Pontificia Universidad
JAVERIANA
Cali

Análisis Predictivo de la Salud Mental en Estudiantes y Colaboradores de una Universidad Privada Colombiana mediante Técnicas de Ciencia de Datos

Autora:

Nini Alejandra Valderrama Moreno

Proyecto Aplicado para optar al título de

Magister en Ciencia de Datos

Director:

Daniel Enrique González Gómez

FACULTAD DE INGENIERÍA Y CIENCIAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI, JUNIO DE 2024

RESUMEN

La salud mental presenta un desafío a nivel mundial con repercusiones negativas en contextos sociales, institucionales, familiares, laborales, educativos, entre otros, este proyecto tuvo como objetivo principal comprender la salud mental de estudiantes y colaboradores de una universidad privada de Colombia, a través de la implementación de técnicas de modelamiento predictivo en Ciencia de Datos, para ello se empleó técnicas avanzadas de análisis de datos, aprendizaje automático y visualización interactiva. En una primera etapa, se realizó una exhaustiva exploración descriptiva de la base de datos, seguidamente, se aplicaron técnicas de reducción de dimensionalidad (PCA, t-SNE, UMAP) y métodos de agrupamiento (KMeans, clustering aglomerativo, GMM) para identificar patrones y posibles subgrupos latentes, aunque las métricas cuantitativas no evidenciaron clústers naturales bien definidos. En la segunda fase, se implementaron modelos de aprendizaje supervisado, incluyendo Regresión Lasso, Random Forest, XGBoost y LightGBM, para predecir variables clave como depresión, ansiedad, estrés, soledad, resiliencia, satisfacción con la vida y recursos psicosociales. Se emplearon técnicas de sobremuestreo (SMOTE) y validación cruzada para asegurar la robustez de los modelos y se analizaron las variables predictoras más relevantes asociadas a cada indicador. Finalmente, se desarrolló una herramienta de visualización interactiva desarrollada en PowerBi, que integra los resultados descriptivos, de clustering y de predicción, permitiendo a usuarios técnicos y no técnicos explorar dinámicamente la estructura y los determinantes del bienestar (Indicadores Positivos) y malestar psicológico (Indicadores Negativos) en la población de Colaboradores, estudiantes de Posgrado y estudiantes de Pregrado. Los hallazgos obtenidos aportan una visión integral y basada en evidencia sobre los factores asociados a la salud mental en cuanto bienestar y malestar en contextos universitarios, y constituyen una base sólida para el diseño de intervenciones focalizadas y futuras investigaciones en salud mental y determinantes sociales.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABLAS

GLOSARIO

PCA: Principal Component Analysis – Análisis de Componente Principales

OMS: Organización Mundial de la Salud

ML: Machine Learning – aprendizaje automático

Dataset: Conjunto de datos

Df: Dataframe (estructura de datos)

IA: Inteligencia Artificial

DASS: Depresión, Ansiedad, Estrés.

EDA: Exploratory Data Analysis (EDA) en inglés, análisis exploratorio de datos

INTRODUCCIÓN

La salud mental es una preocupación cada vez más relevante a nivel mundial debido al aumento de problemas como la depresión y la ansiedad. Según la Organización Panamericana de la Salud, aproximadamente el 25% de la población mundial experimenta algún trastorno mental o del comportamiento en algún momento de su vida [1]. Estas cifras se han visto afectadas negativamente por la pandemia de COVID-19, con un aumento significativo en los trastornos mentales y emocionales [2]. En respuesta a esta alta incidencia, la Pontificia Universidad Javeriana Cali ha lanzado el proyecto Salud y Bienestar, que tiene como objetivo abordar las necesidades de salud mental en su comunidad educativa desde una perspectiva de determinantes sociales [3]. A partir de esta concepción se busca entender el bienestar mental a través del contexto social de las personas, identificando factores estructurales e intermedios que influyen en su salud mental [4].

El presente proyecto se enfoca en el análisis de la encuesta Javeriana de Bienestar y Salud, planteada como herramienta para contribuir desde la perspectiva de la Ciencia de Datos la comprensión de bienestar y salud en la comunidad universitaria (estudiantes, profesores y personal administrativo) de la Universidad Javeriana Cali. La iniciativa, liderada por la vicerrectoría del Medio Universitario y la facultad de Humanidades, tiene como propósito identificar herramientas para fortalecer o replantear programas de bienestar, incluyendo la consejería estudiantil, éxito académico y otros aspectos relevantes [3]. La encuesta consistió en un cuestionario en línea, de participación abierta y opcional, que incluyó preguntas estandarizadas sobre bienestar y la salud mental. Como resultado, se obtuvo un conjunto de datos con 4211 registros [3].

A través de este proyecto se busca fortalecer las políticas de prevención en salud mental, identificando factores clave para el bienestar. Siendo de interés particular el análisis de los determinantes sociales que afectan a la población de estudiantes y colaboradores, a partir de los cuales se puedan construir intervenciones que mejoren áreas deficientes y potencien aquellos factores que impactan positivamente el bienestar. El objetivo del proyecto es comprender la salud mental de estudiantes y colaboradores de una universidad privada de Colombia, a través de la implementación de técnicas de modelamiento predictivo en Ciencia de Datos, lo anterior, con basen en los datos obtenidos por la Encuesta Javeriana de Bienestar y Salud, que permiten identificar condiciones de riesgo de manera anticipada.

1 DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

La salud mental (SM) es un elemento de interés creciente a nivel mundial. Cada año millones de personas se ven afectadas por problemas de salud mental como depresión y ansiedad. El aumento en la visibilidad de la salud mental está ligado con la evolución de las concepciones sobre el bienestar, que han migrado desde una mirada puramente física (e.g. el padecimiento de afecciones físicas) a una más holística (incluyendo la espiritualidad, el manejo de las emociones, entre otras) [4] [3]. Surge entonces la pregunta: ¿por qué es relevante hablar de salud mental?

Una de las posibles razones es la elevada incidencia de problemáticas de SM. Según la organización panamericana de la salud, se estimaba que un 25% de la población padecen algún trastorno mental o del comportamiento a lo largo de su vida [1]. En América Latina y el Caribe el trastorno depresivo, por ejemplo, es del 5% en la población adulta [1]. Estas cifras fueron gravemente afectadas por los efectos de la pandemia originada por el Covid-19. Factores como el aislamiento obligatorio, el miedo a contagiarse de la enfermedad o la perdida de seres queridos han disparado la aparición de trastornos mentales y emocionales [2]. En Estados Unidos, por ejemplo, se reportó un aumento de 36.4% a 41.5% en la ocurrencia de episodios de ansiedad y depresión por efecto de la pandemia [2]. En Colombia, la situación no es muy diferente. Las cifras oficiales del ministerio de salud indican que un 44% de la población de niños y niñas del país tienen indicios de algún problema mental [5]. En los jóvenes el panorama no es mucho más alentador, con una tasa de incidencia de ideación suicida del 6.6% [5]. La pandemia ha tenido efectos negativos no solo por su aumento en la incidencia de trastornos del bienestar, sino también por su efecto en las redes de apoyo. Según el ministerio de salud, hubo un aumento de 34% en la atención de personas por temas de salud mental entre 2017 y 2021 [6].

A nivel laboral existen también otros factores agravantes frente al bienestar y la salud mental. En un estudio de la firma Marsh McLennan realizado en 2022, se encontró que cerca del 64% de las empresas registra agotamiento en su fuerza laboral, además de reportar que los principales riesgos laborales percibidos por los empleados se relacionan con ciberseguridad, temas financieros y ambientales [7]. En Colombia, el ministerio

del trabajo ha reportado que entre un 20% y un 33% de los trabajadores de jornada completa manejan altos niveles de estrés [8]. Entre el 2009 y el 2012, se registró un incremento del 43% en trastornos de salud mental entre los trabajadores de Colombia, principalmente asociados a eventos de ansiedad y de depresión [8].

Frente a la alta incidencia de problemáticas de salud mental, surge la necesidad de diseñar intervenciones para su mejoramiento. Para lograr esto, se debe definir el concepto de salud mental. Según la Organización Mundial de la Salud (OMS), salud mental es "un estado completo de bienestar físico, mental y social, y no solamente la ausencia de afecciones o enfermedades" [9]. De las diferentes concepciones de la salud mental presentadas en el marco teórico la que mejor se adapta al diseño de intervenciones es la de los determinantes sociales. Como se detalla más adelante, la aproximación por determinantes sociales busca entender el bienestar a partir del contexto social de las personas, identificando factores estructurales e intermedios que determinan, fijan o establecen el estado de salud del individuo [3] [10].

Para la Pontificia Universidad Javeriana Cali, la gestión del bienestar en su comunidad es de gran importancia. Desde el año 2022, la universidad desarrolla el proyecto Salud y Bienestar que busca diagnosticar y atender las necesidades de salud mental en la Comunidad Educativa Javeriana desde un enfoque de determinantes sociales [3]. Como parte del proyecto se busca desarrollar programas de promoción de la salud que incluyan a toda la comunidad educativa y que permitan trascender la atención de los casos de riesgo [3], evolucionando de un enfoque reactivo a uno preventivo.

1.2 FORMULACIÓN DEL PROBLEMA

El presente trabajo se enmarca en el proyecto Salud y bienestar en la Comunidad Educativa Javeriana liderado por los investigadores María Teresa Valera, Leonardo Cepeda y Ana Marcela Uribe del grupo de Investigación Salud y Calidad de vida, en alianza con Natalia Cadavid y Jimena botero del Grupo de Investigación Bienestar, Trabajo, Cultura y Sociedad, además del Equipo del Comité Javeriana Saludable, dicho proyecto se lleva a cabo desde el 2022. A nivel global, el macroproyecto busca responder a la pregunta ¿Cómo es la salud mental de la comunidad Educativa Javeriana en Cali y cuáles son sus determinantes sociales? [3] En consecuencia, el macroproyecto pretende obtener información útil para el diseño de programas de promoción del bienestar con un enfoque más preventivo, y menos reactivo.

En particular, este proyecto aplicado tiene la intención de ayudar a resolver la pregunta de investigación dentro de la población de estudiantes y colaboradores de la universidad. Siendo los últimos, una población definida como el grupo de profesores, administrativos y directivos de la Universidad, con contrato durante el periodo académico 2022-2, que incluye un total de 1441 personas [3] y, los estudiantes, aquellos que estuvieron matriculados en el periodo académico 2022-2 en el la totalidad de los programas de pregrado y postrado que voluntariamente aceptaron participar en el proyecto, siendo 6850 estudiantes de pregrado y 1225 de posgrado para una muestra total de 8075 estudiantes. Para este estudio, las investigadoras realizaron la entrega de la base de datos con 4240 registros en un archivo .sav denominado: BD COMUNIDAD JAVERIANA – MCD, el cual se tuvo que leer y transformar en un entorno diferente al software SPSS.

La fuente de datos en la que se apoyaron los análisis de este estudio se generó a través de la Encuesta Javeriana de Bienestar, un formulario de participación voluntaria fue diseñado para evaluar indicadores de salud mental como depresión, ansiedad, resiliencia, entre otros.

A partir de esta fuente, este estudio se plantea las siguientes preguntas: ¿Qué patrones o características comunes asociadas a la salud mental se pueden identificar en la población de estudiantes y colaboradores de la universidad mediante la aplicación de métodos de aprendizaje no supervisado?; ¿Cómo se puede predecir el nivel de riesgo psicosocial en estudiantes y colaboradores universitarios utilizando herramientas estadísticas y de machine learning basadas en los determinantes sociales identificados? Y ¿De qué manera una herramienta de visualización de datos de Ciencia de Datos puede facilitar la toma de decisiones efectivas para mejorar la salud mental de estudiantes y colaboradores de la universidad?, las preguntas permiten identificar una pertinencia en línea con las necesidades planteadas en el problema descrito en este estudio.

2 OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Comprender la salud mental de estudiantes y colaboradores de una universidad privada de Colombia, a través de la implementación de técnicas de modelamiento predictivo en Ciencia de Datos.

2.2 OBJETIVOS ESPECÍFICOS

- Segmentar la población de estudiantes y colaboradores de la universidad, utilizando métodos de aprendizaje no supervisado, para identificar patrones o características comunes asociadas como determinante en salud mental a partir del conjunto de datos
- Modelar, a partir de herramientas estadísticas y de machine learning, el nivel de riesgo psicosocial en estudiantes y colaboradores a partir de los determinantes sociales previamente identificados
- Visualizar los datos a través de una herramienta de Ciencia de Datos que permite la toma de decisiones en beneficio de la salud mental de los estudiantes y colaboradores

3 MARCO TEÓRICO

A continuación, se presentan los conceptos relevantes que contiene el proyecto, información sobre salud mental, determinantes sociales y técnicas de ciencias de datos empleadas.

3.1 DETERMINANTES SOCIALES

Este proyecto se enfoca en el análisis de determinantes sociales del bienestar y la salud mental para los colaboradores y estudiantes de la universidad. Es necesario entonces presentar una definición clara de los determinantes sociales y cómo se relacionan con las necesidades de la población seleccionada para el análisis. El concepto de determinantes sociales es un enfoque de estudio de la salud, mental y física, que busca entender los efectos de las diferentes circunstancias bajo las que se desarrollan los individuos y las comunidades [9] [4] [3]. Según la definición ofrecida por la organización mundial de la salud (OMS), estos determinantes incluyen todas las condiciones asociadas al nacimiento y crecimiento de la persona, las condiciones en la que se vive y trabaja, la disponibilidad de sistemas de salud, educación, las normas sociales y comunitarias en las que se encuentra la persona, y en general todos los sistemas que afectan la vida diaria de los individuos [9] [10] [11]. Como característica particular de este enfoque, se resalta que busca entender aquellos factores de gran escala que determinan el estado de salud de las personas [4]. Algunos ejemplos de determinantes sociales son [10]:

- Niveles de ingreso y acceso a programas de protección social
- Educación
- Desempleo y condiciones de trabajo
- Seguridad alimentaria
- Acceso a vivienda
- Factores medioambientales
- Desarrollo en etapas tempranas
- Discriminación o inclusión social
- Conflictos sociales
- Acceso a servicios de salud

Cabe resaltar que, en general, los determinantes sociales corresponden a factores externos al individuo (i.e. no biológicos o genéticos) que están definidos por el contexto social y ambiental en que viven las personas (por ejemplo, políticas económicas, los sistemas de salud, cambio climático, etc.) [4] [11] [12].

Según la OMS, el impacto de los determinantes sociales en la salud de las personas es muy elevado, incluso mayor al impacto de intervenciones puramente médicas. Según cifras reportadas en algunos estudios, estos factores establecen entre el 30% y 55% de los estados de salud de las personas [9] [10]. El enfoque de determinantes sociales empieza a jugar un papel relevante en la década del 2000, después de que la OMS estableció su comisión de los determinantes sociales de la salud y presentó su reporte final en 2008 [9]. [5]

Como marco de análisis, el enfoque de determinantes sociales busca definir la salud del individuo integrando el contexto socio económico y político a nivel global, regional y local. Este modelo parte del contexto mencionado para definir unos determinantes estructurales e intermedios que repercuten en el estado de bienestar de las personas [3]. La Figura 1 presenta un diagrama de relación para estos factores, en los cuales se aprecian las relaciones causales entre el contexto y la salud del individuo. Una de las premisas de este enfoque indica que las acciones tomadas a nivel macro (por ejemplo, a través de políticas públicas del gobierno, o aquellas de las instituciones) pueden ayudar a mejorar el bienestar de las personas [4] [3] [11].

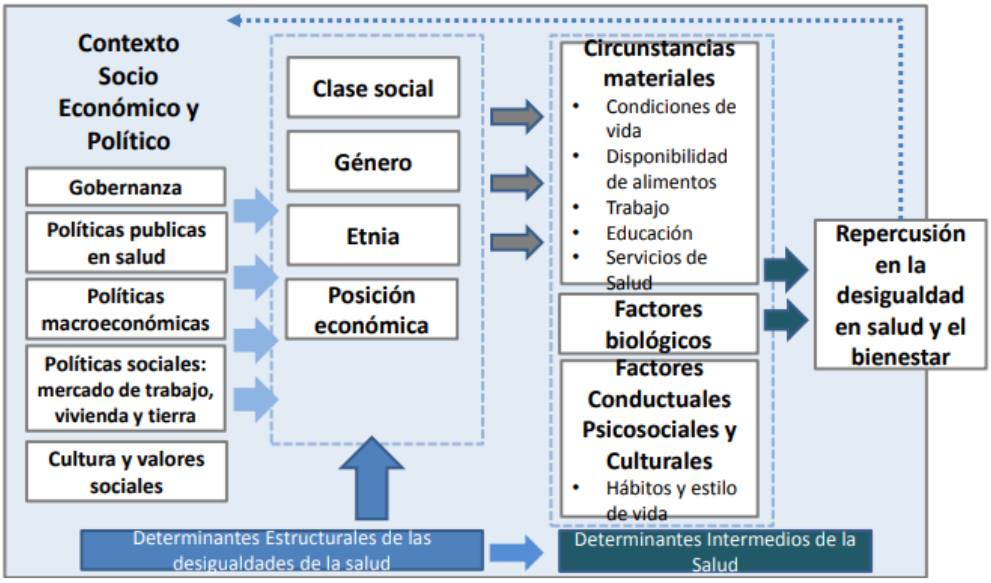


Fig. 1. Enfoque de los determinantes sociales de la salud

Nota: fuente M. Varela, I. Cepeda, A. Uribe, N. Cadavid y J. Botero, «Proyecto Salud y Bienestar en la Comunidad Educativa Javeriana,» Cali, 2023.

3.2 SALUD MENTAL

La salud mental, un concepto fundamental en el bienestar humano, ha sido objeto de diversas definiciones y enfoques. La Ley 1616 de 2013 de Colombia ofrece una perspectiva integral, describiéndola como un "estado dinámico que se expresa en la vida cotidiana a través del comportamiento y la interacción" [5]. Esta definición subraya la naturaleza multifacética de la salud mental, enfatizando su manifestación en las interacciones diarias y su papel crucial en el despliegue de recursos emocionales, cognitivos y mentales de los individuos y colectivos.

Esta conceptualización se alinea con la creciente comprensión de que el bienestar emocional y un contexto que propicie una vida digna son determinantes directos de la salud mental. La Organización Mundial de la Salud (OMS) ha adoptado un enfoque que va más allá de la mera ausencia de trastornos mentales, centrándose en un estado de bienestar que permite a las personas afrontar el estrés cotidiano, desarrollar plenamente sus capacidades y contribuir de manera significativa a su comunidad.

Los determinantes sociales emergen como factores críticos en la comprensión y promoción de la salud mental. Estos abarcan las condiciones en las que las personas nacen, crecen, trabajan y envejecen, influyendo profundamente en sus patrones de pensamiento, sentimientos y comportamientos. La intrincada relación entre estos determinantes y la salud mental subraya la necesidad de un enfoque holístico que considere no solo los aspectos individuales, sino también los contextos sociales, económicos y culturales en los que se desenvuelven las personas.

Esta perspectiva integral no solo enriquece nuestra comprensión de la salud mental, sino que también ofrece vías más efectivas para su promoción y cuidado en diversos ámbitos, incluyendo el entorno universitario.

3.3 VARIABLES DE SALUD MENTAL

El abordaje de la salud mental desde una perspectiva binaria, que contempla aspectos tanto negativos como positivos, (normales y anormales) proporciona a los profesionales de la salud un marco de referencia valioso. Esta dualidad facilita la identificación de factores determinantes, permitiendo una clasificación más precisa de los individuos y, por ende, un estudio más profundo de su condición. Tal enfoque no solo enriquece las estrategias de prevención y promoción, sino que también optimiza los protocolos de tratamiento.

En este contexto, la Encuesta Javeriana de Bienestar y Salud adopta un enfoque integral, evaluando la salud mental a través de Indicadores Positivos y Negativos. Este método reconoce la naturaleza multifactorial de la salud mental, abordando su complejidad mediante el uso de escalas estandarizadas y validadas tanto a nivel internacional como nacional. La aplicación de este cuestionario a colaboradores y estudiantes de la universidad permitió una evaluación comprensiva, capturando la diversidad de experiencias y manifestaciones de la salud mental en la comunidad académica.

3.3.1 Aspectos negativos:

Depresión: La depresión se puede referir a un síntoma o a un trastorno. El síntoma del estado de ánimo depresivo no necesariamente significa que una persona padece un trastorno del estado de ánimo. El Trastorno es caracterizado por la pérdida de la felicidad y desgano, que conlleva a un malestar interior y dificultando la interacción con el entorno. [6]

Ansiedad: Trastorno emocional donde la persona experimenta conmoción, intranquilidad, nerviosismo o preocupación [6]

Estrés: se refiere a nuestra reacción ante una situación que presenta demandas, restricciones u oportunidades y no es normalmente placentero [6]

Soledad: experiencia subjetiva y desagradable que ocurre cuando la red de relaciones sociales de una persona es deficiente de alguna manera importante, ya sea cuantitativa o cualitativamente. [7]

Ideación Suicida: Pensamientos de participar en comportamientos destinados a acabar con la propia vida. Pueden variar en gravedad, desde un deseo vago de muerte hasta planes específicos para suicidarse e intención suicida. [8]

3.3.2 Aspectos positivos:

Resiliencia: La capacidad de adaptarse con éxito frente a la adversidad, el trauma, la tragedia, las amenazas o incluso fuentes significativas de estrés [9]. La resiliencia es un proceso dinámico que implica la capacidad de mantener un funcionamiento adaptativo de los sistemas biológicos y psicológicos ante circunstancias adversas. No es simplemente la ausencia de psicopatología, sino que implica, habilidades de recuperación, sostenibilidad y aprendizaje.

Satisfacción con la vida: Una evaluación global de la calidad de vida de una persona según los criterios elegidos por ella misma. La satisfacción con la vida es un componente clave del bienestar subjetivo, junto con los afectos positivos y negativos. A pesar de ser subjetiva, los criterios de cada persona son basados por aspectos externos, generalmente impuestos por presiones sociales. [10]

Recursos psicológicos: Características positivas de los individuos que pueden ser utilizadas como fortalezas para afrontar desafíos y promover el bienestar. Incluyen aspectos como optimismo, autoeficacia, esperanza y resiliencia. Los recursos psicológicos, también conocidos como capital psicológico, son un conjunto de capacidades psicológicas positivas que contribuyen al desarrollo y éxito personal. [11]
A continuación, se describen las variables establecidas para el análisis en este estudio

Tabla 1. Descripción de variables resultado

Variables Resultado	Indicadores	Etiqueta
Salud mental	**Indicadores Negativos**	
	Depresión	NIVDEP
	Ansiedad	NIVANS
	Estrés	NIVEST
	Soledad	NIVSOLED
	Ideación Suicida	NIVIDEASUIC
	Indicadores positivos	
	Resiliencia	NIVRESIL
	Recursos psicológicos	NIVRECPSIC
	Satisfacción con la vida	NIVSATVIDA

Como variables de exposición, los investigadores del proyecto Salud y bienestar en la Comunidad Educativa Javeriana, establecieron como variables de exposición a los determinantes sociales, así:

Tabla 2. Variables de exposición en estudiantes y colaboradores

Dimensión	Indicadores	Etiqueta
Factores individuales	**Sociodemográficos:** sexo, edad, nivel educativo, estrato socioeconómico, procedencia, residencia rural/urbana, composición familiar.	Género GéneroReco Edad NivEducativo Estratosoc NivelSociec Nacioen ZonaResidencia Vivecon ViveHijos ROLPrincipal RazaEtnia Estadocivil NivEduMadre NivEduPadre Residencia
	Psicosociales: antecedentes de violencia y de abuso sexual, apoyo social, funcionamiento familiar y afrontamiento	FPSantecvio1 FPSantecvio2 FPSantecvio3 FPSantecvio4 FPSantecvio5 FPSantecvio6 FUNCFLIAR APYOSOC FPSafrontam1 FPSafrontam2 FPSafrontam3 FPSafrontam4

		FPSafrontam5
	Estado general de salud: percepción general de saludo, IMC, enfermedades diagnosticadas, condiciones psiquiatrías, dolor, discapacidad	PERCSALUD IMC ENFERMEDAD CONDPSIQU DOLOR DISCAPACIDAD
Aspectos educativos del contexto universitario	Facultad	FACULTODEPEN
Hábitos de salud	Actividad física, sedentarismo y sueño	NIVACTFIS TIEMSED NIVSUEÑO
	Tiempo de ocio	TOCIOrelaj TOCIOartist TOCIOmusic TOCIOmanual TOCIOespirsolit TOCIOespirgrup TOCIOentretsolit TOCIOentretgrup
	Consumo de SPA	SPAalcohol SPAcigarrillo SPAvapeo SPAmarihuana
Alimentación	Prácticas de alimentación	ALIMcafeteriaU ALIMcomerTV ALIMmaquinas

		ALIMtiempo ALIMhoras ALIMcomerotros
	Frecuencia de consumo de alimento	ALIMfrutas ALIMverdur ALIMembutid ALIMpaquetes ALIMcomidrapid ALIMgaseos ALIMdulces ALIMcomidprepar
Sexualidad	Uso del preservatio	Sexpreserv SEXorientasex
Dependencia a tecnología	Uso abusivo de TICS	NIVABUSOTICS
Bienestar	Físico, psicológicos, social, espiritual y ambiental	BIEFISICO BIEPSICO BIESOC BIENESPIR BIEAMBI
Condiciones de vida	Condiciones de vivienda y del barrio, convivencia, acceso a servicios básicos, acceso a espacios, seguridad del barrio, violencia en el barrio, transporte hogar – universidad, seguridad social en salud, ingresos y suficiencia de ingresos del hogar	CVservpub Cvinternet CVzonasocial CVcentrodepor CVtransp CVparques CVcentrossalud CVespaccomunit CVsegurbarrio CVviolenbarrio

		CVTransVehicprop CVTransVehicompar CVTransPublicoMas CVTransPublicoTax CVTransBici CVTransCamina CVingresufic CVingreshogar
Apropiación de la vida universitaria	Conocimiento y acceso a programas de salud y bienestar en la universidad	CONACCPROG1 CONACCPROG2 CONACCPROG3 CONACCPROG4 CONACCPROG5 CONACCPROG6 CONACCPROG7 CONACCPROG8 CONACCPROG9 CONACCPROG10 CONACCPROG11 CONACCPROG12 CONACCPROG13
	Condiciones para la alimentación en la universidad	AMBALIM1 AMBALIM2 AMBALIM3 AMBALIM4 AMBALIM5

El resumen ejecutivo del proyecto Salud y Bienestar en la Comunidad Educativa Javeriana destaca la importancia de incluir variables socioeconómicas y culturales para identificar inequidades en salud dentro de la comunidad educativa, estudiantes y colaboradores

3.4 ANÁLISIS EXPLORATORIO DE DATOS

Es el primer paso de cualquier proyecto de ciencia de datos, se define como un conjunto de procedimientos para sacar conclusiones sobre grandes poblaciones a partir de muestras de pequeño tamaño. [12] En 1962, John W. Tukey, propuso una nueva disciplina científica denominada *análisis de datos,* en donde introdujo la inferencia estadística para representar, resumir y organizar los datos, hoy sus aportes son aún válidos y son parte fundamental de la ciencia de datos.

