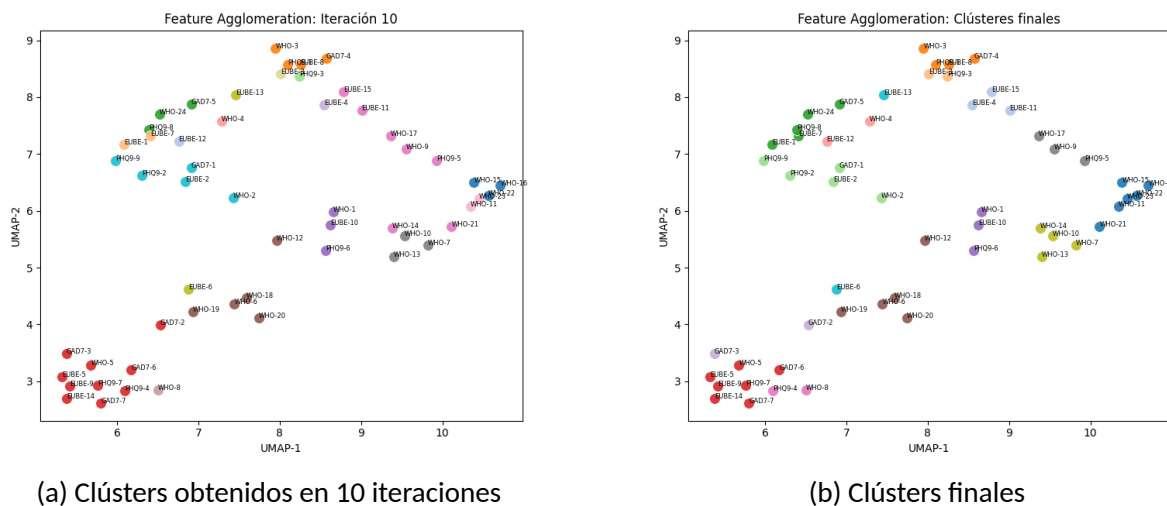


Figura 22: Resultados FA 10 vs. 20 iteraciones
Fuente: Elaboración propia.

4.1.5.2 Método Iterativo

El método iterativo produjo una organización relativamente clara en sus primeras etapas; sin embargo, mostró variaciones importantes entre las iteraciones 10 y 20. Esto sugiere que es un enfoque sensible a pequeños cambios en la geometría del espacio latente y que su convergencia no siempre se traduce en una mejor separación semántica.

Un ejemplo evidente aparece en la zona superior central. En la iteración 10 (Figura 23a), los ítems EUBE-6 y EUBE-10, ambos relacionados con desinterés y cansancio académico, se encuentran agrupados correctamente dentro del mismo clúster. No obstante, en la iteración 20 (Figura 23b) esta relación se rompe: EUBE-10 conserva su posición en el mismo grupo, mientras que EUBE-6 es desplazado hacia un clúster distinto junto a ítems vinculados a ansiedad (GAD-7).

Este cambio, presente incluso en las etapas finales, muestra que el método iterativo puede perder coherencia semántica conforme avanza, lo que dificulta su uso para construir categorías estables en el cuestionario.

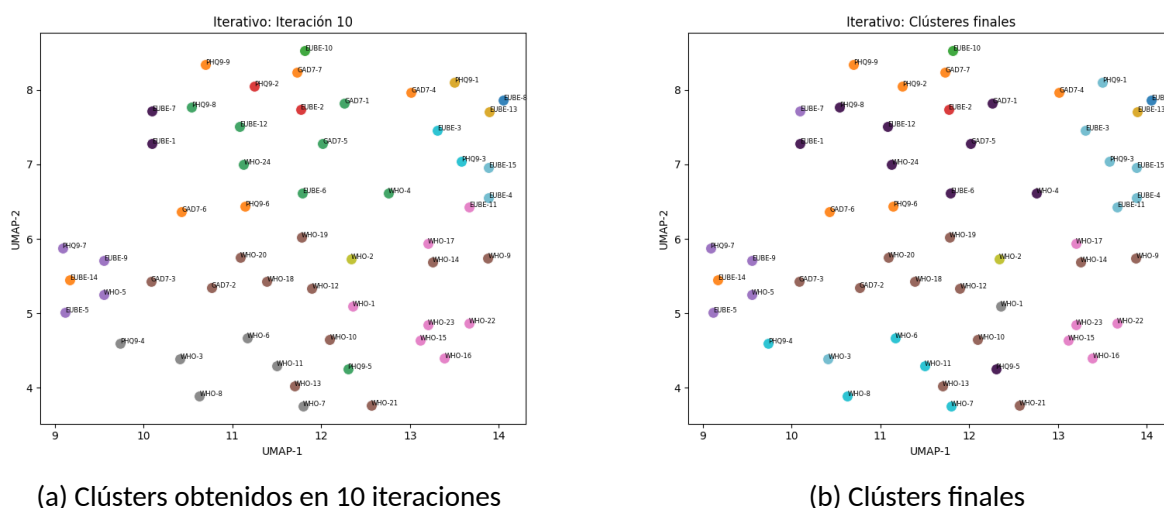


Figura 23: Resultados Método Iterativo
Fuente: Elaboración propia.

4.1.5.3 Método Top-Down

El método *Top-Down* demostró el desempeño más robusto y consistente de los tres enfoques evaluados. A diferencia de la fragmentación observada en *Feature Agglomeration* o la inestabilidad del método *Iterativo*, el enfoque *Top-Down* exhibió una convergencia temprana y estable.

En la Iteración 10 (Figura 24a) ya se aprecia una organización ordenada y fácil de interpretar: los ítems relacionados a los mismos temas tienden a ubicarse cerca entre sí, sin mezclarse con otros dominios. Por ejemplo, los ítems de concentración (EUBE-5, PHQ9-7, WHO-5) aparecen agrupados en una misma zona, mientras que los de anhedonia y fatiga académica forman conglomerados diferenciados. Esto indica que el método capta desde temprano las relaciones conceptuales entre los ítems.

Al terminar en la Iteración 20 (Figura 24b), la estructura se vuelve aún más clara. Los clústeres se estabilizan, las fronteras entre grupos se definen mejor y casi no se observan solapamientos.

tos. Ítems de dominios distintos quedan bien separados, lo que muestra una representación semántica consistente y alineada con la teoría psicológica.

En contraste con los demás métodos, Top-Down no presenta “saltos” inesperados entre iteraciones ni reasignaciones contradictorias. Los grupos formados en iteraciones tempranas se mantienen coherentes en la etapa final, lo que evidencia que el modelo converge sin perder la lógica interna del espacio latente.

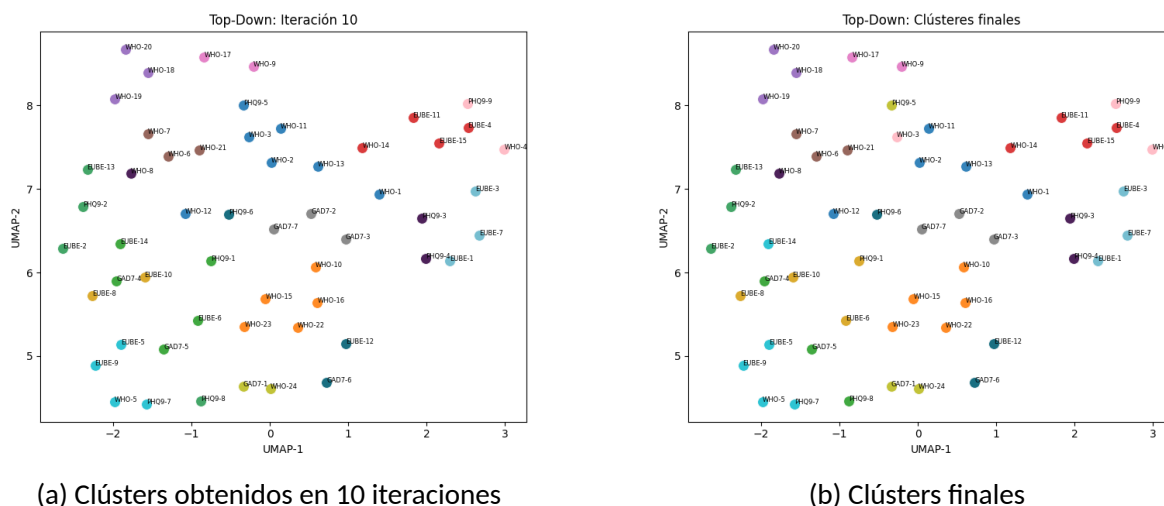


Figura 24: Resultados Método Top Down
Fuente: Elaboración propia.

Más allá de la estabilidad visual, la validación definitiva de este método reside en la coherencia semántica de los grupos resultantes. La Tabla 9 detallan la composición final de los clústeres, destacando la capacidad del modelo para identificar constructos transversales que cruzan las barreras de los instrumentos originales.

Un ejemplo notable es el clúster “Concentración”, que agrupa exitosamente ítems de déficit atencional académico (EUBE), síntomas cognitivos de depresión (PHQ-9) y autopercepción de capacidad (WHO). Del mismo modo, el clúster “Riesgo Suicida” logra asociar semánticamente la ideación suicida (PHQ9-9) con la falta de sentido vital (WHO-3, WHO-4), demostrando que el espacio latente captura matices existenciales complejos.

Por lo tanto, debido a su alta estabilidad inter-iteraciones, la ausencia de solapamientos ambiguos y, fundamentalmente, la interpretabilidad psicométrica de sus agrupamientos, el método *Top-Down* fue seleccionado como la base estructural para el diseño del cuestionario adaptativo. Los clústeres aquí presentados constituyen los nodos de decisión sobre los cuales opera el algoritmo de screening dinámico.

Tabla 9: Clústeres obtenidos en Top-Down.

Clúster	Item_ID	Texto
Inquietud	PHQ9-8	¿Se ha movido o hablado tan lento que otras personas podrían haberlo notado? O lo contrario: muy inquieto(a) o agitado(a), moviéndose mucho más de lo normal.
	GAD7-4	Ha tenido dificultad para relajarse.
	GAD7-5	Se ha sentido tan inquieto(a) que no ha podido quedarse quieto(a).
Agotamiento	EUBE-11	Siento que estudiar me está desgastando físicamente.
	EUBE-15	Creo que estudiar me está desgastando emocionalmente.
	EUBE-4	Creo que estudiar hace que me sienta agotado.
	WHO-14	¿Cómo de satisfecho/a está con su sueño?
Autoestima	WHO-9	¿Es capaz de aceptar su apariencia física?
	WHO-17	¿Cómo de satisfecho/a está de sí mismo?
Preocupación excesiva	GAD7-2	No ha sido capaz de parar o controlar su preocupación.
	GAD7-3	Se ha preocupado demasiado por motivos diferentes.
	GAD7-7	Ha tenido miedo de que algo terrible fuera a pasar.
Condiciones del entorno	WHO-21	¿Cómo de satisfecho/a está de las condiciones del lugar donde vive?
	WHO-6	¿Cuánta seguridad siente en su vida diaria?
	WHO-7	¿Cómo de saludable es el ambiente físico a su alrededor?
Concentración	EUBE-5	Cada vez me es más difícil concentrarme en las clases.
	EUBE-9	Cada vez me cuesta más trabajo ponerle atención al maestro.
	EUBE-14	Siento que tengo más problemas para recordar lo que estudio.
	PHQ9-7	Ha tenido dificultad para concentrarse en ciertas actividades, tales como leer el periódico o ver la televisión.
	WHO-5	¿Cuál es su capacidad de concentración?
Anhedonia	EUBE-6	Me desilusionan mis estudios.
	EUBE-8	No me interesa asistir a clases.
	EUBE-10	El asistir a clases se me hace aburrido.
	PHQ9-1	Poco interés o placer en hacer cosas.