El Exploratory Data Analysis (EDA) en inglés, o Análisis Exploratorio de Datos es un enfoque para analizar conjuntos de datos con el fin de resumir sus principales características, a menudo utilizando métodos visuales. Es un proceso crucial en la comprensión inicial de los datos antes de aplicar técnicas de modelado o pruebas de hipótesis más formales. [12]

3.5 APRENDIZAJE AUTOMÁTICO

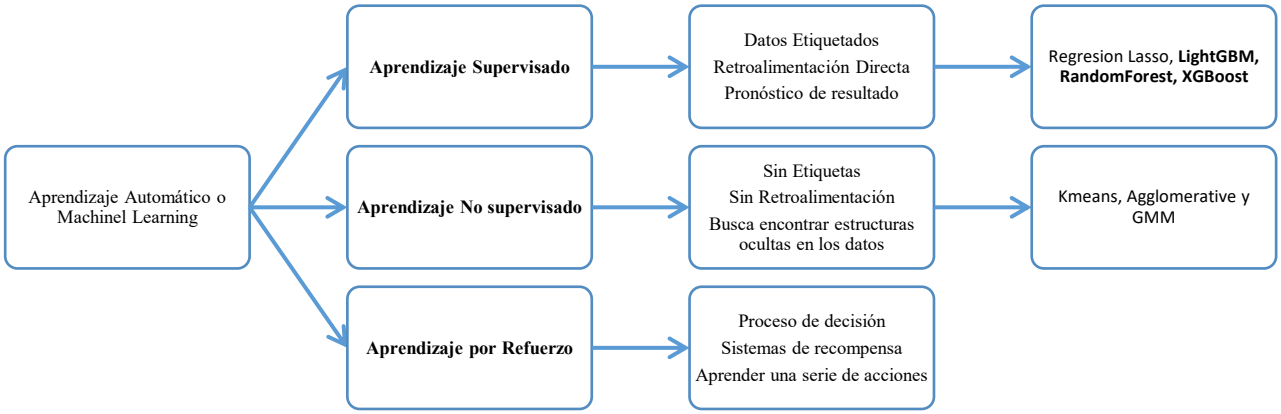
En 1959, el profesor Arthur Samuel planteó que El *aprendizaje automático es el campo de estudio que da al computador la habilidad de aprender sin haber sido explícitamente programado para ello*. [13] actualmente, el aprendizaje automático o Machine Learning (ML), busca que un programa de computar aprenda de un conjunto de datos con los cuales se entrena para buscar identificar patrones para realizar predicciones sobre nuevos datos, en esencia, el aprendizaje automático implica el uso de algoritmos de autoaprendizaje que obtienen conocimiento de los datos con el propósito de hacer predicciones.

En el ML los datos atraviesan por un proceso de entramiento, reglas, la cuales se denominan modelos.

Los modelos son la especificación de una relación matemática (o probabilística) existente entre distintas variables, [14] se pueden comprender como una representación matemática de un fenómeno del mundo real, aprendido a partir de datos. Las estimaciones sobre datos futuros es una de las características de los modelos, es decir, pueden predecir.

El aprendizaje automático se compone de tres tipos:

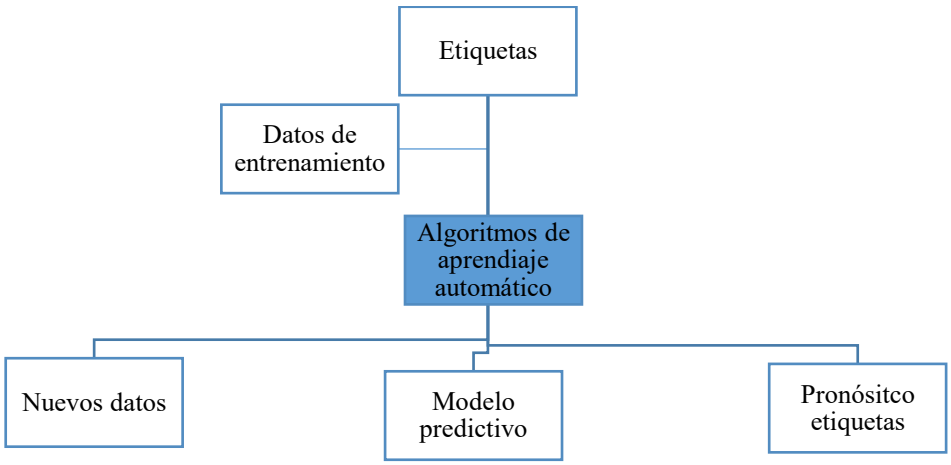
Figura 1. Los tres tipos de aprendizaje automático. Fuente [15]



3.5.1 Aprendizaje supervisado:

Su principal objetivo es descubrir un modelo a partir de datos de entrenamiento etiquetados, que facilite la formulación de pronósticos sobre datos futuros. EL termino supervisado hace referencia a que la entrada y salida de los datos es ya conocida. El aprendizaje supervisado es el proceso de modelar la relación entre las entradas de datos y las etiquetas. [15]

Figura 2. Proceso de aprendizaje supervisado. Fuente [15]



3.5.2 Modelos de Regresión:

En aprendizaje supervisado, la regresión permite predecir un valor continuo, tomando en consideración varias variables de entrada, la regresión busca encontrar la relación entre una variable dependente con algunas variables independientes. [13] Los modelos de regresión es un subcampo del aprendizaje automático que busca predecir valores continuos a partir de ciertas variables descriptivas.

3.5.2.1 Regresión Logística con Regularización LASSO:

Regresion Lasso (Least Absolute Shrinkage and Selection Operator), es un método lineal que aplica una penalización L1 a los coeficientes del modelo, forzando a algunos a ser exactamente cero. Esto permite selecciona automática de variables y es útil para mejorar alta multicolinealidad, es decir un VIF alto. [16]

Una característica importante de la Regresión Lasso es que tiende a eliminar completamente los pesos de las características menos importantes (es decir, a ponerlos a cero). En otras palabras, la regresión de lazo realiza automáticamente la selección de características y genera un modelo disperso (es decir, con pocos pesos de características distintos de cero). [17]

Para el caso de este estudio, se plantearon las variables resultado al inicio del proyecto: Indicadores Negativos e Indicadores Positivos, las cuales fueron tratadas para ajustarse el Modelo de Regresión Lasso.

Tabla 3. Tratamiento de las variables resultado. Fuente Propia

Dimensión	**Variable**	**Tratamiento**	**Nueva Etiqueta**
Indicadores Negativos	NIVDEP	"Normal/leve" pasa a ser 0 "Moderado/severo" y "Extremadamente severo" pasa a ser 1	NIVDEP_bin
	NIVANS		NIVANS_bin
	NIVEST		NIVEST_bin
	NIVSOLED	"Normal" pasa a ser 0 "Moderado" y "Severo" pasa a ser 1	NIVSOLED_bin
Indicadores Positivos	NIVRESIL	"Baja" pasa a ser 0 "Medio y "Alta" para ser 1	NIVRESIL_bin
	NIVSATVIDA		NIVSATVIDA_bin
	NIVRECPSIC		NIVRECPSIC_bin

3.5.2.2 LightGBM -Light Gradient Boosting Machine:

LightGBM es un framework de gradient boosting basado en árboles de decisión que utiliza técnicas como GOSS muestreo de gradientes y EFB agrupamiento de características para mejorar la eficiencia computacional. [18]

3.5.2.3 Random Forest (Bosque Aleatorio):

Un Random Forest es un algoritmo de aprendizaje automático supervisado versátil y potente que se utiliza tanto para tareas de clasificación como de regresión. Pertenece a la familia de los métodos de ensemble, lo que significa que combina múltiples modelos para obtener un mejor rendimiento predictivo que el que se obtendría con un solo modelo. En este caso, un Random Forest construye una colección de árboles de decisión y combina sus predicciones para obtener un resultado final. [19]

3.5.2.4 XGBoost - eXtreme Gradient Boosting:

XGBoost es una implementación optimizada de gradient boosting que incluye regularización (L1/L2), manejo de datos faltantes y paralelización, usa método de ensamble sobre Decision Trees, su diseño esta orientado a la eficiencia, portabilidad y versatilidad.

3.5.3 Aprendizaje no supervisado:

Los datos no cuentan con etiquetas para las muestras de entrenamiento del modelo, el algoritmo actúa directamente sobre los datos de entrada y tiene el propósito de encontrar relaciones entre ellos basándose en características en común. Este tipo de aprendizaje buscar resolver problemas a través del agrupamiento o clustering y la reducción de la dimensionalidad.

Siendo el K-medias (K-mean en inglés) el algoritmo más usado para agrupamiento y en cuanto a reducción de la dimensionalidad el Análisis de Componentes Principales (PCA Prinicpal Component Analysis, en inglés) es la técnica más comúnmente usada en el procesamiento de características para eliminar el ruido de los datos.

3.5.3.1 Kmeans:

La agrupación en clústeres es una técnica para dividir los datos en diferentes grupos, donde los registros de cada grupo son miliares entre sí. Como propósito, este algoritmo busca identificar grupos de datos importantes y con sentido. [12]

K-medias divide los datos en *k* grupos, minimizando la suma de las distancias al cuadrado de cada registro a le media (mean) del grupo que ha sido asignado. K-medias trabaja, identificando la cantidad de *k* de grupos o clústeres a generar, posteriormente, cada grupo tendrá un punto central denominado centroide, finalmente, una vez identificados los centroides, se agrupa en un clúster las muestras del conjunto de datos que comparte el centroide más cercano formando un clúster. [13]

3.5.3.2 Método del Codo (Elbow):

Un aspecto critico en el uso del algoritmo de K-medias es encontrar el número optimo del hiperparámetro de *k,* el método del codo utiliza los valores de inercia, la inercia es la suma de las distancias al cuadrado de cada punto del clúster a su centroide, posteriormente se visualiza este proceso en un gráfico de línea el cual se asemeja a un brazo, se debe revisar el punto donde se note una atenuación (el codo del brazo) en la inercia, lo cual representa el número optimo de clústeres. [13]

3.5.3.3 Análisis de Correspondencia:

El Análisis de Correspondencia (AC) es un método de reducción de dimensionalidad no supervisado diseñado para analizar la asociación entre variables categóricas en una tabla de contingencia. Su objetivo es representar visualmente las relaciones entre las categorías de dos o más variables en un espacio de baja dimensión (generalmente 2D o 3D), facilitando la interpretación de patrones.

3.5.3.4 Análisis de Componente Principales (PCA)

El Análisis de Componentes Principales (PCA) es una técnica de aprendizaje automático no supervisado que se utiliza para reducir la dimensionalidad de un conjunto de datos. Opera sobre variables correlacionadas, buscando identificar patrones y relaciones entre ellas. Para ello, realiza una transformación lineal que rota el sistema de coordenadas original, de manera que las nuevas variables (llamadas componentes principales)

sean ortogonales entre sí, es decir, no estén correlacionadas linealmente. [12]

3.5.3.5 Multicolinealidad:

La colinealidad perfecta representa un escenario límite donde dos variables predictoras exhiben una correlación absoluta, moviéndose conjuntamente de manera idéntica. Sin embargo, en la práctica, la **multicolinealidad** se presenta como una correlación sustancial, aunque no perfecta, entre dos o más variables independientes dentro de un modelo de regresión. Esta interdependencia entre los predictores dificulta la tarea de aislar y cuantificar con precisión el efecto individual de cada variable independiente sobre la variable dependiente.

A diferencia de la colinealidad perfecta, donde la relación lineal es directa y evidente, la multicolinealidad puede manifestarse a través de combinaciones lineales más complejas y menos obvias entre las variables predictoras. Esta sutileza complica su detección a simple vista.

Para identificar y medir el grado de multicolinealidad, una herramienta estadística comúnmente empleada es el Factor de Inflación de la Varianza (VIF). El VIF cuantifica cuánto se incrementa la varianza del coeficiente estimado de una variable predictora debido a su correlación con otras variables predictoras en el modelo. Un valor de VIF de 1 indica la ausencia de correlación con otras variables independientes. A medida que la correlación aumenta, el valor del VIF se eleva, sin límite superior. Valores elevados de VIF sugieren la presencia de multicolinealidad, lo que puede comprometer la estabilidad y la interpretabilidad de los coeficientes de regresión. [20]

3.5.3.6 Análisis de Correspondencias:

Dado que el Análisis de Componentes Principales (PCA) se aplica a variables numéricas, no es adecuado para el análisis de datos categóricos. En este estudio, se ha optado por el Análisis de Correspondencias (AC), una técnica específicamente diseñada para explorar relaciones entre variables categóricas. [12]

El AC permite identificar asociaciones entre diferentes categorías y visualizarlas en un espacio de baja dimensionalidad. Para ello, se construye una tabla de contingencia donde las filas y columnas representan las variables categóricas, y las celdas contienen las frecuencias o conteos de las combinaciones de categorías. A

partir de esta tabla, el AC busca reducir la dimensionalidad de los datos y representar las relaciones entre las categorías de forma gráfica, facilitando la interpretación de los patrones y asociaciones.

En este estudio, se aplicó el AC de forma complementaria al Clustering Jerárquico. Para cada muestra (Colaboradores, Estudiantes de Pregrado y Posgrado), se realizó un análisis de correspondencias posterior a la obtención de los clústeres. La tabla de contingencia utilizada en el AC representó la distribución de las variables "INDICADOR_NEG" e "INDICADOR_POS" dentro de cada clúster. Esto permitió visualizar cómo se relacionan las categorías de estas variables con los clústeres identificados, proporcionando una comprensión más profunda de la estructura de los datos.

3.6 PRUEBA KMO (KAISER-MEYER-OLKIN)

La prueba KMO mide la adecuación de la muestra para el análisis factorial. Evalúa la proporción de varianza entre las variables que podría ser varianza común. El KMO, indica si las variables comparten suficientes factores comunes como para justificar el uso del análisis factorial.

Valores de KMO:

- 0.90 o superior: Excelente
- 0.80 - 0.89: Bueno
- 0.70 - 0.79: Aceptable
- 0.60 - 0.69: Mediocre
- 0.50 - 0.59: Malo
- Inferior a 0.50: Inaceptable

Un valor de KMO bajo indica que las correlaciones entre pares de variables no pueden ser explicadas por otras variables, y por lo tanto, el análisis factorial no sería apropiado. [21] Por lo tanto, es relevante aplicar KMO antes de implementar técnicas de aprendizaje automático.

3.7 FACTOR DE INFLACIÓN DE LA VARIANZA (VIF)

El Factor de Inflación de la Varianza (VIF) es una medida que cuantifica la severidad de la multicolinealidad en un análisis de regresión de mínimos cuadrados ordinarios. Específicamente, estima cuánto se incrementa

la varianza del coeficiente estimado de una variable predictora debido a la correlación con otras variables predictoras en el modelo [22]

No existe un consenso absoluto sobre los umbrales exactos para determinar cuándo el VIF indica un nivel problemático de multicolinealidad. Sin embargo, se han propuesto varias reglas generales basadas en la experiencia y la literatura estadística, en donde:

- **VIF= 1:** Indica que no hay correlación entre la variable predictora y las demás variables predictoras en el modelo. Esto es el escenario **ideal.**
- **1 < VIF < 5:** Generalmente se considera **aceptable.** Este rango sugiere una correlación moderada entre la variable predictora y otras predictoras, pero no lo suficientemente alta como para causar problemas significativos en la estimación de los coeficientes de regresión.
- **VIF ≥ 5:** A menudo se considera una señal de advertencia de multicolinealidad **moderada** a alta.
- **VIF ≥ 10:** Comúnmente se acepta como un umbral que indica multicolinealidad **alta** o severa.

El VIF directamente cuantifica cuánto se infla la varianza del coeficiente de una variable predictora debido a la multicolinealidad. Un VIF de 5 significa que la varianza del coeficiente es cinco veces mayor de lo que sería si esa variable no estuviera correlacionada con las demás predictoras. Un VIF de 10 implica una inflación de la varianza diez veces mayor, lo que reduce significativamente la precisión de la estimación del coeficiente. [20]

La multicolinealidad hace que los coeficientes de regresión sean inestables. Pequeños cambios en los datos pueden llevar a grandes cambios en los coeficientes estimados. Los VIF altos indican una mayor sensibilidad de los coeficientes a las fluctuaciones en los datos debido a la interdependencia entre las variables predictoras. [20]

4 ESTADO DEL ARTE

A continuación se describen algunas investigaciones relacionadas con salud mental y su convergencia con la ciencia de datos, bajo la premisa que las condiciones de depresión y ansiedad representan un desafío creciente para la salud publica a nivel global, la detección temprana y la intervención oportuna, destacan como iniciativas fundamentales para mitigar el impacto de estos trastornos; a su vez presentan una innovadora intervención de la IA (inteligencia artificial) para la intervención con planes de tratamiento personalizados.

Título: Prediction of mental health (Depression) using data science and Machine Learning techniques. [23]
Autor: C. MADHUMITHA Sathybama Institute of Science and Technology
Marzo 2022, Chennai India
Hallazgos: ML detectó patrones tempranos de depresión y ansiedad en registros de un hospital de Chennai, los robustos resultados permitieron perfilar a la población con bases solidasd para la prevención de trastornos en salud mental.

Título: Enchancing mental health with Artificial Intelligence: Currente trends and future prospects [24]
Autores: David B. Olawade, Ojima Z. Wada, Aderonke Odetayo, Aanuoluwapo Clement David- Olawade, Fiyinfoluwa Asaolu, Judith Eberhardt.
Abril 2024, Londres Inglaterra
Hallazgos: Revisión sistemática en PubMed, IEEE Xplore, PsycINFO y Google Scholar que reveló tendencias actuales con evidente potencial transformador de la IA, con aplicaciones como la detección temprana de trastornos de salud mental, planes de tratamiento personalizados y terapeutas virtuales basados en IA

Título: Posibles aplicaciones prácticas del uso de Machine Learning (ML) en la investigación y práctica de la clínica psicológica. [25]
Autores: López Steinmetz Lorena, Godoy Juan Carlos
Octubre 2023, Córdoba, Argentina
Hallazgos: se resaltó cómo los algoritmos de ML permiten el análisis eficiente de grandes conjuntos de datos, la identificación de patrones ocultos y la generación de conocimientos profundos en el estudio de la

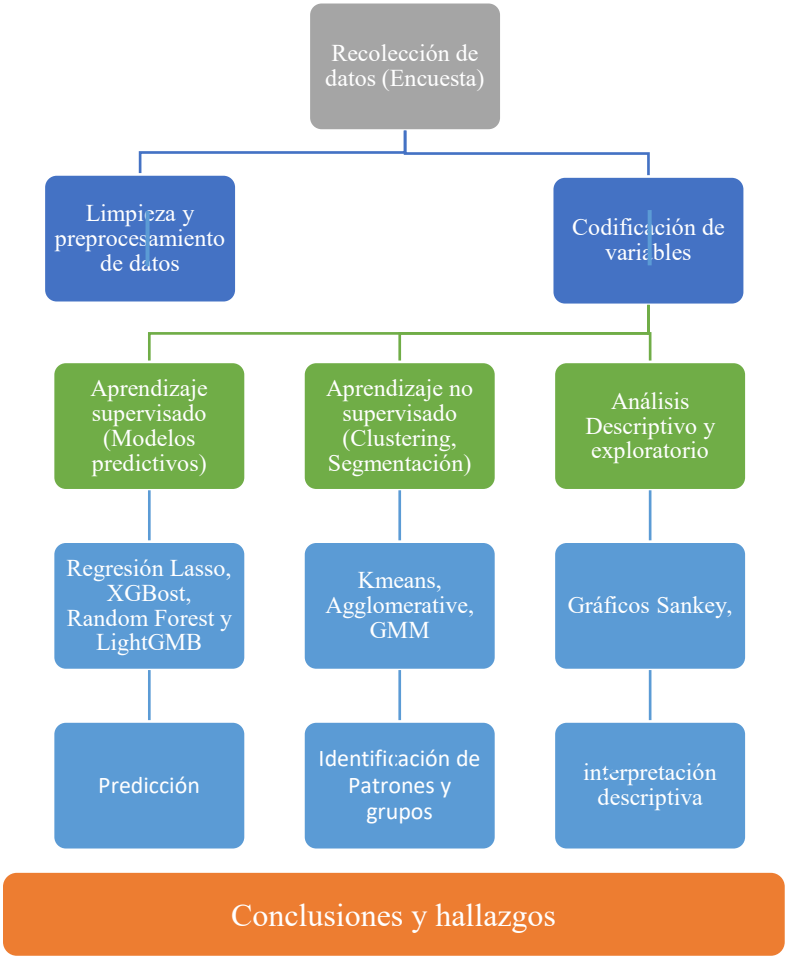
mente humana.

El análisis del estado del arte en la aplicación de la ciencia de datos y el Machine Learning en la predicción de la salud mental subraya el potencial de estas herramientas para revolucionar la psiquiatría y la psicología clínica. Los hallazgos de estudios como los de C. Madhumitha, David B. Olawade et al., y López Steinmetz y Godoy, tienen en común que la idea de que los algoritmos de ML no solo son capaces de detectar patrones tempranos de trastornos como la depresión y la ansiedad, sino que también ofrecen la posibilidad de personalizar tratamientos y generar conocimiento profundo a partir de grandes volúmenes de datos.

5 METODOLOGÍA

Para dar cumplimiento a los objetivos de este estudio, se presenta la metodología utilizada, teniendo en cuenta el enfoque de investigación para proyectos en Ciencia de Datos aprendidos en la maestría, la cual se puede resumir en el siguiente diagrama.

Figura 3. Resumen metodológico del estudio. Fuente: elaboración propia



5.1 ENTENDIMIENTO DE LOS DATOS

Con base en la información suministrada por los investigadores de la universidad, se procedió a realizar un análisis exploratorio de datos así:

Primero, el proceso de limpieza de datos consistió imputando la moda como estrategia para datos faltantes.

5.2 ESTRUCTURA DE LOS DATOS

El conjunto de datos denominado: BD COMUNIDAD JAVERIANA – MCD.sav fue suministrado como información base para el desarrollo de este proyecto, allí se encontró una estructura de datos para lectura en el software SPSS, conformado por 123 variables categóricas codificadas numéricamente y 1 variable cuantitativa continua, denominada “Edad”; todos los datos distribuidos en 4240 registros.

El conjunto de datos registraba las respuestas de una encuesta aplicada en el segundo semestre del 2022 a Colaboradores, estudiantes de Pregrado y Posgrado de la Pontificia Universidad Javeriana Cali con el propósito de caracterizar la salud mental de la comunidad educativa Javeriana e identificar sus determinantes sociales.

En el preprocesamiento de los datos, se tuvieron que codificar las variables cualitativas, en su mayoría en escalado Likert, de conformidad al diccionario de datos aportado por las investigadoras de proyecto, siendo las personas encargadas de la recolección de los datos, así mismo, las variables venían en escalado tipo Likert, pero en algunas variables el escalado estaba invertido, en donde ‘Nunca = 1’ y ‘Siempre = 5’; y en otras ‘Nunca= 5’ y ‘Siempre = 1’.

La tabla 4 describe la estructura de datos original

Tabla 4. Distribución de Variables, dimensiones y No. de ítems. Fuente: propia

Variables	Dimensiones	No de ítems
Salud Mental	Indicadores Negativos	5
	Indicadores positivos	3
Salud y hábitos de salud	Estado general de salud	42
	Hábitos de Salud	
	Ocio	
	Consumo de SPA	
	Alimentación	
	Sexualidad	

	Dependencia a tecnología	
Calidad de Vida	Bienestar	5
Factores individuales	Sociodemográficos	20
	Psicosociales	
Condiciones de vida	Condiciones de la vivienda y del barrio	26
	Movilidad	
	Situación económica del grupo familiar	
Determinantes estructurales	Apropiación de la vida universitaria	18
	Condiciones para la alimentación en la universidad	

El conjunto de datos contenía 124 variables en total, luego del tratamiento de variables siguiendo Agregación por Dimensiones para simplificar el número de datos, se redujo a 38 variables así:

Tabla 5. Nueva distribución de Variables, dimensiones y No. de ítems. Fuente: propia

Variables	**Dimensiones**	**No de ítems**
Salud Mental	Indicadores Negativos	1
	Indicadores positivos	1
Salud y hábitos de salud	Estado general de salud	10
	Hábitos de Salud	
	Ocio	
	Consumo de SPA	
	Alimentación	
	Sexualidad	
	Dependencia a tecnología	
Calidad de Vida	Bienestar	1
Factores individuales	Sociodemográficos	17
	Psicosociales	
Condiciones de vida	Condiciones de la vivienda y del barrio	6
	Movilidad	
	Situación económica del grupo familiar	
Determinantes estructurales	Apropiación de la vida universitaria	2
	Condiciones para la alimentación en la universidad	

En el proceso de entendimiento de datos, se identificaron variables que estaban completamente vacías o con una cantidad insignificante de valores disponibles, otra en particular no contenía información relevante para

la investigación, por lo tanto, debido a su falta de información, estas variables no contribuían al cumplimiento de los objetivos de la investigación, ya que no proporcionaban valor añadido ni permitieron obtener patrones significativos para el análisis. Además, su presencia podría distorsionar los resultados al incrementar la dimensionalidad sin aportar información relevante.

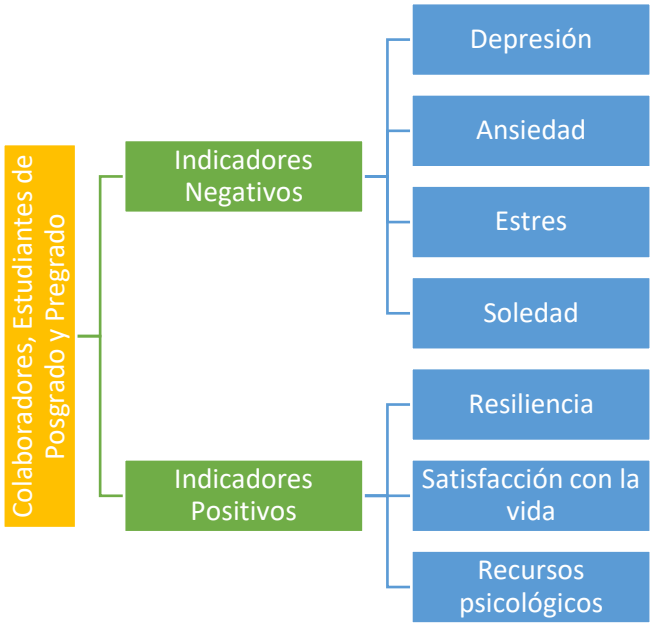
Por lo tanto, la decisión de eliminar estas variables se basó en criterios de calidad de datos, con el fin de mejorar la eficiencia y precisión del análisis, enfocando los esfuerzos en variables que sí tienen datos significativos y que pueden influir en la comprensión de las relaciones estudiadas.

Se eliminaron las variables: 'CODIGO', 'SPAILEGALES', 'ESTPROGRAMAPRE', 'ESTPROGRAMAPOS', 'ESTSEMESTRE', 'ESTBECA' Y 'PROFESORTIPO'

5.3 EXPLORACIÓN DE DATOS

Para conocer la estructura del Dataset, se aplicaron medidas estadísticas para conocer la distribución de lo datos, teniendo en cuenta que parte de los objetivos implícitos del estudio es emitir resultados en las poblaciones de Colaboradores, estudiantes de Posgrado y Pregrado, a su vez, con las variables de interés del estudio, así:

Figura 4. Variables resultado del estudio. Fuente: propia



La exploración y análisis de los datos se llevó a cabo en Python, utilizando el entorno de desarrollo Visual Studio Code. Para el procesamiento y manipulación de datos se emplearon las librerías Pandas y Numpy. La visualización de los resultados se realizó tanto con Matplotlib y Seaborn para gráficos estáticos, como con Plotly para visualizaciones interactivas, incluyendo los diagramas Sankey.

Para el análisis estadístico y la implementación de técnicas de aprendizaje automático, se utilizaron scikit-learn (sklearn), que facilitó la segmentación de la población mediante algoritmos de clustering y la construcción de modelos predictivos.

Adicionalmente, para abordar el desbalance de clases en los datos, se aplicaron técnicas de sobremuestreo como SMOTE a través de la librería imblearn.over_sampling. En etapas complementarias del análisis, se recurrió a Scipy para pruebas estadísticas y manejo de funciones matemáticas avanzadas.

5.4 CONSTRUCCIÓN DE DATOS A TRAVÉS DE LÓGICA DIFUSA

Con el objetivo de simplificar el análisis e interpretación de los resultados, se ha optado por utilizar la lógica difusa para agregar los valores de las dimensiones "Indicadores Positivos" e "Indicadores Negativos" en un

único valor representativo. Se tuvo en cuenta la capacidad de la lógica difusa para manejar la incertidumbre y la vaguedad inherentes a la evaluación de estas variables resultado. La lógica difusa permite modelar grados de pertenencia a diferentes categorías. Esto proporciona una representación más precisa y matizada de la realidad, capturando la naturaleza gradual de muchos fenómenos.

La siguiente tabla detalla cómo se aplicó la lógica difusa a las dimensiones "Indicadores Positivos" e "Indicadores Negativos", especificando las dimensiones, variables, las funciones de pertenencia y las reglas de inferencia utilizadas para obtener un valor único que represente el desempeño de cada dimensión

Tabla 6. Tratamiento Lógica difusa a Indicadores Negativos e Indicadores Positivos. Fuente. propia

Dimensión	Variable	Reglas de inferencia	Tratamiento Lógica Difusa
Indicadores Negativos	NIVDEP NIVANS NIVEST NIVSOLED	Valor máximo en cada variable = 5 Valor mínimo en cada variable = 1 Valor Máximo en la dimensión = 20 Valor Mínimo en la dimensión = 4	Nueva Variable: **INDICADOR_NEG** Rango: • 4– 9 = Normal • 10 – 15 = Moderado • 16 – 20 = Severo
Indicadores Positivos	NIVRESIL NIVSATVIDA NIVRECPSIC	Valor máximo en cada variable = 5 Valor mínimo en cada variable = 1 Valor Máximo en la dimensión = 15 Valor Mínimo en la dimensión = 3	Nueva Variable: **INDICADOR_POS** Rango: • 3 – 7 = Bajo • 8 – 11 = Medio • 12 –15 = Alto

5.4.1 Agregación por Dimensiones:

Siguiendo las premisas de la lógica difusa, dado el volumen de datos, se procedió a agregar por dimensiones debido a la simplicidad en el análisis y las ventajas de tener un conjunto de datos reducido sin perder los constructos teóricos que soportaran la fuerza del estudio.

Occam's razor aplicado a análisis de datos

Simplicidad como signo de verdad es la premisa de la Occam's razor' o navaja de Ockham, en análisis de datos se traduce en que modelos simples deberían ser priorizados sobre los complejos, este principio también conocido como el Principio de Parsimonia, implica que modelos más simples con incluso menos variables son preferibles cuando no pierden capacidad explicativa [26].

Para este estudio, se aplicó dicho principio siguiendo la naturaleza de la estructura de datos sumando las puntuaciones por dimensiones, las cuales ya contaban con una estructuración definida según las dimensiones propuestas por el *Proyecto Salud y Bienestar en la Comunidad Educativa Javeriana,* la disminución de variables o dimensionalidad se hizo así:

Tabla 7. Agregación por Dimensiones. Variables de exposición en estudiantes y colaboradores

Variables Originales	**Denominación original**	**Tratamiento**	**Denominación nueva variable**
Percepción de salud	PERCSALUD	Se queda igual	PERCSALUD
Índice de masa corporal	IMC	Se queda igual	IMC
• Enfermedad • Condición psiquiátrica • Dolor	ENFERMEDAD CONDPSIQU DOLOR	Se transforma sumando cada puntuación. Valor mínimo: 0 Valor máximo: 3	SALUD_GENERAL
• Actividad física • Tiempo sedentario • Calidad de sueño	NIVACTFIS TIEMSED NIVSUEÑO	Se transforma sumando cada puntuación. Valor mínimo: 4 Valor máximo: 12	HABITOS
• Actividades de relajación mental y corporal • Actividades artísticas • Actividades musicales • Actividades artísticas manuales • Actividades espirituales en solitario	• TOCIOrelajreco • TOCIOartistreco • TOCIOmuscireco • TOCIOmanualreco • TOCIOespirsolitreco • TOCIOespirgrupreco • TOCIOentretsolitreco	Se transforma sumando cada puntuación. Valor mínimo: 0 Valor máximo: 8	OCIO

• Actividades espirituales en grupo • Actividades de entretenimiento en solitario • Actividades de entretenimiento en grupo	• TOCIOentretgrupreco		
• Consumo de Alcohol Consumo de Cigarrillo • Consumo de Vapeadores • Consumo de Marihuana	SPAalcohol SPAcigarrillo SPAvapeo SPAmarihuana	Se transforma sumando cada puntuación. Valor mínimo: 4 Valor máximo: 20	SPA
• Frutas enteras o en jugo • Verduras crudas • Embutidos • Alimentos de paquete • Alimentos de comida rápida • Gaseosas té refrescos o jugos • Dulces helados pasteles o tortas • Comida preparada • Cafeterías de la Universidad • Comer mientras se ve televisión • Consumo alimentos de las máquinas expendedoras de la Universidad • Tiempo suficiente para comer • Comer en horas fijas • Comer junto con otras personas en la Universidad	ALIMfrutas ALIMverdur ALIMembutid ALIMpaquetes ALIMcomidrapid ALIMgaseos ALIMdulces ALIMcomidprepar ALIMcafeteriaU ALIMcomerTV ALIMmaquinas ALIMtiempo ALIMhoras ALIMcomerotros	Se transforma sumando cada puntuación. Valor mínimo: 14 Valor máximo: 70	HAB_ALIMENTICIOS

Uso de preservativo	SEXpreserv	Se queda igual	SEXpreserv
Orientación sexual	SEXorientacsex	Se queda igual	SEXorientacsex
Abuso de tics	NIVABUSOTICS	Se queda igual	NIVABUSOTICS
Bienestar físico Bienestar psicológico Bienestar social Bienestar espiritual Bienestar ambiental	BIEFISICO BIEPSICO BIESOC BIENESPIR BIEAMBI	Se transforma sumando cada puntuación. Valor mínimo: 5 Valor máximo:25	BIENESTAR
Género	Genero	Se queda igual	Genero
Género reconocido	GeneroReco		GeneroReco
Edad	Edad		Edad
Estrato socioeconómico	Estratosoc		Estratosoc
Nivel socioeconómico	NivelSocioec		NivelSocioec
Estado civil	Estadocivil		Estadocivil
Autorreconocimiento de raza o etnia	RazaEtnia		RazaEtnia
Lugar de Nacimiento	Nacioen		Nacioen
Nivel educativo de la madre	NivEduMadre		NivEduMadre
Nivel educativo del padre	NivEduPadre		NivEduPadre
¿dónde resides?	Residencia		Residencia
Zona de residencia	ZonaResidencia		ZonaResidencia
Convivencia con	Vivecon		Vivecon
Convivencia con hijos	ViveHijos		ViveHijos
• Concentra sus pensamientos en lo positivo de la situación. • Piensa constantemente en el problema. • Minimiza lo que le molesta con burla o ignorancia. • Habla con otros buscando opiniones, consejos o ideas. • Desarrolla y ejecuta un plan para afrontar	FPSafrontam1 FPSafrontam2 FPSafrontam3 FPSafrontam4 FPSafrontam5	Se transforma sumando cada puntuación. Valor mínimo: 5 Valor máximo:25	AFRONTAMIENTO

eficazmente futuras situaciones			
APOYO SOCIAL FUNCIONAMI ENTO FAMILIAR	APOYOSOC FUNCFLIAR	Se transforma sumando cada puntuación. Valor mínimo: 2 Valor máximo:10	APOYO_SOC_FAM
• Agresiones físicas • Agresiones físicas a terceros • Obligación a tener relaciones sexuales por la fuerza • Obligación a alguien a hacer algo que no quería • Sentimientos de menosprecio por características personales • Has hecho sentir menos a alguien por tus características s personales	FPSantecviol1 FPSantecviol2 FPSantecviol3 FPSantecviol4 FPSantecviol5 FPSantecviol6	Se transforma sumando cada puntuación. Valor mínimo: 6 Valor máximo: 30	ANTECVIOL
• Acceso a Servicios públicos • Acceso a Internet • Acceso a Zona social o recreativa • Acceso a centros deportivos • Transporte público • Parques o zonas verdes • Centros de salud • Espacios comunitarios	CVservpub Cvinternet CVzonasocial CVcentrodepor CVtransp CVparques CVcentrossalud CVespaccomunit	Se transforma sumando cada puntuación. Valor mínimo: 0 Valor máximo: 8	SERVICIOS
Respecto a la seguridad en tu barrio o sector consideras que es	CVsegurbarrio	Se queda igual	CVsegurbarrio
Respecto a la violencia en tu	CVviolenbarrio		CVviolenbarrio

barrio o sector consideras que es			
• Frecuencia de uso de transporte en • Vehículo propio • Vehículo compartido • Transporte público masivo • Transporte publico taxi • Transporte en bicicleta • Transporte caminando	CVTransVehicprop CVTransVehicompar CVTransPublicoMas CVTransPublicoTax CVTransBici CVTransCamina	Se transforma sumando cada puntuación. Valor mínimo: 0 Valor máximo: 8	TRANSPORTE
¿Consideras que tus ingresos o los de tu grupo familiar alcanzan a cubrir todos los gastos del hogar?	CVingresufic	Se queda igual	CVingresufic
¿Cuánto es el ingreso económico en tu hogar?	CVingreshogar		CVingreshogar
• Conoce recursos universitarios para ayuda profesional en salud. • Usa servicio médico universitario al enfermar. • Busca ayuda médica externa al enfermar. • Conoce recursos universitarios para ayuda profesional en salud mental. • Busca ayuda en Bienestar Universitario por malestar emocional. • Consulta Centro Pastoral por malestar emocional. • Busca ayuda de profesional de salud mental externo por malestar emocional.	CONACCPROG1 CONACCPROG2 CONACCPROG3 CONACCPROG4 CONACCPROG5 CONACCPROG6 CONACCPROG7 CONACCPROG8 CONACCPROG9 CONACCPROG10 CONACCPROG11 CONACCPROG12 CONACCPROG13	Se transforma sumando cada puntuación. Valor mínimo: 13 Valor máximo: 52	PROGRAMAS_UNI

• Siente que la Universidad promueve discusión abierta sobre salud mental. • Percibe que la Universidad valora diversidad e inclusión. • Considera suficientes los espacios de descanso en la Universidad. • Puede realizar actividades físicas en la Universidad. • Puede realizar actividades artísticas en la Universidad. • Puede participar en experiencias de crecimiento en la Universidad.			
• Encuentra suficiente variedad de alimentos en la Universidad. • Puede comprar alimentos deseados a precios razonables en la Universidad. • Puede comprar alimentos saludables en la Universidad. • Los espacios para comer en la Universidad son agradables. • Los espacios para comer en la Universidad son suficientes.	AMBALIM1 AMBALIM2 AMBALIM3 AMBALIM4 AMBALIM5	Se transforma sumando cada puntuación. Valor mínimo: 5 Valor máximo: 20	AMBIENTE_UNI

El conjunto de datos pasa de tener 107 variables predictoras o independientes a 36.

5.4.2 Metodología Clustering, aprendizaje no supervisado:

Adicionalmente, para complementar el análisis de agrupamiento correspondiente al objetivo 1, se implementaron diversas técnicas de clustering, incluyendo KMeans, clustering Agglomerativo y Gaussian Mixture Models (GMM). Estas metodologías permitieron explorar la posible existencia de subgrupos o patrones latentes dentro de la población estudiada.

Para facilitar la visualización y la interpretación de los resultados de agrupamiento en un espacio de menor dimensión, se emplearon técnicas avanzadas de reducción de dimensionalidad. En particular, se utilizaron tanto el Análisis de Componentes Principales (PCA), que realiza una transformación lineal maximizando la varianza explicada, como métodos no lineales como t-distributed Stochastic Neighbor Embedding (t-SNE) y Uniform Manifold Approximation and Projection (UMAP).

Mientras que PCA es útil para identificar las direcciones principales de variabilidad en los datos, t-SNE y UMAP están diseñados para preservar la estructura local y la proximidad relativa entre observaciones similares, lo que resulta valioso en contextos donde la estructura de los datos puede ser compleja o no lineal, como ocurre frecuentemente en estudios de salud mental y determinantes sociales. [27]

Aunque las métricas cuantitativas de calidad de los clústers (como Silhouette, Calinski-Harabasz y Davies-Bouldin) calculadas sobre los resultados de KMeans y PCA no evidenciaron la presencia de agrupamientos naturales bien definidos, la aplicación de t-SNE y UMAP se justificó como una herramienta exploratoria. Estas técnicas permitieron identificar visualmente posibles subestructuras, gradientes o solapamientos entre grupos que podrían pasar desapercibidos con métodos lineales o métricas tradicionales.

Es importante destacar que la interpretación cuantitativa de la calidad del agrupamiento se basó exclusivamente en las métricas calculadas en el espacio original estandarizado, ya que las proyecciones generadas por t-SNE y UMAP pueden distorsionar las distancias globales y no son adecuadas para la evaluación formal de la calidad del clustering. En resumen, t-SNE y UMAP se emplearon como apoyo visual y exploratorio, proporcionando una comprensión más profunda de la estructura interna de los datos, pero no como base para la validación cuantitativa de los resultados de agrupamiento.

5.4.3 Modelos predictivos, aprendizaje supervisado

El cumplimiento del objetivo 2, orientado a la predicción y modelado de los indicadores de interés, se implementaron diversas técnicas de aprendizaje supervisado. El proceso metodológico incluyó la selección y preparación de variables predictoras, la partición de los datos en conjuntos de entrenamiento y prueba, y la evaluación rigurosa del desempeño de los modelos.

Entre los algoritmos empleados se encuentran la Regresión Lasso, un método lineal que incorpora regularización L1 para la selección automática de variables y la reducción del sobreajuste, y varios modelos de ensamble y boosting ampliamente reconocidos por su capacidad predictiva: Random Forest, XGBoost y LightGBM. Random Forest se utilizó por su robustez ante datos ruidosos y su habilidad para manejar relaciones no lineales y variables categóricas. XGBoost y LightGBM se seleccionaron por su eficiencia computacional y su capacidad para capturar interacciones complejas entre variables, siendo especialmente útiles en contextos de grandes volúmenes de datos y alta dimensionalidad.

El ajuste y la validación de los modelos se realizaron mediante técnicas de validación cruzada, optimización de hiperparámetros y el uso de métricas apropiadas para el tipo de variable objetivo. Además, se aplicaron técnicas de sobremuestreo como SMOTE para abordar posibles desbalances en las clases, asegurando así una evaluación justa y representativa del desempeño de los modelos.

Finalmente, la interpretación de los resultados incluyó el análisis de la importancia de las variables y la comparación de los modelos en términos de precisión, robustez y capacidad de generalización, permitiendo identificar los factores más relevantes asociados a los indicadores de interés y proponer recomendaciones basadas en evidencia.

5.4.4 Herramienta de visualización:

Para el desarrollo del objetivo 3, se diseñó y construyó una herramienta de visualización interactiva orientada a facilitar la exploración y comunicación de los resultados obtenidos en los análisis previos. Esta herramienta fue implementada en PowerBi, la herramienta permite al usuario explorar de manera dinámica la estructura y segmentación de la población, identificar patrones relevantes y analizar la distribución de los indicadores clave a través de diferentes dimensiones. Además, la interactividad de los gráficos facilita la identificación

de subgrupos minoritarios y la comprensión de la complejidad inherente a la población estudiada, aspectos que serían difíciles de captar mediante visualizaciones estáticas tradicionales.

El desarrollo de la herramienta incluyó la integración de los resultados de los análisis descriptivos, de clustering y de predicción, permitiendo así una visión integral y flexible de los datos.

6 ANTECEDENTES

El primer gran reporte que abordó el problema de los determinantes sociales frente a la salud corresponde al presentado por la OMS en 2008. Dentro de este reporte se hace una evaluación general del estado de la salud y el bienestar a nivel mundial y su relación con múltiples factores sociales [9]. Como parte de este reporte se presenta una serie de recomendaciones que buscan orientar el desarrollo de políticas más efectivas para la conservación del bienestar, resaltando el papel de los gobiernos y las acciones internacionales en esta iniciativa [9]. Las áreas de acción y recomendaciones presentadas en este reporte han sido utilizadas de manera casi universal en estudios posteriores de los determinantes sociales en salud.

A nivel académico, muchos estudios se han enfocado en la cuantificación de los efectos en salud mental de los determinantes sociales. En un artículo de 2019, escrito por Alegría, NeMoyer, Falgas, Wang y Álvarez, se presenta una revisión de trabajos científicos alrededor de este tema [20]. Dentro de este artículo se evidencia que la mayoría de los estudios buscan establecer o confirmar correlaciones entre factores socioeconómicos y la incidencia de trastornos mentales [20]. Aunque estas correlaciones han permitido el diseño de políticas públicas, se observa que la medición del impacto real de estas políticas puede no ser tan clara [20]. Una de las necesidades más frecuentes en los estudios analizados, es la falta de estandarización de la metodología analítica. En muchos casos, las estrategias de colección de datos no han sido implementadas a partir de un diseño experimental y, por lo tanto, pueden presentar altos problemas de sesgo o información incompleta [20]. En este sentido, muchas de las conclusiones presentadas en estos estudios podrían representar relaciones incorrectas entre los determinantes sociales y su impacto en la salud mental. Adicionalmente, los autores resaltan la necesidad de estudiar los impactos negativos que puedan tener las conclusiones presentadas en este tipo de estudios, por ejemplo, con la renuncia a acciones de mitigación por asignación de responsabilidad a factores no controlables [20]. Esta situación es particularmente crítica en aquellos estudios en los que las fuentes de datos presentan sesgos, información incompleta o no cumplen con criterios de aleatoriedad [20].

En el país, la implementación de políticas para la gestión de los determinantes sociales en salud ha sido liderada por el ministerio de salud. En un informe del 2015, titulado “La Equidad en salud para Colombia” se presenta un diagnóstico detallado del estado de algunos determinantes sociales y su impacto en indicadores

de salud en el país. En particular, el estudio se enfoca en la comparación de la desigualdad en los indicadores de salud a nivel territorial y su evaluación entre los años 2005 y 2012 [21]. El estudio resalta una amplia brecha para los indicadores de salud en Colombia frente a los de la OCDE y algunos países de América Latina y el Caribe, siendo la baja inversión per cápita en salud una de las razones principales para esta brecha en el país [21]. Como parte de este estudio se determinó la necesidad de intervenir sobre algunos determinantes puntuales como lo son: el acceso a agua potable y sistemas de evacuación de aguas servidas y de lluvia, los sistemas de disposición de basuras, las condiciones de higiene en las viviendas y sus entornos, la adopción de prácticas de vida y de entretenimiento, entre otras [21]. Finalmente, se resalta la necesidad de extender estos estudios para medir las condiciones de equidad entre grupos poblacionales [21].

Desde la academia se ha participado también en la caracterización de los determinantes sociales en salud para Colombia. En un estudio presentado por Tovar y su equipo en 2018, se realizó un análisis de la encuesta de calidad de vida del DANE en el que se evaluaron diferentes modelos de respuesta cuantitativa para medir el efecto de los determinantes sociales sobre el estado de salud [22]. La fuente de información tuvo una muestra representativa de las regiones del país con un total de 52044 individuos, para quienes se evaluaron algunos determinantes sociales como: el ingreso del hogar, el nivel educativo, la edad, el sexo, la etnia, la afiliación a salud, el desempleo, la zona y la región [22]. Sobre estos datos se ajustaron modelos de regresión probit y logit, utilizando como medida de ajuste el criterio de información de Akaike [22]. El estudio reporta algunos grupos críticos en los que se requiere especial atención, como las mujeres, la población afrodescendiente, y la población del pacífico [22]. Las relaciones reportadas, sin embargo, pueden verse afectadas por la limitada disponibilidad de factores evaluados.

De manera complementaría, Pérez presentó un estudio de correlación entre el índice de condiciones de vida (ICV) y el estado de salud para los habitantes de la ciudad de Bogotá [23]. En este estudio, se utilizaron los datos del ICV evaluado por el DANE en la encuesta de calidad de vida para 2003 y 2007, y en la encuesta multipropósito para 2011. En particular, se utilizó el factor de acceso y calidad de servicios del ICV que evalúa la disponibilidad de servicios de acueducto, aseo y combustible para cocinar [23]. Este estudio se enfocó en el análisis de regresión entre este factor y el ICV a nivel de las localidades, además de un análisis de agrupación (clustering) a partir de los resultados de estas regresiones [23]. Aunque el estudio refleja diferencias en las tendencias entre localidades, los modelos lineales utilizados parecen ser muy limitados

(con factores de ajustes muy bajos) lo que limita la validación de las conclusiones.

Finalmente, vale la pena resalta los esfuerzos desarrollados por la Universidad Javeriana Cali frente al diagnóstico del estado de salud mental en la comunidad. Como parte de su planeación estratégica, la universidad incluye el mega 4 “Vivir la fraternidad en nuestra casa común”, dentro de la cual se hace explícito el interés sobre clima organizacional, salud y bienestar [3]. Es en relación con este interés que se desarrolló la encuesta de Salud y Bienestar que evaluó múltiples indicadores de salud y sus determinantes sociales en la comunidad Javeriana [3]. El estudio tiene un enfoque mixto, con un componente cuantitativo desarrollado a través de una encuesta aplicada en el año 2022. Para la encuesta se incluyeron dos poblaciones: estudiantil y colaboradores. En la encuesta se evaluaron indicadores de salud enfocados en depresión, ansiedad, estrés, soledad, ideación suicida, resiliencia, recursos psicológicos y satisfacción con la vida [3]. Para el conjunto de colaboradores en particular, se midieron determinantes sociales asociados con factores individuales, aspectos del contexto laboral, condiciones de vida y determinantes estructurales [3]. Es a partir de la información recolectada en esta encuesta que se desarrollarán los análisis propuestos en este proyecto.

7 RESULTADOS

Para el procesamiento y análisis de datos, se utilizó **Visual Studio Code (VS Code)**, un editor de código fuente desarrollado por Microsoft, que ofrece soporte para múltiples lenguajes de programación, integración con herramientas de análisis de datos (como Jupyter Notebooks, extensiones para Python/R) y funcionalidades avanzadas de depuración. Su flexibilidad y capacidad para manejar grandes conjuntos de datos lo hicieron adecuado para la implementación de algoritmos de aprendizaje no supervisado en este estudio.

7.1 ENTENDIMIENTO DE LOS DATOS

Con base en la información suministrada por los investigadores de la universidad, se procedió a realizar un análisis exploratorio de datos así:

7.2 ESTRUCTURA DE LOS DATOS

El conjunto de datos denominado: BD COMUNIDAD JAVERIANA – MCD.sav fue suministrado como información base para el desarrollo de este proyecto, allí se encontró una estructura de datos para lectura en el software SPSS, conformado por 123 variables categóricas codificadas numéricamente y 1 variable cuantitativa continua, denominada “Edad”; todos los datos distribuidos en 4240 registros.