Clúster	Item_ID	Texto
Fatiga académica	EUBE-1	El tener que asistir diariamente a clases me cansa.
	EUBE-3	Durante las clases me siento somnoliento.
	EUBE-7	Antes de terminar mi horario de clases ya me siento cansado.
Afecto negativo	GAD7-1	Se ha sentido nervioso(a), ansioso(a) o con los nervios de punta.
	PHQ9-5	Sin apetito o ha comido en exceso.
	WHO-24	¿Con qué frecuencia tiene sentimientos negativos, tales como tristeza, desesperanza, ansiedad, o depresión?
Bienestar general	WHO-1	¿Hasta qué punto piensa que el dolor (físico) le impide hacer lo que necesita?
	WHO-2	¿En qué grado necesita de un tratamiento médico para funcionar en su vida diaria?
	WHO-11	¿Dispone de la información que necesita para su vida diaria?
	WHO-12	¿Hasta qué punto tiene oportunidad de realizar actividades de ocio?
	WHO-13	¿Es capaz de desplazarse de un lugar a otro?
Relaciones personales	WHO-18	¿Cómo de satisfecho está con sus relaciones personales?
	WHO-19	¿Cómo de satisfecho/a está con su vida sexual?
	WHO-20	¿Cómo de satisfecho/a está con el apoyo que obtiene de sus amigos/as?
Energía y sueño	PHQ9-4	Se ha sentido cansado(a) o con poca energía.
	PHQ9-3	Ha tenido dificultad para quedarse o permanecer dormido(a), o ha dormido demasiado.
	WHO-8	¿Tiene energía suficiente para la vida diaria?
Condiciones de vida	WHO-10	¿Tiene suficiente dinero para cubrir sus necesidades?
	WHO-15	¿Cómo de satisfecho/a está con su habilidad para realizar sus actividades de la vida diaria?
	WHO-16	¿Cómo de satisfecho/a está con su capacidad de trabajo?
	WHO-22	¿Cómo de satisfecho/a está con el acceso que tiene a los servicios sanitarios?
	WHO-23	¿Cómo de satisfecho/a está con los servicios de transporte de su zona?
Ánimo deprimido	EUBE-13	No creo terminar con éxito mis estudios.
	EUBE-2	Mis problemas escolares me deprimen fácilmente.
	PHQ9-2	Se ha sentido decaído(a), deprimido(a) o sin esperanzas.

Clúster	Item_ID	Texto
Riesgo suicida	PHQ9-9	Pensamientos de que estaría mejor muerto(a) o de lastimarse de alguna manera.
	WHO-3	¿Cuánto disfruta de la vida?
	WHO-4	¿Hasta qué punto siente que su vida tiene sentido?
Frustración e irritabilidad	EUBE-12	Cada vez me siento más frustrado por ir a la escuela.
	GAD7-6	Se ha molestado o irritado fácilmente.
	PHQ9-6	Se ha sentido mal con usted mismo(a) - o que es un fracaso o que ha quedado mal con usted mismo(a) o con su familia.

A continuación se presenta una evaluación comparativa de los tres enfoques de agrupamiento. La Tabla 10 resume el desempeño observado en función de tres criterios fundamentales para la construcción del cuestionario adaptativo:

- **Estabilidad de convergencia:** Grado en que el método mantiene asignaciones consistentes entre iteraciones sucesivas.
- **Separación entre clústeres:** Capacidad del modelo para separar los clústeres sin que se mezclen entre sí.
- **Coherencia semántica:** Correspondencia entre los clústeres obtenidos y los constructos psicológicos esperados.

Tabla 10: Evaluación comparativa del desempeño de los algoritmos de clustering.

Método	Estabilidad	Nivel de separación	Calidad semántica
FA	Conserva la estructura general con variaciones menores.	Tendencia al solapamiento en zonas de transición.	Agrupar por similitud local, mezclando constructos distintos.
Iterativo	Muestra cambios tardíos en algunos grupos.	Sensible a variaciones geométricas locales.	Pierde coherencia en etapas avanzadas.
Top-Down	Convergencia estable desde etapas tempranas.	Clústeres compactos sin solapamiento ambiguo.	Identifica constructos transversales y coherentes.

4.1.6. Análisis probabilístico

El análisis probabilístico permitió evaluar el grado de coincidencia entre los métodos de agrupamiento aplicados y determinar qué relaciones entre ítems se mantienen estables a través de distintas configuraciones. Este enfoque hizo posible identificar pares y grupos de ítems que aparecen juntos de manera recurrentes, lo que revela redundancias, núcleos semánticos consistentes y subestructuras robustas dentro de cada instrumento psicométrico.

A partir de las co-asociaciones obtenidas se identificaron diversos clústeres estables, que se pueden observar en la Tabla 11, la cual muestra algunos de los núcleos más representativos detectados.

Tabla 11: Ejemplos de clústeres estables detectados en la matriz de co-asociación.

Clúster 1	EUBE-1	EUBE-3	EUBE-7
Clúster 2	EUBE-4	EUBE-11	EUBE-15
Clúster 3	GAD7-2	GAD7-3	
Clúster 4	WHO-18	WHO-19	WHO-20

Estos grupos evidencian que ciertos ítems funcionan de manera altamente redundante o capturan dimensiones muy específicas dentro del mismo dominio psicológico, como fatiga académica, preocupación excesiva y relaciones personales. Esta información resulta fundamental para la selección posterior de ítems representativos en el cuestionario dinámico.

4.1.6.1 Estadísticas por instrumento

Con el fin de comparar la estructura interna de cada escala, se calcularon la co-asociación promedio y su dispersión, cuyos valores se resumen en la Tabla 12.

Tabla 12: Estadísticas de co-asociación por instrumento.

Instrumento	mean_A	std_A
GAD-7	0.174	0.335
EUBE	0.104	0.292
PHQ-9	0.102	0.259
WHO	0.027	0.164

A partir de estos resultados se observa que GAD-7 es el instrumento con mayor cohesión interna. Esto significa que sus ítems tienden a aparecer juntos con mayor frecuencia que los de las otras escalas. Sin embargo, también presenta la mayor variabilidad, lo que revela

que existe un núcleo de ítems muy similares entre sí junto a otros que se conectan más débilmente.

En EUBE y PHQ-9, la co-asociación media es intermedia y con valores muy cercanos. Esto indica que sus ítems se relacionan entre sí de forma moderada, reflejando la presencia de varias subdimensiones internas (fatiga, sueño, ánimo, cognición). La dispersión es algo mayor en EUBE que en PHQ-9, lo que sugiere que sus subgrupos están un poco más diferenciados entre sí.

Por último, WHOQOL-BREF muestra la cohesión más baja y la menor variabilidad. Esto confirma que se trata de una escala muy heterogénea, donde la mayoría de los ítems representan dimensiones distintas y no se agrupan entre sí, salvo por algunos núcleos vinculados a condiciones de vida y acceso a servicios. En consecuencia, la reducción de ítems por redundancia es más limitada en esta escala, y es necesario priorizar la cobertura de dominios clave por sobre la compactación del instrumento.

4.1.6.2 Matrices de co-asociación

La Figura 25 presenta las matrices de co-asociación intra-instrumento. Cada subfigura refleja la densidad de relaciones internas y permite identificar núcleos semánticos robustos.

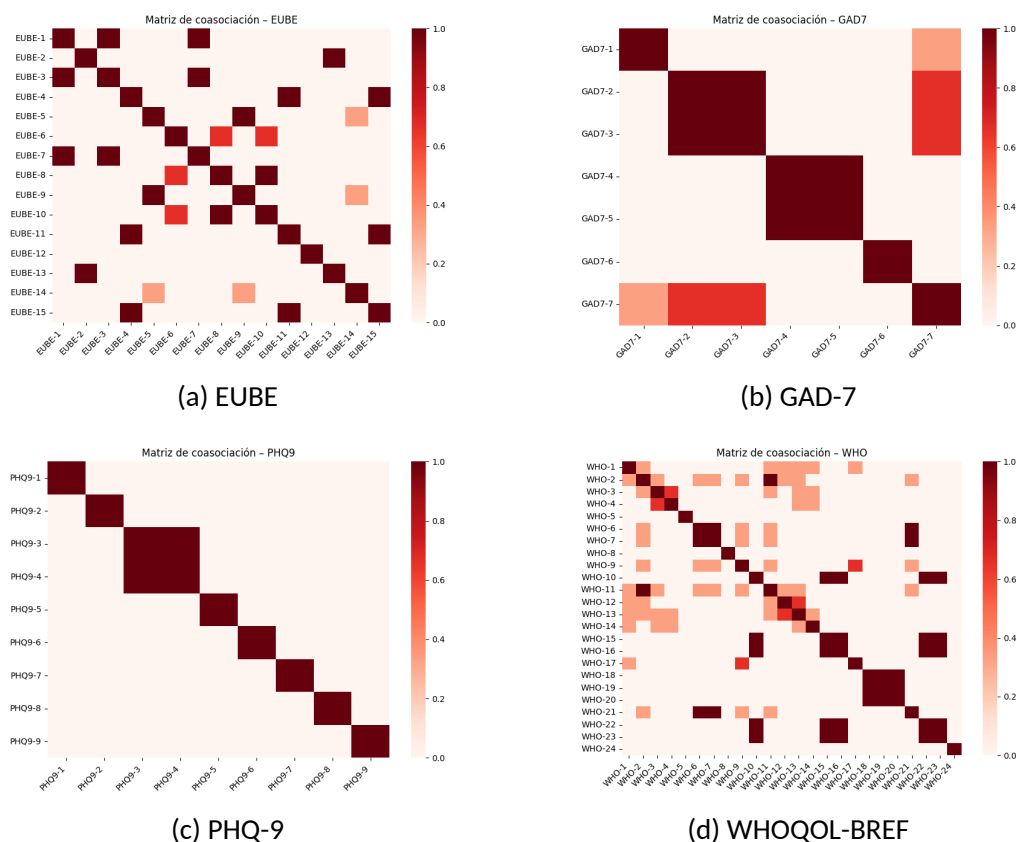


Figura 25: Matrices de co-asociación de cada instrumento psicométrico.

A partir de esta figura se observan los siguientes patrones:

- **EUBE:** Presenta tres grupos bien diferenciados, uno asociado a *fatiga física* (EUBE-1, EUBE-3, EUBE-7), otro a *agotamiento emocional* (EUBE-4, EUBE-11, EUBE-15) y un tercero ligado a la *anhedonia* (EUBE-6, 8, 10). Esto confirma que el burnout estudiantil no es un único bloque, sino que se expresa en varias áreas distintas.
- **GAD-7:** Es el instrumento con mayor consistencia interna. Destaca un núcleo muy fuerte formado por GAD7-2 y GAD7-3 ($A_{ij} = 1.0$), que representan la preocupación excesiva. Además, aparece un segundo grupo estable compuesto por GAD7-4, GAD7-5 y GAD7-6, relacionado con inquietud e irritabilidad.
- **PHQ-9:** Presenta un nivel de interrelación intermedio. Los vínculos más fuertes se concentran en pares como PHQ9-3 y PHQ9-4 (sueño y fatiga) o PHQ9-5 y PHQ9-6 (cognición y autovaloración). El resto de los ítems se organiza en grupos más pequeños y separados, lo que refleja la naturaleza multidimensional del instrumento.
- **WHOQOL-BREF:** Es el instrumento más diverso y con menor coherencia, lo cual coincide con su diseño para evaluar distintas áreas de la calidad de vida.

En conclusión, este análisis probabilístico proporciona la evidencia cuantitativa necesaria para discriminar entre información redundante (núcleos estables de alta co-asociación) e información única. Estos hallazgos fundamentan directamente las reglas de reducción y selección de ítems que se detallan en el diseño del cuestionario adaptativo en la sección siguiente.

4.2. Diseño del cuestionario optimizado

Una vez identificadas las estructuras latentes de los ítems, la coherencia semántica de los clústeres y los patrones probabilísticos de co-asociación, se procedió a consolidar estos resultados en un instrumento aplicado: un **cuestionario dinámico adaptativo**. El objetivo de esta etapa fue transformar la evidencia computacional en un mecanismo de evaluación breve, sensible y capaz de reducir de manera sustantiva la carga de respuesta, manteniendo la cobertura conceptual de los cuatro instrumentos originales (PHQ-9, GAD-7, EUBE y WHOQOL-BREF).

El diseño final se articuló en cinco pasos:

1. Análisis por instrumento.
2. Análisis global de correlación ítem-puntaje total.
3. Análisis respuesta-nivel.

4. Selección de ítems representativos (*ítems puerta*) y definición de umbrales.
5. Construcción del flujo adaptativo y evaluación de la reducción de carga.

A continuación se presentan los principales resultados obtenidos.

4.2.1. Análisis por instrumento: promedios y discriminación

Para cada instrumento se calculó el promedio por ítem según nivel de severidad y la discriminación entre niveles adyacentes. Esto incluyó las diferencias entre los promedios de respuesta del ítem en los niveles N_k^i , las cuales se expresan como:

$$\Delta N_k^{1 \rightarrow 2}, \quad \Delta N_k^{2 \rightarrow 3}, \quad \dots, \quad \Delta N_k^{1 \rightarrow M}$$

con el fin de evaluar cómo progresa la respuesta a medida que aumenta la severidad clínica reportada por el instrumento.

Tabla 13: Promedios por nivel y discriminación por ítem GAD-7

Ítem	N_k^1	N_k^2	N_k^3	N_k^4	$\Delta N_k^{1 \rightarrow 2}$	$\Delta N_k^{1 \rightarrow 4}$
GAD7-1	0.496	1.210	1.866	2.757	0.713	2.261
GAD7-2	0.306	1.194	2.023	2.848	0.888	2.542
GAD7-3	0.482	1.358	2.204	2.888	0.875	2.405
GAD7-4	0.510	1.097	2.000	2.616	0.875	2.405
GAD7-5	0.170	0.579	1.224	2.181	0.409	2.011
GAD7-6	0.333	0.897	1.307	2.151	0.564	1.818
GAD7-7	0.122	0.671	1.110	2.191	0.549	2.069

Finalmente, los resultados obtenidos para el instrumento GAD-7 muestran que ciertas preguntas presentan una alta capacidad para distinguir la transición inicial desde niveles sin sintomatología ansiosa hacia su manifestación leve. En particular, los ítems GAD7-2, GAD7-3 y GAD7-4 destacan por exhibir los mayores valores de $\Delta N_k^{1 \rightarrow 2}$, lo que indica una elevada sensibilidad al primer cambio relevante en el nivel de severidad. Esta característica los posiciona como candidatos apropiados para su uso como ítems puerta dentro del esquema de *screening*.

De manera complementaria, estos ítems mantienen diferencias elevadas entre N1 y N4, lo que sugiere que su comportamiento continúa siendo informativo a medida que la ansiedad aumenta. En contraste, otros ítems presentan variaciones más acotadas en la transición inicial, lo que limita su utilidad como indicadores tempranos. En conjunto, estos resultados

permiten identificar de forma objetiva los ítems más adecuados para activar evaluaciones más profundas únicamente cuando las respuestas iniciales lo justifican.

Antes de presentar los resultados del WHOQOL-BREF, fue necesario ajustar la dirección de algunos de sus ítems para mantener la coherencia interpretativa entre instrumentos (mayor puntaje = mayor bienestar). En particular, los ítems WHO-1 (dolor físico), WHO-2 (necesidad de tratamiento médico) y WHO-24 (frecuencia de sentimientos negativos) están formulados en sentido inverso al resto del instrumento. Por ello, estos valores fueron invertidos previo al cálculo de las diferencias entre niveles. Una vez realizada esta corrección, las progresiones $\Delta N_k^{1 \rightarrow 2}$ y $\Delta N_k^{1 \rightarrow 3}$ mostraron el comportamiento esperado, consistente con una mejora gradual del bienestar.

Tabla 14: Promedios por nivel y discriminación por ítem WHOQOL-BREF

Ítem	N_k^1	N_k^2	N_k^3	$\Delta N_k^{1 \rightarrow 2}$	$\Delta N_k^{1 \rightarrow 3}$
WHO-1	2.617	1.928	1.409	0.688	1.208
WHO-2	2.659	1.777	1.301	0.882	1.358
WHO-3	2.546	3.235	3.869	0.689	1.323
WHO-4	2.277	3.171	4.000	0.894	1.722
WHO-5	2.330	2.840	3.336	0.510	1.006
WHO-6	2.535	2.903	3.482	0.367	0.947
WHO-7	2.795	3.408	3.988	0.612	1.192
WHO-8	2.314	2.966	3.734	0.652	1.420
WHO-9	2.312	3.114	2.967	0.801	1.654
WHO-10	2.719	3.515	4.278	0.795	1.559
WHO-11	3.290	3.857	4.386	0.566	1.095
WHO-12	2.576	3.183	3.596	1.607	1.020
WHO-13	2.471	3.971	4.319	0.500	0.848
WHO-14	1.905	2.573	3.289	0.667	1.383
WHO-15	2.261	2.900	3.813	0.638	1.551
WHO-16	2.308	2.848	3.560	0.539	1.251
WHO-17	2.017	2.726	3.690	0.708	1.672
WHO-18	2.464	3.560	4.329	1.096	1.864
WHO-19	1.754	3.067	4.129	1.312	2.374
WHO-20	3.060	3.871	4.529	0.811	1.468
WHO-21	2.892	3.882	4.687	0.989	1.794
WHO-22	3.377	4.198	4.789	0.821	1.412
WHO-23	2.841	3.484	4.232	0.642	1.391
WHO-24	3.911	3.165	2.380	0.746	1.531

Los resultados obtenidos para WHOQOL-BREF muestran que ciertos ítems presentan una alta capacidad de discriminación entre los niveles N_k^1 y N_k^2 , lo que los convierte en candidatos adecuados para su uso como ítems puerta dentro del esquema de *screening*. En particular, los ítems WHO-12, WHO-18 y WHO-19 destacan por exhibir los mayores valores de $\Delta N_k^{1 \rightarrow 2}$, indicando una transición marcada desde niveles asociados a mayor bienestar hacia estados de deterioro inicial. Esta característica sugiere que dichos ítems son sensibles a la aparición temprana de cambios relevantes en la percepción de calidad de vida.

Adicionalmente, algunos de estos ítems también presentan diferencias elevadas entre N_1 y N_3 , lo que indica que su comportamiento sigue siendo informativo cuando el deterioro es más marcado. Esto sugiere que no solo permiten detectar el primer cambio desde niveles de bienestar, sino que también reflejan adecuadamente escenarios de mayor severidad. En conjunto, estos resultados muestran que ciertos ítems del WHOQOL-BREF pueden utilizarse como puntos de decisión iniciales, permitiendo definir cuándo es necesario profundizar la evaluación sin aplicar el instrumento completo en todos los casos.

Tabla 15: Promedios por nivel y discriminación por ítem EUBE

Ítem	N_k^1	N_k^2	N_k^3	N_k^4	$\Delta N_k^{1 \rightarrow 2}$	$\Delta N_k^{1 \rightarrow 4}$
EUBE-1	1.0	1.953	2.709	3.603	0.953	2.603
EUBE-2	1.2	1.832	2.701	3.672	0.632	2.472
EUBE-3	1.0	1.929	2.596	3.362	0.929	2.362
EUBE-4	1.0	1.914	2.810	3.724	0.914	2.724
EUBE-5	1.0	1.708	2.657	3.534	0.708	2.534
EUBE-6	1.0	1.404	2.342	3.224	0.404	2.224
EUBE-7	1.0	2.062	2.923	3.862	1.062	2.862
EUBE-8	1.0	1.463	2.016	2.758	0.742	1.758
EUBE-9	1.0	1.622	2.362	3.465	0.622	2.465
EUBE-10	1.0	1.719	2.346	3.068	0.719	2.068
EUBE-11	1.0	1.626	2.439	3.672	0.626	2.672
EUBE-12	1.0	1.303	2.294	3.362	0.303	2.362
EUBE-13	1.0	1.268	1.834	2.637	0.268	1.637
EUBE-14	1.0	1.696	2.387	3.413	0.696	2.413
EUBE-15	1.0	1.712	2.745	3.827	0.712	2.827

En el caso del EUBE, el análisis de discriminación por ítem muestra que algunos ítems presentan una alta capacidad para distinguir la transición inicial desde niveles sin burnout hacia su manifestación incipiente. En particular, los ítems EUBE-1, EUBE-3, EUBE-4 y EUBE-7 destacan por exhibir los mayores valores de $\Delta N_k^{1 \rightarrow 2}$, lo que los posiciona como candidatos adecuados a ítems puerta dentro del esquema de *screening*. Estos ítems capturan de manera sensible

el primer cambio relevante en el nivel de agotamiento académico.

Adicionalmente, algunos de estos ítems mantienen diferencias elevadas entre N_k^1 y N_k^4 , especialmente EUBE-7 y EUBE-4, lo que indica que su comportamiento continúa siendo informativo a medida que aumenta la severidad. En contraste, otros ítems presentan diferencias reducidas en la transición inicial, lo que sugiere una menor utilidad para detectar el inicio del burnout. En conjunto, estos resultados permiten identificar de forma objetiva los ítems más apropiados para activar evaluaciones más profundas solo cuando las respuestas iniciales lo justifican.

Tabla 16: Promedios por nivel y discriminación por ítem PHQ-9

Ítem	N_k^1	N_k^2	N_k^3	N_k^4	N_k^5	$\Delta N_k^{1 \rightarrow 2}$	$\Delta N_k^{1 \rightarrow 5}$
PHQ9-1	0.607	1.113	1.564	2.095	2.551	0.506	1.943
PHQ9-2	0.285	0.864	1.250	1.872	2.653	0.579	2.367
PHQ9-3	0.404	1.124	1.653	2.223	2.775	0.719	2.370
PHQ9-4	0.726	1.324	1.974	2.414	2.857	0.598	2.130
PHQ9-5	0.178	0.800	1.506	1.989	2.571	0.621	2.392
PHQ9-6	0.285	0.810	1.358	1.978	2.734	0.525	2.448
PHQ9-7	0.273	0.800	1.333	1.787	2.428	0.526	2.154
PHQ9-8	0.071	0.291	0.666	1.244	2.020	0.220	1.948
PHQ9-9	0.011	0.140	0.442	0.936	1.836	0.128	1.824

Los resultados asociados al PHQ-9 muestran que el análisis de discriminación por ítem evidencia que ciertas preguntas presentan una mayor capacidad para distinguir la transición inicial desde la ausencia de síntomas depresivos hacia su manifestación leve. En particular, los ítems PHQ-2, PHQ-3 y PHQ-5 destacan por exhibir los mayores valores de $\Delta N_k^{1 \rightarrow 2}$, lo que indica una alta sensibilidad al primer cambio relevante en el nivel de severidad. Esta característica los posiciona como candidatos adecuados para su uso como ítems puerta dentro del esquema de *screening*.

Asimismo, varios de estos ítems presentan diferencias elevadas entre N_1 y N_5 , lo que sugiere que su comportamiento continúa siendo informativo a medida que la severidad aumenta. En contraste, ítems como PHQ-8 y PHQ-9 muestran una discriminación reducida en la transición inicial, lo que limita su utilidad como indicadores tempranos. En conjunto, estos resultados permiten seleccionar de forma objetiva los ítems más apropiados para activar evaluaciones más profundas únicamente cuando las respuestas iniciales lo justifican.

4.2.2. Análisis global: correlación ítem-puntaje total

Posteriormente se realizó un análisis de correlación entre cada ítem normalizado y el puntaje total del instrumento correspondiente. Este análisis permitió identificar qué ítems presentan la mayor contribución a la variabilidad del puntaje global, es decir, aquellos que mejor reflejan la severidad general del constructo evaluado por cada instrumento.

Con el fin de focalizar la interpretación y evitar redundancia, en esta subsección se analizan únicamente los tres ítems con mayor correlación ítem-puntaje total para cada instrumento psicométrico. Estos ítems representan los elementos más influyentes dentro de cada escala y constituyen candidatos naturales a *ítems representativos* en el diseño del screening dinámico.

En el caso del WHOQOL-BREF, cuyo instrumento evalúa cuatro dominios diferenciados (Físico, Psicológico, Relaciones Sociales y Ambiente), se aplicó además un análisis desagregado por dominio. Este procedimiento permitió identificar, para cada área del bienestar, los ítems que más aportan a la variabilidad dentro de su propio dominio, lo que habilita una interpretación más precisa de los factores que contribuyen al estado de bienestar de los estudiantes.

Los ítems con correlaciones más altas son especialmente relevantes porque:

- Capturan la mayor parte de la variación del puntaje total del instrumento.
- Reflejan de manera consistente aumentos en la severidad del constructo.
- Poseen un rol central en la estructura interna del instrumento y, por tanto, son potenciales candidatos para ser utilizados como *ítems puerta* en el cuestionario adaptativo.

A continuación se presentan las tablas con los resultados del análisis de correlación para cada instrumento:

Tabla 17: Top 3 ítems con mayor correlación ítem-puntaje total PHQ-9

Ítem	Correlación
PHQ9-2	0.743
PHQ9-6	0.700
PHQ9-4	0.689

En el caso del PHQ-9, el ítem con mayor correlación fue *PHQ9-2*, correspondiente a la afirmación “Se ha sentido decaído(a), deprimido(a) o sin esperanzas”. Este ítem captura de manera directa el síntoma central del constructo depresivo y presenta la mayor asociación con el puntaje total del instrumento, lo que indica que su variación explica gran parte de la severidad reportada.

El segundo ítem más influyente fue *PHQ9-6* (“Se ha sentido mal con usted mismo(a) - o que es un fracaso o que ha quedado mal con usted mismo(a) o con su familia”), el cual representa la dimensión cognitiva negativa de la depresión. Finalmente, *PHQ9-4* (“Se ha sentido cansado(a) o con poca energía”), también mostró una asociación considerable, reflejando la contribución de síntomas fisiológicos al cuadro depresivo global.

Tabla 18: Top 3 ítems con mayor correlación ítem–puntaje total GAD-7

Ítem	Correlación
GAD7-2	0.857
GAD7-3	0.848
GAD7-1	0.812

En el GAD-7, el ítem con mayor correlación fue *GAD7-2* (“No ha sido capaz de parar o controlar su preocupación”), el cual representa el núcleo del trastorno de ansiedad generalizada y refleja la pérdida de control debido al pensamiento repetitivo e intrusivo.