El conjunto de datos registraba las respuestas de una encuesta aplicada en el segundo semestre del 2022 a Colaboradores, estudiantes de Pregrado y Posgrado de la Pontificia Universidad Javeriana Cali con el propósito de caracterizar la salud mental de la comunidad educativa Javeriana e identificar sus determinantes sociales.

La tabla 8 describe la estructura de datos original

Tabla 8. Distribución de Variables, dimensiones y No. de ítems. Fuente: propia

Variables	Dimensiones	No de ítems
Salud Mental	Indicadores Negativos	5
	Indicadores positivos	3
Salud y hábitos de salud	Estado general de salud	42
	Hábitos de Salud	
	Ocio	
	Consumo de SPA	
	Alimentación	
	Sexualidad	
	Dependencia a tecnología	
Calidad de Vida	Bienestar	5
Factores individuales	Sociodemográficos	20
	Psicosociales	
Condiciones de vida	Condiciones de la vivienda y del barrio	26
	Movilidad	
	Situación económica del grupo familiar	
Determinantes estructurales	Apropiación de la vida universitaria	18
	Condiciones para la alimentación en la universidad	

El conjunto de datos contenía 124 variables en total, luego del tratamiento de variables siguiendo Agregación por Dimensiones para simplificar el número de datos, se redujo a 38 variables así:

Tabla 9. Nueva distribución de Variables, dimensiones y No. de ítems. Fuente: propia

Variables	Dimensiones	No de ítems
Salud Mental	Indicadores Negativos	1
	Indicadores positivos	1
Salud y hábitos de salud	Estado general de salud	10
	Hábitos de Salud	
	Ocio	
	Consumo de SPA	
	Alimentación	
	Sexualidad	
	Dependencia a tecnología	

Calidad de Vida	Bienestar	1
Factores individuales	Sociodemográficos	17
	Psicosociales	
Condiciones de vida	Condiciones de la vivienda y del barrio	6
	Movilidad	
	Situación económica del grupo familiar	
Determinantes estructurales	Apropiación de la vida universitaria	2
	Condiciones para la alimentación en la universidad	

En el proceso de entendimiento de datos, se identificaron variables que estaban completamente vacías o con una cantidad insignificante de valores disponibles, otra en particular no contenía información relevante para la investigación, por lo tanto, debido a su falta de información, estas variables no contribuían al cumplimiento de los objetivos de la investigación, ya que no proporcionaban valor añadido ni permitieron obtener patrones significativos para el análisis. Además, su presencia podría distorsionar los resultados al incrementar la dimensionalidad sin aportar información relevante.

Por lo tanto, la decisión de eliminar estas variables se basó en criterios de calidad de datos, con el fin de mejorar la eficiencia y precisión del análisis, enfocando los esfuerzos en variables que sí tienen datos significativos y que pueden influir en la comprensión de las relaciones estudiadas.

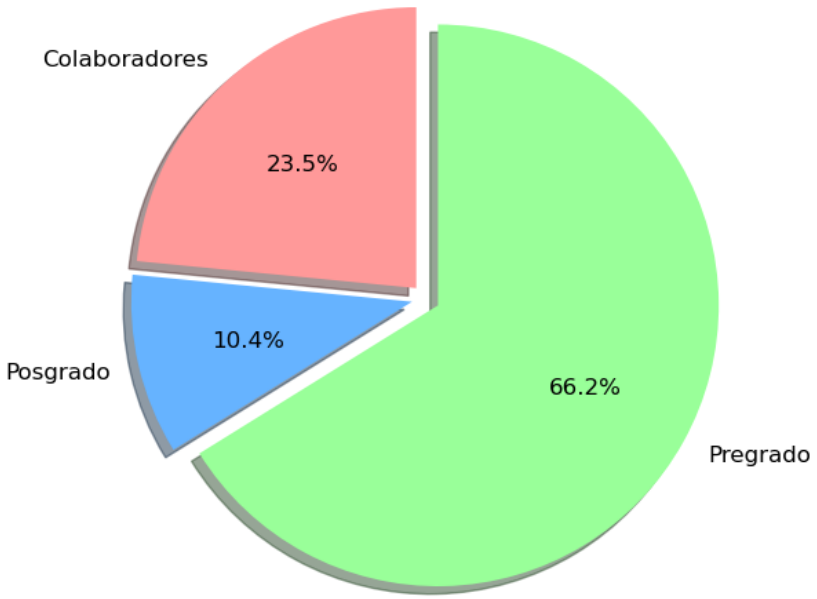
Se eliminaron las variables: 'CODIGO', 'SPAILEGALES', 'ESTPROGRAMAPRE', 'ESTPROGRAMAPOS', 'ESTSEMESTRE', 'ESTBECA' Y 'PROFESORTIPO'

Tabla 10. Distribución de población objeto de estudio. Fuente: propia

Población	**Muestra**	**Porcentaje**
Estudiantes de Pregrado	2786	0.67
Estudiantes de Posgrado	437	0.1
Colaboradores	988	0.23
Total	**4211**	**1**

El gráfico 1 visualiza la distribución, siendo los estudiantes de pregrado quienes ocupan la mayor porción en la distribución de los registros del conjunto de datos.

Gráfico 1 *Distribución de población objeto de estudio.* Fuente. *propia*



7.3 ANALISIS EXPLORATORIO DE DATOS

Posterior a la limpieza de datos, se exportó uno nuevo conjunto de datos con información de calidad, para posteriormente conocer la distribución de los datos previo a la segmentación, se realiza la distribución teniendo en cuenta la estructura de dimensiones propuestas por el *Proyecto Salud y Bienestar en la Comunidad Educativa Javeriana.* En el Anexo 1 de este estudio, se pueden observar las distribuciones por dimensiones del conjunto de datos, el EDA por población se encuentra en un repositorio en Github https://github.com/NiniAlejandra/Proyecto_Salud_Mental_NAVM

Gráfico 2. Matriz de correlaciones - Colaboradores

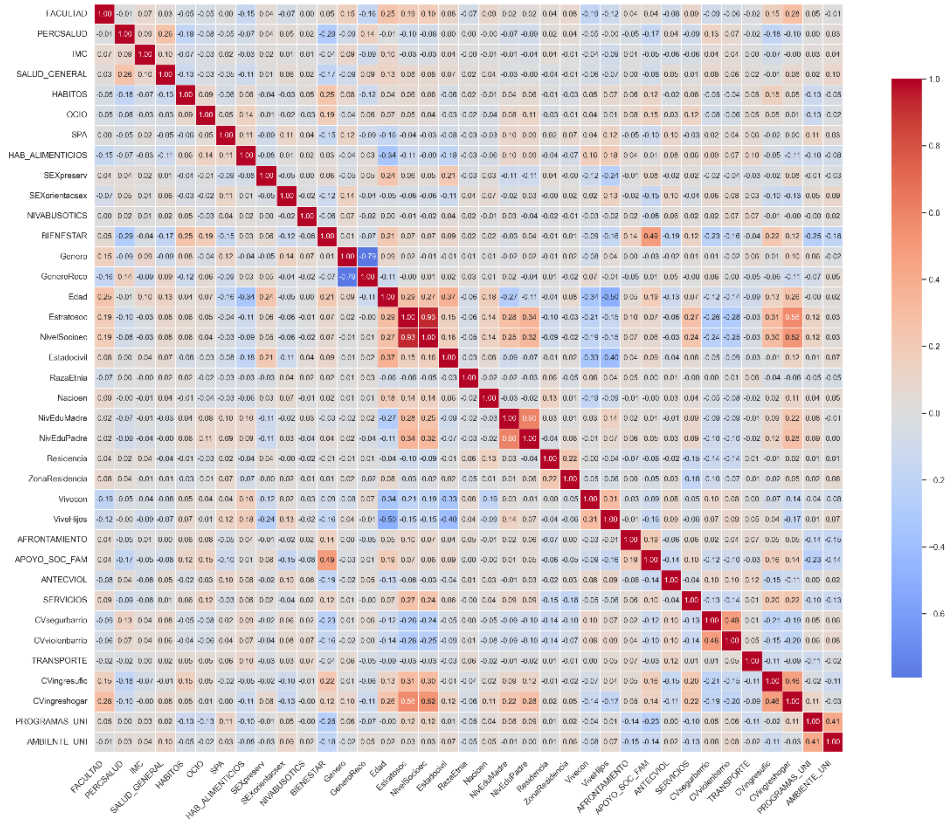


Gráfico 3. Matriz de correlaciones – Estudiantes de Posgrado

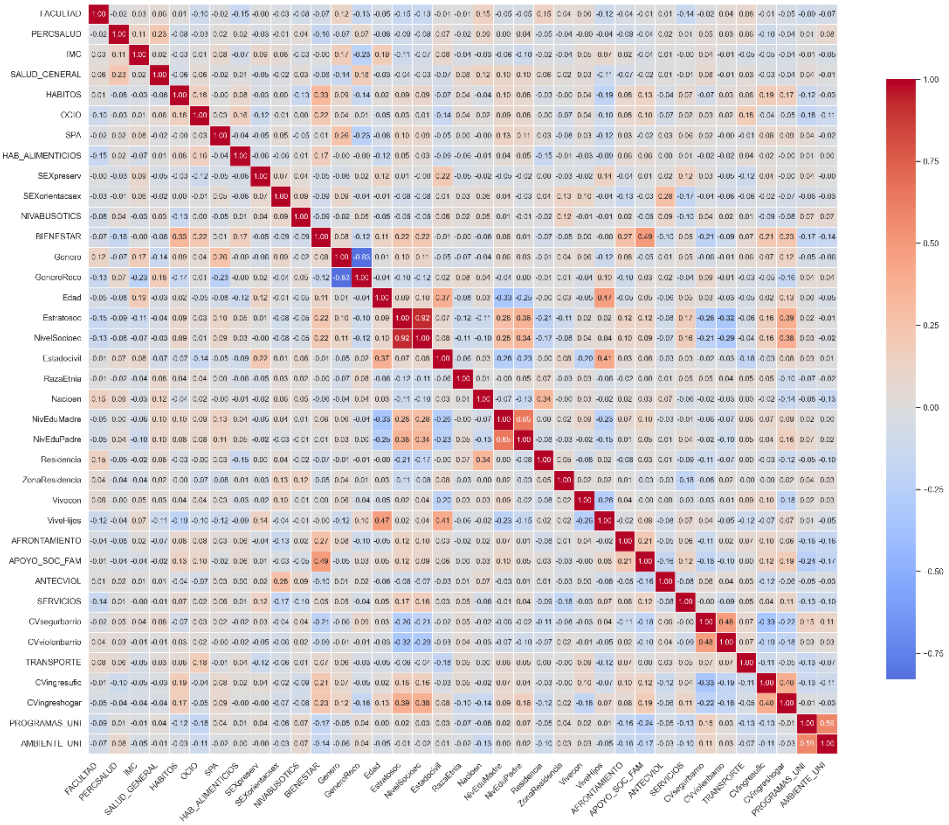
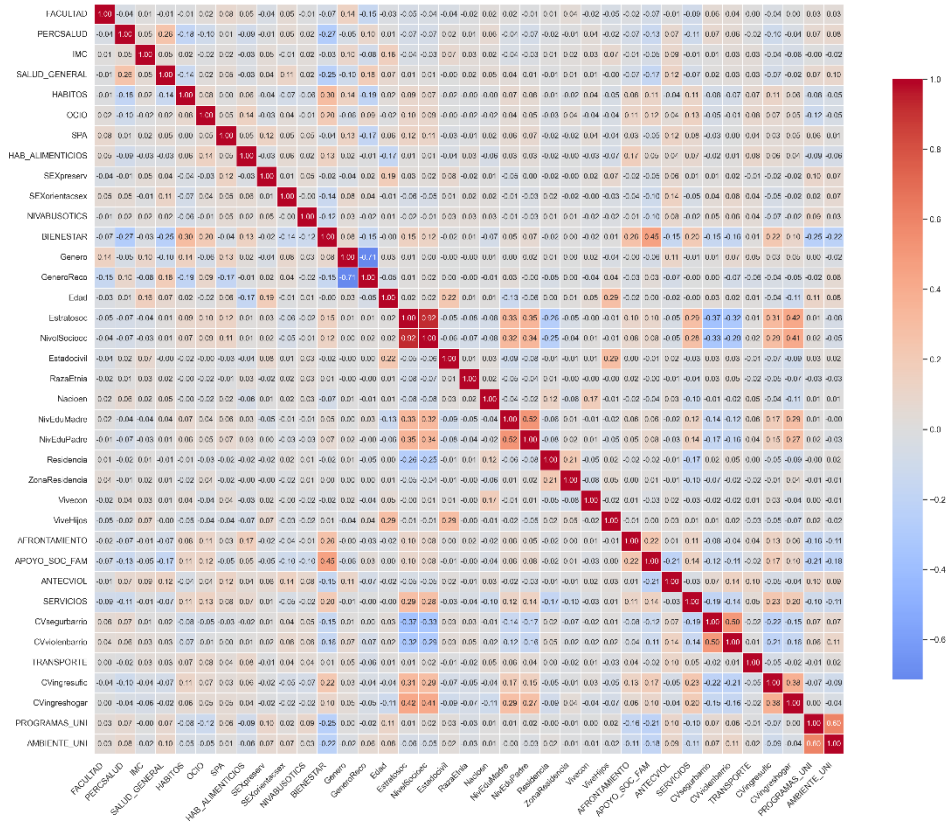


Gráfico 4. Matriz de correlaciones – Estudiantes de Pregrado



7.4 SEGMENTACIÓN APRENDIZAJE NO SUPERVISADO

En Visual Studio Code, se creó directorio de trabajo para cargar conjunto de datos una vez limpio y con las nuevas variables creadas producto del tratamiento a las variables resultado ver tabla 3 y tabla 4.

7.4.1 Aplicación de la Prueba KMO para cada muestra:

Tabla 11. Distribución de los resultados Prueba KMO para cada muestra. Fuente: propia

Rango	Colaboradores	Pregrado	Posgrado
Excelente	0%	0%	0%
Bueno	13.8%	13.9%	0%
Aceptable	36.1%	36.1%	19.5%
Mediocre	33.3%	36.1%	41.6%
Malo	13.8%	13.9%	33.3%
Inaceptable	2.7%	0%	5.6%

El índice KMO (Kaiser-Meyer-Olkin), aunque tradicionalmente utilizado para evaluar la adecuación de datos para análisis factorial, puede ofrecer información preliminar sobre la viabilidad de aplicar técnicas de reducción de dimensionalidad como el Análisis de Componentes Principales (PCA). Sin embargo, es importante destacar que, a diferencia del análisis factorial, el PCA no requiere necesariamente altos valores de KMO, ya que su objetivo es transformar las variables originales en componentes ortogonales que capturen la máxima varianza, independientemente de la estructura de correlación subyacente. [28]

Los resultados del KMO en el presente estudio revelaron que: en los Colaboradores, 50% de las variables en rangos bueno/aceptable (KMO 0.7-0.89), 50% en mediocre/inaceptable (KMO <0.7), para Posgrado, 80% de las variables en rangos mediocre a inaceptable (KMO <0.7), y para Pregrado, distribución similar a colaboradores (50% bueno/aceptable, 50% mediocre/inaceptable).

Estos resultados sugieren que la estructura de correlación entre variables no es óptima para un análisis factorial tradicional, los umbrales clásicos para interpretar el índice KMO (0.50: inaceptable; 0.60-0.70: mediocre; >0.80: bueno) fueron establecidos por Kaiser y son ampliamente recogidos en la literatura metodológica y han sido consistentemente replicados en estudios psicométricos recientes [29] donde se recomienda un KMO global >0.6. No obstante, para el PCA -técnica seleccionada en este estudio este resultado no está critico debido a que:

i. El PCA es una técnica descriptiva que no asume una estructura latente subyacente
ii. Puede aplicarse exitosamente incluso con correlaciones moderadas entre variables
iii. Su objetivo es maximizar la varianza explicada, no identificar factores latentes

7.4.2 Aplicación de la prueba VIF para cada muestra:

Posterior a la aplicación del KMO y sus resultados, se revisaron aspectos del subconjunto de datos y se aplicó el VIF (Variance Inflation Factor) para evaluar el aumento de la varianza de los coeficientes de regresión [20]

Tabla 12. Distribución de los resultados Prueba VIF para cada Muestra. Fuente: propia

Rango	Colaboradores	Pregrado	Posgrado
Ideal	0%	0%	0%
Aceptable	16.7%	19.4%	16.7%
Moderada	13.9%	19.4%	16.7%
Alta	69.4%	61.2%	66.7%

Se encontró en todas las muestras que más del 80% de los datos presentaron un valor de VIF superior a 5, lo que indica una posible multicolinealidad en cada muestra, la mayoría de las variables presentan multicolinealidad "Alta", solo un pequeño porcentaje de variables muestra multicolinealidad "Aceptable" o "Moderada".

7.4.2.1 KMO y VIF para Ciencias Sociales:

Los resultados de las pruebas de adecuación muestral (KMO < 0.60 en la mayoría de las dimensiones) y multicolinealidad (VIF > 5 en múltiples variables) indican que: (i) las variables no comparten suficiente varianza común para un análisis factorial tradicional, y (ii) existen correlaciones significativas entre predictores. Sin embargo, lejos de constituir una limitación metodológica, estos hallazgos reflejan la naturaleza intrínseca de los datos en salud mental y justifican plenamente el uso de técnicas de aprendizaje no supervisado propuestas en el objetivo 1 del estudio.

Esta interpretación se sustenta en que, existe *Multidimensionalidad inherente a los constructos evaluados,* el instrumento recogió información sobre dominios interconectados, pero conceptualmente distintos (hábitos de salud, ocio, factores psicosociales), lo que explica el bajo KMO. Como señala Watkins [29]"*las escalas Likert en salud mental frecuentemente violan los supuestos de normalidad y linealidad, afectando los coeficientes de correlación y la adecuación factorial*" (p. 145) [21]. Este fenómeno es particularmente relevante cuando, como en el presente estudio, se evalúan simultáneamente indicadores positivos y negativos en salud mental [30].

Adicionalmente, es importante resaltar la *Interdependencia real entre variables,* pues los altos valores de VIF no indican redundancia estadística, sino la estructura compleja de la salud mental. La teoría de redes psicopatológicas Borsboom postula que *"los síntomas mentales están causalmente interconectados a través de mecanismos biopsicosociales, formando sistemas no lineales"* [31], este marco explica por qué variables

como 'NIVDEP' (Depresión) y 'Bienestar' -aunque correlacionadas- capturan dimensiones distintas pero relacionadas del fenómeno estudiado.

Para este estudio, las técnicas no supervisadas seleccionadas, se propuso (K-means, PCA, DBSCAN y análisis de correspondencia), contrario a los modelos factoriales, las técnicas propuestas no requieren supuestos estrictos de independencia lineal. Autores como Jolliffe y Cadima [28], expresaron que el PCA es particularmente robusto para manejar multicolinealidad en datos de ciencias sociales, ya que "transforma variables correlacionadas en componentes ortogonales sin perder información estructural" [24].

"En conjunto, estos resultados no constituyen limitaciones metodológicas, sino evidencias empíricas de la complejidad teórica que subyace a los determinantes de salud mental en poblaciones universitaria" (Van Bork et al., 2019, p. 8) [26]. Este enfoque coincide con la tendencia actual en psicometría, donde se reconoce que "*los instrumentos transdiagnósticos frecuentemente presentan baja adecuación factorial sin que esto comprometa su validez ecológica*" [32]

7.4.3 Transformación de variables para el análisis:

Dado que las variables creadas en la Agregación por Dimensiones (ver tabla 5) se obtuvieron sumando diferentes cantidades de ítems tipo Likert, los rangos de las puntuaciones finales varían entre dimensiones. Para evitar que las variables con mayor rango numérico dominen el análisis multivariado, se aplicó una estandarización tipo Z-score (media 0, desviación estándar 1) a todas las variables de exposición antes de realizar los análisis de componentes principales (PCA) y clustering. Esta transformación es recomendada en la literatura para variables de naturaleza ordinal sumadas, tal como recomiendan Jolliffe y Cadima (2016) en [24] para análisis de componentes principales y técnicas de agrupamiento ya que permite centrar y escalar los datos, conservando la forma de la distribución y facilitando la interpretación de los resultados en términos relativos. Además, el uso de Z-score es especialmente adecuado para técnicas basadas en distancias, como Kmeans y PCA, ya que garantiza que todas las variables contribuyan equitativamente al análisis, independientemente de su escala originalReducción de dimensionalidad con PCA y segmentación mediante Kmeans

Se aplicó Análisis de Componentes Principales (PCA) al conjunto de datos original por cada muestra de

población de ***Colaboradores, Posgrado y Pregrado***, identificando que se requieren 13 componentes para explicar el 60% de la varianza acumulada en la muestra de Colaboradores, 14 para Pregrado y 13 para Posgrado. Posteriormente, el algoritmo Kmeans se aplicó sobre los primeros dos componentes principales para permitir visualización bidimensional, a continuación, se presentan los resultados de la visualización con sus métricas respectivamente.

Colaboradores

Gráfico 5. Visualización 2D de Clústers (Kmeans) en espacio PCA- Colaboradores

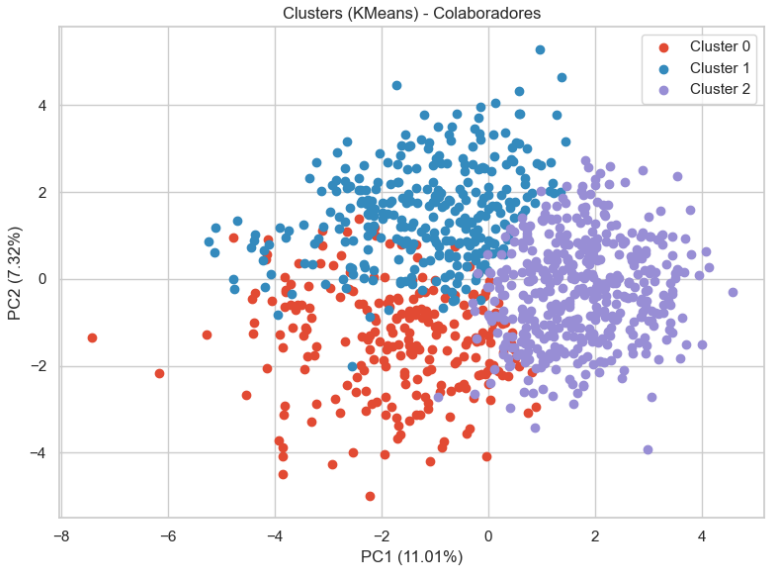


Tabla 13. Métricas de Evaluación Interna para el Agrupamiento Kmeans en Colaboradores. Fuente: propia

Métrica	Resultado	Interpretación
Inertia	31602.466	Muy alta, los puntos están lejos de sus centroides clústers poco compactos.
Silhouette	0.0693	Cercano a 0, los clústers están muy superpuestos no hay separación clara.
Calinski_harabasz	61.799	Relativo, lo que sugiere que existe alguna estructura de agrupamiento discernible en los datos.
Davies_bouldin	3.417	Alto, los clústers están muy cerca entre sí o son de formas irregulares.

Los ejes PC1 y PC2 capturan el 18.26% de la varianza explicada no siendo suficiente información para representar clústers reales, adicionalmente, los resultados de las métricas son pobres, Silhouette bajo y Davies-Bouldin alto sugieren que los clústers son artificiales, se identificó 3 clústers con métricas de calidad bajas (Silhouette=0.069, Davies-Bouldin=3.42), indicando superposición entre grupos. Esto sugiere que la

estructura de los datos en 2D no refleja agrupaciones naturales, sino artefactos de la reducción dimensional.

Estudiantes de Posgrado

Gráfico 6. Visualización 2D de Clústers (Kmeans) en espacio PCA-Estudiantes Posgrado

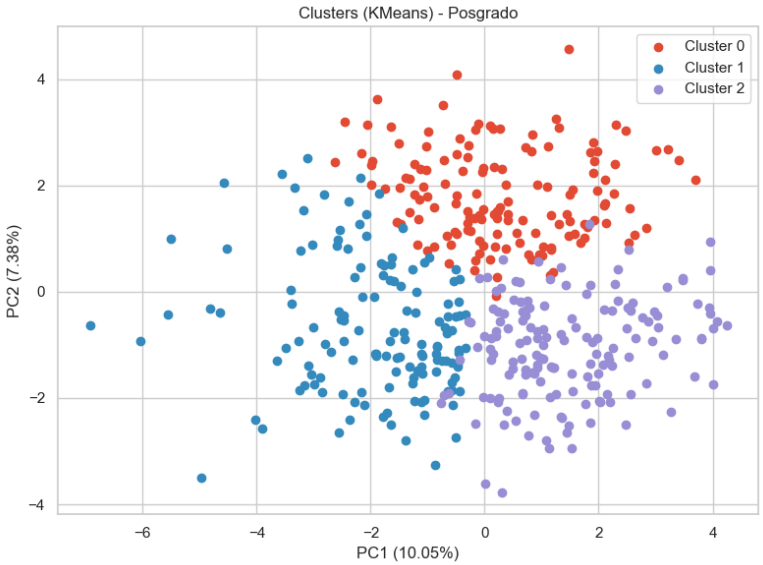


Tabla 14. Métricas de Evaluación Interna para el Agrupamiento Kmeans en estudiantes Posgrado. Fuente: propia

Métrica	Resultado	Interpretación
Inertia	14074.393	Alta (aunque menor que en colaboradores), indica clústers poco compactos.
Silhouette	0.049	Más bajo que en colaboradores (0.069), mayor superposición entre clústers.
Calinski_harabasz	25.557	Muy bajo vs colaboradores (61.80), peor separación entre clústers.
Davies_bouldin	3.315	Similar a colaboradores (3.42), clústers mal definidos y cercanos entre sí.

De conformidad a los resultados de las métricas es posible inferir que los clústers son artificiales y no reflejan agrupaciones reales. En posgrado, Kmeans mostró peores métricas (Silhouette=0.049, Calinski-Harabasz=25.56) que, en colaboradores, confirmando mayor homogeneidad. La inercia (14,074.39) sugiere que los clústers son aún menos compactos, mientras que el Davies-Bouldin (3.32) indica solapamiento severo; en conjunto, las métricas de evaluación interna para los datos de posgrado sugieren que el agrupamiento Kmeans ha identificado una estructura de clústeres, pero su calidad en términos de separación y cohesión no es robusta. El bajo coeficiente de silueta y el alto índice de Davies-Bouldin indican que los clústeres pueden ser poco distintivos, con un grado significativo de superposición entre ellos.