El segundo ítem más asociado al puntaje total fue *GAD7-3* (“Se ha preocupado demasiado por diferentes motivos”), que refuerza la dimensión cognitiva ansiosa. Posteriormente, *GAD7-1* (“Se ha sentido nervioso(a), ansioso(a) o con los nervios de punta”) mostró también una fuerte asociación, vinculada a la activación fisiológica inicial característica de los trastornos ansiosos.

Tabla 19: Top 3 ítems con mayor correlación ítem–puntaje total EUBE

Ítem	Correlación
EUBE-12	0.756
EUBE-15	0.740
EUBE-9	0.729

Para el instrumento EUBE, el ítem con mayor correlación fue *EUBE-12* (“Siento que estudiar me está desgastando físicamente”), el cual refleja de forma directa la dimensión de agotamiento físico asociada al burnout académico.

El ítem *EUBE-15* (“Creo que estudiar me está desgastando emocionalmente”) mostró también un aporte considerable, representando la dimensión emocional del desgaste. Finalmente, *EUBE-9* (“Cada vez me cuesta más trabajo ponerle atención al maestro”) destaca por su relación con la fatiga cognitiva, una manifestación central del burnout en estudiantes.

Tabla 20: Top 3 ítems con mayor correlación ítem-puntaje total WHO-FIS

Ítem	Correlación
WHO-15	0.745
WHO-8	0.679
WHO-14	0.613

En el dominio físico del WHOQOL-BREF, el ítem más relevante fue *WHO-15*, relacionado con la satisfacción respecto a la capacidad para realizar actividades de la vida diaria. Este ítem sintetiza de manera robusta el funcionamiento físico general y mostró la mayor asociación con la puntuación total del dominio.

Le siguen *WHO-8*, que evalúa el nivel de energía disponible para las actividades cotidianas, y *WHO-14*, que mide la calidad del sueño, ambos componentes fundamentales de la percepción de bienestar físico.

Tabla 21: Top 3 ítems con mayor correlación ítem-puntaje total WHO-PSI

Ítem	Correlación
WHO-17	0.823
WHO-3	0.775
WHO-4	0.757

En el dominio psicológico, el ítem con mayor correlación fue *WHO-17*, referido al grado de satisfacción personal consigo mismo. Este ítem captura un componente esencial del bienestar psicológico y mostró la mayor sensibilidad para reflejar variaciones en este dominio.

El ítem *WHO-3* ("¿Cuánto disfruta de la vida?") refuerza la dimensión hedónica, mientras que *WHO-4* ("¿Hasta qué punto siente que su vida tiene sentido?") aporta la perspectiva existencial del bienestar general.

Tabla 22: Top 3 ítems con mayor correlación ítem-puntaje total WHO-SOC

Ítem	Correlación
WHO-18	0.832
WHO-19	0.757
WHO-20	0.693

Dentro del dominio social, el ítem con mayor peso fue *WHO-18*, relacionado con la satisfacción respecto a las relaciones personales. Este ítem representa de manera directa la calidad

del soporte social percibido, mostrando la mayor asociación con la puntuación total del dominio.

El ítem *WHO-19* aborda la vida sexual y *WHO-20* el apoyo recibido por parte de amistades, ambos aspectos complementarios que contribuyen a caracterizar la esfera social del bienestar.

Tabla 23: Top 3 ítems con mayor correlación ítem–puntaje total WHO-AMB

Ítem	Correlación
WHO-21	0.718
WHO-22	0.676
WHO-10	0.647

Finalmente, en el dominio ambiental, el ítem más relevante fue *WHO-21*, que evalúa la satisfacción con las condiciones del lugar de residencia, un componente central del entorno cotidiano.

Le sigue *WHO-22*, relacionado con el acceso a servicios sanitarios, y *WHO-10*, que evalúa la disponibilidad de dinero para cubrir necesidades personales. Estos ítems reflejan elementos críticos del contexto ambiental que influyen directamente en la calidad de vida.

4.2.3. Análisis respuesta–nivel

El análisis respuesta–nivel tuvo como objetivo identificar patrones de co-ocurrencia en los que determinados ítems presentan respuestas altas de manera simultánea en los niveles de mayor severidad de cada instrumento. Este análisis permite observar qué preguntas se activan conjuntamente cuando un estudiante experimenta síntomas de forma intensa, lo cual proporciona evidencia transversal sobre:

- Cuáles ítems caracterizan de manera consistente los perfiles severos.
- Qué instrumentos miden facetas relacionadas del malestar psicológico.
- Cómo los clústeres definidos previamente encuentran respaldo empírico en las respuestas reales.

Los resultados muestran que en cada instrumento existen grupos de ítems que se responden alto en conjunto en los niveles severos de los instrumentos. A su vez, ciertos ítems provenientes de otros instrumentos también aparecen elevados junto a estos grupos, lo que indica la presencia de *síntomas transversales* entre dominios.

Por ejemplo:

- En **EUBE**, los ítems EUBE-7, EUBE-15, EUBE-4 y EUBE-2 se activan conjuntamente, y además co-ocurren con PHQ9-4, PHQ9-1 y GAD7-3, lo cual sugiere una fuerte conexión entre agotamiento, desesperanza y ansiedad fisiológica.
- En **PHQ-9**, los ítems PHQ9-4, PHQ9-3, PHQ9-6 y PHQ9-2 se mantienen altos en depresión severa, acompañados de elementos de GAD-7 (p. ej., GAD7-3 y GAD7-4) y EUBE-2, evidenciando la naturaleza altamente comórbida entre depresión, ansiedad y agotamiento académico.
- En **GAD-7**, los ítems GAD7-3, GAD7-2, GAD7-1 y GAD7-4 se activan juntos en ansiedad severa, co-ocuriendo nuevamente con síntomas de depresión (PHQ9-4) y agotamiento (EUBE-15), confirmando la interacción entre preocupación excesiva, inquietud y afecto negativo.
- En **WHOQOL-BREF**, los perfiles de menor bienestar físico, psicológico, social y ambiental muestran co-activación consistente de ítems como WHO-22, WHO-24, WHO-11 y WHO-13, los cuales también aparecen elevados en estudiantes con altos puntajes en PHQ-9, GAD-7 y EUBE, reflejando que el deterioro del bienestar general emerge en conjunto con sintomatología emocional elevada.

En conjunto, estos patrones de co-ocurrencia respaldan la validez de los clústeres obtenidos mediante agrupamiento latente, ya que los grupos de síntomas que emergen empíricamente al observar las respuestas severas coinciden con las agrupaciones semánticas obtenidas anteriormente.

4.2.4. Selección de ítems puerta

La selección del ítem puerta de cada clúster se basó en dos criterios complementarios:

1. **Discriminación** $\Delta N_k^{1 \rightarrow 2}$: Criterio principal, pues representa el punto donde aparece la primera señal del síntoma (“detección temprana”).
2. **Correlación ítem-puntaje total**: Garantiza que el ítem representa adecuadamente el constructo clínico evaluado.

La selección del ítem representativo (ítem puerta) para cada clúster identificado se basó en un equilibrio entre ambos criterios. Se priorizaron ítems con alta discriminación temprana, ya que estos permiten detectar de forma sensible el primer cambio clínicamente relevante, junto con una correlación suficientemente alta que asegure que el ítem refleja el dominio

evaluado. En aquellos casos donde existían diferencias moderadas entre candidatos, se privilegió además la claridad semántica y la coherencia clínica del enunciado, con el fin de facilitar su interpretación como punto inicial de decisión dentro del cuestionario adaptativo.

Por ejemplo, en el caso del clúster “Inquietud”, el ítem GAD7-4 presenta simultáneamente la mayor correlación con el puntaje total del instrumento ($r = 0.789$) y el mayor valor de discriminación temprana entre los niveles $N_1 \rightarrow N_2$ ($\Delta = 0.587$), superando a los restantes ítems candidatos del clúster. Este comportamiento indica que dicho ítem es especialmente sensible para detectar el primer aumento significativo de sintomatología asociada a inquietud y dificultad para relajarse. La convergencia entre alta representatividad global y buena capacidad de discriminación temprana justifica su elección como ítem puerta, ya que permite activar evaluaciones más profundas únicamente cuando la respuesta inicial sugiere un cambio relevante en el estado del participante.

Tabla 24: Selección de ítem representativo para el clúster Inquietud

Ítem	Correlación	$\Delta_{N1 \rightarrow N2}$	Texto del ítem
PHQ9-8	0.616	0.220	¿Se ha movido o hablado tan lento que otras personas podrían haberlo notado? O lo contrario - muy inquieto(a) o agitado(a) que ha estado moviéndose mucho más de lo normal
GAD7-4	0.789	0.587	Ha tenido dificultad para relajarse
GAD7-5	0.704	0.409	Se ha sentido tan inquieto(a) que no ha podido quedarse quieto(a)

Tabla 25: Selección de ítem representativo para el clúster Agotamiento

Ítem	Correlación	$\Delta_{N1 \rightarrow N2}$	Texto del ítem
EUBE-11	0.711	0.626	Siento que estudiar me está desgastando físicamente
EUBE-15	0.740	0.712	Creo que estudiar me está desgastando emocionalmente
EUBE-4	0.710	0.914	Creo que estudiar hace que me sienta agotado
WHO-14	0.409	0.667	¿Cómo de satisfecho/a está con su sueño?

Tabla 26: Selección de ítem representativo para el clúster Relaciones personales

Ítem	Correlación	$\Delta_{N1 \rightarrow N2}$	Texto del ítem
WHO-18	0.832	1.096	¿Cómo de satisfecho/a está con sus relaciones personales?
WHO-19	0.757	1.012	¿Cómo de satisfecho/a está con su vida sexual?
WHO-20	0.693	0.811	¿Cómo de satisfecho/a está con el apoyo que obtiene de sus amigos/as?

Tabla 27: Selección de ítem representativo para el clúster Condiciones del entorno

Ítem	Correlación	$\Delta_{N1 \rightarrow N2}$	Texto del ítem
WHO-21	0.718	0.989	¿Cómo de satisfecho/a está de las condiciones del lugar donde vive?
WHO-6	0.308	0.367	¿Cuánta seguridad siente en su vida diaria?
WHO-7	0.603	0.612	¿Cómo de saludable es el ambiente físico a su alrededor?

Tabla 28: Selección de ítem representativo para el clúster Autoestima

Ítem	Correlación	$\Delta_{N1 \rightarrow N2}$	Texto del ítem
WHO-9	0.678	0.801	¿Es capaz de aceptar su apariencia física?
WHO-17	0.823	0.708	¿Cómo de satisfecho/a está de sí mismo/a?

Tabla 29: Selección de ítem representativo para el clúster Preocupación excesiva

Ítem	Correlación	$\Delta_{N1 \rightarrow N2}$	Texto del ítem
GAD7-2	0.857	0.888	No ha sido capaz de parar o controlar su preocupación
GAD7-3	0.848	0.875	Se ha preocupado demasiado por motivos diferentes
GAD7-7	0.714	0.549	Ha tenido miedo de que algo terrible fuera a pasar

Tabla 30: Selección de ítem representativo para el clúster Afecto negativo

Ítem	Correlación	$\Delta_{N1 \rightarrow N2}$	Texto del ítem
GAD7-1	0.812	0.713	Se ha sentido nervioso(a), ansioso(a) o con los nervios de punta
PHQ9-5	0.667	0.621	Sin apetito o ha comido en exceso
WHO-24	0.722	0.746	¿Con qué frecuencia tiene sentimientos negativos, tales como tristeza, desesperanza, ansiedad, o depresión?

Tabla 31: Selección de ítem representativo para el clúster Concentración

Ítem	Correlación	$\Delta_{N1 \rightarrow N2}$	Texto del ítem
EUBE-5	0.719	0.708	Cada vez me es más difícil concentrarme en las clases
EUBE-9	0.729	0.622	Cada vez me cuesta más trabajo ponerle atención al maestro
EUBE-14	0.641	0.696	Siento que tengo más problemas para recordar lo que estudio
PHQ9-7	0.619	0.526	Ha tenido dificultad para concentrarse en ciertas actividades, tales como leer el periódico o ver la televisión
WHO-5	0.755	0.510	¿Cuál es su capacidad de concentración?

Tabla 32: Selección de ítem representativo para el clúster Anhedonia

Ítem	Correlación	$\Delta_{N1 \rightarrow N2}$	Texto del ítem
EUBE-6	0.605	0.404	Me desilusionan mis estudios
EUBE-8	0.545	0.463	No me interesa asistir a clases
EUBE-10	0.586	0.719	El asistir a clases se me hace aburrido
PHQ9-1	0.660	0.506	Poco interés o placer en hacer cosas

Tabla 33: Selección de ítem representativo para el clúster Energía y sueño

Ítem	Correlación	$\Delta_{N1 \rightarrow N2}$	Texto del ítem
PHQ9-4	0.689	0.598	Se ha sentido cansado(a) o con poca energía
PHQ9-3	0.628	0.719	Ha tenido dificultad para quedarse o permanecer dormido(a), o ha dormido demasiado
WHO-8	0.679	0.652	¿Tiene energía suficiente para la vida diaria?

Tabla 34: Selección de ítem representativo para el clúster Fatiga académica

Ítem	Correlación	$\Delta_{N1 \rightarrow N2}$	Texto del ítem
EUBE-1	0.657	0.953	El tener que asistir diariamente a clases me cansa
EUBE-3	0.608	0.929	Durante las clases me siento somnoliento
EUBE-7	0.706	1.062	Antes de terminar mi horario de clases ya me siento cansado

Tabla 35: Selección de ítem representativo para el clúster Ánimo deprimido

Ítem	Correlación	$\Delta_{N1 \rightarrow N2}$	Texto del ítem
EUBE-13	0.593	0.268	No creo terminar con éxito mis estudios
EUBE-2	0.703	0.632	Mis problemas escolares me deprimen fácilmente
PHQ9-2	0.743	0.579	Se ha sentido decaído(a), deprimido(a) o sin esperanzas

Tabla 36: Selección de ítem representativo para el clúster Frustración e irritabilidad

Ítem	Correlación	$\Delta_{N1 \rightarrow N2}$	Texto del ítem
EUBE-12	0.756	0.303	Cada vez me siento más frustrado por ir a la escuela
GAD7-6	0.651	0.564	Se ha molestado o irritado fácilmente
PHQ9-6	0.700	0.525	Se ha sentido mal con usted mismo(a) – que es un fracaso o que ha quedado mal con usted mismo(a) o con su familia

Tabla 37: Selección de ítem representativo para el clúster Riesgo suicida

Ítem	Correlación	$\Delta_{N1 \rightarrow N2}$	Texto del ítem
PHQ9-9	0.625	0.128	Pensamientos de que estaría mejor muerto(a) o de lastimarse de alguna manera
WHO-3	0.775	0.689	¿Cuánto disfruta de la vida?
WHO-4	0.757	0.894	¿Hasta qué punto siente que su vida tiene sentido?

El clúster “Riesgo suicida” constituye un caso particularmente relevante, ya que ilustra la diferencia entre relevancia clínica directa y capacidad discriminativa temprana. Si bien el ítem *PHQ9-9* aborda de manera explícita la presencia de ideación suicida, su bajo valor de discriminación entre N_1 y N_2 ($\Delta = 0.128$) indica que no es el más sensible para detectar transiciones iniciales desde la ausencia de sintomatología. En contraste, los ítems *WHO-4* y *WHO-3* presentan mayores valores de discriminación temprana y correlación, capturando cambios graduales en el sentido de vida y disfrute, que suelen preceder a manifestaciones más severas. Este resultado refuerza la lógica del enfoque adaptativo, donde el ítem puerta no se selecciona únicamente por su contenido explícito, sino por su capacidad estadística para anticipar escenarios de mayor riesgo, permitiendo una activación progresiva y controlada del cuestionario.

Tabla 38: Selección de ítem representativo para el clúster Bienestar general

Ítem	Correlación	$\Delta_{N1 \rightarrow N2}$	Texto del ítem
WHO-1	0.493	0.688	¿Hasta qué punto piensa que el dolor (físico) le impide hacer lo que necesita?
WHO-2	0.525	0.882	¿En qué grado necesita de un tratamiento médico para funcionar en su vida diaria?
WHO-11	0.584	0.566	¿Dispone de la información que necesita para su vida diaria?
WHO-12	0.491	0.607	¿Hasta qué punto tiene oportunidad de realizar actividades de ocio?
WHO-13	0.613	0.500	¿Es capaz de desplazarse de un lugar a otro?

Tabla 39: Selección de ítem representativo para el clúster Condiciones de vida

Ítem	Correlación	$\Delta_{N1 \rightarrow N2}$	Texto del ítem
WHO-10	0.647	0.795	¿Tiene suficiente dinero para cubrir sus necesidades?
WHO-15	0.745	0.638	¿Cómo de satisfecho/a está con su habilidad para realizar sus actividades de la vida diaria?
WHO-16	0.602	0.539	¿Cómo de satisfecho/a está con su capacidad de trabajo?
WHO-22	0.676	0.821	¿Cómo de satisfecho/a está con el acceso que tiene a los servicios sanitarios?
WHO-23	0.718	0.642	¿Cómo de satisfecho/a está con los servicios de transporte de su zona?

A diferencia del resto de los clústeres, en “Bienestar general” y “Condiciones de vida” no se seleccionó un ítem puerta. Si bien algunos ítems presentan buena correlación con el puntaje total y diferencias claras entre niveles de severidad, la heterogeneidad semántica de estos clústeres (dolor, movilidad, acceso a servicios, recursos económicos, entre otros) impide que un único ítem represente adecuadamente el constructo completo. Por esta razón, estos clústeres fueron incorporados al cuestionario sin asignar un ítem representativo, por lo que se deben preguntar cada uno de ellos en todos los casos.

4.2.5. Definición de umbrales de activación

Los umbrales de activación del flujo adaptativo se definieron a partir del ítem puerta de cada clúster. Para cada uno de estos ítems se calcularon las medias de los niveles N1 (ausencia de síntomas) y N2 (síntomas leves), obtenidas en el análisis respuesta-nivel. A partir de ellas, el umbral se definió como el punto medio entre ambas medias:

$$\tau_i = \frac{\mu_{N1} + \mu_{N2}}{2}.$$

Los valores de umbral obtenidos se expresan en la escala original de respuesta de cada instrumento (PHQ-9 y GAD-7: 0-3; EUBE: 1-4; WHOQOL-BREF: 1-5), lo que permite interpretar directamente los criterios de activación sin necesidad de transformaciones adicionales. De este modo, el umbral define el punto a partir del cual la respuesta a un ítem puerta deja de ser compatible con la ausencia de sintomatología y pasa a indicar la primera manifestación relevante del constructo evaluado, activando la exploración posterior del clúster correspondiente.

Tabla 40: Umbrales de activación definidos para cada clúster

Clúster	Umbral de activación
Inquietud	$GAD7-4 \geq 1$
Agotamiento	$EUBE-15 \geq 2$
Relaciones personales	$WHO-18 \leq 3$
Condiciones del entorno	$WHO-21 \leq 3$
Autoestima	$WHO-17 \leq 2$
Afecto negativo	$GAD7-1 \geq 1$
Preocupación excesiva	$GAD7-2 \geq 1$
Concentración	$EUBE-5 \geq 2$
Anhedonia	$PHQ9-1 \geq 1$
Energía y sueño	$PHQ9-4 \geq 2$
Fatiga académica	$EUBE-7 \geq 2$
Ánimo deprimido	$PHQ9-2 \geq 1$
Riesgo suicida	$WHO-4 \leq 2$
Frustración e irritabilidad	$EUBE-12 \geq 1$

En síntesis, la definición de los umbrales de activación permitió transformar la evidencia cuantitativa obtenida en reglas claras de decisión para el cuestionario adaptativo. Este proceso aseguró que la apertura de cada rama del screening se active únicamente cuando existen indicios suficientes de un cambio clínicamente relevante. Con estos umbrales definidos, fue posible construir la estructura final del cuestionario dinámico optimizado, presentada en la siguiente sección.

4.3. Estructura Final del Cuestionario Adaptativo

A partir del análisis probabilístico, la discriminación por ítems y la co-asociación entre constructos, se definió la estructura final del **cuestionario adaptativo optimizado**. Este instrumento se organiza mediante un **árbol de decisión**, en el cual cada ítem funciona como *ítem puerta* y determina la activación o no de un clúster de preguntas asociado a un dominio específico.

Cada uno de estos ítems representativos cuenta con un umbral previamente definido a partir del análisis estadístico. Cuando la respuesta entregada por el participante supera dicho umbral, el clúster asociado se activa y se despliegan sus ítems complementarios. En caso contrario, el sistema continúa hacia el siguiente punto de decisión sin abrir preguntas adicionales. La Figura 26 presenta el flujo completo del cuestionario adaptativo resultante.

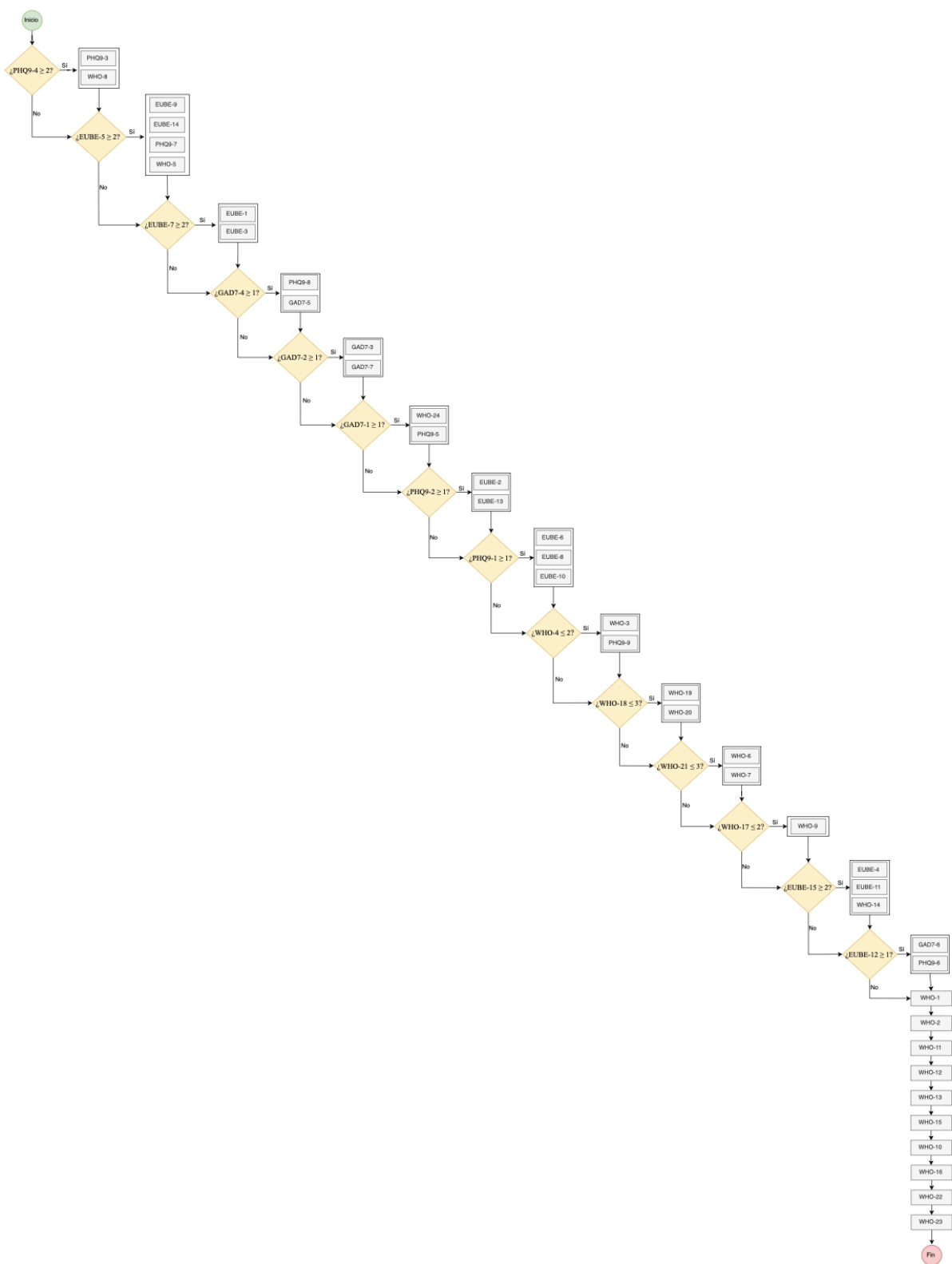


Figura 26: Árbol de decisión del cuestionario adaptativo.

Reducción de longitud

El cuestionario original incluía un total de:

$$9(\text{PHQ-9}) + 7(\text{GAD-7}) + 15(\text{EUBE}) + 26(\text{WHOQOL-BREF}) = 57 \text{ ítems.}$$

Con la estructura adaptativa propuesta, si solo se presentan los ítems puerta en el escenario donde ningún clúster es activado, la carga mínima de respuesta es:

$$\text{Carga mínima} \approx 16 \text{ ítems} \Rightarrow \text{reducción aproximada del } 72 \%.$$

La Figura 27 muestra la secuencia mínima que enfrentaría un participante que no activa ningún clúster. En este escenario, el estudiante responde únicamente 16 ítems en lugar de los 57 del cuestionario original, lo que representa una disminución sustancial en el tiempo de aplicación y en la demanda cognitiva. Esta reducción es especialmente relevante en personas sin sintomatología significativa, para quienes responder la totalidad de los instrumentos tradicionales puede resultar innecesario y generar fatiga o pérdida de concentración.