Estudiantes de Pregrado

Gráfico 7. Visualización 2D de Clústers (Kmeans) en espacio PCA- estudiantes Pregrado

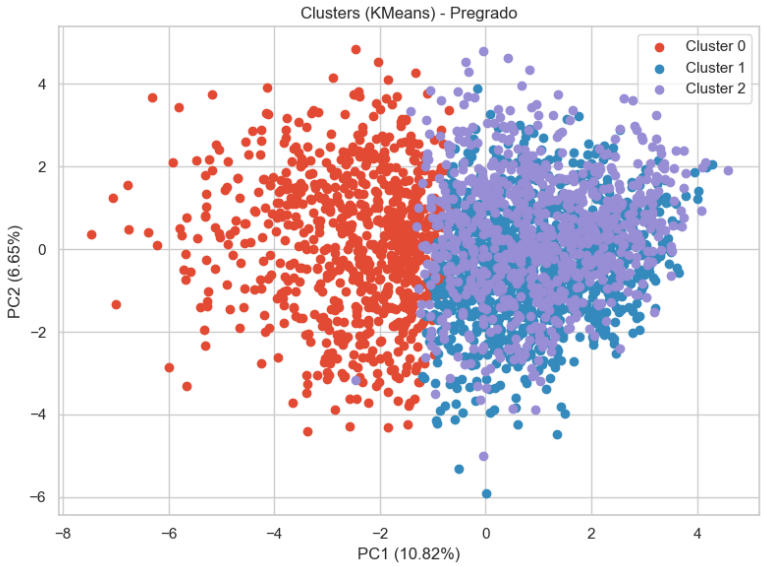


Tabla 15. Métricas de Evaluación Interna para el Agrupamiento Kmeans en estudiantes Pregrado. Fuente: propia

Métrica	Resultado	Interpretación
Inertia	89553.453	Extremadamente alta, reflejando una gran dispersión de los puntos dentro de los clústers
Silhouette	0.046	Muy bajo (cercano a 0), indicando superposición casi total entre clusters
Calinski_harabasz	166.919	Valor relativamente alto comparado con otras poblaciones, pero engañoso por el tamaño muestral, pues Pregrado cuenta con el 66,2% de la muestra.
Davies_bouldin	3.499	Alto, mostrando mala separación entre clústers

Se observa además que, la inercia extremadamente alta (89k vs 31k en colaboradores y 14k en posgrado) se debe principalmente al mayor tamaño muestral de pregrado, sin embargo, el Silhouette consistentemente bajo (0.046) confirma que los clústers no tienen estructura interna clara. El muy bajo coeficiente de silueta y el alto índice de Davies-Bouldin indican una notable falta de separación y cohesión en los clústeres. La alta inercia refuerza la idea de clústeres difusos. Aunque el índice de Calinski-Harabasz es el más alto, lo que podría sugerir cierta estructura entre clústeres, las otras métricas predominan en indicar un agrupamiento con un alto grado de solapamiento y poca distinción clara entre los grupos.

Los análisis de clustering Kmeans con PCA aplicados a las tres muestras de las poblaciones de

Colaboradores, Posgrado y Pregrado mostraron resultados consistentemente débiles en todas las métricas de evaluación, el Silhouette Score cercano a 0 (entre 0.046–0.069), indica que los clústers están superpuestos y no hay separación clara entre grupos. Adicionalmente, la Inercia elevada (desde 14k hasta 89k), refleja alta dispersión de los datos dentro de cada clúster; el Davies-Bouldin alto (> 3.3 en todos los casos), confirma que los clústers están mal definidos y son poco diferenciables.Esto sugiere que no existen agrupaciones naturales basadas en los determinantes sociales y variables de salud mental analizadas.

7.4.4 Reducción de dimensionalidad con UMAP y t-SNE-Segmentación con Kmeans, Agglomerativo y GMM:

Dado que PCA es una técnica lineal que solo captura relaciones lineales entre variables, es posible que considerando la naturaleza compleja y multidimensional de los datos de salud mental recolectados en la encuesta no haya logrado separar adecuadamente los clústers, incluso cuando estos podrían existir en el espacio original. Se usó **UMAP y t-SNE** como técnicas de reducción no lineal, las cuales están específicamente diseñadas para visualizar estructuras complejas en espacios de baja dimensionalidad (2D o 3D).

Su aplicación al objetivo específico 1 de este estudio el cual es, *segmentar la población universitaria mediante aprendizaje no supervisado para identificar patrones asociados a determinantes de salud mental* permite, evidenciar si los clusters están solapados o bien diferenciados, además de identificar estructuras no lineales que PCA no detecta; como apunte metodológico, aunque las métricas de evaluación (Inercia, Calinski-Harabasz, Silueta y Davies-Bouldin) no fueron favorables en la segmentación con PCA y Kmeans, es importante señalar que UMAP/t-SNE no mejoran directamente la calidad del clustering, su rol es exploratorio y visual.

Se aplicaron otros algoritmos de clustering adicionales a ***Kmeans, Agglomerative y GMM*** combinados con las técnicas de reducción de dimensionalidad UMAP y t-SNE en cada muestra de Colaboradores, Posgrado y Pregrado.

Colaboradores

Figura 5. Segmentación con Kmeans, Agglomerative y GGM con proyección UMAP Y t-SNE en Colaboradores

La Figura 5, presenta una comparación visual de los resultados de agrupamiento obtenidos mediante tres algoritmos distintos: ***Kmeans, Agglomerativo y Modelos de Mezcla Gaussiana (GMM).*** Como se mencionó líneas arriba, para facilitar la visualización en dos dimensiones, los datos fueron reducidos utilizando dos técnicas de reducción de dimensionalidad no lineal, UMAP (Uniform Manifold Approximation and Projection), mostrada en la columna izquierda, y t-SNE (t-Distributed Stochastic Neighbor Embedding), en la columna derecha, cada punto representa un colaborador y su color indica el clúster al que fue asignado.

En todas las combinaciones de métodos, se observa que los clústers identificados por los algoritmos de segmentación no corresponden a agrupamientos naturales bien definidos. Tanto en UMAP como en t-SNE, los puntos de ambos clústers aparecen ampliamente solapados y dispersos a lo largo del espacio, sin fronteras claras ni conglomerados compactos exclusivos de un solo grupo. Aunque existen pequeñas regiones donde predomina un color, la mezcla generalizada de los clústers indica que la estructura subyacente de los datos es difusa y continua.

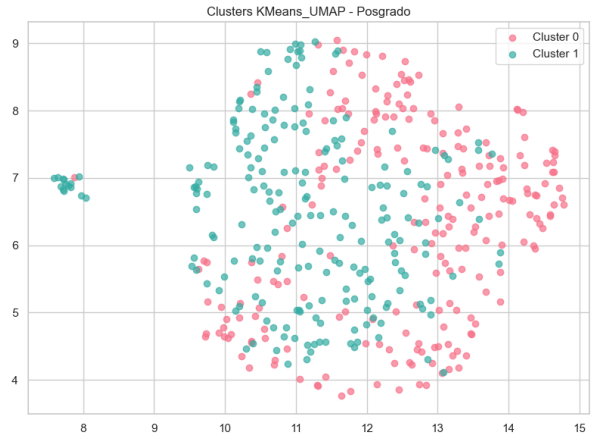
Este patrón se repite de manera consistente en los tres algoritmos evaluados (Kmeans, Agglomerativo y GMM), lo que refuerza la conclusión de que los determinantes sociales y variables de salud mental analizadas en la población de colaboradores no presentan segmentaciones discretas robustas. Las visualizaciones confirman los resultados de las métricas cuantitativas de calidad de clustering, que también fueron desfavorables, y justifican la necesidad de recurrir a enfoques descriptivos complementarios para explorar la heterogeneidad de la población.

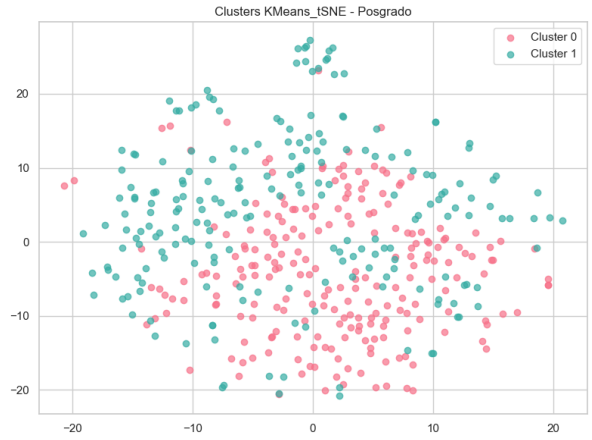
La evidencia visual derivada de estas proyecciones bidimensionales es coherente con los resultados de las métricas cuantitativas de calidad de clustering previamente discutidas (como los bajos coeficientes de silueta y los altos índices de Davies-Bouldin obtenidos para los datos de colaboradores, que indicaban una calidad subóptima del agrupamiento). Estas visualizaciones reafirman que los determinantes sociales y las variables de salud mental analizadas en la muestra de Colaboradores no presentan segmentaciones discretas y robustas.

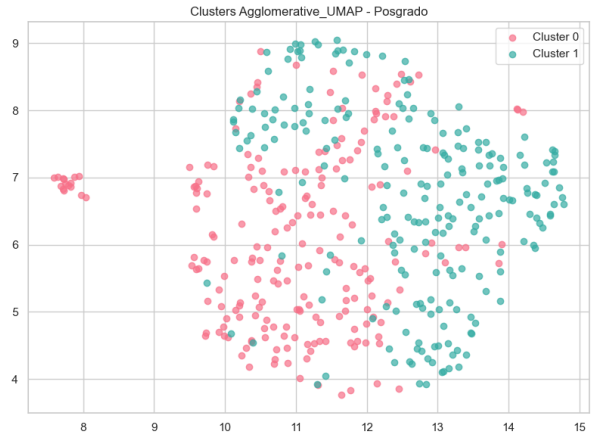
En cambio, la estructura subyacente de los datos parece ser difusa, continua y multidimensional, sin divisiones naturales claras que permitan una categorización estricta de la población en grupos bien diferenciados.

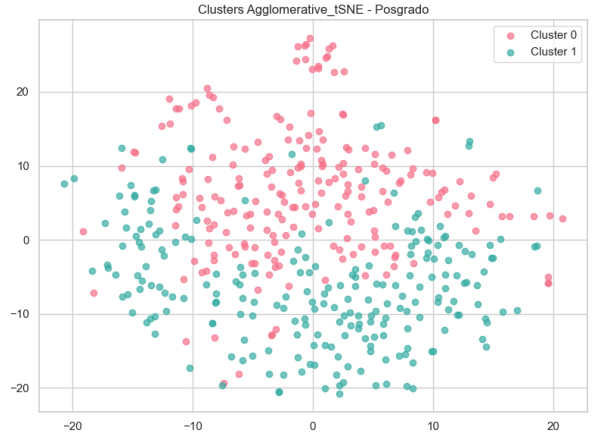
Estudiantes de Posgrado

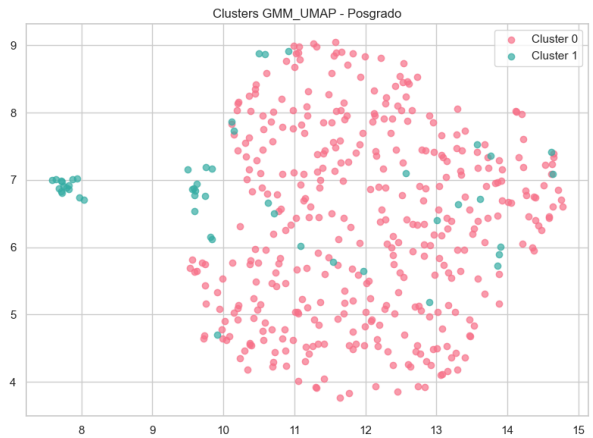
Figura 6. Segmentación con Kmeans, Agglomerativo y GGM con proyección UMAP Y t-SNE en estudiantes Posgrado

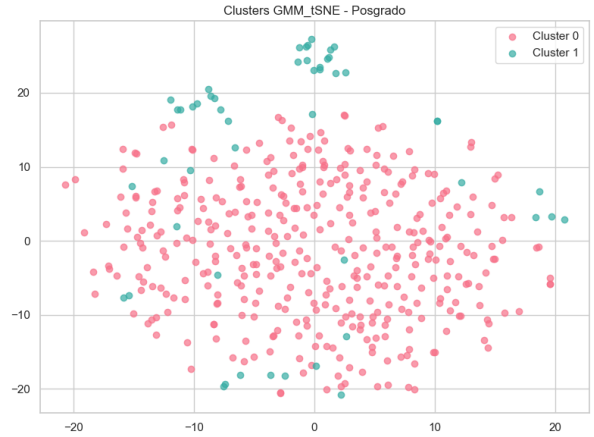












La Figura 6, de igual forma que en Colaboradores, presenta la comparación visual de los resultados de agrupamiento obtenidos mediante tres algoritmos distintos, Kmeans, Agglomerativo y GMM, un examen detallado de las visualizaciones en revela un patrón consistente en todas las configuraciones de algoritmos de agrupamiento y técnicas de reducción de dimensionalidad aplicadas a la muestra de Posgrado, en todas las combinaciones de métodos, los clústeres identificados exhiben un alto grado de solapamiento y una marcada dispersión. Los puntos correspondientes a los dos clústeres se distribuyen de manera entremezclada a lo largo de todo el espacio proyectado. No se observan agrupamientos compactos ni fronteras claras y discernibles entre los grupos. Aunque en algunas regiones puede haber una ligera predominancia de un color sobre otro, la tendencia general es la coexistencia y mezcla de ambos clústeres en la mayoría de las áreas del gráfico. Esto sugiere una baja cohesión interna de los clústeres y una limitada separación entre ellos.

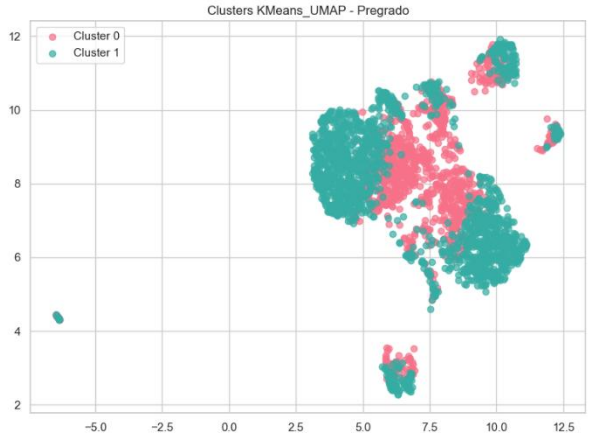
Se observa un patrón de difusión y falta de separación, la cual se mantiene de forma consistente en los tres algoritmos de agrupamiento evaluados, lo que refuerza la conclusión de que la estructura subyacente de los datos de Posgrado es difusa y no susceptible de ser segmentada en grupos discretos bien definidos utilizando estos enfoques. La concordancia entre UMAP y t-SNE también valida esta observación, dado que ambas técnicas se enfocan en preservar diferentes aspectos de la estructura de proximidad de los datos.

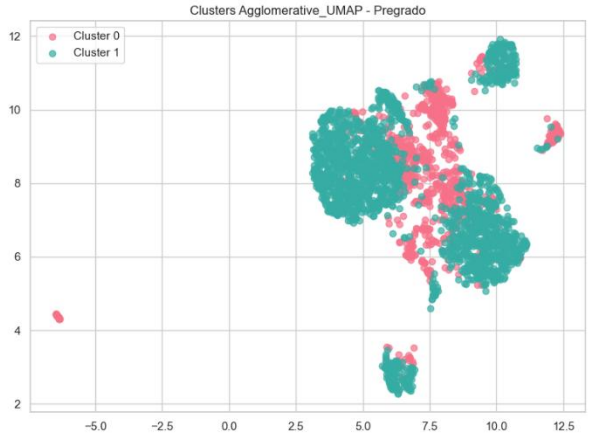
Estas visualizaciones refuerzan de manera significativa los resultados de las métricas cuantitativas de calidad de clustering previamente obtenidas para la muestra de posgrado, pues líneas arriba se logró identificar un bajo coeficiente de silueta de 0.049 y un alto índice de Davies-Bouldin de 3.32. Dichas métricas indicaron una calidad subóptima del agrupamiento, lo cual es plenamente confirmado por la inspección visual que muestra la ausencia de clústeres naturales robustos y bien diferenciados.

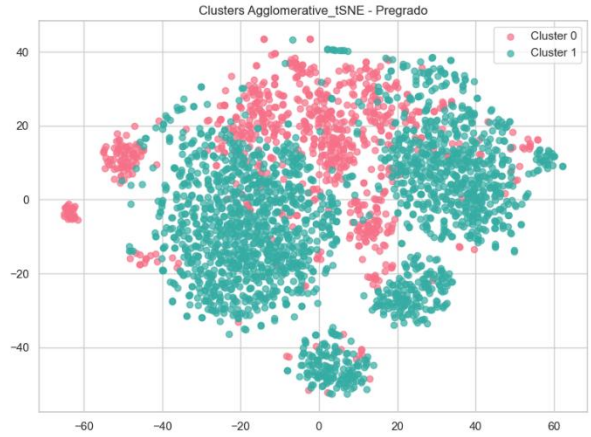
En síntesis, los resultados visuales y cuantitativos convergen para sugerir que, en la muestra de Posgrado, los determinantes sociales y las variables de salud mental analizadas no presentan agrupamientos naturales discretos. La mezcla y dispersión de los clústeres en todas las proyecciones confirman la ausencia de divisiones claras, lo que implica que la población de posgrado es inherentemente heterogénea pero no segmentable en grupos discretos a partir de las variables consideradas.

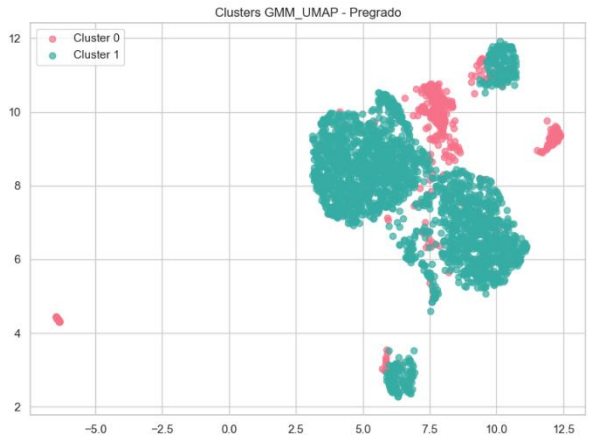
Estudiantes de Pregrado

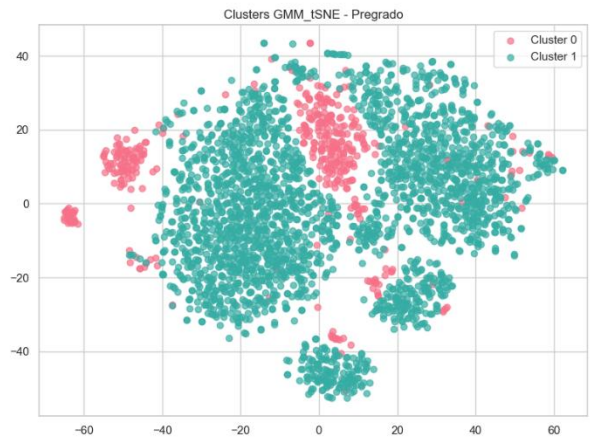
Figura 7. Segmentación con Kmeans, Agglomerativo y GGM con proyección UMAP Y t-SNE en estudiantes Pregrado











Continuando con el análisis e interpretación ahora en la Pregrado, la Figura 7, revela un patrón consistente en todas las configuraciones de algoritmos de agrupamiento y técnicas de reducción de dimensionalidad aplicadas a la población de pregrado, en donde se puede inferir alta superposición y dispersión generalizada, pues en todas las combinaciones de métodos, los clústeres identificados (representados por los colores rosado y verde) presentan un alto grado de solapamiento. Los puntos correspondientes a los dos clústeres se distribuyen de manera ampliamente mezclada a lo largo de la mayor parte del espacio proyectado.

Adicionalmente, no se observan conglomerados compactos ni fronteras bien definidas y nítidas entre los grupos. Si bien en algunas áreas del gráfico se percibe una ligera tendencia a la formación de agrupaciones locales o una predominancia de un color, la característica dominante es la coexistencia significativa y la mezcla de ambos clústeres en la mayoría de las regiones, lo que indica una baja cohesión interna y una separación limitada entre los grupos. Se presenta también un patrón de falta de separación y cohesión se mantiene de forma consistente en los tres algoritmos de agrupamiento, lo que refuerza la conclusión de que la estructura subyacente de los datos de pregrado es inherentemente difusa y no susceptible de ser segmentada en grupos discretos robustos mediante los enfoques de clustering aplicados.

Al igual que en Colaboradores y Posgrado, estas visualizaciones refuerzan de manera decisiva los resultados de las métricas cuantitativas de calidad de clustering previamente obtenidas para la muestra de pregrado. El muy bajo coeficiente de silueta (0.046), el alto índice de Davies-Bouldin (3.499), y la alta inercia (89553.45) indicaron una calidad subóptima del agrupamiento. Aunque el índice de Calinski-Harabasz (166.92) fue el más alto de los tres grupos muestrales, su interpretación en este contexto, junto con las otras métricas y la evidencia visual, sugiere que, si bien puede haber alguna varianza entre los centroides, la cohesión interna y la separación global de los clústeres son insuficientes para definir agrupamientos robustos.

En síntesis, los resultados visuales y cuantitativos convergen para indicar que, en la muestra de Pregrado, los determinantes sociales y las variables de salud mental analizadas no presentan agrupamientos naturales robustos. La mezcla y dispersión de los clústeres en todas las proyecciones confirman la ausencia de divisiones claras, lo que implica que la población de pregrado es heterogénea, pero no segmentable en grupos discretos y bien definidos a partir de las variables consideradas.

7.4.5 Análisis de métricas de Clustering:

Después de aplicar técnicas de reducción de dimensionalidad como PCA (lineal) y UMAP y t-SNE (no lineal) a los algoritmos Kmeans, Agglomerativo y GMM, se presentan las métricas utilizadas Silhouette, Calinski-Harabasz y Davies-Bouldin, las cuales permiten evaluar la calidad de los clústers generados. A continuación, se presenta una interpretación detallada de los resultados:

Tabla 16. Evaluación Comparativa de Técnicas de Clustering en de Colaboradores: Impacto del Balanceo de Datos con y sin SMOTE. Fuente: propia

Métrica	*Sin SMOTE*			*Con SMOTE*		
	Kmeans	**Agglomerativo**	**GMM**	**Kmeans**	**Agglomerativo**	**GMM**
No de clústers	3	2	2	10	10	10
Silhouette	0,08129059	0,038997571	0,13502719	0,12524475	0,122121404	0,10529546
Calinski harabasz	80,2864263	44,47178636	34,8973429	163,891767	148,7690935	130,14242
Davies Bouldin	3,43435297	4,546637689	4,00024865	2,5724654	2,203122203	2,95536828

El análisis comparativo de métricas de calidad de clústers revela que la aplicación de SMOTE produce cambios significativos pero ambivalentes en el rendimiento de los algoritmos de clustering. La técnica de sobremuestreo incrementa el número de clústers identificados de 2-3 a 10, generando particiones más granulares que, si bien mejoran las métricas cuantitativas, podrían representar estructuras artificiales; la homogeneidad inherente de los datos se mantiene, evidenciada por coeficientes de Silhouette inferiores a 0.2, lo que sugiere limitaciones en la capacidad discriminativa de las variables analizadas. En consecuencia, aunque SMOTE incrementa artificialmente la calidad aparente de los clústers, no proporciona evidencia convincente de patrones naturales subyacentes en los datos.

Tabla 17. Evaluación Comparativa de Técnicas de Clustering en estudiantes de Posgrado: Impacto del Balanceo de Datos con y sin SMOTE. Fuente: propia

Métrica	*Sin SMOTE*			*Con SMOTE*		
	Kmeans	**Agglomerativo**	**GMM**	**Kmeans**	**Agglomerativo**	**GMM**
No de clústers	2	**2**	**2**	10	**10**	**10**
Silhouette	0,06426386	0,036107658	0,18080257	0,21113125	0,200710599	0,19604427
Calinski harabasz	30,1495008	16,55028758	13,9205606	95,0882914	89,91115638	84,7994979
Davies Bouldin	3,7129556	5,0085737	3,97121188	2,25068225	1,956976238	2,42900696

En Posgrado, se produce patrones similares a los observados en colaboradores, evidenciando que SMOTE genera mejoras métricas significativas pero cuestionables desde una perspectiva interpretativa. La técnica de sobremuestreo incrementa dramáticamente el número de clústers identificados de 2 a 10, creando una estructura más compleja que, aunque estadísticamente superior, plantea interrogantes sobre su validez conceptual y utilidad práctica. GMM emerge como el algoritmo más competente en condiciones de desbalance, alcanzando un coeficiente de Silhouette de 0.181, lo que sugiere cierta capacidad inherente para identificar patrones en datos heterogéneos. Sin embargo, el rendimiento general permanece deficiente, con valores de Silhouette inferiores a 0.2 indicando agrupaciones superpuestas, índices Davies-Bouldin superiores a 3.7 señalando separación inadecuada entre clústers, y métricas Calinski-Harabasz menores a 30 reflejando compactación insuficiente.

Aunque SMOTE produce mejoras métricas notables, estas podrían representar optimizaciones artificiales que no corresponden a estructuras naturales en los datos estudiantiles.

Tabla 18. Evaluación Comparativa de Técnicas de Clustering en estudiantes de Pregrado: Impacto del Balanceo de Datos con y sin SMOTE. Fuente: propia

Métrica	Sin SMOTE			Con SMOTE		
	Kmeans	**Agglomerativo**	**GMM**	Kmeans	**Agglomerativo**	**GMM**
No de clústers	2	**2**	**2**	2	**2**	**2**
Silhouette	0,09322594	0,098698791	0,15742541	0,17521257	0,173864236	0,17313337
Calinski harabasz	213,202896	110,6874902	102,958005	718,007677	717,4734383	714,948283
Davies Bouldin	3,45061268	4,577296522	3,90314464	1,1891456	1,175780137	1,17326389

El análisis en estudiantes de Pregrado, los patrones distintivos que contrastan significativamente con los observados en otras Colaboradores y Posgrado. La característica más notable es la estabilidad estructural manifestada en el mantenimiento constante de dos clústers antes y después de la aplicación de SMOTE, sugiriendo una arquitectura subyacente más robusta y homogénea en esta población. Esta estabilidad podría atribuirse tanto a la mayor homogeneidad demográfica y académica de los estudiantes de pregrado como al efecto estabilizador del tamaño muestral, que representa el 66.2% del total de participantes.

La mejora consistente en Davies-Bouldin sugiere que SMOTE efectivamente contribuyó a separar centros de clústers y reducir solapamientos inter-grupos, logrando valores cercanos al rango aceptable. Sin embargo,

la persistencia de coeficientes de Silhouette en categorías bajas indica superposición significativa entre agrupaciones y fronteras difusas, señalando posibles limitaciones en las variables discriminativas empleadas.

7.4.6 Análisis de Correspondencia:

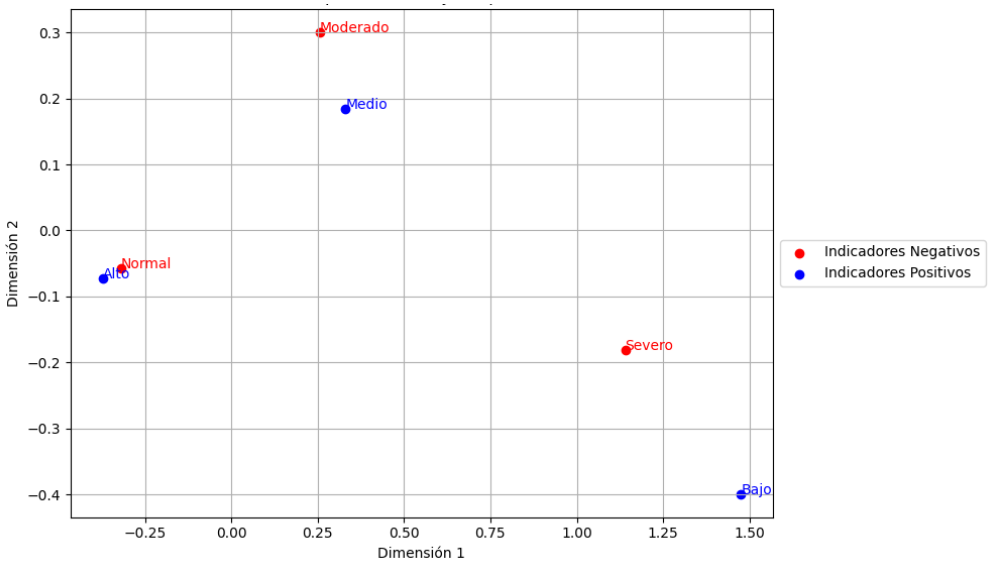
Ante la ausencia de clústers bien definidos y con métricas de calidad, la cual es coherente con la naturaleza compleja y multidimensional de los datos de salud mental y determinantes sociales, donde las diferencias entre individuos tienden a distribuirse de manera continua y no discreta, se avanza con el ***Análisis de Correspondencia*** como herramienta de segmentación descriptiva, a través de esta técnica fue posible visualizar y describir las asociaciones entre los diferentes niveles de Indicadores Positivos y Negativos en salud mental en cada muestra de Colaboradores, Posgrado y Pregrado.

El Análisis de Correspondencia permite explorar y visualizar las asociaciones entre categorías de variables nominales o discretizadas, en este caso, los niveles de Indicadores Positivos y Negativos de salud mental (ver tabla 4), proporcionando una representación gráfica de los perfiles y relaciones presentes en la población.

Esta técnica es especialmente útil cuando los métodos de clustering no logran identificar grupos naturales, ya que facilita la identificación de patrones de proximidad, oposición o gradientes entre categorías, enriqueciendo la comprensión de la heterogeneidad de la muestra desde una perspectiva descriptiva y relacional.

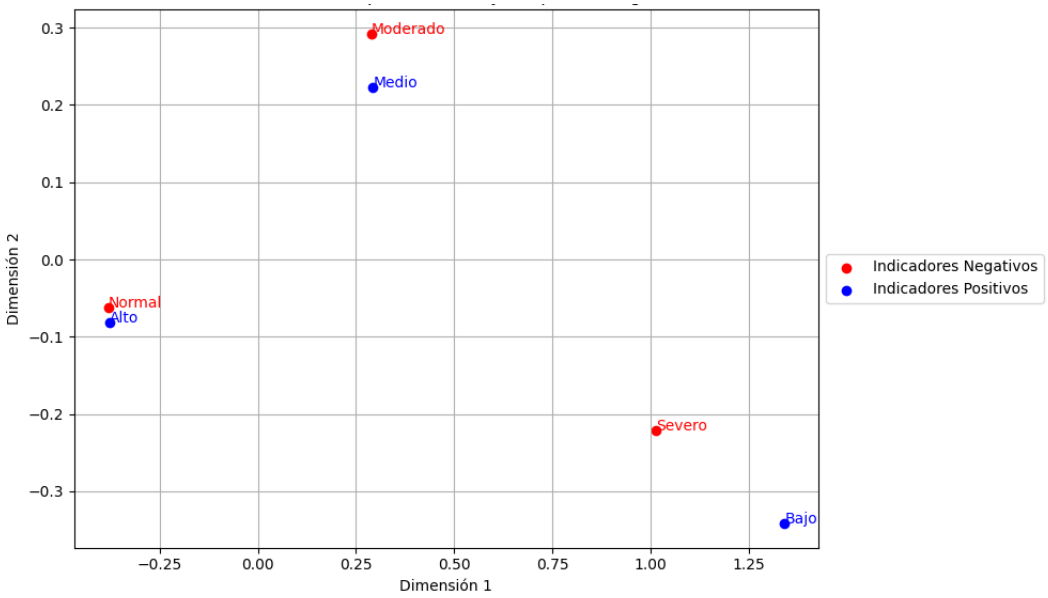
Los parámetros utilizados fue aplicar Clustering Jerárquico (AgglomerativeClustering), fueron, segmentación con 3 clústers, método de enlace por defecto (Ward) y con una estandarización previa de las variables mediante Z-score, el Clustering Jerárquico (agglomerative) permite explorar la estructura de los datos a diferentes niveles de agrupamiento, generando un dendrograma que facilita la visualización de relaciones jerárquicas entre observaciones. No requiere especificar el número de clústers a priori (aunque en este caso se fijó en 3 para comparación directa).

Gráfico 8. Análisis de Correspondencia- Colaboradores



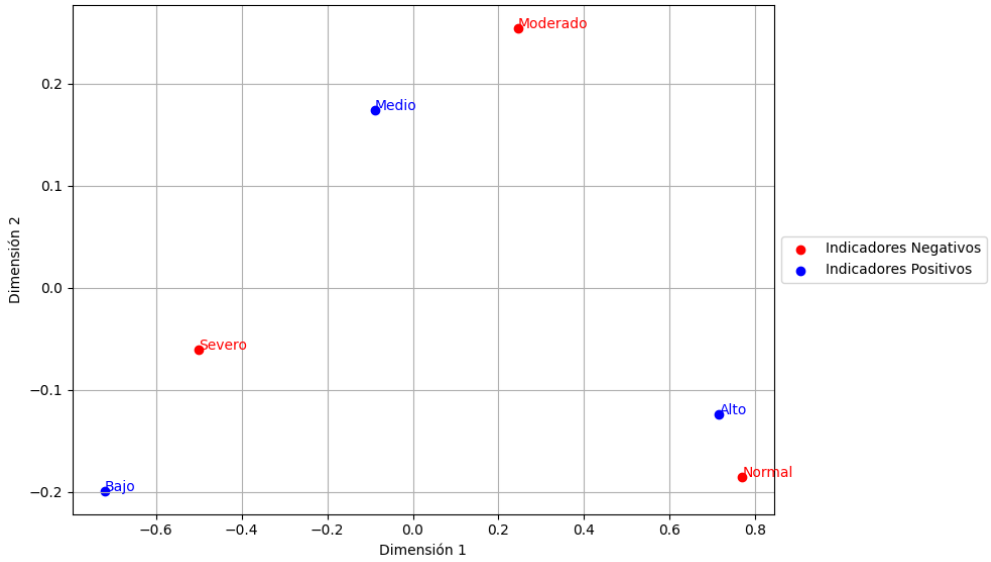
En la gráfica 8 de Colaboradores, se observa que las categorías “Alto” de los indicadores positivos (azul) y “Normal” de indicadores negativos (rojo) se encuentran próximas entre sí, lo que sugiere una asociación entre altos niveles de bienestar y la ausencia de indicadores negativos. Por otro lado, la categoría “Bajo” de los indicadores negativos se ubica en el extremo opuesto, próxima a “Severo” de los positivos, indicando que quienes reportan menor bienestar tienden a presentar mayor severidad en indicadores negativos. La categoría “Moderado” de los negativos y “Medio” de los positivos ocupan posiciones intermedias, reflejando perfiles mixtos o transicionales.

Gráfico 9. Análisis de Correspondencia- estudiantes Posgrado



El análisis de correspondencia en estudiantes de Posgrado reflejó un patrón similar que, en Colaboradores, “Normal” y “Alto” en negativos y positivos respectivamente se agrupan cerca, mientras que “Bajo” en positivos se encuentra alejado y próximo a “Severo” en negativos. “Moderado” y “Medio” se sitúan en posiciones intermedias, lo que sugiere la existencia de un gradiente de bienestar/malestar. Este resultado refuerza la idea de que, en esta población, los perfiles de salud mental tienden a distribuirse de manera continua, con asociaciones claras entre altos niveles de bienestar y bajos niveles de malestar, y viceversa.

Gráfico 10. Análisis de Correspondencia -estudiantes de Pregrado



En estudiantes de Pregrado, se observa una mayor dispersión, pero el patrón general se mantiene, “Normal” y “Alto” en positivos y negativos están próximos y alejados de “Bajo” en positivos, que se acerca a “Severo” en negativos. “Moderado” y “Medio” ocupan posiciones intermedias, lo que indica la presencia de perfiles mixtos. La mayor dispersión puede reflejar una mayor heterogeneidad en esta población, pero la relación entre extremos de bienestar y malestar sigue siendo evidente.

Sobre el análisis de correspondencia jerárquico realizados en las tres muestras permiten inferir un patrón consistente, en el que existe una clara asociación entre los niveles altos de indicadores positivos de salud mental y los niveles bajos de indicadores negativos, y viceversa. Las categorías intermedias (“Medio”, “Moderado”) reflejan perfiles mixtos o transicionales, lo que sugiere que la salud mental en la población universitaria se distribuye a lo largo de un continuo más que en grupos discretos. Estos resultados complementan los hallazgos de los análisis de Clustering, reforzando la idea de que no existen clústers naturales bien definidos, sino más bien gradientes o perfiles que se relacionan de manera continua. El análisis de correspondencia, por tanto, aporta una visión descriptiva y relacional útil para comprender la heterogeneidad y los patrones de bienestar y malestar en la población estudiada.

7.4.7 Gráficos Sankey:

Los gráficos de Sankey son especialmente útiles en este contexto, ya que muestran de manera clara cómo se distribuyen y conectan las distintas poblaciones a través de varios niveles, como el rol, el género y los indicadores negativos y positivos, variables de interés de este estudio.

Gráfico 11. Gráfico Sankey de Indicadores Negativos por Rol - Colaboradores, Estudiantes de Posgrado y Posgrado. Fuente: propia

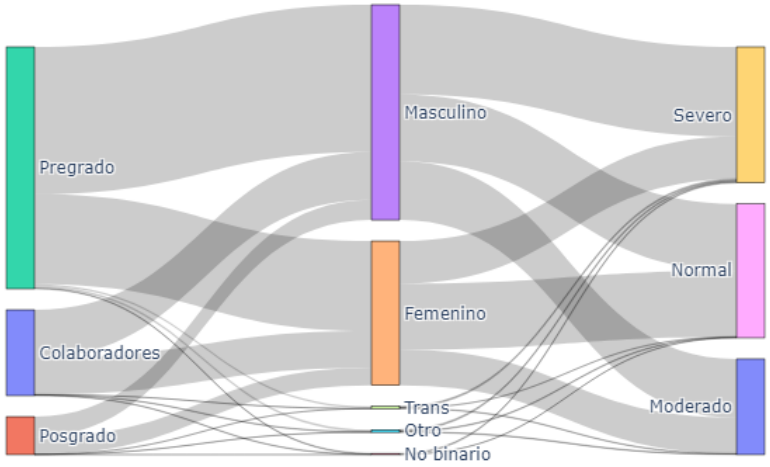
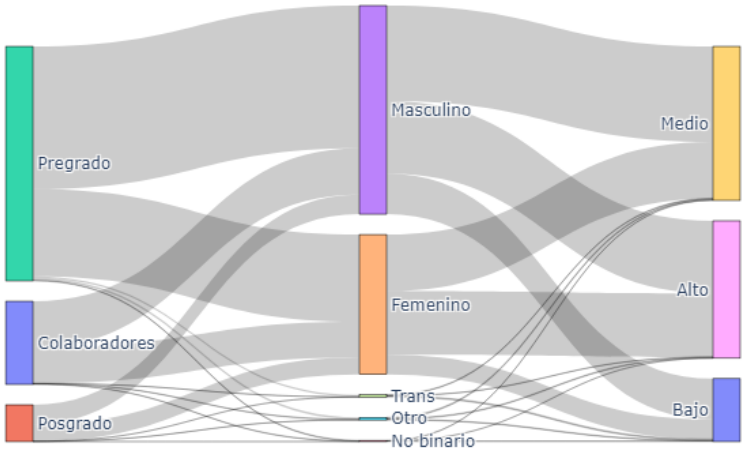


Gráfico 12. Gráfico Sankey de Indicadores Positivos por Rol - Colaboradores, Estudiantes de Posgrado y Posgrado. Fuente: propia



El uso de gráficos Sankey en este análisis permite representar de manera intuitiva y visual la complejidad de una población difícil de clusterizar con aprendizaje no supervisado. La visualización facilita la identificación

de patrones, la visibilidad de minorías y la comprensión de cómo se relacionan diferentes atributos entre sí, aportando valor tanto para el análisis descriptivo como para la comunicación de resultados a audiencias no técnicas.

7.5 MODELADO Y MACHINE LEARNING NIVEL DE RIESGO PSICOSOCIAL

7.5.1 Regresión Lasso:

La Regresión Lasso (*Least Absolute Shrinkage and Selection Operator)* es una técnica de modelado estadístico que combina la regresión lineal con un término de regularización L1. Su principal ventaja es que no solo predice la variable de interés, sino que también selecciona automáticamente las variables más relevantes, eliminando aquellas que no aportan información significativa al modelo [16]

Esto la hace especialmente útil en contextos como el actual donde se dispone de muchas variables predictoras y se busca un modelo simple, interpretable y robusto. En el objetivo 2 de este estudio, la evaluación de riesgo psicosocial en estudiantes y colaboradores universitarios, la Regresión Lasso permite identificar qué factores (variables de exposición ver tabla 5.) están realmente asociados a los indicadores negativos o positivos, facilitando la toma de decisiones informadas y la implementación de intervenciones focalizadas.

7.5.1.1 Descripción del Modelo:

La regresión logística con regularización L1 (Lasso) se implementó para identificar los factores más relevantes asociados con los indicadores en salud mental, este enfoque permite que la selección automática de variables mediante la penalización L1, manejar ya la mencionada multicolinealidad entre predictores e identificar los factores más influyentes en cada indicador.

7.5.1.2 Procesamiento de datos – División de datos:

- Conjunto de entrenamiento: 70% para ajustar los parámetros del modelo
- Conjunto de validación; 15% para ajustar hiperparámetros y evitar sobreajuste
- Conjunto de prueba: 15% para evaluar el desempeño final del modelo sobre datos no vistos.

La división se realizó de manera estratificada, asegurando que la proporción de clases se mantuviera en cada subconjunto.

7.5.1.3 Manejo de Desbalance de Clases:

Se implementó ***RandomOverSampler***, se compararon los resultados con y sin balanceo, para evaluar el impacto del rendimiento del modelo.

7.5.1.4 Preprocesamiento de datos:

Se aplico StandarScaler de scikit-learn, que implementa la estandarización Z-score con la siguiente formula:

$$z = \frac{\times - \mu}{\sigma}$$

donde x es el valor original, μ es la media de la variable y σ es la desviación estándar de la variable. En las características importantes de la implementación, se aplicó solo a las variables numéricas a los datos de entrenamiento usando fit_transform, los mismos parámetros (media y desviación estándar) se aplicaron a los conjuntos de validación y prueba (transform), lo anterior para evitar el *data leakage,* ya que no que se usa información de los conjuntos de validación y prueba para la estandarización.

7.5.1.5 Hiperparámetros del Modelo Lasso:

- penalty: 'l1'
- solver: 'liblinear'
- Cs: np.logspace(-4, 4, 20)
- class_weight: None
- cv: 5
- scoring: 'f1'
- max_iter: 1000
- random_state: 42
- refit: True

7.5.1.6 Validación cruzada:

Durante el entrenamiento, se utilizó validación cruzada de 5 pliegues (cv=5) en el conjunto de entrenamiento para seleccionar el hiperparámetro de regularización óptimo (C) del modelo Lasso. La métrica utilizada para la selección fue el F1-score, que balancea precisión y recall, especialmente relevante en contextos de desbalance de clases como este caso.

El desempeño del modelo se evaluó en el conjunto de prueba utilizando las siguientes métricas:

- **Accuracy:** Proporción de predicciones correctas.
- **F1-score:** Media armónica entre precisión y recall.
- **ROC AUC:** Área bajo la curva ROC, que mide la capacidad de discriminación del modelo.

Se compararon los resultados del modelo entrenado con los datos originales y con datos balanceados mediante ***RandomOverSampler.***

Gráfico 13. Comparación de métricas de desempeño antes y después del balanceo para indicadores Negativos en Colaboradores

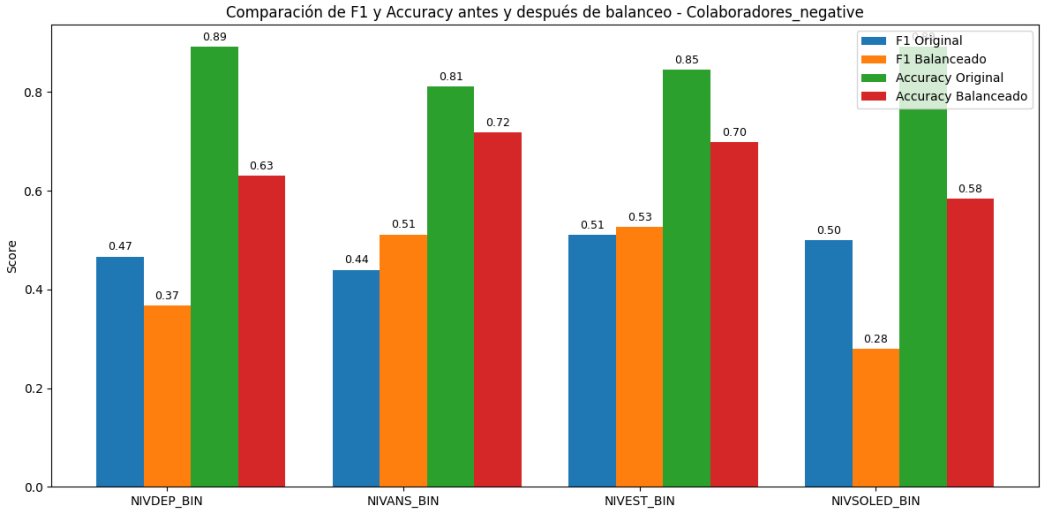
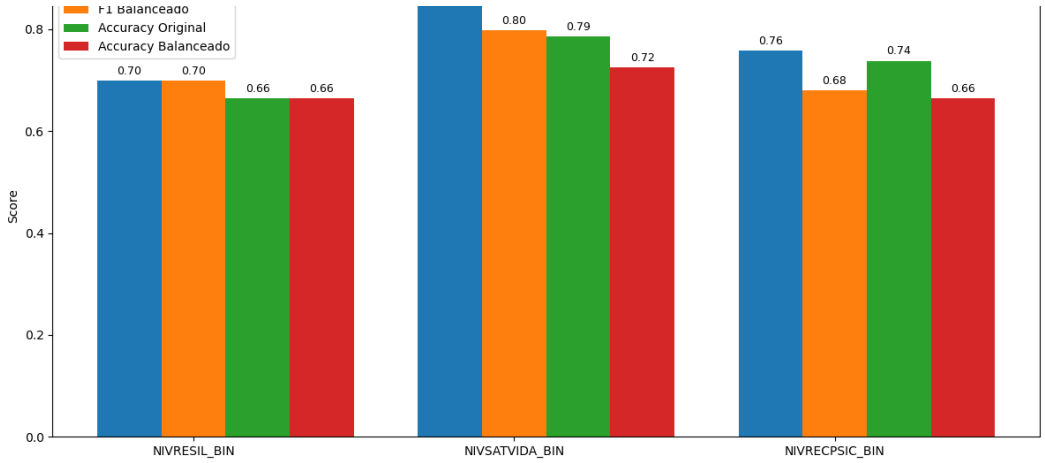


Gráfico 14. Comparación de métricas de desempeño antes y después del balanceo para indicadores Positivos en Colaboradores



En Colaboradores, el balanceo permitió mejorar el F1-score en algunos indicadores, reflejando una mayor sensibilidad para identificar casos positivos, aunque a costa de una menor precisión global, el balanceo de clases tiene un impacto variable según el tipo de indicador. En los indicadores negativos, el balanceo pudo mejorar la detección de casos positivos (F1-score), aunque a costa de una menor precisión global (Accuracy). En los indicadores positivos, el balanceo no aporta mejoras significativas, lo que sugiere que el modelo ya es competente en la identificación de estos casos.

Gráfico 15. Comparación de métricas de desempeño antes y después del balanceo para indicadores Negativos en estudiantes de Posgrado

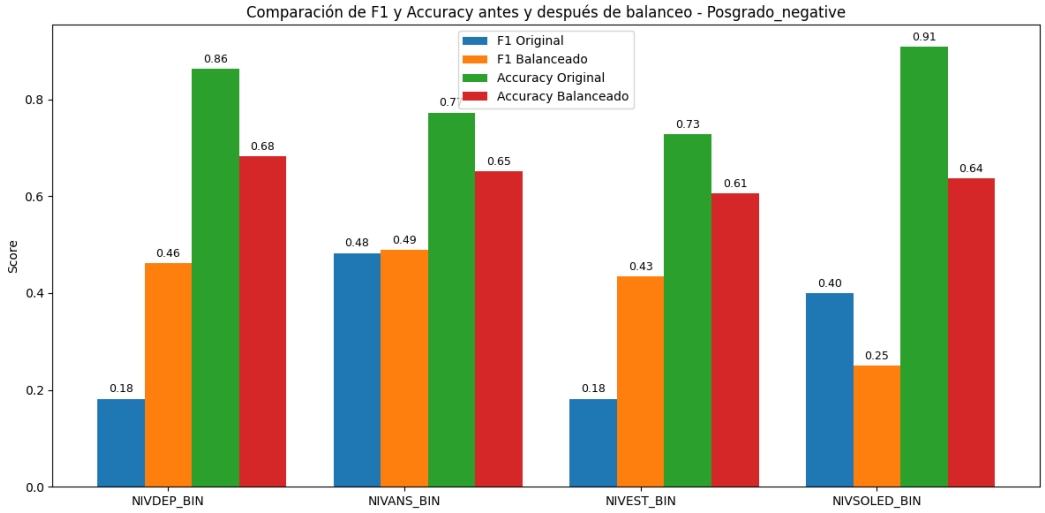
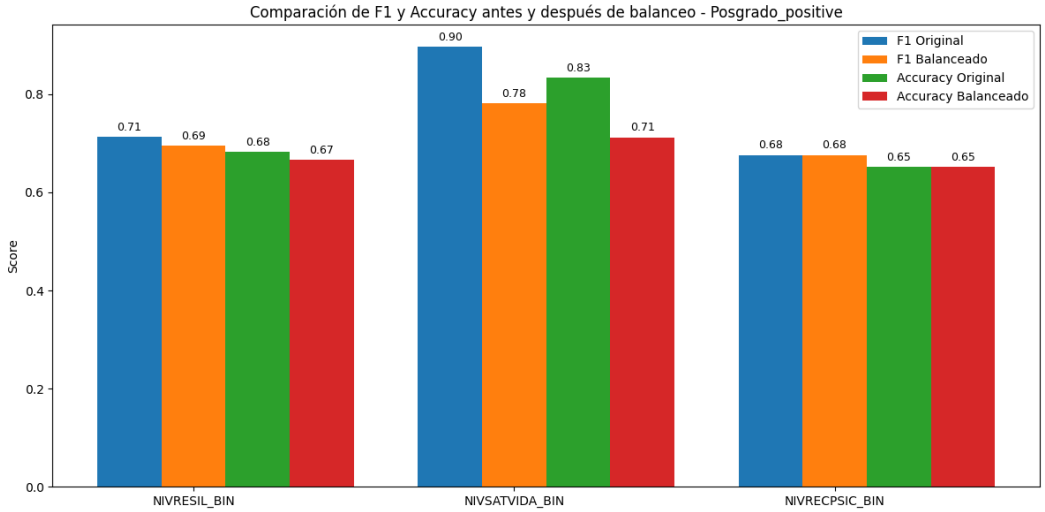


Gráfico 16. Comparación de métricas de desempeño antes y después del balanceo para indicadores Positivos en estudiantes de Posgrado



En estudiantes de Posgrado, el balanceo de clases mediante RandomOverSampler mejora notablemente la capacidad del modelo para identificar casos positivos en algunos indicadores negativos, como depresión y estrés, reflejado en el aumento del F1-score. Sin embargo, esta mejora se acompaña de una reducción en la precisión global (Accuracy), lo que es esperable al priorizar la detección de la clase minoritaria.