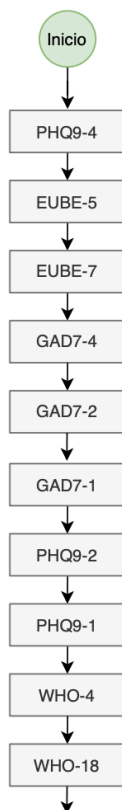


Figura 27: Secuencia mínima del cuestionario adaptativo (Parte 1).

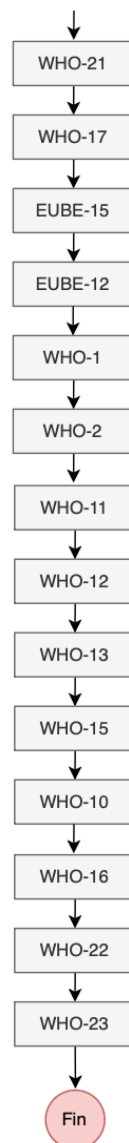


Figura 28: Secuencia mínima del cuestionario adaptativo (Parte 2).

En el escenario contrario, donde todos los ítems puerta superan su respectivo umbral, el participante responde la totalidad de los ítems complementarios del cuestionario. No obstante, a diferencia del instrumento tradicional, este recorrido no corresponde a una secuencia lineal predefinida, sino a un orden reestructurado que agrupa los ítems según los clústeres identificados en el análisis de co-asociación. Esta organización favorece una mayor coherencia temática entre preguntas consecutivas, ya que los ítems se agrupan en función de dominios psicológicos que presentan relaciones empíricas entre sí, evitando saltos abruptos entre constructos no relacionados. Como consecuencia, el participante permanece durante más tiempo en un mismo marco cognitivo y emocional, lo que reduce la fragmentación del proceso de respuesta y favorece una mayor consistencia interna en las respuestas entregadas.

das. En este sentido, el cuestionario adaptativo no solo optimiza la longitud del instrumento, sino que introduce una estructura de aplicación más ordenada y congruente con la forma en que los síntomas se manifiestan y se experimentan en la práctica, incluso en escenarios de alta sintomatología.

Comparación estructural entre el cuestionario original y el adaptativo

La Tabla 41 resume las principales diferencias entre el cuestionario original y la versión adaptativa. Mientras el instrumento tradicional presenta una estructura lineal y fija, el cuestionario optimizado introduce un mecanismo secuencial que ajusta la longitud del recorrido según las respuestas del participante.

Tabla 41: Comparación entre el cuestionario original y el adaptativo.

Aspecto	Cuestionario original	Cuestionario adaptativo
Estructura	Lineal, fija para todos los casos	Ramificada, dependiente de umbrales
Número de ítems	57, siempre respondidos	Entre ~16 y 57, según activaciones
Orden	Secuencia única predeterminada	Orden dinámico según ítems puerta
Redundancia	Mayor presencia de ítems correlacionados	Reducción mediante activación selectiva
Personalización	No presenta	Adaptación por patrón de respuesta
Carga de respuesta	Igual para todos los participantes	Variable, menor en la mayoría de los casos

4.4. Resumen del capítulo

Este capítulo presentó los resultados del proceso de construcción y validación interna de la solución propuesta. A partir del análisis probabilístico, la discriminación por ítems y la co-asociación entre constructos, se definieron los clústeres e ítems puerta utilizados para estructurar el cuestionario adaptativo. Posteriormente, se construyó el árbol de decisión que regula la activación de cada clúster y se evaluó su efecto en la reducción de la carga de respuesta, mostrando una disminución significativa respecto del cuestionario original. Finalmente, se compararon ambas versiones, destacando que el enfoque adaptativo permite organizar el instrumento de manera más eficiente, reduciendo redundancia y ajustando la extensión del recorrido según el patrón de respuestas de cada participante.

CAPÍTULO 5

CONCLUSIONES

Las evaluaciones de salud mental en estudiantes universitarios suelen requerir la aplicación de varios instrumentos psicométricos, por lo que en la construcción de un cuestionario que incluya todas las preguntas de cada instrumento, genera una carga de respuesta elevada, introduce redundancias y dificulta su uso sostenido en contextos donde el tiempo disponible es limitado y la participación de los estudiantes puede variar. Esta situación plantea la necesidad de explorar alternativas que permitan mantener la cobertura conceptual de los instrumentos tradicionales, pero mediante estructuras más eficientes y acordes a las necesidades operativas de los equipos de bienestar estudiantil.

En respuesta a esta problemática, este trabajo propone el diseño de un cuestionario integrador y adaptativo construido a partir de técnicas de representación semántica, modelamiento en espacios latentes y análisis estructural de ítems. La propuesta reorganiza cuatro instrumentos tradicionales en una sola estructura optimizada, capaz de ajustar su extensión según las respuestas iniciales, reduciendo redundancias y priorizando los ítems con mayor capacidad informativa para mejorar la eficiencia del proceso de evaluación.

A partir de esta propuesta, los resultados obtenidos permiten evaluar de manera crítica la validez y el cumplimiento de los objetivos formulados, así como los alcances, limitaciones y proyecciones de la solución desarrollada. En relación con el objetivo general, los resultados obtenidos permiten confirmar la factibilidad de diseñar un cuestionario optimizado mediante técnicas de inteligencia artificial. El estudio también demuestra que es viable reorganizar cuatro escalas tradicionales de instrumentos psicométricos existentes en una estructura adaptativa eficiente, con una reducción significativa de la cantidad de preguntas en escenarios de baja sintomatología.

Respecto a los objetivos específicos, todos fueron alcanzados satisfactoriamente. En primer lugar, la aplicación de métodos de representación semántica permitió caracterizar los ítems en un espacio vectorial entrenado mediante modelos basados en Transformers, una arquitectura de aprendizaje profundo diseñada para interpretar el significado contextual de palabras y frases mediante el uso de mecanismos de atención. A partir de estas representaciones, el Autoencoder Variacional permitió obtener un espacio latente de 16 dimensiones, dentro del cual se modelaron relaciones no lineales entre ítems pertenecientes a diversas escalas. Este espacio de mayor dimensionalidad capturó patrones estructurales profundos y coherentes entre síntomas, permitiendo observar similitudes conceptuales que no son visibles con métodos tradicionales.

Para su análisis e interpretación, dichas 16 dimensiones fueron posteriormente proyectadas mediante UMAP a un espacio bidimensional, lo que facilitó la visualización de agrupamientos naturales y zonas de solapamiento semántico. Estos solapamientos reflejan que dominios que usualmente se evalúan por separado, como ansiedad, depresión y burnout, comparten

dimensiones subyacentes comunes en la población universitaria, manifestándose en regiones del espacio latente donde los límites entre ellos no son estrictos. Esto sugiere que ciertos síntomas son experimentados de manera conjunta, es decir, que la presencia de un síntoma puede estar asociada a la aparición de otros relacionados, lo que justifica la presencia de redundancias cuando se aplican múltiples instrumentos tradicionales de forma acumulativa.

En contraste, otros dominios asociados a la calidad de vida se distribuyeron de manera más diferenciada dentro del espacio latente, formando agrupamientos más separados y definidos. Esta separación indica que dichos dominios capturan constructos conceptualmente distintos y menos solapados con las escalas de sintomatología emocional. La distinción observada entre estos grupos es importante porque confirma que el modelo no solo detecta similitudes, sino que también preserva diferencias conceptuales relevantes, lo que aporta evidencia adicional sobre la coherencia de la estructura latente obtenida.

En segundo lugar, la implementación de algoritmos de inteligencia artificial y el uso comparado de tres enfoques de clustering (Top-Down, Feature Agglomeration e Iterativo) permitieron identificar redundancias, similitudes y dependencias entre los ítems con un alto grado de consistencia metodológica. La convergencia entre estos métodos es un resultado relevante, ya que, a pesar de sus diferencias conceptuales y computacionales, todos ellos señalan que la estructura detectada no depende de una técnica específica, sino que refleja patrones subyacentes reales en los datos. A partir de este proceso se definieron 16 clústeres que agrupan ítems con comportamientos estadísticos coherentes y reproducibles. Esta organización interna evidencia que las escalas tradicionales, aunque fueron diseñadas como instrumentos separados, comparten dimensiones comunes que emergen cuando se analizan en conjunto. Además, permitió distinguir qué grupos de ítems aportan información adicional y cuáles presentan altos niveles de solapamiento, lo que hizo posible reducir redundancias de manera informada. Esta claridad estructural fue esencial para justificar la reorganización del cuestionario, ya que permitió definir de forma objetiva qué dominios debían mantenerse separados, cuáles podían integrarse y en qué casos resultaba adecuado activar preguntas adicionales solo cuando las respuestas iniciales lo requerían.

En tercer lugar, el análisis probabilístico permitió profundizar en el rol de cada ítem dentro de su clúster y seleccionar aquellos que presentan la mayor capacidad discriminativa. La identificación de estos ítems puerta no solo redujo el número de preguntas necesarias en la etapa inicial, sino que también entregó un criterio objetivo para definir umbrales de activación basados en diferencias significativas entre niveles de respuesta. La estimación de estos umbrales permitió establecer reglas claras y reproducibles que determinan cuándo un dominio debe ser evaluado en mayor profundidad.

Por último, la integración de todos los elementos anteriores permitió diseñar un cuestionario adaptativo cuya extensión puede variar entre aproximadamente 16 y 57 ítems. En escenarios de baja sintomatología, la reducción potencial cercana al 72 % representa uno de los resultados más relevantes del estudio, ya que demuestra que es posible disminuir de manera considerable la carga de respuesta. Al mismo tiempo, la estructura del cuestionario mantiene la capacidad de ampliar su alcance cuando las respuestas lo requieren, ajustando la

profundidad de la evaluación sin recurrir a la aplicación completa del instrumento en todos los casos.

Adicionalmente, la estructura adaptativa propuesta introduce un valor agregado que va más allá de la reducción de la carga de respuesta. Al organizar los ítems según clústeres empíricamente relacionados, el cuestionario favorece una mayor coherencia temática entre preguntas consecutivas, evitando alternancias abruptas entre dominios no relacionados. Esto permite que el participante permanezca dentro de un mismo marco cognitivo y emocional durante cada etapa del recorrido, lo que contribuye a una experiencia de respuesta más consistente y potencialmente más fiable. De este modo, incluso en escenarios donde se requiere evaluar múltiples dominios en profundidad, el instrumento mantiene una lógica interna que acompaña la forma en que los síntomas se manifiestan y se experimentan en la práctica.

A partir de estos resultados, el alcance de la propuesta radica en demostrar que es posible reorganizar instrumentos psicométricos extensos de una manera más eficiente, utilizando herramientas de análisis que permiten identificar similitudes entre ítems, reducir redundancias y concentrarse en las preguntas que aportan información realmente útil. Gracias a este enfoque, se construyó un cuestionario adaptativo capaz de ajustar su extensión según las respuestas iniciales, lo que permite presentar solo la cantidad de preguntas necesarias en cada caso.

Además, la propuesta muestra que los instrumentos originales, aun cuando provienen de escalas distintas, comparten patrones estructurales que pueden aprovecharse para optimizar su uso. Esto permite mantener la cobertura general de los dominios evaluados, pero con una organización más compacta y flexible. En conjunto, los resultados indican que el enfoque desarrollado no solo simplifica el proceso de evaluación, sino que también mejora su eficiencia y facilita su aplicación en contextos donde el tiempo y la adherencia de los participantes son factores críticos.

No obstante, el trabajo presenta limitaciones que deben ser reconocidas. La estructura de clústeres y los umbrales definidos dependen de una muestra específica, por lo que su estabilidad podría variar en otras poblaciones. Asimismo, la solución propuesta reorganiza la estructura de aplicación del instrumento, pero no constituye aún una validación psicométrica formal ni un estudio clínico comparativo. La interpretabilidad del espacio latente, aunque informativa, depende de modelos de aprendizaje profundo cuya naturaleza no siempre permite una explicación intuitiva de cada dimensión. Por estas razones, se recomienda complementar el enfoque con evaluaciones externas y participación de expertos clínicos.

A pesar de estas limitaciones, el trabajo realizado ofrece contribuciones concretas y de alto valor práctico. Se presenta una forma de reorganizar instrumentos psicométricos extensos mediante análisis de similitud entre ítems, agrupamientos y reglas de activación, lo que constituye un enfoque replicable en otras evaluaciones que enfrentan problemas de redundancia o exceso de ítems. Además, la propuesta tiene un impacto directo tanto para las unidades de bienestar estudiantil, que disponen de una herramienta más eficiente para realizar tamizajes y priorizar casos, como para los estudiantes, quienes se benefician de una menor

carga de respuesta, mayor comodidad al responder y la posibilidad de integrar evaluaciones periódicas sin sobrecargar su ya exigente calendario académico.