Para los indicadores positivos, el balanceo no genera mejoras sustanciales, ya que el modelo presenta un buen desempeño incluso sin balanceo.

Gráfico 17. Comparación de métricas de desempeño antes y después del balanceo para indicadores Negativos en estudiantes de Pregrado

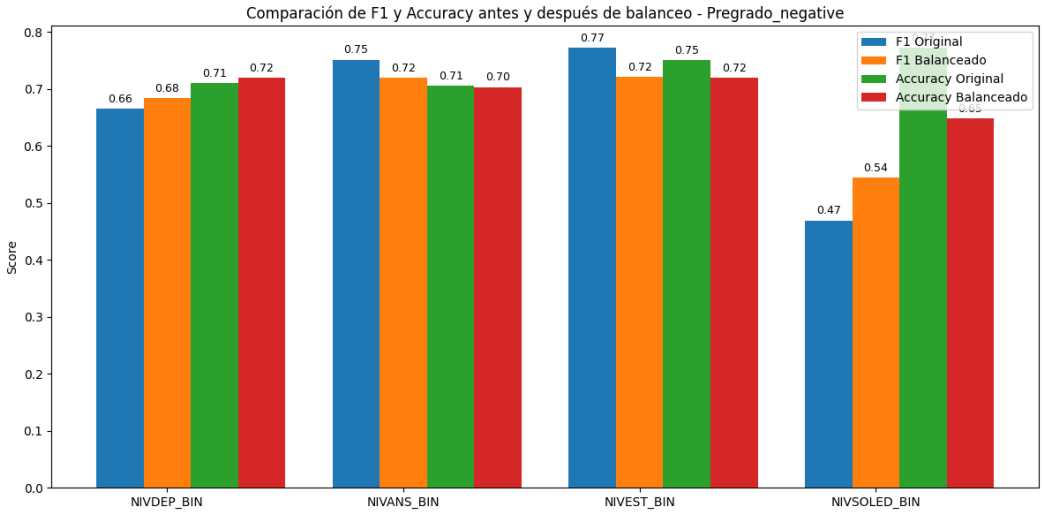
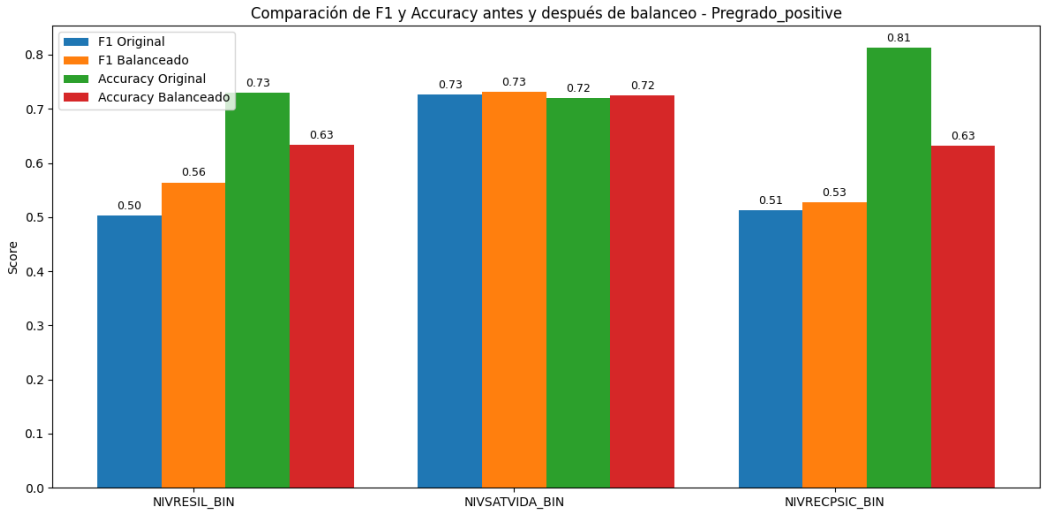


Gráfico 18. Comparación de métricas de desempeño antes y después del balanceo para indicadores Positivos en estudiantes de Pregrado



En estudiantes de Pregrado, el balanceo de clases mediante RandomOverSampler tiene un efecto neutro en la capacidad del modelo para identificar la clase positiva, especialmente en los indicadores de soledad, resiliencia y recuperación psicológica, como se refleja en el aumento del F1-score. Sin embargo, este beneficio puede ir acompañado de una ligera disminución en la precisión global (Accuracy), particularmente en los indicadores donde la clase positiva es más difícil de predecir

7.5.1.7 Resultados Regresión Lasso:

Se presentan los resultados del modelo de regresión logística con regularización Lasso, comparando el desempeño antes y después del balanceo de clases para cada población y tipo de indicado

Tabla 19. Comparación de métricas de desempeño del Modelo Lasso para Colaboradores. Fuente: propia

Indicador	**Métricas originales**			**Métricas balanceadas**		
	Accuracy	**F1**	**ROC_AUC**	**Accuracy**	**F1**	**ROC_AUC**
Depresión	0,89	0,47	0,84	0,63	0,37	0,84
Ansiedad	0,81	0,44	0,76	0,72	0,51	0,77
Estrés	0,85	0,51	0,85	0,70	0,53	0,84
Soledad	0,89	0,50	0,86	0,58	0,28	0,89
Resiliencia	0,66	0,70	0,69	0,66	0,70	0,69
Satisfacción con la vida	0,79	0,86	0,83	0,72	0,80	0,84
Recursos Psicológicos	0,74	0,76	0,81	0,66	0,68	0,80

Tabla 20. Comparación de métricas de desempeño del Modelo Lasso para Estudiantes de Posgrado. Fuente: propia

Indicador	Métricas originales			Métricas balanceadas		
	Accuracy	F1	ROC_AUC	Accuracy	F1	ROC_AUC
Depresión	0,86	0,18	0,89	0,68	0,46	0,87
Ansiedad	0,77	0,48	0,69	0,65	0,49	0,69
Estrés	0,73	0,18	0,61	0,61	0,43	0,64
Soledad	0,91	0,40	0,74	0,64	0,25	0,70
Resiliencia	0,68	0,71	0,75	0,67	0,69	0,74
Satisfacción con la vida	0,83	0,90	0,87	0,71	0,78	0,80
Recursos Psicológicos	0,65	0,68	0,67	0,65	0,68	0,67

Tabla 21. Comparación de métricas de desempeño del Modelo Lasso para Estudiantes de Pregrado. Fuente: propia

Indicador	Métricas originales			Métricas balanceadas		
	Accuracy	F1	ROC_AUC	Accuracy	F1	ROC_AUC
Depresión	0,71	0,66	0,80	0,72	0,68	0,80
Ansiedad	0,71	0,75	0,80	0,70	0,72	0,79
Estrés	0,75	0,77	0,81	0,72	0,72	0,82
Soledad	0,77	0,47	0,78	0,65	0,54	0,78
Resiliencia	0,73	0,50	0,74	0,63	0,56	0,74
Satisfacción con la vida	0,72	0,73	0,80	0,72	0,73	0,80
Recursos Psicológicos	0,81	0,51	0,83	0,63	0,53	0,83

Figura 8. Curvas ROC comparativas - Indicadores Negativos y Positivos en Colaboradores

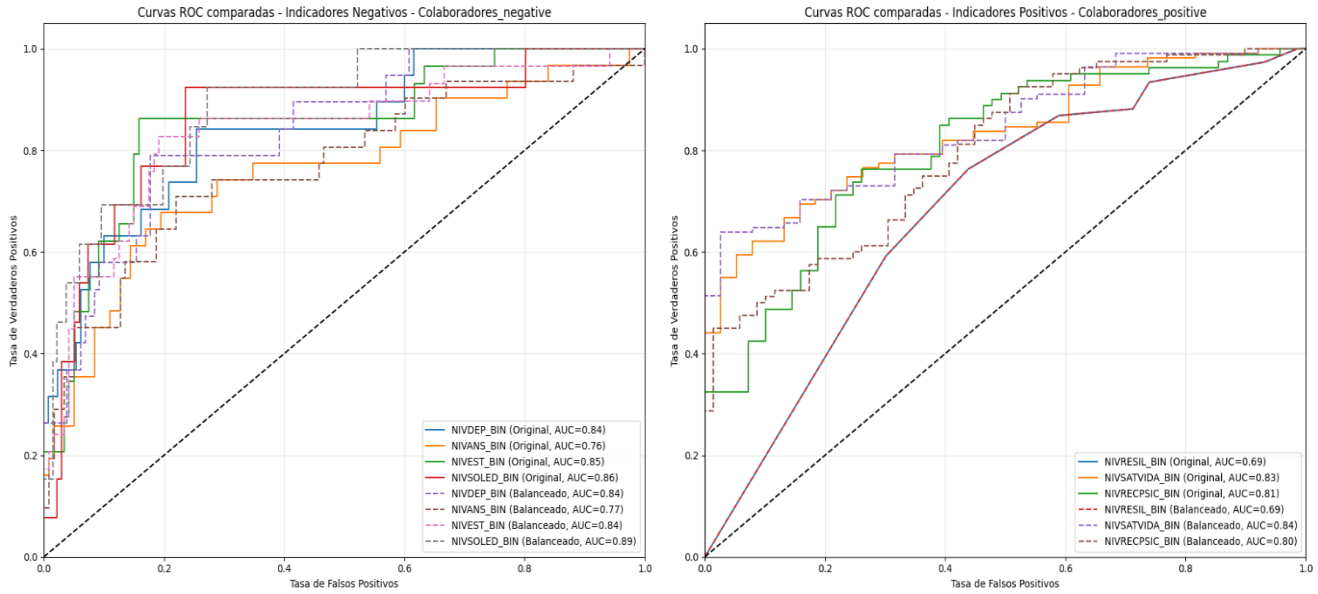


Figura 9. Curvas ROC comparativas - Indicadores Negativos y Positivos en Estudiantes de Posgrado

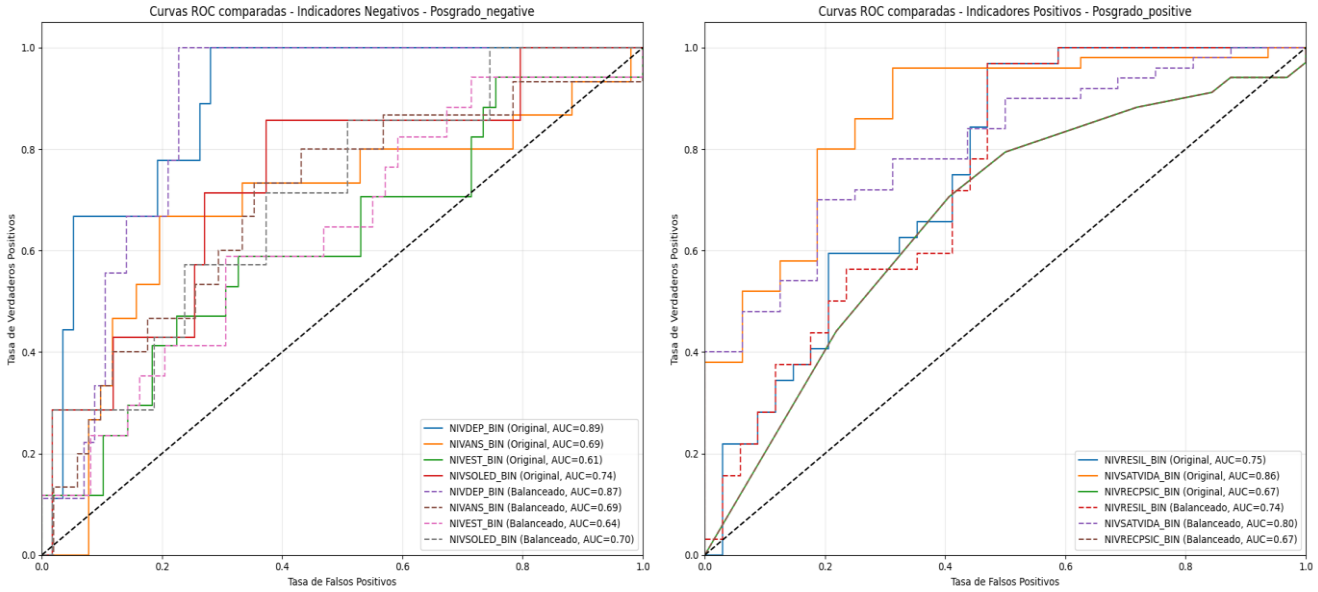
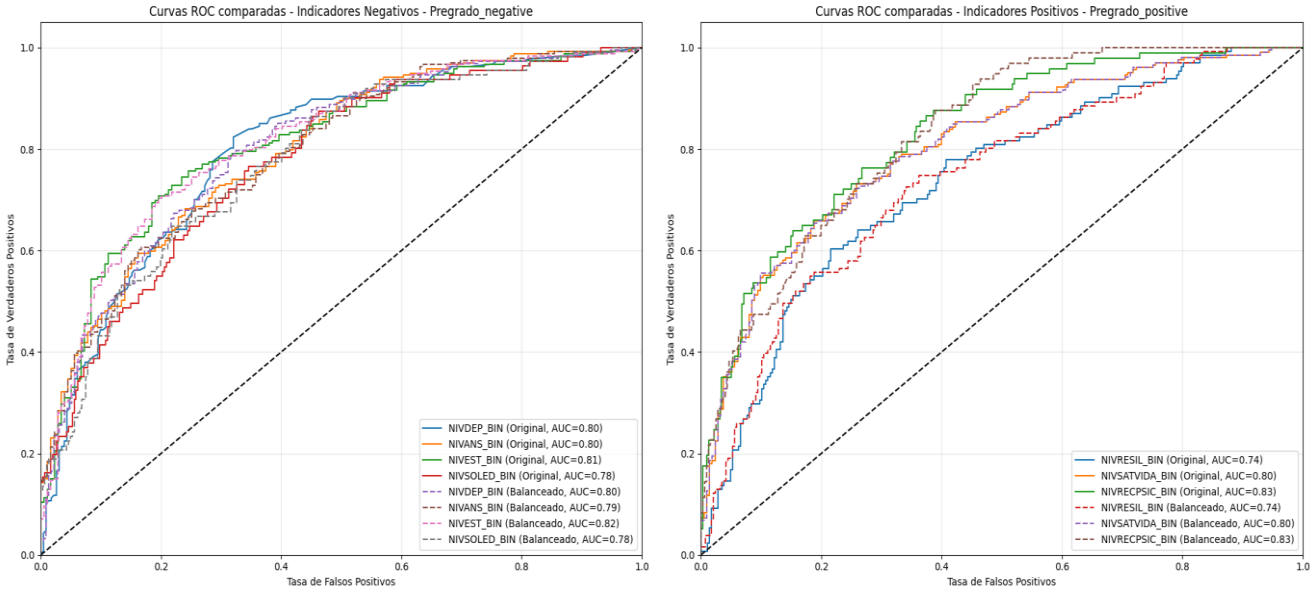


Figura 10. Curvas ROC comparativas - Indicadores Negativos y Positivos en Estudiantes de Pregrado



A continuación, se presentan las tablas con los coeficientes estimados por el Modelo Lasso para cada variable predictora e indicador Positivo o negativo por Colaboradores, estudiantes de Posgrado y Pregrado, los coeficientes distintos de cero indican variables seleccionadas como relevantes por el modelo. Muchas variables tienen coeficiente cero en varios indicadores, lo que indica que no aportan información relevante para la predicción de ese resultado específico según el modelo Lasso.

Tabla 22. Coeficientes estimados por el modelo Lasso para cada variable predictora y cada indicador de riesgo o bienestar psicosocial en Colaboradores. Fuente: propia

Variable predictora	Indicadores Negativos				Indicadores Positivos		
	Depresión	Ansiedad	Estrés	Soledad	Resiliencia	Satisfacción con la vida	Recursos Psicológicos
AFRONTAMIENTO	**-0,102**	**0,235**	**-0,047**	**0,062**	**0,000**	**0,344**	**0,158**
AMBIENTE_UNI	-0,082	-0,082	0,039	0,067	0,000	0,000	0,000
ANTECVIOL	0,181	0,229	0,294	0,366	0,000	-0,279	0,000
APOYO_SOC_FAM	**-0,008**	**-0,342**	**-0,033**	**-0,298**	**0,000**	**0,002**	**0,251**
BIENESTAR	**-1,139**	**-0,767**	**-1,256**	**-1,089**	**0,327**	**0,858**	**0,645**
CVingreshogar	0,000	0,104	0,486	-0,339	0,000	0,202	0,000
CVingresufic	-0,210	0,024	-0,166	0,084	0,000	0,361	0,000
CVsegurbarrio	-0,207	0,031	-0,111	-0,484	0,000	-0,183	0,000
CVviolenbarrio	0,020	-0,174	0,027	-0,001	0,000	0,189	0,000
Edad	-0,180	-0,253	-0,416	-0,083	0,000	0,000	0,115
Estadocivil	-0,353	-0,031	-0,121	-0,165	0,000	-0,030	0,000
Estratosoc	-0,141	-0,435	-0,289	-0,285	0,000	0,000	0,000
FACULTAD	0,096	-0,077	0,151	0,227	0,000	0,221	0,000
Genero	0,224	0,016	-0,334	0,223	0,000	0,000	0,000
GeneroReco	0,000	0,280	0,080	0,604	0,000	0,141	0,000
HAB_ALIMENTICIOS	0,000	0,037	0,227	-0,014	0,000	0,000	0,000
HABITOS	0,019	-0,126	-0,246	-0,241	0,000	0,109	0,000
IMC	0,000	0,054	0,263	0,530	0,000	0,001	0,000
Nacioen	-0,081	-0,014	0,006	0,087	0,000	-0,017	0,000
NIVABUSOTICS	0,119	0,264	0,314	-0,273	0,000	0,000	0,000
NivEduMadre	0,000	0,205	-0,047	-0,312	0,000	0,000	0,000
NivEduPadre	0,213	-0,190	0,040	0,466	0,000	-0,055	0,000
NivelSocioec	0,000	0,579	0,431	-0,016	0,000	0,000	0,000
OCIO	0,185	0,190	0,114	-0,076	0,000	0,051	0,000
PERCSALUD	0,134	0,303	0,091	0,218	0,000	-0,453	0,000
PROGRAMAS_UNI	0,111	0,045	0,008	0,518	0,000	-0,112	-0,010
RazaEtnia	0,092	-0,087	-0,137	-0,064	0,000	0,000	0,000
Residencia	-0,103	-0,084	0,113	-0,170	0,000	0,000	0,000
SALUD_GENERAL	0,257	0,345	0,187	0,283	0,000	-0,022	0,000
SERVICIOS	0,072	-0,079	0,025	0,168	0,000	0,000	0,000
SEXorientacsex	0,000	0,292	0,119	-0,022	0,000	0,000	0,000
SEXpreserv	0,076	0,023	-0,321	-0,327	0,000	0,000	0,000
SPA	0,096	0,172	0,052	0,096	0,000	-0,115	0,000
TRANSPORTE	0,116	0,187	0,392	-0,107	0,000	0,000	0,000
Vivecon	0,046	-0,146	-0,025	0,127	0,000	-0,202	0,000
ViveHijos	-0,056	-0,103	-0,097	-0,265	0,000	-0,098	0,000
ZonaResidencia	-0,023	-0,052	0,093	-0,865	0,000	0,000	0,000

Tabla 23. Coeficientes estimados por el modelo Lasso para cada variable predictora y cada indicador de riesgo o bienestar psicosocial en Estudiantes de Posgrado. Fuente: propia

Variable	Indicadores Negativos				Indicadores Positivos		
	Depresión	**Ansiedad**	**Estrés**	**Soledad**	**Resiliencia**	**Satisfacción con la vida**	**Recursos Psicológicos**
AFRONTAMIENTO	0,000	0,024	0,038	0,079	0,107	0,072	0,000
AMBIENTE_UNI	-0,322	0,221	-0,148	-0,623	0,000	0,000	0,000
ANTECVIOL	0,315	0,136	0,466	0,534	0,000	-0,090	0,000
APOYO_SOC_FAM	**-0,358**	**-0,273**	**0,014**	**-0,234**	**0,000**	**0,281**	**0,000**
BIENESTAR	**-1,027**	**-0,987**	**-1,031**	**-1,425**	**0,532**	**0,633**	**0,637**
CVingreshogar	-0,162	-0,064	0,152	-0,006	0,052	0,077	0,000
CVingresufic	-0,240	-0,353	-0,127	-0,212	0,000	0,032	0,000
CVsegurbarrio	-0,072	0,148	0,494	0,000	0,097	0,000	0,000
CVviolenbarrio	0,231	0,064	-0,060	-0,063	0,000	0,000	0,000
Edad	-0,136	-0,718	-0,285	0,000	0,000	0,000	0,000
Estadocivil	-0,091	-0,245	-0,186	0,000	0,000	0,000	0,000
Estratosoc	0,000	0,143	0,517	0,156	0,000	0,000	0,000
FACULTAD	0,000	0,327	-0,284	0,281	0,000	0,000	0,000
Genero	0,000	0,650	-1,004	0,054	0,211	0,000	0,000
GeneroReco	-0,118	0,719	-1,029	-0,210	-0,068	0,000	0,000
HAB_ALIMENTICIOS	-0,164	0,379	0,291	-0,205	0,000	0,000	0,000
HABITOS	-0,048	-0,013	-0,136	0,024	0,136	0,007	0,000
IMC	0,000	0,057	0,267	-0,832	0,000	0,000	0,000
Nacioen	-0,024	0,713	0,108	0,028	0,000	0,000	0,000
NIVABUSOTICS	-0,125	0,029	0,387	0,000	0,000	0,000	0,000
NivEduMadre	0,000	0,410	0,012	-0,424	0,000	0,000	0,000
NivEduPadre	0,445	-0,571	-0,123	0,057	0,000	0,000	0,000
NivelSocioec	-0,029	0,323	-0,142	-0,298	0,000	0,000	0,000
OCIO	0,103	0,124	-0,167	0,311	0,000	0,000	0,000
PERCSALUD	0,000	0,127	-0,060	0,194	0,000	0,000	0,000
PROGRAMAS_UNI	0,191	0,216	0,344	0,571	0,000	0,000	0,000
RazaEtnia	0,000	0,311	0,010	0,024	0,000	0,000	0,000
Residencia	0,000	0,234	0,224	-0,036	-0,006	0,000	0,000
SALUD_GENERAL	0,038	0,437	0,264	0,024	0,000	-0,084	0,000
SERVICIOS	0,142	-0,444	-0,117	0,297	0,000	0,000	0,000
SEXorientacsex	0,000	-0,073	-0,256	-0,659	0,000	0,000	0,000
SEXpreserv	-0,047	0,357	0,281	-0,587	0,000	0,000	0,000
SPA	-0,028	0,144	-0,063	0,211	0,000	0,000	0,000
TRANSPORTE	-0,074	-0,189	-0,309	0,187	0,006	0,000	0,000
Vivecon	-0,188	-0,364	0,203	0,524	0,025	0,000	0,000
ViveHijos	-0,077	0,402	-0,098	-0,277	0,000	0,000	0,000
ZonaResidencia	0,000	-0,063	0,146	0,022	0,000	0,000	0,000

Tabla 24. Coeficientes estimados por el modelo Lasso para cada variable predictora y cada indicador de riesgo o bienestar psicosocial en Estudiantes de Pregrado. Fuente: propia

Variable	Indicadores Negativos				Indicadores Positivos		
	Depresión	**Ansiedad**	**Estrés**	**Soledad**	**Resiliencia**	**Satisfacción con la vida**	**Recursos Psicológicos**
AFRONTAMIENTO	0,000	0,118	0,096	0,018	0,170	0,171	0,392
AMBIENTE_UNI	0,000	-0,039	0,000	-0,045	0,006	0,000	-0,061
ANTECVIOL	0,083	0,315	0,262	0,126	0,041	0,000	-0,036
APOYO_SOC_FAM	-0,217	-0,222	-0,100	-0,465	0,171	0,329	0,272
BIENESTAR	**-0,898**	**-0,687**	**-0,858**	**-0,945**	**0,670**	**1,008**	**1,209**
CVingreshogar	0,000	0,037	0,000	-0,044	0,029	0,000	0,005
CVingresufic	0,000	-0,021	0,000	0,000	-0,050	0,000	0,028
CVsegurbarrio	0,000	-0,102	0,000	-0,034	0,006	0,000	-0,090
CVviolenbarrio	0,000	0,116	0,000	0,000	-0,061	0,000	0,175
Edad	0,000	-0,182	0,000	-0,165	0,151	0,000	0,104
Estadocivil	0,000	0,075	0,000	-0,024	0,042	0,000	-0,043
Estratosoc	0,000	0,067	0,000	-0,004	0,101	0,000	0,072
FACULTAD	0,000	-0,084	0,000	-0,034	-0,034	-0,056	0,000
Genero	0,000	-0,074	0,000	0,001	-0,027	-0,006	-0,117
GeneroReco	0,000	0,174	0,292	0,093	-0,237	0,025	-0,215
HAB_ALIMENTICIOS	0,000	0,071	0,016	-0,040	0,053	0,000	0,127
HABITOS	0,000	-0,288	-0,174	-0,013	0,190	0,000	0,201
IMC	0,000	0,044	0,000	-0,078	-0,003	0,000	0,144
Nacioen	0,000	0,025	0,000	0,000	0,000	0,000	-0,026
NIVABUSOTICS	0,061	0,164	0,119	-0,009	0,001	-0,018	-0,024
NivEduMadre	0,000	-0,034	0,000	0,021	0,000	0,009	0,111
NivEduPadre	0,000	-0,086	0,000	0,007	0,140	0,000	-0,035
NivelSocioec	0,000	0,000	0,000	0,062	-0,101	0,000	0,000
OCIO	0,000	0,097	0,000	-0,027	-0,066	0,000	0,105
PERCSALUD	0,000	0,144	0,008	0,070	-0,009	-0,124	-0,044
PROGRAMAS_UNI	0,000	0,000	0,000	0,070	-0,195	-0,063	-0,087
RazaEtnia	0,000	-0,019	0,000	-0,015	-0,203	0,000	-0,049
Residencia	0,000	-0,029	-0,021	0,043	0,031	0,000	0,020
SALUD_GENERAL	0,000	0,271	0,148	0,067	-0,133	0,000	-0,146
SERVICIOS	0,000	-0,124	0,000	-0,011	0,034	0,009	-0,013
SEXorientacsex	0,000	0,160	0,025	0,000	0,014	-0,090	-0,066
SEXpreserv	0,000	-0,058	0,000	0,000	0,046	0,000	0,079
SPA	0,000	0,082	0,031	0,176	-0,021	-0,098	-0,074
TRANSPORTE	0,000	0,212	0,000	0,000	-0,108	0,000	-0,097
Vivecon	0,000	0,033	0,000	0,159	-0,039	0,000	0,001
ViveHijos	0,000	-0,193	-0,007	-0,039	0,111	0,000	0,133
ZonaResidencia	0,000	0,013	0,000	-0,001	-0,033	0,000	-0,060

En las tres poblaciones, (Colaboradores, estudiantes de Pregrado y Posgrado), variables como BIENESTAR, APOYO_SOC_FAM, y en menor medida SALUD_GENERAL y AFRONTAMIENTO, aparecen consistentemente con coeficientes distintos de cero y, en muchos casos, con valores altos en valor absoluto. Esto indica que son factores relevantes tanto para los indicadores negativos (riesgo) como positivos

(bienestar), se observa que BIENESTAR y APOYO_SOC_FAM suelen tener coeficientes negativos en los indicadores negativos, es decir, actúan como factores protectores frente a depresión, ansiedad, estrés y soledad y positivos en los indicadores de bienestar resiliencia, satisfacción con la vida y recursos psicológicos. Adicionalmente, ANTECVIOL (antecedentes de violencia) y Estratosoc (estrato socioeconómico) muestran efectos variables según el indicador y la población, pero en general, antecedentes de violencia se asocian a mayor riesgo (coeficientes positivos en negativos).