En conjunto, el trabajo permite identificar líneas claras de trabajo a futuro. Un paso esencial consiste en validar empíricamente el cuestionario adaptativo mediante estudios comparativos con instrumentos tradicionales, evaluando su desempeño en contextos reales de aplicación y determinando su sensibilidad y especificidad frente a distintos niveles de sintomatología. Esta validación debe incluir un proceso formal de psicometría, con participación de expertos clínicos y profesionales del área, a fin de revisar la coherencia conceptual de los clústeres, verificar la pertinencia de los ítems seleccionados y asegurar que la estructura adaptativa refleja adecuadamente los constructos que pretende medir. Asimismo, será necesario aplicar el cuestionario en nuevas muestras para evaluar si es capaz de reproducir resultados diagnósticos equivalentes a los obtenidos mediante las escalas originales, lo que permitiría determinar su grado de concordancia y estabilidad.

Otra línea prometedora es avanzar hacia una versión aún más dinámica del instrumento, capaz de ajustar no solo la activación inicial de cada clúster, sino también la profundidad de las preguntas dentro de cada nivel de severidad, permitiendo explorar con mayor precisión la progresión de cada condición. También sería relevante estudiar la estabilidad de los clústeres en poblaciones diversas para evaluar su generalización y detectar posibles diferencias estructurales entre grupos.

Finalmente, los resultados obtenidos permiten concluir que la solución presentada no solo ofrece un cuestionario adaptativo preliminar, sino que además entrega una base conceptual y técnica que abre el camino hacia nuevas formas de evaluación más eficientes, personalizadas y coherentes con las necesidades actuales de las instituciones de educación superior.

ANEXOS

Ítems originales de los instrumentos utilizados

EUBE – Escala Unidimensional de Burnout Estudiantil

- EUBE-1** El tener que asistir diariamente a clases me cansa.
- EUBE-2** Mis problemas escolares me deprimen fácilmente.
- EUBE-3** Durante las clases me siento somnoliento.
- EUBE-4** Creo que estudiar hace que me sienta agotado.
- EUBE-5** Cada vez me es más difícil concentrarme en las clases.
- EUBE-6** Me desilusionan mis estudios.
- EUBE-7** Antes de terminar mi horario de clases ya me siento cansado.
- EUBE-8** No me interesa asistir a clases.
- EUBE-9** Cada vez me cuesta más trabajo ponerle atención al maestro.
- EUBE-10** El asistir a clases se me hace aburrido.
- EUBE-11** Siento que estudiar me está desgastando físicamente.
- EUBE-12** Cada vez me siento más frustrado por ir a la escuela.
- EUBE-13** No creo terminar con éxito mis estudios.
- EUBE-14** Siento que tengo más problemas para recordar lo que estudio.
- EUBE-15** Creo que estudiar me está desgastando emocionalmente.

GAD-7 – Generalized Anxiety Disorder

- GAD7-1** Se ha sentido nervioso(a), ansioso(a) o con los nervios de punta.
- GAD7-2** No ha sido capaz de parar o controlar su preocupación.
- GAD7-3** Se ha preocupado demasiado por motivos diferentes.
- GAD7-4** Ha tenido dificultad para relajarse.
- GAD7-5** Se ha sentido tan inquieto(a) que no ha podido quedarse quieto(a).
- GAD7-6** Se ha molestado o irritado fácilmente.
- GAD7-7** Ha tenido miedo de que algo terrible fuera a pasar.

WHOQOL-BREF

- WHO-1** ¿Hasta qué punto piensa que el dolor (físico) le impide hacer lo que necesita?
- WHO-2** ¿En qué grado necesita de un tratamiento médico para funcionar en su vida diaria?
- WHO-3** ¿Cuánto disfruta de la vida?
- WHO-4** ¿Hasta qué punto siente que su vida tiene sentido?
- WHO-5** ¿Cuál es su capacidad de concentración?
- WHO-6** ¿Cuánta seguridad siente en su vida diaria?
- WHO-7** ¿Cómo de saludable es el ambiente físico a su alrededor?
- WHO-8** ¿Tiene energía suficiente para la vida diaria?
- WHO-9** ¿Es capaz de aceptar su apariencia física?
- WHO-10** ¿Tiene suficiente dinero para cubrir sus necesidades?
- WHO-11** ¿Dispone de la información que necesita para su vida diaria?
- WHO-12** ¿Hasta qué punto tiene oportunidad de realizar actividades de ocio?
- WHO-13** ¿Es capaz de desplazarse de un lugar a otro?
- WHO-14** ¿Cómo de satisfecho/a está con su sueño?
- WHO-15** ¿Cómo de satisfecho/a está con su habilidad para realizar sus actividades de la vida diaria?
- WHO-16** ¿Cómo de satisfecho/a está con su capacidad de trabajo?
- WHO-17** ¿Cómo de satisfecho/a está de sí mismo?
- WHO-18** ¿Qué tan satisfecho/a está con sus relaciones personales?
- WHO-19** ¿Cómo de satisfecho/a está con su vida sexual?
- WHO-20** ¿Cómo de satisfecho/a está con el apoyo que obtiene de sus amigos/as?
- WHO-21** ¿Cómo de satisfecho/a está de las condiciones del lugar donde vive?
- WHO-22** ¿Cómo de satisfecho/a está con el acceso que tiene a los servicios sanitarios?
- WHO-23** ¿Cómo de satisfecho/a está con los servicios de transporte de su zona?
- WHO-24** ¿Con qué frecuencia tiene sentimientos negativos, tales como tristeza, desesperanza, ansiedad, o depresión?

PHQ-9 – Patient Health Questionnaire

- PHQ9-1** Poco interés o placer en hacer cosas.
- PHQ9-2** Se ha sentido decaído(a), deprimido(a) o sin esperanzas.
- PHQ9-3** Ha tenido dificultad para quedarse o permanecer dormido(a), o ha dormido demasiado.
- PHQ9-4** Se ha sentido cansado(a) o con poca energía.
- PHQ9-5** Sin apetito o ha comido en exceso.
- PHQ9-6** Se ha sentido mal con usted mismo(a) - o que es un fracaso o que ha quedado mal con usted mismo(a) o con su familia.
- PHQ9-7** Ha tenido dificultad para concentrarse en ciertas actividades, tales como leer el periódico o ver la televisión.
- PHQ9-8** ¿Se ha movido o hablado tan lento que otras personas podrían haberlo notado? O lo contrario - muy inquieto(a) o agitado(a) que ha estado moviéndose mucho más de lo normal.
- PHQ9-9** Pensamientos de que estaría mejor muerto(a) o de lastimarse de alguna manera.

Ítems reformulados utilizados en el análisis semántico

GAD-7 – Generalized Anxiety Disorder

- GAD7-1** Últimamente me he sentido inquieto y con el cuerpo tenso, como si tuviera los nervios a flor de piel o no pudiera relajarme.
- GAD7-2** He tenido una preocupación excesiva y persistente que me resulta difícil de controlar, pensando demasiado en diferentes motivos.
- GAD7-3** He estado preocupándome demasiado por diferentes motivos y de forma excesiva.
- GAD7-4** He sentido una fuerte incapacidad para relajarme, teniendo una dificultad significativa para ello.
- GAD7-5** He sentido una fuerte incapacidad para relajarme, sintiéndome tan inquieto(a) que no he podido quedarme quieto(a).
- GAD7-6** Se ha molestado o irritado fácilmente.
- GAD7-7** He sentido una ansiedad intensa, manifestada como un miedo persistente de que algo terrible iba a ocurrir.

EUBE - Escala Unidimensional de Burnout Estudiantil

- EUBE-1** Me siento con fatiga física y cansancio (somnolencia) solo por tener que asistir a mis actividades diarias o clases.
- EUBE-2** Últimamente me he sentido abatido y sin ánimo por mis problemas académicos, todo me parece pesado y sin sentido.
- EUBE-3** Durante las clases me siento somnoliento.
- EUBE-4** He sentido agotamiento y desgaste por mis estudios: sintiéndome completamente agotado(a).
- EUBE-5** He tenido una baja capacidad para mantener la atención y la concentración en mis clases.
- EUBE-6** Me siento desmotivado y vacío cuando pienso en mis estudios, ya no encuentro el mismo interés ni propósito que antes.
- EUBE-7** Me siento con fatiga física y cansancio (somnolencia) antes de terminar mis actividades diarias o clases.
- EUBE-8** No me interesa asistir a clases.
- EUBE-9** He tenido una baja capacidad para mantener la atención y la concentración en mis actividades académicas.
- EUBE-10** El asistir a clases se me hace aburrido.
- EUBE-11** He sentido agotamiento y desgaste por mis estudios: manifestándose como un desgaste físico.
- EUBE-12** Cada vez me siento más frustrado por ir a la escuela.
- EUBE-13** No creo terminar con éxito mis estudios.
- EUBE-14** Siento que tengo más problemas para recordar lo que estudio.
- EUBE-15** He sentido agotamiento y desgaste por mis estudios: manifestándose como un desgaste emocional.

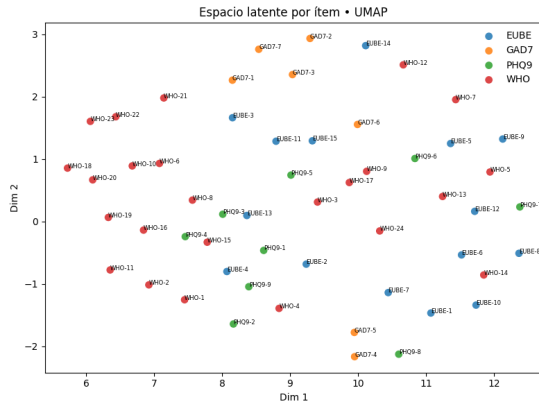
WHOQOL-BREF

- WHO-1** Piensa que el dolor físico le impide hacer lo que necesita.
- WHO-2** ¿En qué grado necesita de un tratamiento médico para funcionar en su vida diaria?
- WHO-3** Falta de disfrute de la vida.
- WHO-4** He sentido una desesperanza severa, sintiendo que mi vida no tiene sentido y pensando que estar muerto sería un alivio.
- WHO-5** He tenido una baja capacidad para mantener la atención y la concentración.
- WHO-6** Falta de seguridad en su vida diaria.
- WHO-7** Tengo sentimientos positivos frecuentes.
- WHO-8** ¿Cómo de saludable es el ambiente físico a su alrededor?
- WHO-9** Evaluación sobre la capacidad de aceptar la propia apariencia física, a nivel de autoestima.
- WHO-10** ¿Tiene suficiente dinero para cubrir sus necesidades?
- WHO-11** Dispone de la información que necesita para su vida diaria.
- WHO-12** ¿Hasta qué punto tiene oportunidad de realizar actividades de ocio?
- WHO-13** ¿Es capaz de desplazarse de un lugar a otro?
- WHO-14** He sentido dificultades e insatisfacción con mi sueño.
- WHO-15** Insatisfacción con su habilidad para realizar sus actividades de la vida diaria.
- WHO-16** Insatisfacción con su capacidad de trabajo.
- WHO-17** ¿Cuál es su nivel de satisfacción consigo mismo? Evaluar la autoestima y la capacidad de aceptar su apariencia física.
- WHO-18** Insatisfacción con sus relaciones personales.
- WHO-19** ¿Cuál es su nivel de satisfacción con las relaciones personales? Evaluar el apoyo social de amigos y la vida sexual.
- WHO-20** Insatisfacción con el apoyo que obtiene de sus amigos, es decir, parte de sus relaciones personales.
- WHO-21** Insatisfacción de las condiciones del lugar donde vive.
- WHO-22** Insatisfacción con el acceso que tiene a los servicios sanitarios.
- WHO-23** Insatisfacción con los servicios de transporte de su zona.
- WHO-24** He notado que tengo con frecuencia sensaciones de malestar emocional, como tristeza, ansiedad o una tensión constante que no desaparece.