Si bien existen diferencias en la magnitud y relevancia de otras variables según la población y el indicador, el modelo Lasso permitió identificar de manera robusta los factores clave, descartando aquellos de menor importancia. Estos hallazgos subrayan la necesidad de fortalecer el bienestar y el apoyo social en las intervenciones universitarias, adaptando las estrategias a las características específicas de cada grupo poblacional.

7.5.2 Modelos con LightGMB, Random Forest y XGBoost:

Seguido de Regression Lasso, se evaluaron tres modelos de clasificación supervisada para la predicción del riesgo psicosocial, utilizando algoritmos de machine learning robustos y ampliamente aceptados en la literatura y la industria. Se evaluaron y compararon los modelos Random Forest, XGBoost y LightGBM para los Indicadores Positivos y Negativos, en las poblaciones de Colaboradores, estudiantes de Posgrado y Pregrado.

7.5.2.1 Parámetros y configuración de los modelos:

Random Forest:

n_estimators=100

max_depth=7

random_state=42

n_jobs=-1

XGBoost:

n_estimators=100

max_depth=7

use_label_encoder=False

eval_metric='logloss'

random_state=42

n_jobs=-1

LightGBM:

n_estimators=100

max_depth=7

random_state=42

n_jobs=-1

7.5.2.2 Preprocesamiento:

Escalado de variables numéricas con StandardScaler, codificación de variables categóricas con OneHotEncoder y pipeline de preprocesamiento y modelo para evitar data leakage.

7.5.2.3 División de datos y validación:

Los datos se dividieron en conjunto de entrenamiento (80%) y prueba (20%) usando train_test_split con estratificación por la variable objetivo para mantener la proporción de clases, sobre la validación, el desempeño reportado corresponde al conjunto de prueba, que no fue visto por el modelo durante el entrenamiento. No se utilizó validación cruzada exhaustiva para priorizar eficiencia computacional, pero la estratificación y el uso de un conjunto de prueba independiente aseguran una evaluación robusta.

7.5.2.4 Métricas:

- **Accuracy:** Proporción de predicciones correctas sobre el total de casos.
- **Precision**: Proporción de verdaderos positivos entre los predichos como positivos.
- **Recall (Sensibilidad):** Proporción de verdaderos positivos correctamente identificados.
- **F1-score:** Media armónica entre precisión y recall, útil en contextos de desbalance de clases.
- **ROC AUC:** Área bajo la curva ROC, mide la capacidad de discriminación del modelo.

- **Average Precision:** Precisión promedio bajo la curva Precision-Recall, relevante en desbalance de clases.

Para interpretar el modelo, un valor alto de F1-score y ROC AUC indica que el modelo es capaz de identificar correctamente los casos positivos manteniendo un buen equilibrio entre precisión y sensibilidad. La métrica de Average Precision complementa la evaluación en escenarios de desbalance de clases.

Tabla 25. Desempeño de modelos supervisados para la predicción de Indicadores en salud mental en Colaboradores. Fuente: propia

	Variable	Modelo	Accuracy	Precisión	Recall	F1-score	ROC AUC	Average Precision
Indicadores negativos	Depresión	RandomForest	0,848	0,432	0,640	0,516	0,849	0,486
		XGBoost	0,803	0,360	0,720	0,480	0,858	0,430
		LightGBM	0,848	0,432	0,640	0,516	0,841	0,453
	Ansiedad	RandomForest	0,773	0,444	0,390	0,416	0,768	0,508
		XGBoost	0,793	0,500	0,439	0,468	0,769	0,490
		LightGBM	0,788	0,486	0,415	0,447	0,762	0,475
	Estrés	RandomForest	0,833	0,563	0,692	0,621	0,824	0,610
		XGBoost	0,843	0,583	0,718	0,644	0,824	0,608
		LightGBM	0,833	0,575	0,590	0,582	0,847	0,604
	Soledad	RandomForest	0,924	0,556	0,588	0,571	0,920	0,446
		XGBoost	0,854	0,313	0,588	0,408	0,901	0,478
		LightGBM	0,833	0,318	0,824	0,459	0,900	0,499
Indicadores Positivos	Resiliencia	RandomForest	0,722	0,718	0,740	0,729	0,775	0,774
		XGBoost	0,652	0,635	0,730	0,679	0,731	0,740
		LightGBM	0,687	0,670	0,750	0,708	0,732	0,711
	Satisfacción con la vida	RandomForest	0,773	0,811	0,905	0,855	0,838	0,936
		XGBoost	0,793	0,840	0,891	0,865	0,821	0,928
		LightGBM	0,813	0,840	0,925	0,880	0,822	0,919
	Recursos psicológicos	RandomForest	0,742	0,762	0,755	0,758	0,784	0,789
		XGBoost	0,753	0,777	0,755	0,766	0,797	0,809
		LightGBM	0,747	0,764	0,764	0,764	0,797	0,814

Tabla 26. Desempeño de modelos supervisados para la predicción de Indicadores en salud mental en estudiantes de Posgrado. Fuente: propia

	Variable	Modelo	Accuracy	Precision	Recall	F1-score	ROC AUC	Average Precision
Indicadores negativos	Depresión	RandomForest	0,852	0,400	0,167	0,235	0,868	0,404
		XGBoost	0,841	0,417	0,417	0,417	0,843	0,430
		LightGBM	0,773	0,278	0,417	0,333	0,833	0,414
	Ansiedad	RandomForest	0,761	0,481	0,650	0,553	0,827	0,633
		XGBoost	0,773	0,500	0,600	0,545	0,804	0,604
		LightGBM	0,705	0,400	0,600	0,480	0,801	0,513
	Estrés	RandomForest	0,750	0,500	0,455	0,476	0,729	0,460
		XGBoost	0,727	0,458	0,500	0,478	0,736	0,417
		LightGBM	0,693	0,391	0,409	0,400	0,731	0,452
	Soledad	RandomForest	0,886	0,444	0,444	0,444	0,805	0,326
		XGBoost	0,886	0,400	0,222	0,286	0,771	0,310
		LightGBM	0,852	0,250	0,222	0,235	0,747	0,252
Indicadores Positivos	Resiliencia	RandomForest	0,727	0,688	0,786	0,733	0,804	0,720
		XGBoost	0,682	0,621	0,857	0,720	0,688	0,604
		LightGBM	0,670	0,618	0,810	0,701	0,714	0,619
	Satisfacción con la vida	RandomForest	0,795	0,853	0,879	0,866	0,793	0,921
		XGBoost	0,739	0,841	0,803	0,822	0,782	0,915
		LightGBM	0,761	0,808	0,894	0,849	0,749	0,912
	Recursos psicológicos	RandomForest	0,693	0,721	0,674	0,697	0,775	0,747
		XGBoost	0,659	0,667	0,696	0,681	0,713	0,681
		LightGBM	0,648	0,660	0,674	0,667	0,711	0,682

Tabla 27. Desempeño de modelos supervisados para la predicción de Indicadores en salud mental en estudiantes de Pregrado. Fuente: propia

	Variable	Modelo	Accuracy	Precision	Recall	F1-score	ROC AUC	Average Precision
Indicadores negativos	Depresión	RandomForest	0,719	0,685	0,683	0,684	0,782	0,730
		XGBoost	0,701	0,674	0,639	0,656	0,778	0,738
		LightGBM	0,706	0,664	0,691	0,677	0,776	0,710
	Ansiedad	RandomForest	0,731	0,739	0,818	0,777	0,801	0,842
		XGBoost	0,704	0,721	0,787	0,753	0,781	0,829
		LightGBM	0,706	0,726	0,781	0,752	0,782	0,826
	Estrés	RandomForest	0,737	0,774	0,762	0,768	0,802	0,834
		XGBoost	0,711	0,753	0,737	0,745	0,794	0,843
		LightGBM	0,735	0,765	0,774	0,769	0,796	0,834
	Soledad	RandomForest	0,708	0,462	0,608	0,525	0,756	0,557
		XGBoost	0,722	0,481	0,595	0,532	0,775	0,591
		LightGBM	0,726	0,486	0,588	0,532	0,775	0,590
Indicadores Positivos	Resiliencia	RandomForest	0,713	0,541	0,560	0,551	0,726	0,539
		XGBoost	0,679	0,489	0,531	0,510	0,700	0,476
		LightGBM	0,690	0,505	0,560	0,531	0,691	0,504
	Satisfacción con la vida	RandomForest	0,724	0,713	0,729	0,721	0,813	0,814
		XGBoost	0,717	0,720	0,689	0,704	0,787	0,783
		LightGBM	0,728	0,722	0,722	0,722	0,798	0,789
	Recursos psicológicos	RandomForest	0,794	0,559	0,546	0,553	0,812	0,582
		XGBoost	0,765	0,496	0,462	0,478	0,800	0,536
		LightGBM	0,778	0,523	0,523	0,523	0,799	0,525

Para cada población y variable objetivo, se entrenaron los modelos Random Forest, XGBoost y LightGBM utilizando el 80% de los datos para entrenamiento y el 20% para prueba. El preprocesamiento de los datos se realizó mediante pipelines que incluyeron escalado y codificación de variables. Los modelos fueron entrenados con hiperparámetros estándar, priorizando la interpretabilidad y la eficiencia computacional. El desempeño se evaluó en el conjunto de prueba, asegurando que los resultados reportados reflejan la capacidad de generalización de los modelos.

Los tres modelos aplicados *Random Forest, XGBoost, LightGBM* muestran un desempeño robusto, con valores de ROC AUC generalmente superiores a 0.80 en la mayoría de los indicadores, lo que indica una excelente capacidad de discriminación. En cuanto a los Indicadores negativos en ***Colaboradores*** (Depresión,

Ansiedad, Estrés, Soledad), El F1-score y la sensibilidad (recall) son moderados, lo que es esperable en contextos de desbalance de clases, ampliamente demostrado en este estudio, adicionalmente, Random Forest y LightGBM tienden a obtener los mejores valores de F1-score y ROC AUC, especialmente en depresión y estrés. XGBoost destaca en recall para depresión y resiliencia, lo que sugiere que es más sensible a la detección de casos positivos. Sobre los Indicadores positivos en (Resiliencia, Satisfacción con la vida, Recursos psicológicos), los modelos presentan altos valores de precisión y F1-score, especialmente en satisfacción con la vida y recursos psicológicos. La métrica de Average Precision es alta, lo que refuerza la capacidad de los modelos para identificar correctamente los casos positivos.

Hay que destacar que, en los estudiantes de ***Posgrado***, el desempeño es ligeramente inferior al de colaboradores, pero los modelos siguen mostrando valores aceptables de ROC AUC (mayor a 0.80 en varios indicadores), así mismo, en Indicadores negativos de los estudiantes de Posgrado, el F1-score es más bajo, especialmente en depresión y soledad, lo que puede deberse a un mayor desbalance de clases o a una menor cantidad de casos positivos. Sin embargo, la precisión y el ROC AUC se mantienen en niveles aceptables, lo que indica que los modelos no están sobreajustando. Los modelos muestran un mejor desempeño en resiliencia y satisfacción con la vida, con F1-score y ROC AUC altos. XGBoost y Random Forest tienden a ser los modelos más robustos en esta población.

El desempeño de los modelos en estudiantes de ***Pregrado*** es muy sólido, con valores de F1-score y ROC AUC consistentemente altos en todos los indicadores, para los Indicadores negativos, el F1-score y la precisión son superiores a los observados en posgrado, lo que sugiere que los modelos logran un mejor equilibrio entre la identificación de casos positivos y negativos. Random Forest y LightGBM destacan en la mayoría de los indicadores. En cuanto Indicadores positivos, los valores de F1-score, ROC AUC y Average Precision son especialmente altos en satisfacción con la vida y recursos psicológicos, lo que indica una excelente capacidad predictiva. La consistencia de los resultados sugiere que los determinantes sociales seleccionados son altamente informativos para esta población.

7.6 HERRAMIENTA DE VISUALIZACIÓN

Para cumplir el objetivo 3, se desarrolló un tablero interactivo en Power BI que permite visualizar y explorar los resultados del análisis y modelado del riesgo psicosocial. El tablero incluye visualizaciones de las métricas de desempeño de los modelos, la importancia de los determinantes sociales para cada indicador y población, y herramientas de filtrado para facilitar la toma de decisiones y la identificación de grupos de riesgo.

Acceso: https://app.powerbi.com/links/pO7cS852kk?ctid=16af6b45-00c5-4ba3-9d4c-8bfa16e43603&pbi_source=linkShare&bookmarkGuid=50375fc7-63ea-4990-8523-fc4269a813f1

8 CONCLUSIONES

El presente estudio abordó de manera integral el análisis de una población compleja y difícil de segmentar, pues desde el inicio se detectó limitaciones con el conjunto de datos, no obstante, se procedió con un combinando de técnicas avanzadas de análisis de datos, machine learning y visualización interactiva. A través de los tres objetivos propuestos, se logró no solo describir y caracterizar la diversidad interna de la muestra, sino también identificar patrones latentes y construir modelos predictivos robustos para los indicadores de interés.

En el primer objetivo, la aplicación de análisis descriptivos y técnicas de reducción de dimensionalidad, junto con métodos de agrupamiento como Kmeans, clustering aglomerativo y GMM, permitió explorar la estructura interna de los datos. Aunque las métricas cuantitativas no evidenciaron la existencia de clústers naturales bien definidos, el uso de visualizaciones avanzadas como t-SNE y UMAP facilitó la identificación de posibles subgrupos y gradientes, los cuales fueron complementados con la visualización Sankey.

Para el segundo objetivo, la implementación de modelos de aprendizaje supervisado incluyendo Regresión Lasso, Random Forest, XGBoost y LightGBM posibilitó la predicción de las variables resultado (INDICADORES NEGATIVOS E INDICADORES POSITVOS) asociadas al 'BIENESTAR', 'APOYO_SOC_FAM', 'SALUD_GENERAL' y 'AFRONTAMIENTO', siendo un hallazgo de interés para los encargados de las políticas de permanencia estudiantil y de convivencia laboral, para establecer estrategias que busquen el fortalecimiento de estos aspectos en Colaboradores y estudiantes. El uso de técnicas de sobremuestreo y validación cruzada garantizó la robustez y generalización de los modelos, aportando evidencia sólida para la toma de decisiones.

Finalmente, en el tercer objetivo, el desarrollo de una herramienta de visualización interactiva basada en PowerBi permitió comunicar de manera clara y accesible la complejidad de los hallazgos, facilitando la exploración dinámica de los datos y la identificación de patrones relevantes por parte de usuarios técnicos y no técnicos.

9 REFERENCIAS BIBLIOGRÁFICAS

[1] Organización Panamericana de la Salud, «Día Mundial de la Salud Mental: la depresión es el trastorno mental más frecuente,» 9 Octubre 2012. [En línea]. Available: https://www3.paho.org/hq/index.php?option=com_content&view=article&id=7305:2012-dia-mundial-salud-mental-depresion-trastorno-mental-mas-frecuente&Itemid=0&lang=es#gsc.tab=0. [Último acceso: 6 Junio 2023].

[2] Organización Panamericana de la Salud, «El Impacto de la pandemia COVID-19 en la salud mental de la población,» [En línea]. Available: https://www.paho.org/es/boletin-desastres-n131-impacto-pandemia-covid-19-salud-mental-poblacion. [Último acceso: 6 Junio 2023].

[3] M. Varela, I. Cepeda, A. Uribe, N. Cadavid y J. Botero, «Proyecto Salud y Bienestar en la Comunidad Educativa Javeriana,» Cali, 2023.

[4] J. Allen, R. Balfour, R. Bell y M. Marmot, «Social determinants of mental health,» *International Review of Psychiatry,* vol. 26, nº 4, pp. 392-407, 2014.

[5] Ministerio de Salud y Protección Social;, «Análisis de Situación de la Salud Mental con Énfasis en Determinantes Sociales.,» Ministerio de Salud y Protección Social, Bogotá, 2024.

[6] I. G. Sarason y B. R. Sarason, Psicopatología. Psicología anormal: el problema de la conducta inadaptada, México: Pearson Educación, 2006.

[7] D. Perlman y L. A. Peplau, «Toward a Social Psychology of Loneliness,» *Academic Press,* p. 13, 1981.

[8] M. K. Nock, G. Borges, E. J. Bromet, C. B. Cha, R. C. Kessler y S. Lee, «Suicide and Suicidal Behavio,» *Epidemiol Rev,* vol. 30, nº 1, pp. 133-154, 2008.

[9] R. Newman, «The road to resilience,» *American Psychological Association,* vol. 33, nº 9, p. 62, 2002.

[10] E. Diener, R. A. Emmons, R. J. Larsen y S. Griffin, «The Satisfaction With Life Scale,» *PubMed,* vol. 49, nº 1, p. 71, 1985.

[11] F. Luthans, C. M. Youssef y B. J. Avolio, Psychological capital: Developing the human competitive edge, Oxford: Oxford University Press, 2007.

[12] P. Bruce, A. Bruce y P. Gedeck, Estadística práctica para ciencia de datos con R y Python, Bogotá: Marcombo, S. L., 2022.

[13] C. M. Pineda Pertuz, Aprendizaje Automático y profundo en Python, Bogotá: Ediciones de la U, 2021.

[14] J. Grus, Ciencia de datos desde cero, Madrid: O'REILLY, 2019.

[15] S. Raschka, Y. H. Liu y V. Mirjalili, Machine Learning con Pytorch y Scikit-Learn, Madrid: Marcombo, 2023.

[16] R. Tibshirani, «Regression Shrinkage and Selection via the Lasso,» *Royal Statistical Society,* vol. 58, nº 1, pp. 267-288, 1996.

[17] A. Géron, Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow, Sebastopol: O'REILLY, 2019.

[18] LightGBM, Read the docs;, «lightgbm.readthedocs.io,» Microsoft Corporation, [En línea]. Available: https://lightgbm.readthedocs.io/en/stable/index.html. [Último acceso: 01 Mayo 2025].

[19] L. Breiman, «Random Forest,» *Machine Learning,* vol. 45, nº 1, pp. 5- 32, 2001.

[20] R. Ferrero, «Maxima Formación,» Maxima Formación, 10 Enero 2022. [En línea]. Available: https://www.maximaformacion.es/blog-ciencia-datos/que-es-la-multicolinealidad-y-por-que-es-un-problema/. [Último acceso: 15 Abril 2025].

[21] J. F. Hair, W. C. Black, B. J. Babin y R. E. Anderson, Multivariate Data Analysis, New York: Pearson , 2010.

[22] J. F. Hair, W. C. Black, B. J. Babin y R. E. Anderson, Multivariate Data Analysis, Andover: CENGAGE, 2019.

[23] C. Madhumitha, «Prediction of mental health (depression) using data science and machine learning techniques,» Sathybama Institue of Science and Technology, Chennai, India, 2022.

[24] D. Olawade, W. Ojima, A. Odetayo, D. Aanouluwapo Clement, A. Fiyinfoluwa y J. Eberhardt, «Enchancing mental health with Artificial Intelligence: Currente trends and future prospects,» *Journal of Medicine, Surgery, and Public Health,* vol. 3, nº 100099, 2024.

[25] L. C. López Steinmetz y J. C. Godoy, «Posibles aplicaciones prácticas del uso de Machine Learning (ML) en la investigación y práctica de la clínica psicológica.,» *Acta Psiquiátrica y Psicológica de América Latina,* vol. 69, nº 4, pp. 266-272, 2023.

[26] G. Cevolani, F. J. Bargagli Stoffi y G. Gnecco, «Simple Models in Complex Worlds: Occam's Razor and Statistical Learning Theory,» *Minds and Machine,* vol. 32, pp. 13-42, 2022.

[27] B. G. Tabachnick y L. S. Fidell, Using Multivariate Statistics, California: Pearson, 2019.

[28] I. T. Jolliffe y J. Cadima, «Principal component analysis: a review and recent developments,» *The Royal Society Publishing,* vol. 374, nº 2065, 2016.

[29] M. W. Watkins, «Exploratory Factor Analysis: A Guide to Best Practice,» *Black Psychology,* vol. 44, nº 33, 2018.

[30] C. L. Keyes, «Mental Illness and/or Mental Health?,» *Journal of Consulting and Clinical Psychology,* vol. 73, nº 3, pp. 539-548, 2005.

[31] D. Borsboom, «A network theory of mental disorders,» *PubMed,* vol. 16, nº 1, pp. 5-13, 2017.

[32] R. Van Bork, M. Rhentulla, L. J. Waldorp, J. Kruis, S. Rezvanifar y D. Borsboom, «Latent Variable Models and Networks: Statistical Equivalence and Testability,» *Multivariate Behavioral Research,* vol. 56, nº 2, pp. 175-198, 2021.

[33] Ministerio de Salud y Protección Social, «Salud mental: asunto de todos,» 10 Octubre 2022. [En línea]. Available: https://www.minsalud.gov.co/Paginas/Salud-mental-asunto-de-todos.aspx. [Último acceso: 6 Junio 2023].

[34] Ministerio de Salud y Protección Social, «Las cifras de la salud mental en pandemia,» 15 Julio 2021. [En línea]. Available: https://www.minsalud.gov.co/Paginas/Las-cifras-de-la-salud-mental-en-pandemia.aspx. [Último acceso: 6 Junio 2023].

[35] Marsh McLennan, «Riesgos de personas 2022,» 25 Julio 2022. [En línea]. Available: https://www.marsh.com/mx/risks/people-risk/insights/the-five-pillars-of-people-risk.html. [Último acceso: 6 Junio 2023].

[36] Ministerio del trabajo, «Bienestar y salud mental: un compromiso de MinTrabajo y el Sector Público,» 24 Julio 2019. [En línea]. Available: https://www.mintrabajo.gov.co/prensa/comunicados/2019/julio/bienestar-y-salud-mental-un-

compromiso-de-mintrabajo-y-el-sector-publico. [Último acceso: 6 Junio 2023].

[37] Commission on Social Determinants of Health, «Closing the gap in a generation: health equity through action on the social determinants of health,» World Health Organization, Ginebra, 2008.

[38] World Health Organization, «Social determinants of health,» [En línea]. Available: https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1. [Último acceso: 6 Junio 2023].

[39] World Health Organization, «Social determinants of health: Key concepts,» 7 Mayo 2013. [En línea]. Available: https://www.who.int/news-room/questions-and-answers/item/social-determinants-of-health-key-concepts. [Último acceso: 6 Junio 2023].

[40] R. Shim y M. Comptom, «Addressing the Social Determinants of Mental Health: If Not Now, When? If Not Us, Who?,» *Psychiatric Services,* vol. 69, nº 8, p. 844–846, 2018.

[41] Ministerio de Salud y protección social, «www.minsalud.gov.co,» 12 05 2013. [En línea]. Available: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/ley-1616-del-21-de-enero-2013.pdf. [Último acceso: 23 08 2024].

[42] M. Alegría, A. NeMoyer, I. Falga, Y. Wang y K. Alvarez, «Social Determinants of Mental Health: Where We Are and Where We Need to Go,» *Current Psychiatry Reports,* vol. 20, nº 95, 2018.

[43] Ministerio de Salud y Protección Social, «La Equidad en salud para Colombia,» Bogotá, 2015.

[44] L. Tovar, L. Perea, J. Tovar y C. Zúñiga, «Determinantes sociales de la salud autorreportada: Colombia después de una década,» *O Mundo da Saúde,* vol. 42, nº 1, pp. 230-247, 2018.

[45] G. Pérez, «DETERMINANTES SOCIALES DE LA SALUD EN BOGOTA DC: EL CASO DEL ACCESO Y CALIDAD DE LOS SERVICIOS 2003, 2007 Y 2011,» Bogotá, 2013.

[46] T. Hastie, R. Tibshirani y J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, New York: Springer, 2009.

[47] B. s. Matemáticas, «Universo Mass effect,» [En línea]. Available: https://www.universomasseffect.es/descubriendo-el-valor-de-p-una-guia-practica/. [Último acceso: 25 11 2024].

[48] V. D. Muñoz Jaramillo, «Evaluación de Modelos de Machine Learning para la Predicción de Crímenes en la Ciudad de Medellín,» Universidad Nacional de Colombia, Medellín, 2021.

[49] R. Sambandam, «trcmarkertresearch.com,» Marketing Research, Abril 2003. [En línea]. Available: https://trcmarketresearch.com/whitepaper/cluster-analysis-gets-complicated/. [Último acceso: 01 Mayo 2025].