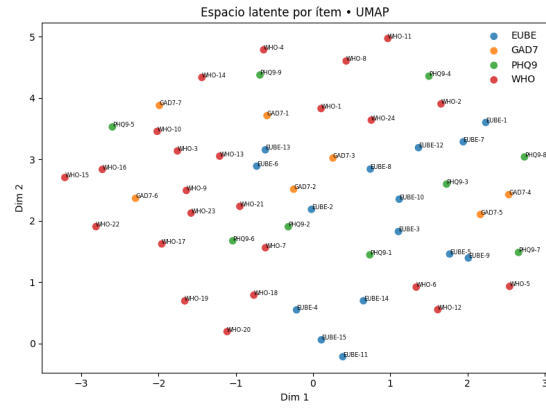
PHQ-9 – Patient Health Questionnaire

- PHQ9-1** He sentido una pérdida de interés o placer en hacer las cosas que antes disfrutaba, incluyendo mis responsabilidades académicas.
- PHQ9-2** Me he sentido sin esperanza y con un cansancio emocional profundo, como si nada tuviera valor ni dirección.
- PHQ9-3** He tenido dificultad para quedarme dormido, mantenerme dormido, o he dormido demasiado.
- PHQ9-4** Siento una falta de energía vital y fatiga generalizada, un cansancio que afecta mi funcionamiento en la vida diaria.
- PHQ9-5** Sin apetito o ha comido en exceso, síntomas relacionados a la ansiedad, o el sentirse ansioso o nervioso.
- PHQ9-6** Se ha sentido mal con usted mismo(a) o que es un fracaso o que ha quedado mal con usted mismo(a) o con su familia.
- PHQ9-7** He tenido una baja capacidad para mantener la atención y la concentración en tareas cotidianas, como leer o ver TV.
- PHQ9-8** He estado muy inquieto(a) o agitado(a), moviéndome mucho más de lo normal, y no he podido relajarme.
- PHQ9-9** He sentido una desesperanza severa, con pensamientos de muerte o de lastimarme, sintiendo que la vida no vale la pena.

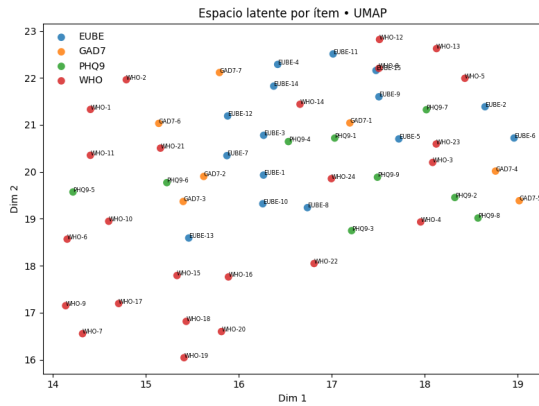
Espacios latentes (VAE)



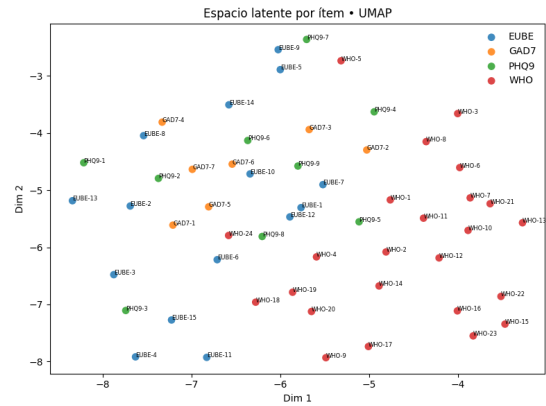
(a) Espacio latente VAE con 6 dimensiones.



(b) Espacio latente VAE con 12 dimensiones.



(c) Espacio latente VAE con 16 dimensiones y dropout = 0.



(d) Espacio latente VAE con 16 dimensiones y $C_{\max} = 3$.

REFERENCIAS BIBLIOGRÁFICAS

- [1] OMS, "Informe sobre la salud mental en el mundo: Transformar la salud mental para todos," [En línea], Jun. 2022, disponible en: <https://www.who.int/publications/i/item/9789240050860>.
- [2] F. C. Solar *et al.*, "Problemas de salud mental en estudiantes de la universidad de concepción," *Terapia psicológica*, vol. 25, no. 2, pp. 105–112, Dec. 2007.
- [3] G. Larraín *et al.*, "Using a classifier ensemble for preventing burnout in university students: A study case in valparaíso," *IEEE Access*, vol. 11, Oct. 2023.
- [4] Z. Antúnez *et al.*, "Impacto de problemas de salud mental y apoyo social en la calidad de sueño: Seguimiento longitudinal antes y durante la pandemia de covid-19 en universitarios chilenos," *Anales de Psicología*, vol. 39, no. 2, pp. 188–196, Apr. 2023.
- [5] I. N. de la Juventud (INJUV), "9.ª encuesta nacional de juventud 2018," [En línea], Dec. 2019, disponible en: <https://www.injuv.gob.cl/9encuestanacionaldejuventud>.
- [6] P. Martínez, A. Jiménez-Molina, S. Mac-Ginty, V. Martínez, and G. Rojas, "Salud mental en estudiantes de educación superior en Chile: una revisión de alcance con meta-análisis," *Terapia Psicológica*, vol. 39, no. 3, pp. 405–426, Dec. 2021.
- [7] C. G. Pinto-Cortez *et al.*, "Prevalencia y características sociodemográficas asociadas con la depresión, la ansiedad y el estrés en estudiantes universitarios," *LRIFP*, vol. 19, no. 1, Oct. 2024.
- [8] A. Barraza, "Validación psicométrica de la escala unidimensional del burnout estudiantil," *Rev. Intercontinental de Psicología y Educ.*, vol. 13, no. 2, pp. 51–74, May 2011.
- [9] M. E. B. Alarcón, P. G. T. Castro, A. V. S. Barradas, and J. L. González, "Burnout estudiantil en universitarios veracruzanos," *RIDE Revista Iberoamericana Para La Investigación Y El Desarrollo Educativo*, vol. 7, no. 14, Feb. 2017.
- [10] H. Rahman, M. M. Tohan, A. A. Easha, and N. Fatema, "Depression symptoms among university students in the khulna region of bangladesh," *PLOS Mental Health*, vol. 2, no. 4, Apr. 2025.
- [11] M. Posso-Yépez, V. León-Ron, C. Guzmán-Torres, and C. Bastidas, "Agotamiento (burnout) en estudiantes universitarios: análisis desde la resiliencia, sexo, edad y el nivel de estudios," *Formación Universitaria*, vol. 18, no. 2, pp. 59–70, Apr. 2025.
- [12] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The phq-9: validity of a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, Sep. 2001.

- [13] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: the gad-7," *Archives of Internal Medicine*, vol. 166, no. 10, pp. 1092–1097, May 2006.
- [14] THE WHOQOL GROUP, "Development of the world health organization whoqol-bref quality of life assessment," *Psychological Medicine*, vol. 28, no. 3, pp. 551–558, 1998.
- [15] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, 2nd ed. Boston, MA, USA: Pearson Addison Wesley, 2019.
- [16] Z. Keita, "Clasificación en machine learning: Guía para principiantes," [En línea], Mar. 2024, disponible en: <https://www.datacamp.com/es/blog/classification-machine-learning>.
- [17] A. K. Hartmann, "Practical introduction to clustering data," arXiv (Cornell University), pp. 1–21, Jul. 2021.
- [18] D. Bergmann, "Aprendizaje semisupervisado," [En línea], Dec. 2023, disponible en: <https://www.ibm.com/es-es/think/topics/semi-supervised-learning>.
- [19] Y.-Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Archives of Psychiatry*, vol. 27, no. 2, pp. 130–135, Jun. 2015.
- [20] G. Larraín, "Estudio de un problema de clasificación sobre la salud mental en estudiantes universitarios," Memoria de título, Departamento de Informática, Universidad Técnica Federico Santa María, Valparaíso, Chile, 2022.
- [21] M. Y. Khan *et al.*, "Automated prediction of good dictionary examples (gdex): A comprehensive experiment with distant supervision, machine learning, and word embedding-based deep learning techniques," *Complexity*, vol. 2021, Sep. 2021.
- [22] IBM, "¿qué es support vector machine?" [En línea], Dec. 2023, disponible en: <https://www.ibm.com/mx-es/think/topics/support-vector-machine>.
- [23] G. Gupta, G. Khatri, and C. Sharma, "Artificial neural networks," *Pratibodh: A Journal for Engineering, NCICT*, Apr. 2025.
- [24] C. M. Bishop and H. Bishop, *Deep Learning*. Springer, 2023.
- [25] A. S. Saud, S. Shakya, and B. Neupane, "Analysis of depth of entropy and gini index based decision trees for predicting diabetes," *Indian Journal of Computer Science*, vol. 6, no. 6, Jan. 2021.
- [26] V. Kumar, "Feature selection: A literature review," *The Smart Computing Review*, vol. 4, no. 3, Jun. 2014.
- [27] Y. Lyu, Y. Feng, and K. Sakurai, "A survey on feature selection techniques based on filtering methods for cyber attack detection," *Information*, vol. 14, no. 3, pp. 2–19, Mar. 2023.

- [28] GeeksforGeeks, "Wrapper methods for feature selection in machine learning," [En línea], disponible en: <https://www.geeksforgeeks.org/machine-learning/wrapper-methods-feature-selection/>.
- [29] Codefinity, "Sequential backward and forward selection," [En línea], Jan. 2024, disponible en: <https://codefinity.com/blog/Sequential-Backward-and-Forward-Selection>.
- [30] G. Yadav and M. U. Bokhari, "Hybrid classifier for optimizing mental health prediction: feature engineering and fusion technique," *International Journal of Mental Health and Addiction*, Jun. 2024.
- [31] J. Murel and E. Kavlakoglu, "Dimensionality reduction," [En línea], Jan. 2024, disponible en: <https://www.ibm.com/think/topics/dimensionality-reduction>.
- [32] F. Luna, "UMAP: an Alternative dimensionality Reduction Technique," Medium, May 2023. [Online]. Available: <https://medium.com/mcd-unison/umap-an-alternative-dimensionality-reduction-technique-7a5e77e80982>
- [33] D. Bergmann and C. Stryker, "Autocodificador variacional," [En línea], Jun. 2024, disponible en: <https://www.ibm.com/es-es/think/topics/variational-autoencoder>.
- [34] P. Li, Y. Pei, and J. Li, "A comprehensive survey on design and application of autoencoder in deep learning," *Applied Soft Computing*, vol. 138, Mar. 2023.
- [35] J. Montes, J. Pérez, and R. Perera, "Damage indicators for structural monitoring of fiber-reinforced polymer-strengthened concrete structures based on manifold invariance defined on latent space of deep autoencoders," *Applied Sciences*, vol. 15, no. 11, May 2025.
- [36] H. Abdelmotaleb, C. Mcneile, and M. Wojtyś, "A comparative study of word embedding techniques for classification of star ratings," *Expert Systems With Applications*, Jul. 2025.
- [37] Y. Han, "Automatic pipeline for information of curve graphs in papers based on deep learning," Research Square [Preprint], Jan. 2024.
- [38] T. T. Erguzel, C. Tas, and M. Cebi, "A wrapper-based approach for feature selection and classification of major depressive disorder–bipolar disorders," *Computers in Biology and Medicine*, vol. 64, pp. 127–137, Jul. 2015.
- [39] D. Glavin et al., "Patient health questionnaire-9 item pairing predictiveness for prescreening depressive symptomatology: Machine learning analysis," *JMIR Mental Health*, vol. 10, Aug. 2023.
- [40] F. Wang, Y. Wu, S. Wang, Z. Du, and Y. Wu, "Development of an optimal short form of the gad-7 scale with cross-cultural generalizability based on riskslim," *General Hospital Psychiatry*, vol. 87, pp. 33–40, Jan. 2024.

- [41] M. Noetel, J. Ciarrochi, B. Sahdra, and C. Lonsdale, "Using genetic algorithms to abbreviate the mindfulness inventory for sport: A substantive-methodological synthesis," *Psychology of Sport and Exercise*, vol. 45, p. 101545, Jan. 2019.
- [42] D. Colledani, C. Barbaranelli, and P. Anselmi, "Fast, smart, and adaptive: using machine learning to optimize mental health assessment and monitor change over time," *Scientific Reports*, vol. 15, no. 1, Feb. 2025.
- [43] N. Kuleindiren *et al.*, "Optimizing existing mental health screening methods in a dementia screening and risk factor app: Observational machine learning study," *JMIR Formative Research*, vol. 6, no. 3, Dec. 2021.
- [44] A. Tallroth and M. Ålander, "The use of machine learning algorithms for adaptive question selection in questionnaire-based mental health data collection apps," Tesis de máster, Dept. of Biomedical Engineering, Lund University, Lund, Suecia, 2016.
- [45] M. T. Tabares, C. V. Álvarez, J. B. Salcedo, and S. M. Rendón, "Anxiety in young people: Analysis from a machine learning model," *Acta Psychologica*, vol. 248, Jul. 2024.
- [46] M. G. Pereira, M. Santos, R. Magalhães, C. Rodrigues, O. Araújo, and D. Durães, "Burnout risk profiles in psychology students: An exploratory study with machine learning," *Behavioral Sciences*, vol. 15, no. 4, p. 505, Apr. 2025.
- [47] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, 2023.
- [48] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.
- [49] K. Kroenke, R. L. Spitzer, J. B. Williams, and B. Löwe, "An ultra-brief screening scale for anxiety and depression: The phq-4," *Psychosomatics*, vol. 50, no. 6, pp. 613–621, December 2009.